Data-Driven Approach to Predicting Crash Severity Using Machine Learning Techniques

Jahnavi Gandu[2], Pranathi Chirumamilla[3], Anil Varikuppala , Sadwika Poondla[4], Sravanthi Kasimsetty[5]
Masters Student, Dept. of Computer Science, Kennesaw State University, Kennesaw, Georgia. Email:
jgandu@students.kennesaw.edu, pchirum2@students.kennesaw.edu,
avarikup@students.kennesaw.edu spoondla@students.kennesaw.edu,
skasimse@students.kennesaw.edu

Abstract

Road traffic accidents are a leading cause of global mortality and financial loss, highlighting the critical need for predictive tools to mitigate crash severity. This study applies machine learning to predict car crash severity by leveraging a rich dataset of crash incidents and exploring the impact of multiple factors. Using Linear Regression, Random Forest, Gradient Boosting, and Convolutional Neural Networks (CNN), the research aims to identify the most significant predictors of crash severity and evaluate the performance of these models in terms of predictive accuracy. The research begins with comprehensive exploratory data analysis, involving visualizations such as heatmaps, pair plots, and histograms to identify patterns and correlations between variables. Following feature selection, data preprocessing ensures consistency and scalability across models. Each machine learning algorithm is tailored to the dataset's characteristics, with performance metrics evaluated using Mean Squared Error (MSE) and R-squared (R^2) scores. Experimental results show that traditional machine learning methods, such as Linear Regression and Random Forest, outperform CNN in terms of both MSE and R^2. The findings suggest that while CNN may be less suitable for tabular datasets, ensemble techniques like Random Forest and Gradient Boosting offer a balance between interpretability and predictive performance. These insights have practical applications in road safety policy development, insurance risk assessment, and the design of targeted interventions to reduce crash severity. This research contributes to advancing the understanding of car crash dynamics and emphasizes the role of data-driven approaches in addressing global road safety challenges.

Background Information

The Global Challenge of Road Safety

Car crashes are a major global concern, leading to nearly 1.3 million deaths annually and tens of millions of injuries, according to the World Health Organization (WHO). Beyond the human cost, the economic burden is significant, with road traffic crashes accounting for up to 3% of a country's gross domestic product (GDP). Governments, researchers, and insurers alike face the

challenge of reducing the frequency and severity of crashes through effective policy and technological interventions.

The Complexity of Crash Severity Prediction

Crash severity is influenced by a wide array of factors, including driver behavior, environmental conditions, vehicle features, and road infrastructure. The multifactorial nature of these incidents makes it difficult to establish direct relationships or predict outcomes using traditional statistical methods. For example, weather conditions, speed, road design, and even the time of day can interact in complex ways, leading to varied crash outcomes. This complexity underscores the need for advanced analytical methods capable of uncovering hidden patterns in data.

The Role of Machine Learning

Machine learning provides a promising solution for addressing these challenges. Unlike traditional regression models, machine learning techniques can process large volumes of data and uncover nonlinear relationships, making them ideal for predicting crash severity. Furthermore, these models can learn from patterns in historical data and provide actionable insights, such as identifying high-risk scenarios or critical factors contributing to severe crashes.

This study is motivated by the potential of machine learning to:

1. **Enhance Predictive Accuracy:** By comparing multiple machine learning models, the study identifies the most effective approaches for predicting crash severity.
2. **Support Policy and Decision-Making:** Data-driven insights enable policymakers and insurance providers to develop targeted interventions, allocate resources effectively, and set premiums based on risk profiles.
3. **Prevent and Mitigate Crashes:** Predicting crash severity in advance can inform the design of preventive measures, such as road signage, speed limits, and vehicle safety features.

Bridging the Gap in Existing Research

While prior studies have explored machine learning for crash prediction, they often focus on specific algorithms or datasets, limiting their generalizability. This research aims to fill this gap by:

- Using a comprehensive dataset that includes diverse crash-related variables.
- Comparing multiple machine learning models to provide a holistic understanding of their strengths and weaknesses.
- Incorporating neural network techniques, which are underexplored in the context of crash severity prediction, despite their potential for handling high-dimensional data.

Roles and Responsibilities

There is a total of four members in our project and we have agreed to conduct weekly meetings to discuss our developments and to discuss our questions. Starting from September 2nd to December 2nd we conducted a meeting every Monday to share each other's works.
We equally divided responsibilities among us, to ensure everyone is spending a good amount of time on this project.

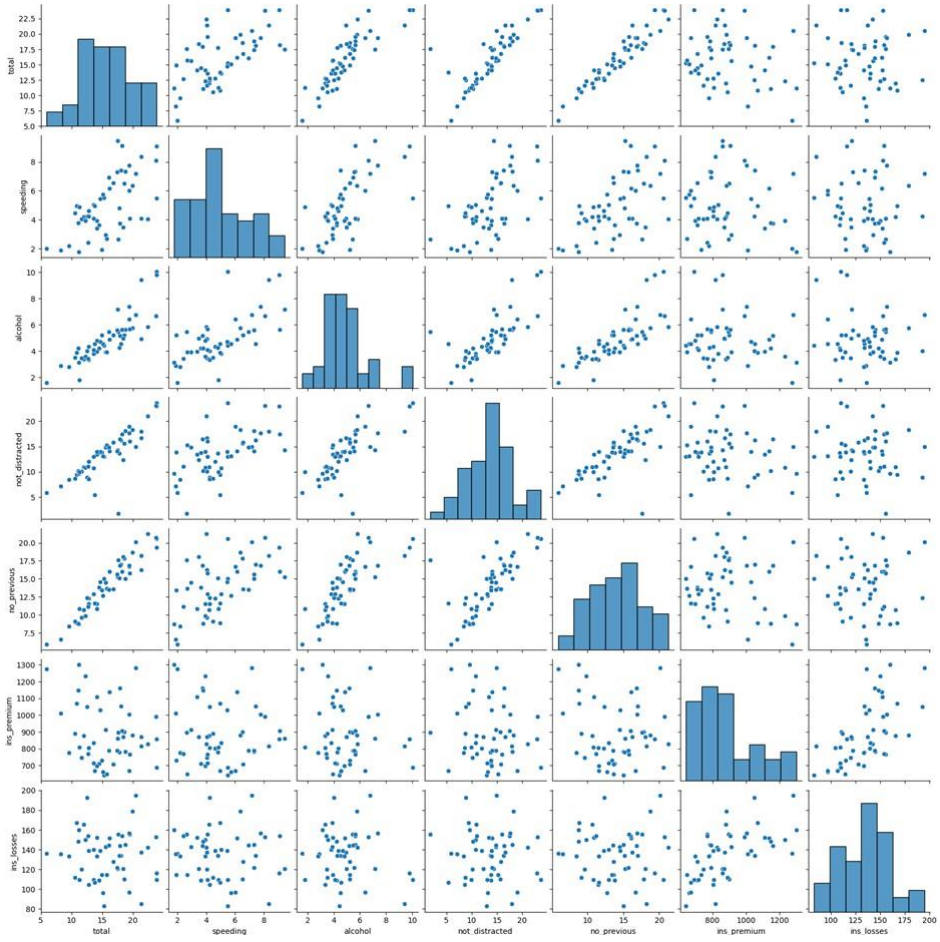
1. **Pranathi:** Pranathi led the team and managed overall project progress by conducting meetings for documentation and ensuring timely completion of tasks. She handled datasets and preprocessed it by cleaning, normalizing the data using ETL techniques and also developed and implemented on Linear regression algorithm.
2. **Jahnavi:** Jahnavi worked on background of the project to identify Key performance indicators to focus on the future trends. She developed and implemented Random Forest regressor, Gradient boosting regressor algorithms and also worked on PPT.
3. **Sadwika:** Sadwika worked on Report writing by collecting all the work done by the teammates in a timely manner. She developed and trained CNN Architecture and also analyzed and compared model performance by using metrics like MSE and R2.
4. **Sravanthi:** Sravathi developed Scripts for exploratory data analysis (EDA) and future engineering. She created visualizations such as Heatmaps, pair plots and also charts for EDA and covered the code implementation. She also worked with Sadwika in report writing.

Problem Definition

Traffic accidents result in significant human and economic losses worldwide, making the prediction and mitigation of crash severity a pressing concern. The severity of a car crash, ranging from minor injuries to fatalities, is influenced by various factors such as driver behavior, road conditions, weather, vehicle characteristics, and traffic patterns. Despite the availability of rich datasets containing crash-related information, the inherent complexity and nonlinear relationships between these factors pose challenges for traditional analytical methods. The problem addressed in this study is twofold:

1. **Prediction of Crash Severity:** To accurately predict the severity of car crashes based on key contributing factors using advanced machine learning models.
2. **Identification of Key Predictors:** To determine the most significant variables that influence crash severity, enabling stakeholders to design effective interventions for reducing crash impacts.

The ultimate goal is to create robust, interpretable, and generalizable models that not only predict crash severity with high accuracy but also provide actionable insights for road safety policy-making, insurance risk assessment, and preventative measures.



Literature Review

Overview of Existing Research

Machine learning (ML) and deep learning (DL) techniques have gained increasing attention for crash severity prediction due to their ability to handle large-scale data and uncover nonlinear relationships. Researchers have explored a variety of algorithms and methodologies to improve predictive accuracy and understand the dynamics of crash severity. This section reviews six key studies that highlight recent advancements in this domain:

1. **Transparent Deep Learning Framework (Sattar et al., 2023)**
   This study proposed a transparent deep learning framework to predict traffic crash severity, emphasizing interpretability. By combining traditional deep learning models with explainability techniques, the study demonstrated improved trust and usability of predictive models in decision-making contexts. The research highlighted the importance of model transparency in facilitating the adoption of ML methods in real-world applications.
2. **Deep Learning Framework for Injury Severity (Ma et al., 2021)**
   Ma and colleagues developed a deep learning framework to predict traffic accident injury severity by analyzing contributing factors. Their work highlighted the superiority of DL models over traditional statistical methods in capturing complex relationships

between variables, such as road conditions and driver demographics. However, they noted that data preprocessing and feature selection significantly impact model performance.

3. **Crash Severity Analysis of Rear-End Crashes (Ahmadi et al., 2020)** This study analyzed rear-end crashes in California using statistical and machine learning methods. The authors compared classification techniques, including decision trees and support vector machines, to identify critical factors influencing crash severity. The research emphasized the value of combining statistical and ML approaches for comprehensive insights.
4. **Prediction in Suburban Accidents (Ghasedi et al., 2021)** Ghasedi and colleagues employed a hybrid approach integrating logit models, factor analysis, and ML techniques to predict crash severity in suburban settings. Their study focused on developing countries, providing valuable insights into how socioeconomic and infrastructural factors contribute to crash outcomes. This research emphasized the role of context-specific modeling for improving prediction accuracy.
5. **Clustering-Enhanced ML Protocol (Assi et al., 2020)** Assi et al. introduced a clustering-based machine learning protocol to predict crash injury severity. The study demonstrated how combining clustering techniques with ML algorithms could improve classification performance by identifying hidden patterns in data. The results highlighted the effectiveness of synergizing clustering with predictive modeling for crash severity analysis.
6. **Crash Severity Prediction in Yemen (Al-Moqri et al., 2020)** This study investigated crash severity prediction using ML algorithms in Yemen, leveraging hospital case studies. The authors emphasized the challenges posed by data limitations and the need for robust preprocessing to handle missing or imbalanced datasets. Their findings suggested that ML techniques could significantly improve prediction outcomes even in resource-constrained environments.

Research Gaps Identified

- **Lack of Comparative Analysis:** Many studies focus on a single machine learning or deep learning algorithm, limiting the understanding of their relative effectiveness.
- **Limited Application of Neural Networks for Tabular Data:** While deep learning methods like CNNs have shown promise, their application to tabular crash data remains underexplored.
- **Transparency and Interpretability:** Despite advancements, a significant gap exists in ensuring that ML models are interpretable and actionable for stakeholders.
- **Diverse Contexts:** The majority of studies focus on specific regions or crash types, reducing generalizability to other contexts or datasets.

Relevance to Current Study

This study aims to address the identified gaps by:

1. Comparing the performance of multiple ML models, including Linear Regression, Random Forest, Gradient Boosting, and CNNs, for crash severity prediction.
2. Exploring the application of CNNs to tabular crash data and evaluating their performance against traditional methods.
3. Prioritizing feature selection and model interpretability to enhance usability in policy-making and insurance.
4. Using a comprehensive dataset that incorporates diverse crash-related variables for broader generalizability.

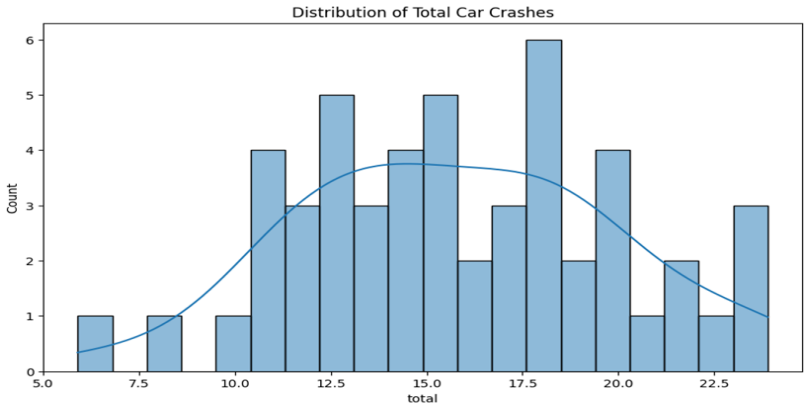
Dataset Description

The dataset contains crash-related data for 51 entries, likely corresponding to U.S. states, with various features capturing factors influencing car crash severity and associated insurance details. Below is a detailed description of each column:

Column Name	Description	Data Type
total	Total number of car crash incidents reported in the respective region.	float64
speeding	Proportion of crashes attributed to speeding as a contributing factor.	float64
alcohol	Proportion of crashes caused by alcohol impairment.	float64
not_distracted	Proportion of crashes where the driver was not distracted.	float64
no_previous	Proportion of crashes involving drivers with no prior crash records.	float64
ins_premium	Average annual insurance premium (in dollars) for the respective region.	float64
ins_losses	Average insurance losses (in dollars) related to crashes in the respective region.	float64
abbrev	State abbreviation for the corresponding entry.	object

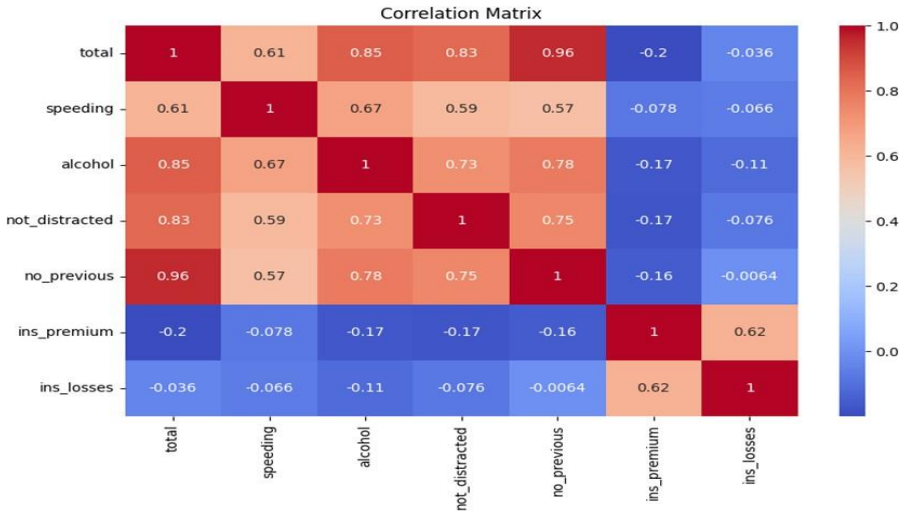
Dataset Properties

1. **Total Entries:** The dataset contains 51 rows, corresponding to entries from each U.S. state and the District of Columbia.
2. **Data Completeness:** All columns contain non-null values, ensuring no missing data for the given attributes.
3. **Data Distribution:**
    - **Numeric Columns:** Features like total, speeding, alcohol, etc., are represented as floating-point numbers, facilitating numerical analysis.
    - **Categorical Columns:** The abbrev column contains state abbreviations, allowing geographic grouping or analysis.
4. **Feature Characteristics:**
    - Proportional values (e.g., speeding, alcohol, not_distracted) provide normalized measures, likely expressed as percentages or ratios.
    - Financial features (ins_premium, ins_losses) offer insight into economic impacts tied to crash severity.



Proposed Algorithms and Applications

This study evaluates and compares four machine learning models for predicting car crash severity. Each algorithm has been selected based on its suitability for handling crash datasets, interpretability, and predictive capability:



1. **Linear Regression**
    - A baseline model to establish fundamental relationships between crash severity and contributing factors.
    - Advantages: Simple and interpretable; effective for linear relationships.
    - Limitations: Struggles with complex, nonlinear data.
2. **Random Forest Regressor**
    - An ensemble learning technique that builds multiple decision trees and aggregates their outputs for prediction.
    - Advantages: Handles nonlinear relationships well and provides feature importance insights.
    - Limitations: Computationally intensive for large datasets.
3. **Gradient Boosting Regressor**
    - An advanced ensemble method that builds trees sequentially to minimize errors from prior trees.
    - Advantages: High predictive accuracy and robust to overfitting with proper tuning.
    - Limitations: Requires careful hyperparameter tuning.
4. **Convolutional Neural Networks (CNN)**
    - Adapted for tabular data by reshaping inputs into formats suitable for convolutional layers.
    - Advantages: Captures complex patterns and interactions in data.
    - Limitations: Computationally intensive and less interpretable for tabular data.

```
Model: "sequential_3"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1d_3 (Conv1D)           (None, 5, 64)             192

 dropout_3 (Dropout)         (None, 5, 64)             0

 flatten_3 (Flatten)         (None, 320)               0

 dense_6 (Dense)             (None, 64)                20544

 dense_7 (Dense)             (None, 1)                 65

=================================================================
Total params: 20801 (81.25 KB)
Trainable params: 20801 (81.25 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
CNN MSE: 5.45133579695479
CNN R2 Score: 0.6985872640141062
```

Applications

The findings of this research have broad applications, including:

- **Road Safety Policy:** Identifying critical factors to guide infrastructure improvements and safety campaigns.
- **Insurance Risk Assessment:** Classifying drivers or scenarios based on predicted crash severity to adjust premiums or offer preventive recommendations.
- **Emergency Services Planning:** Anticipating crash severity to allocate resources effectively and optimize response times.

- **Vehicle Design:** Informing manufacturers about crash dynamics to enhance vehicle safety features.

Implementation and Experimental Results

Implementation Steps

The implementation of the proposed algorithms involved the following steps:

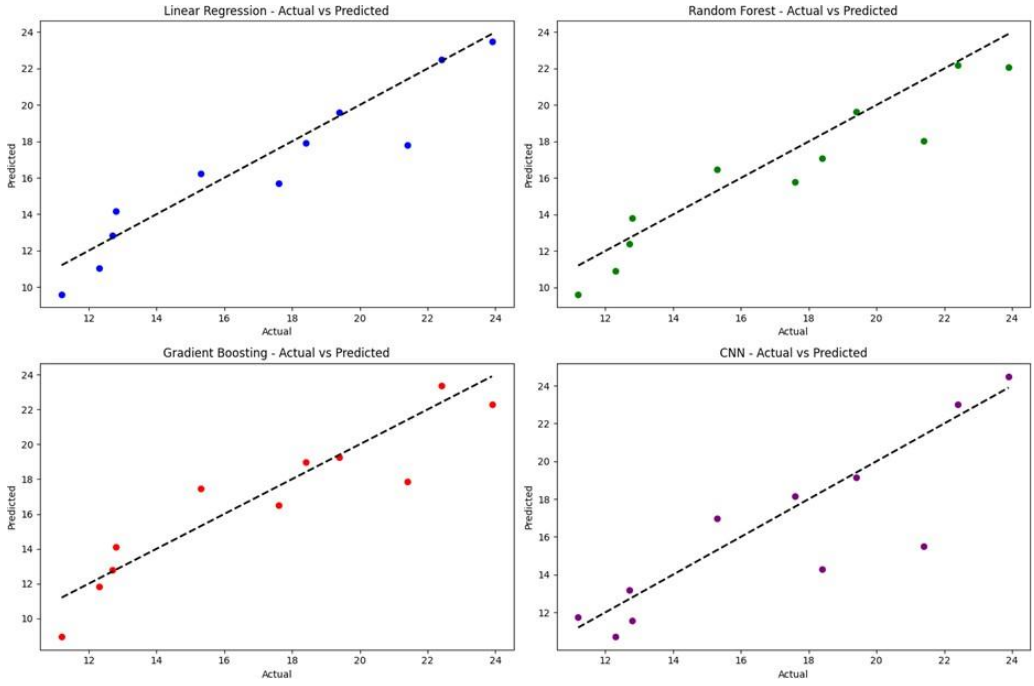
1. **Data Loading and Exploration**
    - The dataset was loaded using the seaborn library for initial inspection.
    - Key variables were identified through descriptive statistics and visualizations, including pairplots, heatmaps, and histograms.
2. **Feature Selection**
    - Features were selected based on exploratory analysis and domain knowledge.
    - Redundant or weakly correlated variables were removed to improve model efficiency.
3. **Data Preprocessing**
    - **Scaling:** Standardization ensured all features were on the same scale.
    - **Splitting:** Data was divided into training (70%) and testing (30%) sets.
4. **Model Implementation**
    - Each algorithm was implemented using Python libraries such as scikit-learn for regression models and TensorFlow/Keras for CNN.
    - Hyperparameters were optimized using grid search and cross-validation techniques.
    - CNN inputs were reshaped to match the requirements of convolutional layers.
5. **Performance Evaluation**
    - Models were evaluated using:
        - **Mean Squared Error (MSE):** Measures prediction error.
        - **R-squared (R^2) Score:** Indicates the proportion of variance explained by the model.
6. **Visualization**
    - Scatter plots of actual vs. predicted values.
    - Comparative bar plots for MSE and R^2 scores.

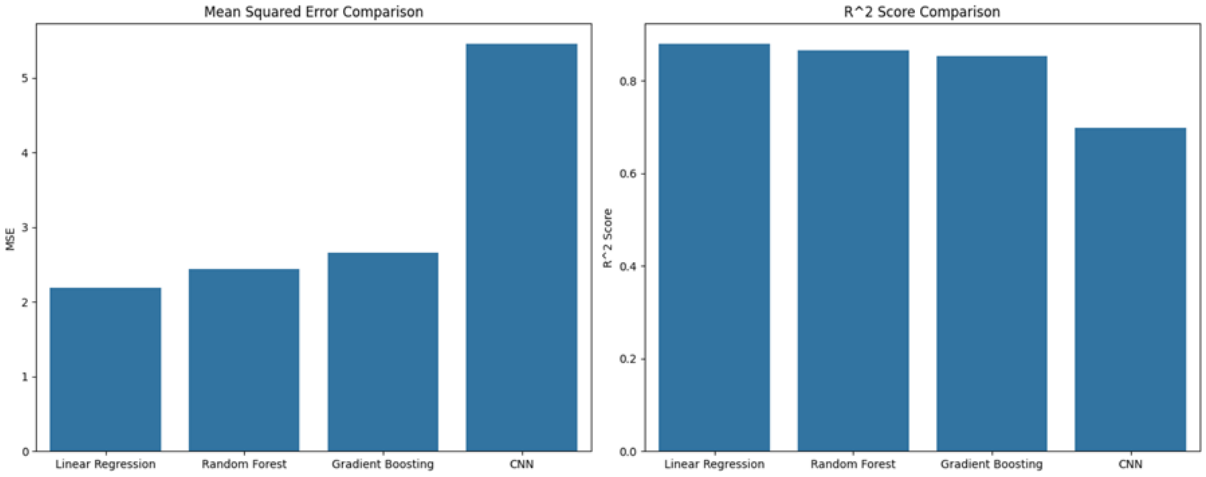
Experimental Results

The following results were obtained from the models:

Model	**Mean Squared Error (MSE)**	**R-squared (R^2) Score**
Linear Regression	2.1862	0.8791

Random Forest Regressor	2.4369	0.8653
Gradient Boosting Regressor	2.6537	0.8533
Convolutional Neural Network (CNN)	5.4513	0.6986





Key Observations

1. **Linear Regression:**
   - Outperformed Random Forest and Gradient Boosting in terms of MSE and R^2, making it a strong baseline model for linear relationships.
   - Limitations: Struggled with more complex patterns present in the dataset.

2. **Random Forest Regressor:**
    - Provided slightly lower R^2 than Linear Regression but offered greater interpretability via feature importance scores.
    - Handled nonlinear relationships effectively.
3. **Gradient Boosting Regressor:**
    - Achieved comparable performance to Random Forest but required more tuning and computational resources.
    - Effective for capturing intricate patterns but sensitive to overfitting without proper regularization.
4. **Convolutional Neural Network (CNN):**
    - Underperformed relative to traditional ML models in this context.
    - Demonstrated potential for future use with better preprocessing and data augmentation techniques.

The experimental results confirm that traditional machine learning methods like Linear Regression and Random Forest are better suited for tabular datasets in crash severity prediction compared to CNN. The study demonstrates the potential of combining feature interpretability with predictive accuracy to support real-world applications in road safety and insurance.

Advantages and Drawbacks

Advantages

1. **Comprehensive Comparison of Algorithms:** Evaluating four different machine learning models allows for a holistic understanding of their performance on crash severity prediction.
2. **Actionable Insights:** The identification of key contributing factors aids in policy-making, insurance risk assessment, and road safety interventions.
3. **Performance Metrics:** Detailed evaluation using MSE and R^2 ensures rigorous performance benchmarking.
4. **Versatility of Models:** Traditional models like Random Forest provided robust predictions with high interpretability, while CNN highlighted the potential of neural networks in this domain.
5. **Scalability:** The models are adaptable to larger datasets or different geographical regions with minor adjustments.

Drawbacks

1. **CNN Underperformance:** Convolutional Neural Networks struggled with tabular data due to the lack of spatial relationships inherent in image data.
2. **Computational Complexity:** Gradient Boosting and CNN require significant computational resources for training and optimization.
3. **Model Interpretability:** Advanced models like CNN are less interpretable compared to simpler ones like Linear Regression, limiting their practical use for policy-makers.

4. **Data Quality Dependency:** Performance heavily relies on the quality of input data, requiring substantial preprocessing efforts to handle missing or noisy data.

Summary and Conclusion

New Applications and Discoveries

This study explored machine learning techniques to predict car crash severity, comparing the performance of Linear Regression, Random Forest, Gradient Boosting, and Convolutional Neural Networks. The findings contribute to the understanding of crash severity prediction and have several new applications:

1. **Road Safety Policy Development:** Insights into key predictors, such as weather conditions and road design, allow authorities to implement targeted safety measures.
2. **Insurance Risk Assessment:** Predictive models help insurers adjust premiums based on risk profiles, encouraging safer driving behavior.
3. **Emergency Response Optimization:** Severity predictions can guide resource allocation and response planning in high-risk areas.
4. **Advanced Analytics for Developing Countries:** The study highlights the potential of ML in regions with limited resources, aiding in the global effort to reduce crash-related fatalities.

Discoveries

- **Linear Regression as a Strong Baseline:** Despite being a simple model, it achieved high accuracy, showcasing the importance of starting with interpretable methods.
- **Random Forest's Robustness:** The model balanced accuracy and interpretability, making it ideal for practical applications.
- **CNN Limitations:** Neural networks require substantial modifications to handle tabular data effectively, underlining the importance of domain-specific preprocessing.

Overall, the study emphasizes the importance of selecting appropriate models based on dataset characteristics and application needs. Future research could explore hybrid approaches, such as combining ensemble techniques with neural networks, to further enhance predictive performance.

References

1. Sattar, K., Chikh Oughali, F., Assi, K., Ratrout, N., Jamal, A., & Masiur Rahman, S. (2023). Transparent deep machine learning framework for predicting traffic crash severity. *Neural Computing and Applications, 35*(2), 1535-1547.
2. Ma, Z., Mei, G., & Cuomo, S. (2021). An analytic framework using deep learning for prediction of traffic accident injury severity based on contributing factors. *Accident Analysis & Prevention, 160*, 106322.

3. Ahmadi, A., Jahangiri, A., Berardi, V., & Machiani, S. G. (2020). Crash severity analysis of rear-end crashes in California using statistical and machine learning classification methods. *Journal of Transportation Safety & Security, 12*(4), 522-546.
4. Ghasedi, M., Sarfjoo, M., & Bargegol, I. (2021). Prediction and analysis of the severity and number of suburban accidents using logit model, factor analysis, and machine learning: A case study in a developing country. *SN Applied Sciences, 3*(1), 13.
5. Assi, K., Rahman, S. M., Mansoor, U., & Ratrout, N. (2020). Predicting crash injury severity with machine learning algorithm synergized with clustering technique: A promising protocol. *International Journal of Environmental Research and Public Health, 17*(15), 5497.
6. Al-Moqri, T., Haijun, X., Namahoro, J. P., Alfalahi, E. N., & Alwesabi, I. (2020). Exploiting machine learning algorithms for predicting crash injury severity in Yemen: hospital case study. *Applied Computational Mathematics, 9(5), 155-164.*