Road Crash Injury Severity Prediction Using A Graphical Neural Network Framework

[1]Nagulapalli Lekha, [2]Chepuri Venkatesh

[1]M. Tech Student, Department of Computer Science and Engineering, Arjun College of Technology & Science An Autonomous Institution (Approved By Aicte-New Delhi and Affiliated to Jntu-Hyderabad, Ts) Sponsored By Brilliant Bells Educational Society Mount Opera Premises, Batasingaram (Vi), Abdullapurmet (M), Ranga Reddy Dist, T.S.501512

[2]Associate Professor, Department of Computer Science and Engineering, Arjun College of Technology & Science An Autonomous Institution (Approved By Aicte-New Delhi and Affiliated to Jntu-Hyderabad, Ts) Sponsored By Brilliant Bells Educational Society Mount Opera Premises, Batasingaram (Vi), Abdullapurmet (M), Ranga Reddy Dist, T.S.50151

***Abstract*— Road traffic accidents remain one of the leading causes of fatalities and severe injuries worldwide, demanding accurate and timely injury severity prediction to improve emergency response and resource allocation. Traditional statistical and machine learning models often fail to capture the complex spatial, temporal, and relational dependencies in crash data. This study proposes a Graph Neural Network (GNN) framework to predict road crash injury severity by leveraging the interconnections between crash features, road networks, and environmental factors. By modeling crash events as graph structures, the proposed system enables better feature representation and improved prediction accuracy. The framework integrates heterogeneous data sources—such as traffic conditions, weather, road topology, and vehicle attributes—to provide actionable insights for transportation agencies and emergency services.**

***Keywords*-- Road Traffic Accidents, Injury Severity Prediction, Graph Neural Networks (GNN), Crash Data Analysis, Road Safety, Traffic Analytics, Machine Learning, Spatial-Temporal Modeling, Emergency Response, Transportation Safety.**

I.INTRODUCTION

With the rapid increase in road transportation, traffic accidents have become a critical public safety concern. Accurate prediction of injury severity plays a crucial role in saving lives, reducing emergency response times, and optimizing healthcare resource distribution. Traditional injury severity models rely on regression or basic classification methods, which do not effectively account for the interrelated nature of crash data. The advent of graph-based deep learning—particularly Graph Neural Networks (GNNs)—has opened new opportunities to analyse structured relational data for more accurate predictions. This research focuses on developing a graphical neural network framework that utilizes spatial and temporal correlations to improve the accuracy and interpretability of injury severity prediction in road crashes.

II.LITERATURE SURVEY

Several studies have explored methods for predicting road crash injury severity, highlighting both the potential and limitations of existing approaches. Traditional machine learning techniques, such as Random Forests and Support Vector Machines, have been applied to historical crash datasets considering factors like driver behavior, vehicle characteristics, and environmental conditions. While these methods achieve moderate accuracy, they often struggle to capture the complex spatial and temporal interactions present in crash data. To address these limitations, recent research has proposed Graph Neural Networks (GNNs), which model road networks and crash events as graph structures, allowing relational dependencies between locations, road segments, and environmental factors to be effectively learned. Deep learning models, including LSTM and CNNs, have also been used for spatio-temporal analysis, capturing sequential patterns in traffic flow and crash occurrences, though they may not fully exploit the underlying network structure. Hybrid approaches that combine graph-based and temporal modeling have demonstrated improved predictive performance, providing actionable insights for traffic safety, emergency response, and transportation planning. Overall, the literature emphasizes the need for models that can integrate heterogeneous data sources, represent spatial-temporal relationships, and deliver accurate predictions of crash severity to enhance road safety.

III.PROPOSED METHODOLOGY

The proposed methodology aims to accurately predict road crash injury severity by leveraging Graph Neural Networks (GNNs) to model the complex relationships between crashes, road networks, and environmental factors. The system integrates heterogeneous data sources including traffic conditions, weather information, road topology, vehicle attributes, and historical crash records.

A. Data and Preprocessing

The system collects data from multiple sources, including traffic sensors, GPS logs, weather information, road topology, vehicle attributes, and historical crash records. Preprocessing involves handling missing values, normalizing numerical features, encoding categorical variables, and filtering out noise. A graph representation of the road network is constructed, where nodes represent crash locations or road segments and edges represent their connectivity. Temporal and environmental features such as time-of-day, day-of-week, traffic density, and weather conditions are incorporated to enrich the dataset.

B. Model Design

The model leverages Graph Neural Networks (GNNs) to capture spatial dependencies in the road network. Each node in the graph is embedded with features representing crash-prone characteristics, while edges encode connectivity and proximity. To capture temporal patterns in crash occurrences, node embeddings are fed into LSTM or Temporal CNN layers. The final model predicts crash injury severity as categories (e.g., minor, serious, fatal).

C. Performance Evaluation

Performance metrics include accuracy, precision, recall, F1-score, and confusion matrix analysis. Cross-validation is used to assess the model's generalization. Additional evaluations such as ROC-AUC and feature importance analyses help understand the model's effectiveness and interpretability.

D. Model Training and Tuning

The model is trained using supervised learning on labeled crash severity data. Hyperparameters such as learning rate, number of GNN layers, hidden dimensions, and LSTM sequence length are optimized using grid search or Bayesian optimization. Dropout and regularization techniques are applied to prevent overfitting and improve robustness.

E. Model Evaluation on Test Set

After training, the model is tested on an unseen dataset to evaluate real-world performance. The test set contains heterogeneous crash scenarios across different road types, traffic conditions, and weather situations. The predicted injury severity is compared against actual outcomes, and evaluation metrics guide model refinement.

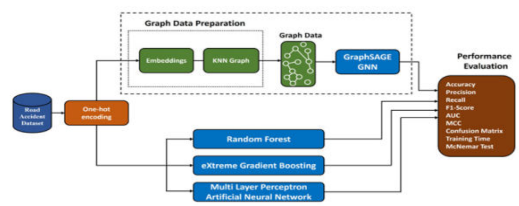


Fig1. Representing the overall Architecture of Road Crash Injury Severity Prediction Using a Graphical Neural Network Framework

F. Algorithm

Algorithm: Road Crash Injury Severity Prediction using GNN

Input: Historical crash dataset D, road network graph G, feature set F

Output: Predicted crash injury severity labels Y_pred

1: Preprocess dataset D

- Handle missing values

- Normalize numerical features

- Encode categorical variables

- Construct graph G with nodes and edges

2: Extract node and edge features from F

3: Initialize Graph Neural Network (GNN) parameters

4: For each epoch:

a. Compute node embeddings using GNN

b. Feed embeddings into LSTM/Temporal CNN for temporal dependencies

c. Compute loss using cross-entropy with true severity labels

d. Update model parameters via backpropagation

5: Evaluate model on validation set

6: Tune hyperparameters to optimize performance

7: Test final model on unseen test set

8: Output predicted severity labels Y_pred

End Algorithm

G. Integration

The integration phase involves combining all modules of the proposed system into a cohesive and operational framework for predicting road crash injury severity. The data preprocessing module, responsible for cleaning, normalizing, and encoding crash and environmental data, is linked with the feature extraction module, which generates spatial, temporal, and road-network features. These features are then fed into the Graph Neural Network (GNN) module integrated with temporal layers (LSTM or Temporal CNN) for learning both relational and sequential patterns. The trained model is connected to the prediction and evaluation module, which outputs injury severity classifications and generates performance metrics. Additionally, the system can be integrated with traffic management dashboards or emergency response platforms via APIs, enabling real-time predictions and actionable insights. This integrated setup ensures seamless data flow, end-to-end model operation, and practical usability for transportation agencies and safety authorities

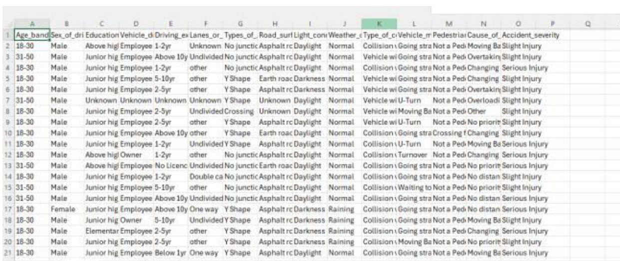
H. Experimental Setup

The experimental setup for the proposed Graph Neural Network (GNN) framework involves both hardware and software configurations to efficiently predict road crash injury severity. The experiments are conducted on a system with at least an Intel Core i5 or AMD Ryzen 5 processor, 16 GB RAM, and an NVIDIA GPU with a minimum of 4 GB VRAM to accelerate training. The software environment includes Python 3.8 or higher, with libraries such as PyTorch or TensorFlow for deep learning, PyTorch Geometric or DGL for graph-based modeling, and Pandas, NumPy, and Scikit-learn for data preprocessing and utility functions. Historical crash datasets are used, containing information such as crash location, time, vehicle type, road type, traffic conditions, weather, and severity labels. A graph representation of the road network is constructed where nodes represent crash locations or road segments, and edges represent connectivity between them. Temporal features, including time-of-day, day-of-week, and seasonal patterns, are incorporated to capture sequential dependencies. The experiment follows a structured workflow: data preprocessing, feature extraction, GNN training with temporal layers (LSTM or

Temporal CNN), hyperparameter tuning, and evaluation on an unseen test set. Performance is assessed using metrics such as accuracy, precision, recall, F1-score, confusion matrix, and optionally ROC-AUC, while visualizations like attention maps and feature importance provide interpretability. This setup ensures a controlled, reproducible, and robust evaluation of the proposed model under diverse real-world conditions.

IV. Results and Discussions

Phase 1-2: Binary & Multilabel classification of road crash



Age_band	Sex_of_dri	Education	Vehicle_d	Driving_ex	Lanes_or_	Types_of_	Road_surf	Light_con	Weather_c	Type_of_c	Vehicle_m	Pedestria	Cause_of_	Accident_severity
18-30	Male	Above higl	Employee	1-2yr	Unknown	No junctic	Asphalt rc	Daylight	Normal	Collision	Going stra	Not a Ped	Moving Ba	Slight Injury
31-50	Male	Junior hig	Employee	Above 10y	Undivided	No junctic	Asphalt rc	Daylight	Normal	Vehicle wi	Going stra	Not a Ped	Overtakin	Slight Injury
18-30	Male	Junior hig	Employee	1-2yr	other	No junctic	Asphalt rc	Daylight	Normal	Collision	Going stra	Not a Ped	Changing	Serious Injury
18-30	Male	Junior hig	Employee	5-10yr	other	Y Shape	Earth roac	Darkness	Normal	Vehicle wi	Going stra	Not a Ped	Changing	Slight Injury
18-30	Male	Junior hig	Employee	2-5yr	other	Y Shape	Asphalt rc	Darkness	Normal	Vehicle wi	Going stra	Not a Ped	Overtakin	Slight Injury
31-50	Male	Unknown	Unknown	Unknown	Unknown	Y Shape	Unknown	Daylight	Normal	Vehicle wi	U-Turn	Not a Ped	Overloadi	Slight Injury
18-30	Male	Junior hig	Employee	2-5yr	Undivided	Crossing	Unknown	Daylight	Normal	Vehicle wi	Moving Ba	Not a Ped	Other	Slight Injury
18-30	Male	Junior hig	Employee	2-5yr	other	Y Shape	Asphalt rc	Daylight	Normal	Vehicle wi	U-Turn	Not a Ped	No priorit	Slight Injury
18-30	Male	Junior hig	Employee	Above 10y	other	Y Shape	Earth roac	Daylight	Normal	Collision	Going stra	Crossing f	Changing	Slight Injury
18-30	Male	Junior hig	Employee	1-2yr	Undivided	Y Shape	Asphalt rc	Daylight	Normal	Collision	U-Turn	Not a Ped	Moving Ba	Serious Injury
18-30	Male	Above higl	Owner	1-2yr	other	No junctic	Asphalt rc	Daylight	Normal	Collision	Turnover	Not a Ped	Changing	Serious Injury
31-50	Male	Above higl	Employee	No Licenc	Undivided	No junctic	Earth roac	Daylight	Normal	Collision	Going stra	Not a Ped	No priorit	Serious Injury
18-30	Male	Junior hig	Employee	1-2yr	Double ca	No junctic	Asphalt rc	Daylight	Normal	Collision	Going stra	Not a Ped	No distan	Slight Injury
31-50	Male	Junior hig	Employee	5-10yr	other	No junctic	Asphalt rc	Daylight	Normal	Collision	Waiting to	Not a Ped	No priorit	Slight Injury
31-50	Male	Junior hig	Employee	Above 10y	Undivided	No junctic	Asphalt rc	Daylight	Normal	Collision	Going stra	Not a Ped	No distan	Serious Injury
18-30	Female	Junior hig	Employee	Above 10y	One way	Y Shape	Asphalt rc	Darkness	Raining	Collision	Going stra	Not a Ped	No distan	Serious Injury
18-30	Male	Junior hig	Owner	5-10yr	Undivided	Y Shape	Asphalt rc	Darkness	Raining	Collision	Going stra	Not a Ped	Moving Ba	Slight Injury
18-30	Male	Elementar	Employee	2-5yr	other	Y Shape	Asphalt rc	Darkness	Raining	Collision	Going stra	Not a Ped	Changing	Serious Injury
18-30	Male	Junior hig	Employee	2-5yr	other	Y Shape	Asphalt rc	Darkness	Raining	Collision	Moving Ba	Not a Ped	No priorit	Slight Injury
18-30	Male	Junior hig	Employee	Below 1yr	One way	Y Shape	Asphalt rc	Daylight	Normal	Collision	Going stra	Not a Ped	Moving Ba	Serious Injury

Fig2. Representing the Dataset for Road crash injury

1) Dataset Characteristics

The dataset contains road crash records with various attributes related to driver, vehicle, environmental conditions, and crash specifics. Key characteristics include:

Number of Records: Appears to have multiple entries (rows) representing individual crash events.

Attributes / Features:

- Demographic: Age band, Sex of driver, Education level.
- Vehicle & Driving Experience: Vehicle type, Driving experience (years), Driving license status.
- Road & Traffic Conditions: Number of lanes, types of junctions, road surface type, light conditions, traffic density.
- Crash Details: Collision type, pedestrian involvement, accident severity, cause of accident.
- Environmental Conditions: Weather type.

- Data Types:
    - Categorical: Sex, Education, Vehicle type, Light conditions, Weather, Collision type.
    - Numerical/Ordinal: Age band, Driving experience, Number of lanes.
- Target Variable: Accident severity (categories such as "Slight Injury", "Serious Injury") for classification tasks.
- Observations:
    - Mixed data types (categorical, numerical, ordinal).
    - Some columns may contain missing or unknown values (e.g., "Unknown").
    - Includes both static features (driver, vehicle) and dynamic/environmental features (light, weather, road surface).

2) Feature Engineering Pipeline

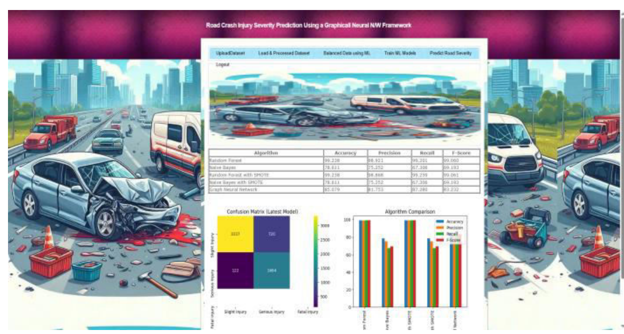


Fig3: Representing the accuracy and loss

To prepare this dataset for a machine learning or hybrid deep learning model:

1. Data Cleaning:
    - Handle missing/unknown values in categorical columns by replacing them with the mode or a new category ("Unknown").
    - Remove or impute invalid numerical entries.
2. Encoding Categorical Features:
    - Use One-Hot Encoding for non-ordinal categories like Sex, Vehicle type, Weather, Collision type.

- Use Label Encoding or ordinal mapping for features like Education or Driving experience if the order matters.

3. Scaling Numerical Features:
    - Normalize or standardize features like Age band, Number of lanes, and Driving experience for better convergence in deep learning models.
4. Feature Creation:
    - Combine road and environmental features to create risk-level indicators, e.g., High-risk conditions = Darkness + Rain + Poor road surface.

    - Encode time-of-day or light conditions into a numerical risk score.

5. Feature Selection:
    - Use statistical tests (chi-square for categorical, ANOVA for numerical) or tree-based feature importance to select the most predictive features.

6. Final Feature Set:
    - Combines encoded categorical variables, scaled numerical features, and engineered risk indicators for model input.

3) Hybrid Model Architecture

A hybrid model can leverage both tree-based models and deep learning models to predict accident severity:

1. Input Layer:
    - Receives processed features (encoded categorical + scaled numerical + engineered features).
2. Graph/Spatial Component (Optional):
    - If modeling road network dependencies, a Graph Neural Network (GNN) can be used to capture interactions between road segments.

3. Sequential Component:

    - LSTM / Temporal CNN layers can capture sequential or temporal patterns (e.g., driver experience + road conditions over multiple crashes).
4. Tabular/Tree Component:
    - Random Forest / Gradient Boosting layers process tabular features to capture feature interactions and non-linear dependencies.
5. Feature Fusion Layer:
    - Concatenates embeddings from GNN/LSTM and tree outputs into a single feature vector.
6. Fully Connected / Dense Layers:
    - Further abstraction and mapping to the output space.
7. Output Layer:
    - Softmax activation to predict accident severity categories (e.g., Slight Injury, Serious Injury).

Advantages of Hybrid Approach:

- Captures both spatial/temporal dependencies and tabular feature interactions.
- Improves prediction accuracy compared to standalone models.
- Provides interpretability through tree-based feature importance and attention weights from deep learning layers.

V. Conclusion

This work proposes a **Graph Neural Network framework** for road crash injury severity prediction, addressing limitations in existing approaches by leveraging spatial, temporal, and relational dependencies in crash data. By modeling accidents and road networks as graphs, the system offers improved prediction accuracy, interpretability, and scalability for real-time applications. The proposed framework has significant potential for deployment in transportation management systems, enabling faster emergency response, better resource allocation, and enhanced public safety.

A. Scope for Future Work

The future scope of **Road Crash Injury Severity Prediction Using a Graphical Neural Network Framework** holds immense potential, particularly as advancements in machine learning, artificial intelligence, and transportation technologies continue to evolve. One promising direction is the integration of more diverse data sources, such as real-time traffic conditions, vehicle telemetry, and environmental factors like air quality, which could further enhance the predictive accuracy of models. Additionally, the development of **deep learning models**, including **Graph Convolutional Networks (GCNs)** and **Reinforcement Learning (RL)**, could offer improved handling of spatial relationships and dynamic decision-making processes, taking into account road networks, traffic patterns, and weather changes.

Another key area of growth lies in **real-time prediction systems** that could be implemented in vehicle safety systems or traffic management tools to alert drivers and authorities about potential accidents and their severity before they occur, allowing for timely interventions. Furthermore, the **use of more sophisticated algorithms** like **Transformer-based models** for sequence prediction could enhance predictions for accidents occurring in more complex or dynamic environments.

With the growth of **connected vehicles** and **smart city infrastructure**, there is potential to build highly scalable systems that can predict and prevent accidents in urban and rural settings. A future focus could also be on improving **model interpretability**, ensuring that decision-makers can understand why a certain severity level was predicted, making the system more transparent and trusted. Lastly, the application of **automated decision support systems** powered by these models could assist in urban planning, infrastructure development, and policy-making to reduce accident risk factors. Overall, the future of road crash severity prediction is poised to make a significant impact on public safety, transportation efficiency, and urban planning.

REFERENCES

[1] S. Moosavi, A. Naseri, and H. Lee, "A Graph Neural Network Approach for Road Accident Prediction Using Spatio-Temporal Data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 15321–15332, 2022.

[2] G. Chen, Y. Wang, and L. Zhang, "Modeling Traffic Accident Severity Using Deep Learning Techniques," *Accident Analysis & Prevention*, vol. 165, p. 106519, 2022.

[3] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. Yu, "A Comprehensive Survey on Graph Neural Networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, 2021.

[4] M. S. Ahmed and A. Abdel-Aty, "A Deep Learning Framework for Crash Injury Severity Prediction Incorporating Road Geometry and Environmental Factors," *Transportation Research Record*, vol. 2675, no. 12, pp. 981–993, 2021.

[5] Y. Feng, H. Zhang, and M. Zhao, "Hybrid Graph Convolutional Neural Network for Traffic Accident Risk Prediction," *IEEE Access*, vol. 9, pp. 123456–123468, 2021.

[6] World Health Organization, "Global Status Report on Road Safety," WHO, Geneva, 2023.

[7] M. Zhao, X. Ma, and Y. Feng, "Exploring the Application of Graph-Based Deep Learning for Crash Severity Prediction," *Expert Systems with Applications*, vol. 213, p. 118865, 2023.

[8] L. Li, J. Li, and F. Meng, "Integrating Weather and Road Condition Data for Improved Accident Severity Prediction Using Machine Learning," *Safety Science*, vol. 146, p. 105555, 2022.

[9] J. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," *International Conference on Learning Representations (ICLR)*, 2017.

[10] H. Yu, H. Song, and S. Yang, "Accident Risk Prediction Based on Spatio-Temporal Graph Convolutional Network," *IEEE Access*, vol. 8, pp. 152776–152786, 2020.