Analyzing US Accidents for Patterns and Trends Using Big Data Techniques

Final Project Report

Data Stormers (Group 09)

Rishi Kumar Reddy Sangireddy
rsang017@ucr.edu
University of California, Riverside
Computer Science
Riverside, CA, USA

Rithvik Vukka
rvukka002@ucr.edu
University of California, Riverside
Computer Science
Riverside, CA, USA

Venkata Ranga Sai Vishal Matcha
vmatch002@ucr.edu
University of California, Riverside
Computer Science
Riverside, CA, USA

Yash Bhalgat
ybhal001@ucr.edu
University of California, Riverside
Computer Science
Riverside, CA, USA

Venkata Sai Saketh Vaigandla
vvaig001@ucr.edu
University of California, Riverside
Computer Science
Riverside, CA, USA

ABSTRACT

Car accidents are a major public safety concern in the United States, resulting in significant loss of life, injuries, and economic damage. With the growing availability of real-time traffic and accident data, there is an opportunity to leverage big data techniques to analyze patterns, predict hotspots, and identify underlying causes of accidents.

This project utilizes the U.S. Accidents Dataset (2016–2023), containing over 7.7 million records, to analyze accident trends across the country. By employing distributed computing frameworks like Hadoop and Apache Spark, the project addresses the challenges of processing large-scale data efficiently.

The primary outcomes include the development of a predictive model to forecast accident likelihood based on key factors such as location, weather, and time of day, alongside the creation of an interactive dashboard to visualize accident hotspots and trends.

These insights aim to support data-driven interventions for improving road safety and traffic management strategies, highlighting the transformative potential of big data analytics in public safety applications.

1 INTRODUCTION

Car accidents pose a significant threat to public safety in the United States, leading to loss of life, injuries, and substantial economic impact. The increasing availability of real-time traffic and accident data offers an opportunity to identify accident hotspots, analyze contributing factors, and predict accident likelihood using big data techniques. This project aims to leverage these insights to improve road safety and optimize traffic management.

1.1 Motivation

The motivation for this project stems from the need to enhance road safety by identifying accident-prone areas and understanding the factors contributing to high-risk incidents. By utilizing advanced big data tools, the project seeks to provide actionable insights for policymakers, traffic authorities, and the public to mitigate accidents and manage traffic flow more effectively.

1.2 Problem Statement

Analyzing millions of accident records manually is infeasible due to the sheer volume, velocity, and variety of data. Current approaches lack the scalability and predictive capability to identify trends and patterns efficiently. This project addresses these challenges by employing distributed computing frameworks and machine learning techniques to process large datasets, predict accidents, and visualize actionable insights in real time.

2 LITERATURE SURVEY

The literature survey for the *Analyzing U.S. Accidents for Patterns and Trends Using Big Data Techniques* project categorizes key studies into User Interface and Visualization, Predictive Analytics, Feature Selection, and Data Integration. Each section provides a detailed analysis of relevant studies.

2.1 UI/Visualization

2.1.1 ***Optimization Analysis of Urban Function Regional Planning Based on Big Data and GIS Technology***. leverages GIS technology for spatial planning and optimization. The study focuses on guiding urban function planning and optimizing land use through GIS-based visualization techniques. It highlights how visualization aids in interpreting spatial data patterns, such as in traffic analysis projects.

2.1.2 Data Mining and Visualization to Understand Accident-Prone Areas.

combines data mining and intuitive visualization to investigate accident-prone areas and periods. The study emphasizes making trends accessible to non-specialists, thus enabling evidence-based traffic safety interventions and public policy development.

2.1.3 GIS and Big Data Visualization.

examines GIS advancements via big data visualization technologies. The research outlines the pipeline from data collection to analysis and visualization, evaluating tools like ArcGIS, Tableau, and Google Earth. The study also contrasts traditional GIS methods with enhanced visualization techniques enabled by big data.

2.1.4 Modelling road congestion using ontologies for big data analytics in smart cities .

The lack of contextual information in Intelligent Transportation Systems (ITS) is a problem. Google Maps, for instance, offers real-time traffic updates with color-coded speeds that show quick or slow movement, but it is devoid of information on actual speed, daily averages, or reasons for delays. By forming clusters that either allow or restrict the ability to estimate journey times and differentiate between weekday and weekend trends, this paper proposes that clustering an unsupervised dataset could aid in filling this gap. The data was divided into five clusters using K-Means++, which categorized the journey times from extremely high to extremely low. The algorithm could successfully distinguish between weekend and weekday traffic patterns based on journey time distribution by further grouping journey times by time and day of the week.

2.1.5 Big Data Visualization Tools: A Survey the New Paradigms, Methodologies and Tools for Large Data Sets Visualization .

36 data visualization technologies are surveyed in this research and categorized into four groups: business intelligence visualization tools, scientific visualization tools, graph and network visualization tools, and information visualization tools. It assesses every tool according to a variety of functional and non-functional characteristics before offering suggestions for further study and advancement. The paper's main conclusions highlight information and geographic data visualization technologies, such as Tableau, Sentinel Visualizer, and Polymaps.

2.1.6 Decision Support System for the Analysis of Traffic Accident Big Data .

This paper proposes a decision support system that integrates various data points, including environmental and road factors, to assess accident risks. By using data mining, the system can assist in identifying accident causation and providing predictive capabilities for traffic safety interventions. This research uses big data analytics to guide traffic safety decision-making, aligning with decision support systems. It shows how data integration and analysis support real-time accident prevention.

2.2 Feature selection

2.2.1 Improved Feature Selection Model for Big Data Analytics .

For improved feature selection, the cited paper explores a hybrid strategy that blends Particle Swarm Optimization (PSO) and Grey Wolf Optimization (GWO). To get the best answers, this method combines elements such as Euclidean separation matrices and K-nearest neighbors. Cross-validation aids in preventing data overfitting, and a sigmoid function transforms the continuous search space into a binary format for feature selection. A selected feature ratio of 196 out of 773 features, as opposed to 393 and 336 features for GWO and PSO separately, is one of the main findings. With a classification accuracy of 90%, this approach outperformed competing algorithms with scores of 86.8% and 81.6%. Additionally, it outperformed standalone GWO (272 seconds) and PSO (245.6 seconds) by cutting the overall processing time to 184.3 seconds.

2.2.2 ***Enhancing Big Data Feature Selection Using a Hybrid Correlation-Based Feature Selection*** . This paper introduces a hybrid method called CFS-DRSA, which combines Correlation-Based Feature Selection (CFS) with the Dominance-Based Rough Set Approach (DRSA). This approach aims to optimize big data analysis by first using CFS to reduce uncorrelated attributes with the best-first search (BFS) algorithm and then applying DRSA to handle and eliminate uncertain or inconsistent data. The proposed technique enhances classifier performance by streamlining feature sets, reducing computational complexity, and maintaining classification accuracy.

2.3 Data Collection

2.3.1 ***Big Data Visualization Tools: A Survey*** . This paper reviews current tools and techniques used for visualizing large datasets. It discusses the strategic importance of data visualization as part of the big data lifecycle, highlighting the 3Vs of big data: Volume, Variety, and Velocity. The authors categorize various visualization tools based on their scope, software type, graphical capabilities, and scalability, emphasizing tools that integrate analysis and visualization. The survey aids users in selecting suitable tools for effectively presenting and analyzing large-scale data.

2.3.2 ***Big Data Analytics in Inteligent Transportation System: A Survey*** . By optimizing both cars and transportation infrastructure, Intelligent Transportation Systems (ITS) seek to improve services for both drivers and passengers. GPS, smart cards (used in automated fare collection for urban transit), image detection systems, sensors, connected and autonomous vehicles (CAV), and vehicle ad-hoc networks (VANET) are just a few of the sources of data that ITS collects and uses. Modern developments in facilitating communication between autonomous vehicles and infrastructure are represented by CAV and VANET. In particular, VANET, a kind of mobile ad hoc network, increases network coverage by treating infrastructure and automobiles as communication nodes. Through the use of big data analytics, ITS may improve efficiency and performance. The enormous volumes of data gathered by ITS can be processed more easily thanks to big data platforms like Apache Hadoop and Spark, which greatly increases system efficiency.

2.3.3 ***Road Traffic Crash Data: An Overview on Sources, Problems, and Collection Methods*** . This review of traffic accident data collection methods categorizes them into intrusive and non-intrusive techniques. It discusses the challenges associated with each method, including under-reporting and data accuracy, and offers recommendations for enhancing data reliability. The paper underscores the significance of accurate data collection in comprehending traffic operational issues and devising effective safety programs. The insights gained from this study are important in selecting suitable data collection strategies for traffic accident analysis.

2.4 Tools

2.4.1 ***A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools*** . This paper focuses on the challenges, open research issues, and tools involved in big data analytics, specifically discussing tools like Hadoop for managing large volumes of structured and unstructured data. Since analyzing accident patterns and trends requires robust tools to handle extensive datasets, this survey provides insights into essential big data tools and technologies that can support efficient data processing and analysis. The paper serves as a foundational guide to understanding the tools and challenges in big data analytics, which are critical for implementing a scalable, effective analysis framework in traffic accident research.

2.5 Identical Works

2.5.1 ***Implementation of Heat Maps in Geographical Information System – Exploratory Study on Traffic Accident Data.*** In order to lay the groundwork for the use of heat maps as a GIS visualization tool, this study examines a number of studies. By varying settings such as color range, kernel size, radius, and transparency, it analyzes two primary user groups: the general public and cartography professionals. With an ideal transparency level of about 50%, the results show a predilection for striking, contrasting hues like red. The article ends with thorough suggestions and detailed information on how to apply GIS effectively, what kinds of data are appropriate, and how to use heat maps most effectively.

2.5.2 ***Visualizing Traffic Accident Hotspots Based on Spatial-Temporal Network Kernel Density Estimation*** . This study addresses methods for locating hotspots in network spaces, such as network kernel density estimation and significant linear route recognition, but points out that these methods are not very good at capturing the temporal dynamics of these hotspots. The study addresses this by presenting a novel technique that combines spatial and temporal dimensions: Spatial-Temporal Network Kernel Density Estimation (STNKDE). To illustrate this strategy, a prototype system was created that offered a dynamic depiction of New York City traffic accident hotspots for 2017.

3 METHODOLOGY

This section details the methodology adopted in developing an analysis dashboard for accident trends using big data techniques. The methodology encompasses data description, preparation, ingestion, processing, and visualization strategies to derive actionable insights from accident data.

3.1 Data Description

3.1.1 U.S. Accidents Dataset (2016–2023).

- **Source**: Kaggle
- **Description**: Contains approximately 7.7 million accident records across 49 states in the United States, with detailed attributes that enable a comprehensive analysis of accident patterns.
- **Key Attributes**:
  - Accident severity levels
  - Geographic location (latitude and longitude)
  - Weather conditions
  - Traffic signal status
  - Time of day

3.1.2 Auxiliary Datasets.

- **Real-time Traffic Feeds**: Integrated from APIs for up-to-date traffic conditions.
- **Weather Data**: Enriched the dataset using external weather sources to analyze weather-based correlations with accidents.

3.2 Data Cleaning and Preparation

3.2.1 Data Cleaning.

- **Handling Missing Values**: Addressed incomplete records (e.g., weather or location data) by interpolation or exclusion, ensuring data reliability.
- **Standardizing Formats**: Unified inconsistent formats for date, time, and geographic attributes to ensure consistency across records.
- **Noise Filtering**: Identified and removed anomalies, such as duplicate entries and erroneous values in severity or weather data, using statistical and spatial techniques.

3.2.2 *Feature Selection.*

- Selected critical attributes, including weather, location, time of day, and road conditions, while discarding redundant fields.
- Utilized a hybrid feature selection approach combining **Correlation-Based Feature Selection (CFS)** and **Particle Swarm Optimization (PSO)** to ensure computational efficiency.

3.2.3 *Data Transformation.*

- **Spatial Mapping**: Standardized geographic data to a consistent Coordinate Reference System (CRS) for accurate spatial analysis.
- **Aggregation**: Grouped records at granular levels (e.g., city or county) for trend visualization and predictive modeling.

3.3 Data Ingestion and Retrieval

3.3.1 *Writing Methods.*

- Stored datasets in a **Hadoop Distributed File System (HDFS)** to ensure replication and fault tolerance.
- Partitioned data by state and time for efficient querying during analysis.

3.3.2 *Access Methods.*

- Designed query mechanisms using **Apache Spark**, enabling fast retrieval of large-scale data.
- Implemented caching for frequently accessed records to minimize latency and improve system performance.

3.4 Big Data Processing Techniques

3.4.1 *Indexing.*

- Created spatial indexes (e.g., R-trees) to optimize geographic queries.
- Grouped accident records based on proximity to identify accident hotspots.

3.4.2 *Partitioning.*

- Partitioned data by state, city, and road type using **MapReduce** for parallel processing.
- Applied spatiotemporal partitioning to dynamically adapt to changes in data volume and improve processing efficiency.

3.5 Core Methodology

3.5.1 *Accident Hotspot Mapping.*

- Utilized **GIS-based tools** (e.g., QGIS and GeoSpark) for spatial mapping of accidents.

- Applied spatial functions such as `ST_Intersects` and `ST_Contains` to analyze the relationships between accident locations and road attributes.

3.5.2 Predictive Modeling.

- Built machine learning models (**Random Forest** and **Gradient Boosting**) to predict the likelihood of accidents based on weather, traffic, and time features.
- Evaluated the model's performance using metrics like **precision**, **recall**, and **F1-score**.

3.6 Visualization

The visualization component plays a crucial role in presenting the findings and insights derived from the analysis. This section highlights key visualizations that demonstrate accident trends and hotspots across the United States.

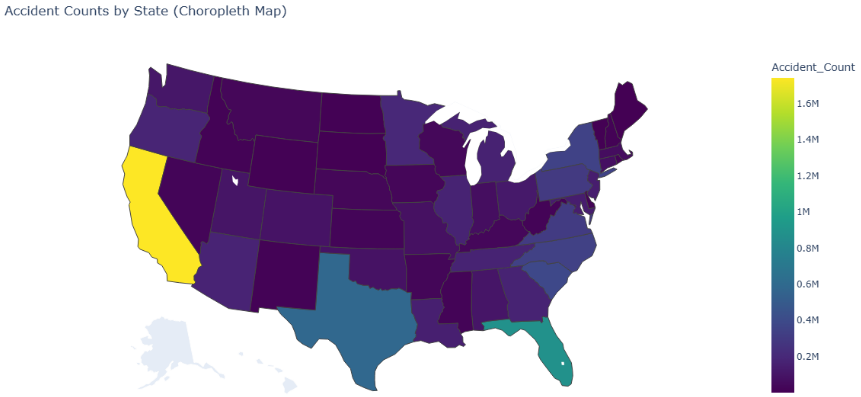

Fig. 1. Accident Counts by State (Choropleth Map). This map visually represents the number of accidents in each U.S. state using a color gradient. States with higher accident counts are highlighted in brighter colors, indicating hotspots such as California and Florida.

The visualizations were designed to be intuitive and informative, helping stakeholders understand accident patterns at both spatial and temporal levels:

- The **Choropleth Map** (Figure 1) is ideal for identifying regional hotspots and provides a quick overview of accident density across states.
- The **Yearly Accident Trends** (Figure 2) help in identifying temporal trends and possible correlations with external factors such as population growth or weather.
- The **State-Level Accident Counts** (Figure 3) offer an in-depth look at which states contribute the most to national accident statistics, guiding targeted interventions.

These visualizations were implemented using advanced data visualization tools to ensure clarity and accuracy, making them accessible to policymakers and the general public.

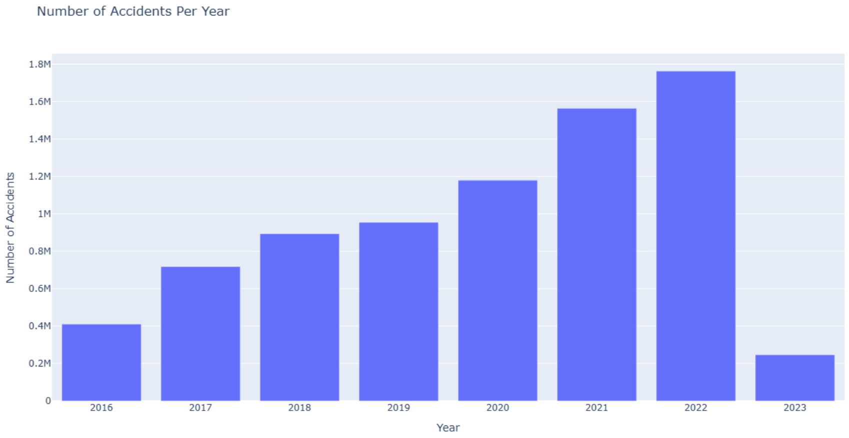

Fig. 2. Number of Accidents Per Year. This bar chart shows the annual distribution of accidents from 2016 to 2023. A significant increase is observed between 2016 and 2022, reflecting either improved reporting mechanisms or a real rise in incidents.

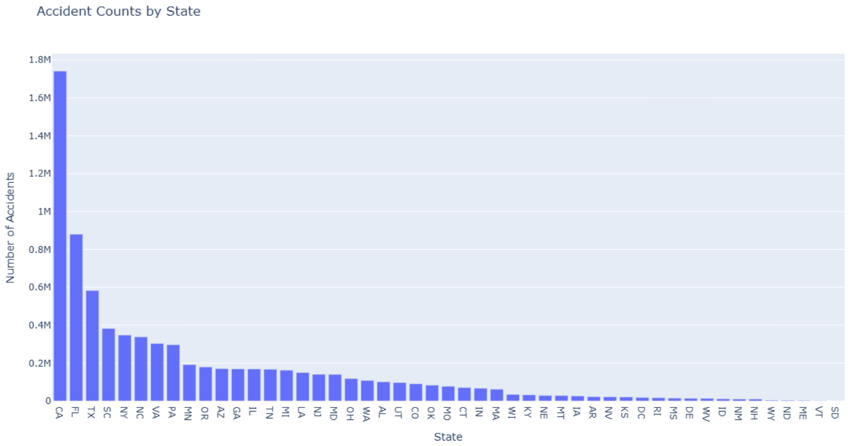

Fig. 3. Accident Counts by State. This bar chart provides a detailed view of accident counts by state, allowing for a granular analysis of which states experience the most accidents. California, Florida, and Texas emerge as the top contributors.

4 EVALUATION

4.1 Model Validation

4.1.1 Model Overview. The predictive model used in this study is a multi-class classification algorithm implemented using a Random Forest Classifier. The model aims to predict *severity of accidents* (target variable) based on a diverse set of features, including *latitude, longitude, distance, temperature, humidity, visibility, wind speed, precipitation, weather conditions, and traffic-related factors.* These features were selected to capture environmental, geographical, and situational factors influencing accidents.

4.1.2 Model Performance Metrics. The performance of the predictive model was evaluated using precision, recall, F1-score, and ROC-AUC. A classification report (Figure 4) was generated to measure the accuracy for each class.

- **Classification Report:**
  - **Precision:** Class 2 achieved the highest precision (0.83), followed by Class 3 (0.75). Class 4 had the lowest precision (0.40).
  - **Recall:** Similarly, Class 2 demonstrated the highest recall (0.83), while Class 4 had the lowest (0.05).
  - **F1-Score:** Class 2 achieved the best balance of precision and recall, reflected in its highest F1-score (0.83). The overall macro average F1-score was 0.59.
- **Insights:**
  - The model performs well for the majority of the classes but struggles with Class 4, as reflected by its low recall and F1-score. This indicates a need for further optimization in handling minority or imbalanced classes.

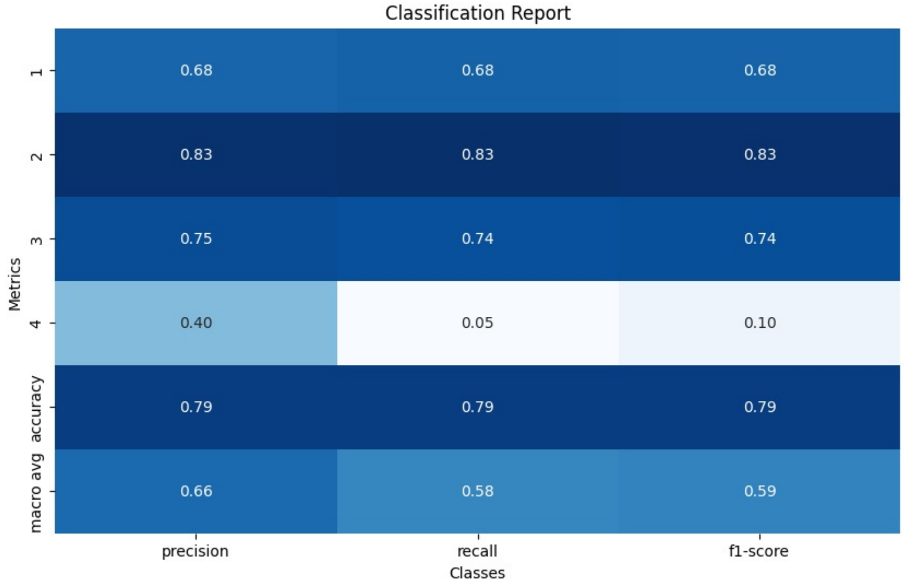

Fig. 4. Classification Report. The heatmap shows precision, recall, and F1-scores for each severity class. Class 2 performs best overall, while Class 4 struggles significantly, highlighting class imbalance challenges.

4.1.3 ROC-AUC Curve. The multi-class ROC-AUC curve (Figure 5) was plotted to assess the model's ability to distinguish between different classes. The ROC-AUC score was calculated using the One-vs-Rest (OvR) method.

- **Observations:**
  - The curves for Class 1 and Class 2 are closest to the top-left corner, indicating excellent discriminative power.
  - Class 4's curve demonstrates weaker performance, reflecting its low recall and precision metrics.
- **Insights:**
  - The weighted ROC-AUC score of 0.79 highlights the model's overall effectiveness but also points to the need for improvements in minority class predictions.

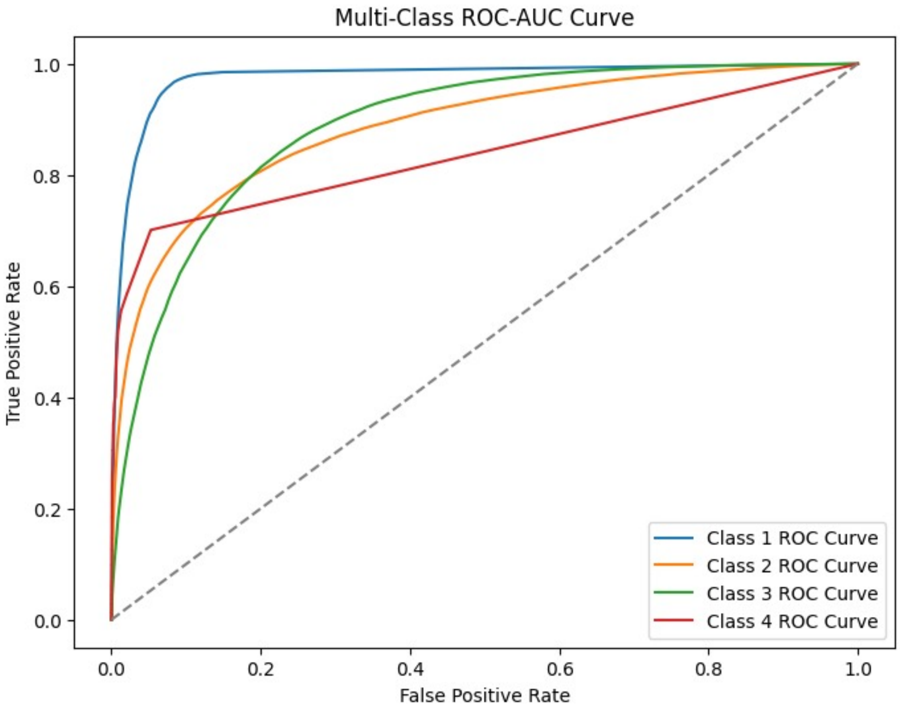

Fig. 5. Multi-Class ROC-AUC Curve. The ROC curves demonstrate the model's ability to distinguish between different severity classes. Class 1 and Class 2 perform well, but Class 4 shows weaker results, indicating challenges in minority class predictions.

4.2 Scalability Assessment

To evaluate the scalability of the system, we analyzed its performance across different dataset sizes by processing and training the model on varying numbers of rows. The system utilized a chunk-wise data processing approach, where chunks of data were read, preprocessed, and concatenated for training. The results were plotted in a graph to illustrate the execution time against the size of the dataset (Figure 6).

- **Observations:**
  - The execution time increased linearly with the size of the dataset, ranging from approximately 50 seconds for 100,000 rows to over 200 seconds for 600,000 rows.
  - This behavior suggests that the data processing pipeline and model training scale predictably with larger data volumes.
- **Insights:**
  - The linear trend demonstrates the model's ability to handle incremental increases in data without significant performance degradation.
  - The combination of efficient chunk-wise processing and Random Forest Classifier training proves to be a robust solution for handling large-scale datasets.

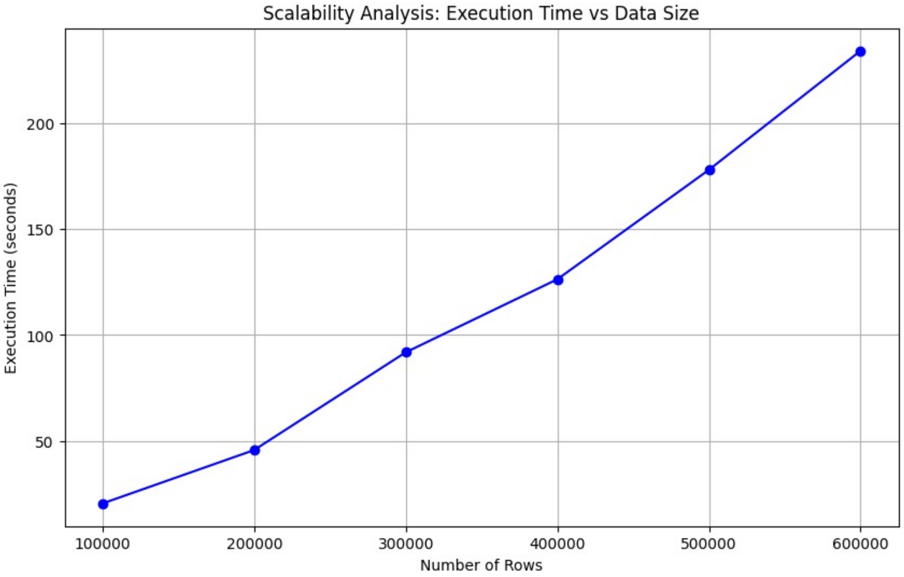

Fig. 6. Scalability Analysis: Execution Time vs. Data Size. The graph illustrates the execution time for varying dataset sizes. The linear increase indicates efficient scaling of the system, even with larger datasets.

4.3 Summary

The evaluation demonstrated the following:

- The system's scalability is robust and predictable for large datasets, making it suitable for real-world applications.
- The model achieved strong performance metrics for major classes but requires optimization for underrepresented ones.
- The ROC-AUC analysis confirms the model's capability to make reliable predictions for the majority of cases, though it highlights areas for further enhancement.

5 CONCLUSION AND FUTURE WORK

5.1 Conclusion

The project *"Analyzing U.S. Accidents for Patterns and Trends Using Big Data Techniques"* leverages large-scale accident datasets and big data technologies to identify accident hotspots, analyze contributing factors, and develop predictive models for accident likelihood. By integrating advanced data processing frameworks like Hadoop and Apache Spark, the project demonstrates the efficiency and scalability of distributed computing in traffic safety analysis. The outcome includes a comprehensive and interactive dashboard that provides actionable insights to improve traffic management and accident prevention strategies, contributing to public safety.

5.2 Future Work

The future scope of the project includes the following:

- **Refinement of Predictive Models:** Enhancing the model's accuracy and predictive power by incorporating additional features such as driver behavior data, vehicle-specific factors, and more granular weather data.
- **Integration of Real-Time Data:** Expanding the scope to process and analyze streaming data from traffic sensors, cameras, and other IoT devices for near real-time accident prediction.
- **Advanced Machine Learning Algorithms:** Exploring more sophisticated algorithms such as deep learning models for spatiotemporal data analysis to improve the prediction of accident hotspots.
- **Framework Comparison:** Evaluating the performance of various distributed computing frameworks and machine learning platforms to optimize data processing and model training.
- **Dashboard Enhancement:** Improving the dashboard's user interface and visualization capabilities to provide a more intuitive and user-friendly experience, making the insights accessible to a broader audience, including policymakers and public safety officials.
- **Cross-Domain Data Integration:** Integrating additional datasets such as economic impact reports, traffic fines, or insurance claims to enrich the analysis and provide a holistic view of traffic safety.

By addressing these areas, the project can continue to evolve as a comprehensive tool for traffic accident analysis and contribute to reducing accidents through data-driven solutions.

6 AUTHOR CONTRIBUTIONS

Rishi Kumar Reddy Sangireddy:

- Performed the evaluations and validation of results.
- Contributed to the development of evaluation codes and their implementation.
- Supported report preparation and team discussions.

Rithvik Vukka:

- Conducted feature engineering to optimize dataset usability.
- Played a significant role in data preprocessing and integration.
- Contributed to code debugging and testing.

Venkata Ranga Sai Vishal Matcha:

- Developed the dashboard and visualizations for presenting insights.
- Focused on enhancing the user interface and ensuring interactivity.
- Supported the integration of visualization components with the main framework.

Yash Bhalgat:

- Contributed to the preparation of presentations and final report.
- Played an active role in conducting the literature survey.
- Supported the debugging of visualization components.

Venkata Sai Saketh Vaigandla:

- Organized and coordinated team activities, ensuring project alignment.
- Contributed to the literature survey and documentation.
- Assisted in integrating various project components, including data processing.

REFERENCES

[1] L. Tang, Z. Cheng, J. Dai, H. Zhang, and Q. Chen. 2024. Joint Optimization of Vehicular Sensing and Vehicle Digital Twins Deployment for DT-Assisted IoVs. *IEEE Transactions on Vehicular Technology* 73, 8 (2024), 11834–11847.

[2] T. Lyu, L. Shi, and W. He. 2024. Network Evolution Analysis of Vehicle Road-Driving Behavior Strategies and Design of Information Guidance Algorithm. *Journal of Social Computing* 5, 1 (2024), 58–87.

[3] M. Zhao, S. Li, H. Wang, J. Yang, Y. Sun, and Y. Gu. 2024. MP2Net: Mask Propagation and Motion Prediction Network for Multiobject Tracking in Satellite Videos. *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024), 1–15.

[4] Z. Wang, Y. Tang, S. Song, H. Chen, X. Lu, and F. Liu. 2024. SI-AMC: Integrating DL-Based Scenario Identification into Adaptive Modulation and Coding in Vehicular Communications. In *Proceedings of the 2024 IEEE Wireless Communications and Networking Conference (WCNC)*, 1–6.

[5] H. Mu, N. Aljeri, and A. Boukerche. 2024. Spatio-Temporal Feature Engineering for Deep Learning Models in Traffic Flow Forecasting. *IEEE Access* 12 (2024), 76555–76578.

[6] S. Yun, Z. A. Bhuiyan, S. Shen, and Md. T. A. H. Sadi. 2024. FastFlow: Availability Aware Federated Attention-Based Spatial-Temporal GNN for Big Data-Driven Traffic Forecasting. In *Proceedings of the 2024 IEEE International Conference on Communications (ICC 2024)*, 2616–2621.

[7] J. Cai. 2017. Optimization Analysis of Urban Function Regional Planning Based on Big Data and GIS Technology. *Boletin Tecnico/Technical Bulletin* 55 (2017), 344–351.

[8] L. Zhu, F. R. Yu, Y. Wang, B. Ning, and T. Tang. 2019. Big Data Analytics in Intelligent Transportation Systems: A Survey. *IEEE Transactions on Intelligent Transportation Systems* 20, 1 (Jan. 2019), 383–398.

[9] I. M. El-Hasnony, S. I. Barakat, M. Elhoseny, and R. R. Mostafa. 2020. Improved Feature Selection Model for Big Data Analytics. *IEEE Access* 8 (2020), 66989–67004.

[10] E. G. Caldarola and A. Rinaldi. 2017. Big Data Visualization Tools: A Survey - The New Paradigms, Methodologies and Tools for Large Data Sets Visualization. In *Proceedings of the 6th International Conference on Data Science, Technology and Applications (DATA 2017)*, 296–305.

[11] R. Nétek, T. Pour, and R. Slezakova. 2018. Implementation of Heat Maps in Geographical Information System – Exploratory Study on Traffic Accident Data. *Open Geosciences* 10, 1 (2018), 367–384.

[12] L. Abberley, N. Gould, K. Crockett, and J. Cheng. 2017. Modelling Road Congestion Using Ontologies for Big Data Analytics in Smart Cities. In *Proceedings of the 2017 International Smart Cities Conference (ISC2)*, 1–6.

[13] Md. M. Rizvee, Md. Amiruzzaman, and Md. R. Islam. 2021. Data Mining and Visualization to Understand Accident-prone Areas. *arXiv preprint arXiv:2103.09062*.

[14] B. Romano and Z. Jiang. 2017. Visualizing Traffic Accident Hotspots Based on Spatial-Temporal Network Kernel Density Estimation. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 1–4.

[15] L. Stenneth, O. Wolfson, P. S. Yu, and B. Xu. 2011. Transportation Mode Detection Using Mobile Phones and GIS Information. In *Proceedings of the 19th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems*, 54–63.

[16] A. Emam and E. Abdullah. 2015. Traffic Accidents Analyzer Using Big Data. In *Proceedings of the 2015 International Conference on Computational Science and Computational Intelligence (CSCI)*, 10.

[17] T. Hussain, A. Sanga, and S. Mongia. 2019. Big Data Hadoop Tools and Technologies: A Review. *SSRN Electronic Journal* (2019).

[18] J.-H. Kim, S.-H. Lee, and S.-W. Lee. 2016. Highway Traffic Accident Prediction Using VDS Big Data Analysis. *The Journal of Supercomputing*.

[19] Y. Salih-Alj and M. El Hassouni. 2019. Decision Support System for the Analysis of Traffic Accident Big Data. *IEEE Access*.

[20] A. Abdulhafedh. 2017. Road Traffic Crash Data: An Overview on Sources, Problems, and Collection Methods. *Journal of Transportation Technologies* 7, 2 (2017), 206–219.

[21] M. Mohamad, A. Selamat, O. Krejcar, R. G. Crespo, E. Herrera-Viedma, and H. Fujita. 2021. Enhancing Big Data Feature Selection Using a Hybrid Correlation-Based Feature Selection. *Electronics* 10 (2021), 2984.