World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Advanced analytics for predicting traffic collision severity assessment

Mohammad Fokhrul Islam Buian [1], Ramisha Anan Arde [2], Md Masum Billah [3], Amit Debnath [3] and Iqtiar Md Siddique [4,*]

[1] *Department of Mechanical Engineering, Lamar University, Beaumont, Texas, US.*
[2] *Department of Computer Science and Engineering, Dhaka City College, Dhaka-6408, Bangladesh.*
[3] *Department of Electrical and Computer Engineering, Lamar University, Beaumont, Texas, US.*
[4] *Department of Industrial, manufacturing and Systems Engineering, University of Texas at El Paso, US.*

World Journal of Advanced Research and Reviews, 2024, 21(02), 2007–2018

Publication history: Received on 06 January 2024; revised on 27 February 2024; accepted on 29 February 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.21.2.0704

Abstract

Accurate prediction of accident risks plays a crucial role in proactively implementing safety measures and allocating resources effectively. This paper introduces an innovative approach aimed at improving accident risk prediction by harnessing unique data sources and extracting insights from diverse yet sparse datasets. Traditional models often face limitations due to a lack of diversity and scope in the available data, which hinders their predictive capabilities. In response to this challenge, our study integrates a broad spectrum of heterogeneous data, encompassing traffic flow, weather conditions, road infrastructure details, and historical accident records. To overcome the difficulties associated with sparse data, we employ advanced data science techniques such as feature engineering, imputation, and machine learning. The paper introduces a novel dataset that amalgamates various data types, establishing a robust foundation for our predictive model. Through meticulous analysis, we derive valuable insights from these diverse sources, significantly enhancing our ability to assess accident risks. The proposed approach offers numerous advantages, including the capacity to predict accidents in areas that were previously underrepresented and under varying conditions. We rigorously evaluate the model's performance through extensive experimentation and validate its accuracy using real-world accident data. Our results indicate substantial improvements in prediction accuracy compared to conventional models. This research contributes significantly to the field of accident risk prediction by highlighting the potential benefits of integrating heterogeneous sparse data and leveraging advanced data science techniques. The study underscores the importance of tapping into novel data sources and extracting concealed patterns and insights to promote safety and optimize resource allocation in accident-prone regions, fostering more secure environments.

Keywords: Accidents; Visualization; Machine Learning; Risk Assessment; Road Safety.

1. Introduction

Reducing traffic accidents is a critical global challenge, with 1.35 million annual fatalities worldwide. The limitations of traditional approaches stem from their reliance on structured and often siloed data sources. Accidents are multifaceted events influenced by a myriad of factors, including traffic conditions, weather, road infrastructure, and driver behavior. Predictive models based solely on historical accident data may struggle to capture the dynamic and interconnected nature of these factors, leading to suboptimal predictions. Moreover, sparse data often plague accident prediction efforts, particularly in regions with limited data collection infrastructure or for emerging types of accidents (e.g., those involving autonomous vehicles). It ranks as the eighth leading cause of death globally and is the leading cause of death for young people aged 5-29. Accidents claim more lives than HIV/AIDS. We explore US road accidents using the "US Accidents" dataset, aiming to uncover patterns, contributing factors, and safety insights. This analysis benefits public

* Corresponding author: Iqtiar Md Siddique

safety, transportation planning, and policy development, providing valuable insights for various stakeholders. Whether we are a data scientist or a road safety advocate, we can use this data-driven journey to understand and improve road safety in the US. Our research premise is that within these seemingly sparse and heterogeneous data lie valuable insights and correlations waiting to be discovered. We believe that the integration of diverse data sources, when coupled with advanced data science techniques, can illuminate the intricate relationships among various factors influencing accident occurrence. In this paper, we embark on a journey to unlock the potential of such data by proposing a novel accident risk prediction model. This paper unfolds in a structured manner, beginning with a comprehensive review of the state of the art in accident risk prediction models, highlighting their strengths and limitations. We then delve into the methodology section, where we introduce the novel dataset, we have compiled, detailing the types of data sources included and the challenges of working with sparse data. The methodology section also outlines the advanced data science techniques employed, providing readers with a clear understanding of our approach. Subsequently, we present the results of our experimentation, showcasing the improvements achieved in accident risk prediction accuracy. The discussion section interprets these results, emphasizing the insights gained from the analysis of heterogeneous sparse data. We conclude by summarizing the contributions of our research, its potential implications for accident prevention and safety measures, and avenues for future research in this critical domain. However, existing studies exhibit certain limitations that need addressing. These limitations include reliance on small-scale datasets with limited geographical coverage (e.g., a restricted number of road segments or focus on a single city) [1, 13, 2, 10,11], dependence on a wide array of data attributes that may not be universally accessible (e.g., satellite imagery, traffic volume data, and road network properties) [5, 12], limited applicability to real-time applications due to modeling constraints and data prerequisites (e.g., predictions over extended time intervals like a day or a week, or the need for extensive datasets) [3, 3, 4, 5], and the use of overly simplistic methodologies for traffic accident prediction [3, 13, 16]. This research underscores the pressing need for a transformative shift in accident risk prediction, emphasizing the adoption of data-rich, multidimensional approaches grounded in advanced data science principles. In an era marked by the increasing complexity of urban environments and transportation systems, accident prevention and mitigation remain pivotal concerns for governments, municipalities, and the broader society. Accurate prediction of accident risk is instrumental in the formulation of proactive safety measures, resource allocation, and the overall enhancement of public well-being. To address these critical issues, this paper introduces an innovative approach that harnesses the power of advanced data science and novel data sources to significantly improve accident risk prediction. Accidents on roadways, whether they involve automobiles, pedestrians, or cyclists, result in substantial human and economic costs. The quest for effective accident risk prediction models has been ongoing for decades, with researchers and policymakers continuously striving to develop better tools to anticipate and prevent accidents. Traditional models have often relied on historical accident data, road attributes, and traffic volume information. While these sources have contributed to our understanding of accident patterns, they have inherent limitations.

Our research addresses these limitations head-on by pioneering an approach that integrates diverse and heterogeneous data sources to develop a robust accident risk prediction model. We recognize that the potential for improved prediction lies not only in expanding the types of data sources but also in the sophisticated analysis of these heterogeneous data. In this endeavor, we employ advanced data science techniques, such as feature engineering, imputation, and machine learning, to extract hidden patterns and insights. The term "heterogeneous sparse data" encapsulates the essence of our approach. Heterogeneous data refers to information from a variety of sources, including traffic flow data, weather conditions, road infrastructure, vehicle telemetry, and historical accident records. Unlike traditional models that focus primarily on structured and uniform data, we embrace the complexity of real-world data landscapes by integrating these diverse sources. Sparse data, on the other hand, alludes to the gaps and limitations inherent in many datasets. Sparse data challenges are particularly prevalent in areas with underdeveloped data collection infrastructure or for specific accident types that are infrequent but of high importance.

Throughout this Research, we support for a fundamental shift in the approach to accident risk prediction. We propose a departure from the conventional, data-scarce, and isolated methods towards a data-rich, multidimensional exploration rooted in the principles of advanced data science. We firmly assert that this transition is not just timely but also imperative in the pursuit of safer roadways and the reduction of the societal toll exacted by accidents. The challenge of reducing traffic accidents is of paramount importance on a global scale. A comprehensive report on traffic safety worldwide [16] reveals a staggering 1.25 million traffic-related deaths in 2013 alone, with fatalities increasing in 68 countries compared to 2010. Accurate accident prediction plays a pivotal role in optimizing public transportation, facilitating safer route planning, and cost-effectively enhancing transportation infrastructure, all aimed at creating safer road environments. The field of accident analysis and prediction has garnered significant research attention in recent decades. This research has encompassed diverse areas such as evaluating the influence of environmental factors (e.g., road network characteristics, weather conditions, and traffic dynamics) on traffic accident occurrences [6,7,8], forecasting the likelihood of accidents within specific geographic areas [1, 2, 3, 4, 5], and predicting accident risks [9]. Noman et al. (2018) introduces a modified approach of retrieving data from the World Health Organization (WHO)

database which is helpful for our research [37,38]. Rahman et. al (2023) uses the machine learning algorithm in tensor flow and python environment which is very useful for this study specifically for the prediction of the accident at United states. Here the author uses big data tools for Apace and MapReduce which idea will be beneficial to calculate the accident [17,18,19,28]. The global magnitude of traffic accidents, with millions of lives at stake annually, underscores the urgency of proactive measures. Our study highlights the shortcomings of existing methodologies, notably their limitations in data availability, geographic scope, and real-time applicability. Ahmmed et al. (2023) describes brain tumor imaging using deep learning algorithm from which we have adopted the programming concept for developing high concept in future research [29]. By advocating for a more comprehensive and sophisticated approach, we aim to enhance road safety, optimize transportation systems, and reduce the devastating societal impact of accidents. Embracing these advancements promises to usher in a new era of accident prediction with far-reaching implications for public safety worldwide. Jamil et. al. (2024) and Mustofa (2020) describes supply chain strategy minimization and bullwhip effect and integrate it with industry 5.0 that works greatly with machine learning and robot with human intelligence interaction [45,46]. Analyzing accidents and predicting the severity of accidents is of paramount importance for various reasons, including public safety, transportation efficiency, and business solutions. Let's delve deeper into these aspects: Addressing the multifaceted challenges posed by traffic accidents involves a comprehensive approach spanning various domains. Foremost among these is the overarching commitment to public safety and the preservation of human lives. Ullah et al. (2023) and Shakil et al. (2013) provide insights into job shop production scenarios, potentially aiding in mitigating state-by-state accident rates [20,21,22,25]. Through in-depth research and analysis, the aim is to unravel accident patterns, contributing factors, and temporal trends, thereby informing the development of targeted interventions to minimize both the frequency and severity of accidents. Beyond the realm of safety, traffic accident analysis contributes significantly to transportation efficiency by identifying congestion points and optimizing logistics, promoting a seamless transportation experience. The impact extends to business solutions, enabling cost reduction strategies and economic efficiency. Effective resource allocation, guided by insights into accident-prone areas, ensures the strategic deployment of resources such as law enforcement and emergency services. Furthermore, infrastructure planning benefits from an understanding of high-risk zones, leading to the development of safer roads and intersections. In the realm of insurance and risk assessment, accurate data analysis enhances the formulation of policies that align with real-world accident scenarios. The advent of autonomous vehicles and advanced driver assistance systems is intricately linked to accident analysis, influencing the trajectory of future transportation. In shaping policy and legislation, evidence-based insights guide regulatory frameworks that promote road safety. Finally, the ability to predict accident severity through historical data analysis allows for proactive measures, timely responses, and improved medical preparedness, completing the holistic approach to enhancing road safety and transportation systems.

2. Methodology

This dataset encompasses traffic accident records nationwide, covering all 49 states of the United States and spanning from February 2016 to March 2023. The data collection process involved the utilization of multiple APIs that provide real-time streaming traffic incident data. These APIs aggregate information from various sources, including the US and state departments of transportation, law enforcement agencies, and data from traffic cameras and sensors embedded within road networks. Currently, the dataset comprises around 7.7 million accident records, and additional details about it can be found in the associated documentation. The dataset's creation involved real-time access to multiple Traffic APIs, specifically focusing on accident data across the Contiguous United States during the period from February 2016 to March 2023. Our research draws inspiration from various studies. Hossain et al. (2023) explore electricity generation from motors and associated risk factors, laying the groundwork for our risk analysis [26]. Fayshal et al. (2023) consider environmental factors and safety risk assessment, significantly influencing our study [23,24]. Kamal et al. (2019) present empirical evidence on RFID technology for warehouse management, impacting accident detection across various sectors [30]. Parvez et al. (2022) engage in a comprehensive discussion on ergonomics for students, offering insights into human working posture and efficiency, with relevance to our research [31,32]. Building upon these related works, we aim to address diverse circumstances in our approach to accident reduction. Mustaquim et al. (2024) provide valuable insights into the utilization of machine learning and Python for data analysis and remote sensing, serving as significant inspiration for our research [35] [36]. Iqtiar et al. (2023) investigate renewable energy sources for Bitcoin, aligning with our future research plans [27].

2.1. Coding and Library

The code begins by importing essential data processing and analysis libraries, including nltk, which is widely used for natural language processing tasks, and re and string for regular expressions and string manipulation, respectively. The code then brings in the foundational data manipulation libraries pandas and numpy, offering powerful tools for working with structured data. Moving on to visualization, the code imports matplotlib, a versatile plotting library, with pyplot

for creating static, animated, and interactive plots. It also leverages seaborn for statistical data visualization based on matplotlib. Additionally, plotly is imported for creating interactive plots, graphs, and dashboards. For geospatial analysis and mapping, the code utilizes the geopandas library, which extends pandas to enable spatial operations on geometric types. geoplot is imported for high-level geospatial plotting, and the Nominatim geocoder from geopy for obtaining location data based on place names. To handle potential warnings in the code, the warnings library is imported and configured to filter them out. Overall, this comprehensive set of libraries provides a robust toolkit for data analysis, visualization, and geospatial tasks in Python.

2.2. Description of Data

This Data comprises of 46 columns and 7,728,394 rows. The dataset includes several key attributes providing comprehensive information on accident records. The "ID" serves as a unique identifier for each incident, facilitating efficient record management. The "Source" attribute denotes the origin of the raw accident data, offering insights into the data's provenance. Severity, represented by a numerical scale from 1 to 4, signifies the extent of impact on traffic, aiding in categorizing incidents based on their severity. Temporal details, encompassing "Start Time" and "End Time," specify when accidents occurred and when their impact on traffic concluded, respectively. Geographic coordinates, "Start_Lat," "Start_Lng," "End_Lat," and "End_Lng," provide precise location data using GPS coordinates. The "Distance(mi)" attribute quantifies the extent of road affected by the accident. A human-provided narrative of the incident is encapsulated in the "Description" field, adding qualitative context. Address-related attributes such as "Street," "City," "County," "State," "Zipcode," and "Country" offer detailed location information. "Timezone" identifies the temporal zone based on accident location, enhancing temporal context. Weather-related attributes, including "Airport_Code," "Weather_Timestamp," "Temperature(F)," "Wind_Chill(F)," "Humidity (%)," "Pressure(in)," "Visibility(mi)," "Wind_Direction," "Wind_Speed(mph)," "Precipitation(in)," and "Weather_Condition," provide comprehensive meteorological insights. Points of interest annotations such as "Amenity," "Bump," "Crossing," "Give_Way," "Junction," "No_Exit," "Railway," "Roundabout," "Station," "Stop," "Traffic Calming," "Traffic_Signal," and "Turning_Loop" highlight specific elements in the vicinity. Finally, attributes like "Sunrise_Sunset," "Civil_Twilight," "Nautical_Twilight," and "Astronomical_Twilight" shed light on the time of day based on different twilight phases, contributing to a holistic understanding of accident contexts. [14,15,44]

2.3. Approach

The central objective of our research initiative is to gain a comprehensive understanding of traffic accidents, aiming to enhance overall road safety. Our primary focus is on unraveling the complexities within accident data to extract valuable insights into the contributing factors, as well as the spatial and temporal patterns associated with accidents, and their resulting consequences. This foundational understanding forms the basis for the development of robust accident prevention strategies and safety measures. Our overarching goal is to decrease both the frequency and severity of traffic accidents, ultimately fostering safer road environments. This reduction in accidents not only aims to minimize human casualties but also seeks to mitigate economic losses, thereby enhancing the well-being of communities and society. Building on the works of Nazma et al. (2014) and Rahman (2015), which discuss the impact of supplier selection on industry authorities and propose a weightage method using machine learning for setting priorities and rankings for different suppliers [34], our research aligns with the broader goal of improving safety across diverse sectors.

2.4. Data Knowledge

The dataset under examination is a valuable resource containing a wealth of information pertaining to traffic accidents. It encompasses a wide array of attributes, including details about weather conditions, accident severity, geographical locations, and more. However, to streamline our analysis and storytelling process, certain columns deemed less relevant to our specific research questions have been strategically dropped from the dataset using coding techniques. Our analysis is driven by a set of carefully determined questions of interest that encapsulate the key aspects we aim to explore and understand regarding traffic accidents. These questions serve as guiding pillars for our investigative journey: Most Common Weather Conditions for Accidents: We aim to uncover the prevailing weather conditions during accidents. This knowledge is instrumental in identifying weather-related risk factors and can contribute to the development of weather-specific safety measures. Time of Day and Accident Frequencies: By examining the relationship between the time of day and the number of accidents, we seek to discern patterns and trends that shed light on when accidents are most likely to occur. This insight can inform the allocation of resources for enhanced safety during peak accident times. Correlation Between Temperature and Accident Severity: We aim to investigate whether there is a statistically significant correlation between temperature and the severity of accidents. This analysis can provide insights into how temperature variations might influence accident outcomes and the subsequent implications for road safety strategies.

States with the Highest Accident Rates: Identifying states with the highest accident rates is pivotal for focusing preventive measures in regions with elevated risk. This analysis will spotlight areas requiring targeted interventions and safety campaigns and they discussed land sensing method with solar cap integration that is the future measure control for this paper [39,41. Spatial Patterns in Accidents: Exploring spatial patterns in accidents by latitude and longitude allows us to visualize and understand geographic concentrations of accidents. This information is invaluable for pinpointing high-risk zones and tailoring safety measures accordingly.

Cities and Streets with the Most Accidents: To enhance localized accident prevention efforts, we will identify the cities and specific streets within those cities where accidents are most frequent. This level of granularity is essential for city planners and traffic authorities to prioritize safety initiatives effectively. In summary, our data understanding phase involves a purposeful reduction of the dataset to focus on pertinent attributes aligned with our questions of interest. By addressing these questions, we endeavor to unravel the intricate facets of traffic accidents, enabling us to formulate data-driven recommendations and strategies aimed at enhancing road safety and reducing accidents' societal impact.

2.5. Data Processing and Cleaning preparation

In this crucial phase of our research, we meticulously curate and refine the raw accident data to ensure its quality, integrity, and suitability for analysis. Data preparation encompasses several essential steps, including data extraction, transformation, and loading (ETL). We retrieve the accident dataset from the source, ensuring that it is complete and up to date. Next, we perform data cleaning procedures to address missing values, outliers, and inconsistencies. This involves imputing or removing missing data points, identifying, and correcting data anomalies, and standardizing data formats. Additionally, we normalize and scale relevant features to maintain consistency and facilitate accurate modeling. Data preparation is pivotal as it lays the foundation for robust and meaningful analysis, enabling us to derive reliable insights and make informed decisions regarding accident risk prediction and safety measures. Our commitment to data quality ensures the credibility and validity of our research findings [40,42,43].

2.6. Techniques Evaluation and Visualization

2.6.1. Map Analysis

Accident Rates by State. The choropleth map highlights accident rates by state, with darker colors indicating higher accident counts. It shows spatial patterns in accident distribution, with some states having significantly higher accident rates than others. This information can guide state-level road safety initiatives, such as increased enforcement or infrastructure improvements in high-risk areas. The below figure (fig:1) shows the choropleth Map for the desired visualization:

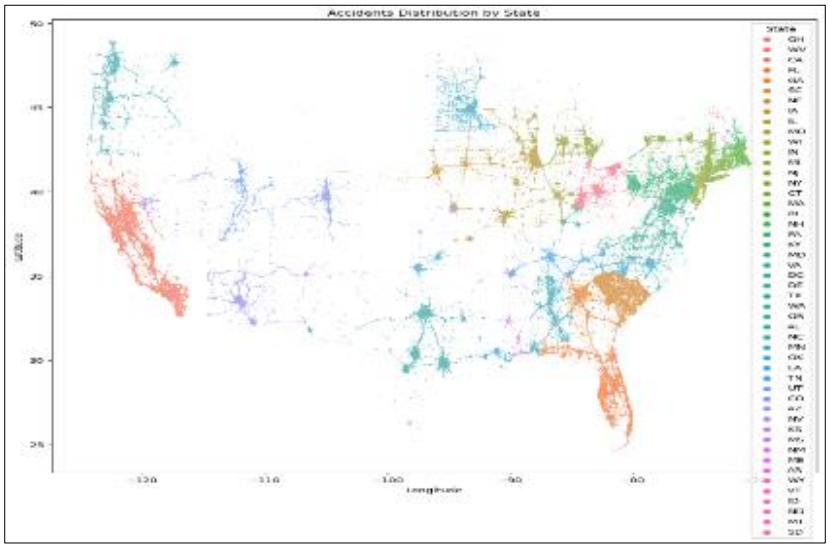

Figure 1 Accident Rates by State

2.7. The Pie slice Analysis

This shows the number of accidents monthly and weekly. The provided code snippet serves as a pivotal component of our research paper's data visualization and analysis process. In this section, we employ Python's Matplotlib library to

generate a stacked bar chart that effectively illustrates the distribution of accidents across different U.S. timezones. This visualization aids in understanding accident patterns concerning geographical regions and time of occurrence [33]. The stacked bar chart, configured with a distinctive color palette, offers a clear and visually engaging representation of accident data. Each bar in the chart corresponds to a specific U.S. time zone, allowing for a straightforward comparison of accident frequencies across these regions. The bars are stacked, with each segment representing a particular accident severity level or another categorical variable, providing insights into the composition of accidents within each time zone. To enhance the readability of the chart, we have set the x-axis labels to display the U.S. time zones, ensuring that they are easily interpretable. Additionally, we have rotated the labels by 45 degrees for better visualization alignment. The y-axis denotes the number of accidents, and the accompanying labels are both clear and prominent, ensuring that the reader can readily discern accident counts. Furthermore, we have employed a custom formatter to present the y-axis labels in a more reader-friendly manner. This format ensures that the accident counts are displayed with appropriate thousand separators, enhancing the chart's comprehensibility.

Overall, this code snippet plays a vital role in our research paper by offering an effective and visually compelling means of conveying the distribution of accidents across U.S. time zones. This visualization empowers researchers and stakeholders to identify regional accident trends and make informed decisions regarding safety measures, resource allocation, and accident prevention strategies tailored to specific geographical areas and timeframes. The below figure (fig:2) shows the Line plot for the desired visualization:

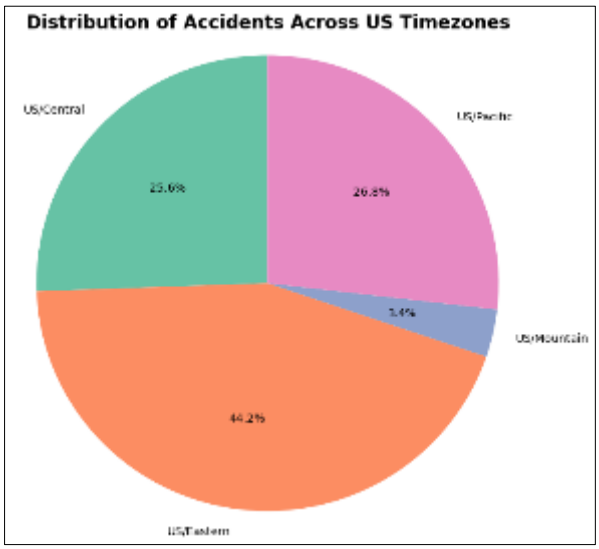

Figure 2 Accident Rates based on time-zones

2.7.1. Bar Plot

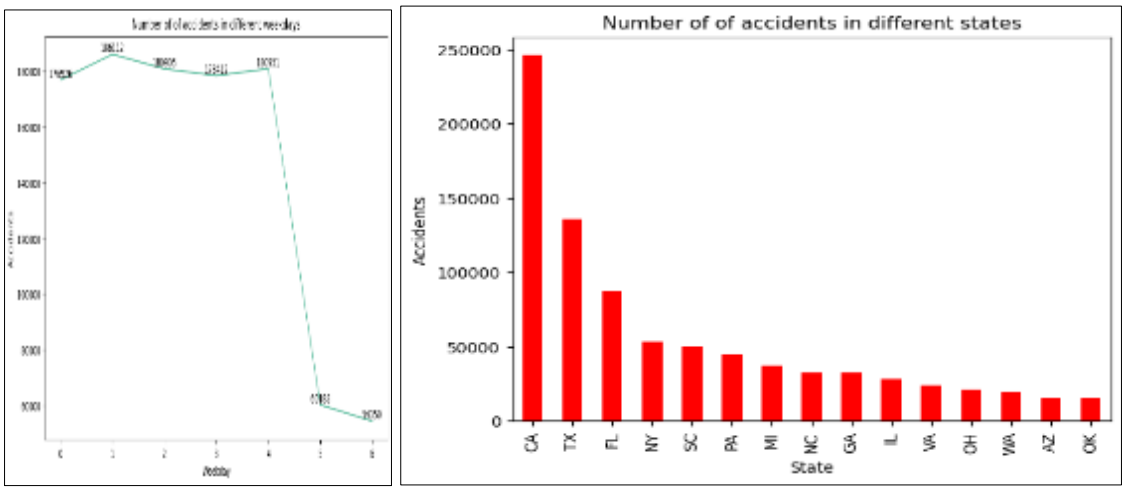

Figure 3 Number of accidents in weekdays and state wise

Clear and fair conditions result in the highest accidents. Cloudy weather contributes significantly, accounting for around 1.35 million accidents. Rain-related conditions cause about 100000 accidents. Adverse Conditions (Snow, fog, haze) cause roughly 200000 accidents. Here, the below figure (fig:3) shows the Bar plot for the desired visualization of the accident conditions.

2.7.2. Line analysis over time

A scatter plot visualization is created to illustrate the geographical distribution of accidents across different locations. The dataset, represented by the DataFrame 'df', consists of randomly generated longitude and latitude coordinates along with a categorical classification of locations (West, Middle, East) based on predefined state groupings. The visualization employs the Seaborn library to enhance the aesthetic appeal and interpretability of the plot. The 'deep' color palette is chosen to ensure a visually pleasing and distinct representation of the categorized locations. The 'whitegrid' style provides a clean and unobtrusive background for the scatter plot. The scatter plot itself is constructed using the 'sns.scatterplot' function, with the x-axis representing longitude, the y-axis representing latitude, and the data points differentiated by color based on their respective locations. The alpha parameter is set to 0.7 to introduce a level of transparency, aiding in the visualization of overlapping data points. Additionally, markers are outlined in white ('edgecolor') to enhance visibility, and marker size ('s') is adjusted for clarity. Labels for the x-axis, y-axis, and plot title are incorporated to provide context and facilitate interpretation. The legend indicates the color-coding scheme for the different geographical locations, enhancing the plot's informativeness. This scatter plot visualization serves as a valuable exploratory tool for researchers examining the spatial distribution of accidents across various regions. It enables a quick and intuitive assessment of whether there are any discernible patterns or concentrations of accidents in specific geographical areas, contributing to a comprehensive understanding of the dataset's spatial characteristics.

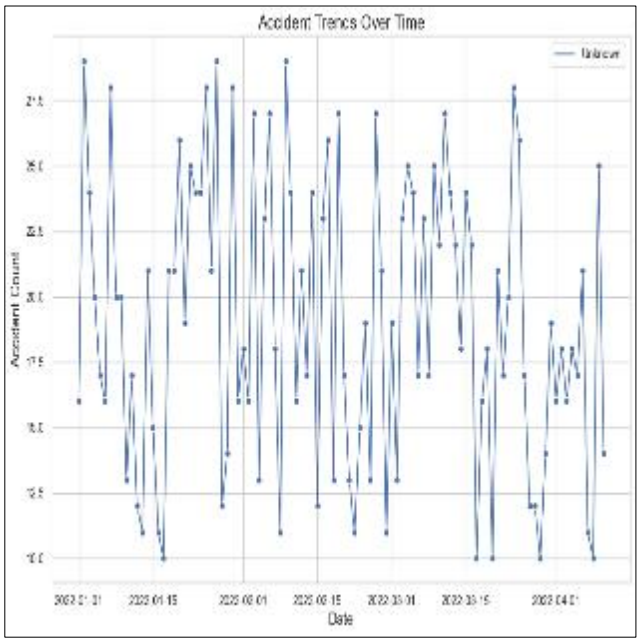

Figure 4 Line analysis

2.7.3. Pie Slices

January exhibits the highest ratio of weekend accidents compared to weekday accidents, standing at 24.36%. In contrast, March displays the lowest proportion of weekend accidents, with only 16.48%. Notably, December records the highest overall number of accidents, while October registers the lowest. The provided code snippet furnishes a detailed visual examination of the "Is Weekend" attribute's distribution across each month of the year. It constructs a 4x3 grid of pie charts, with each chart dedicated to a specific month, offering a nuanced exploration of how weekends ("True") and weekdays ("False") influence accident occurrences. The vibrant pie charts vividly depict the percentage of accidents on weekends versus weekdays for each month. A thorough analysis of these charts can unveil patterns and fluctuations in accident frequency throughout the year. For instance, it aids in identifying whether weekends witness a surge in accidents, possibly due to heightened recreational activities or a lax adherence to traffic regulations. The use of exploded slices accentuates the "True" (weekend) category, ensuring it stands out visually and facilitates the interpretation of weekend-specific trends. Furthermore, the inclusion of labels containing percentages and absolute case counts provides

both relative and absolute context, augmenting the clarity of the charts. Furthermore, the month-specific titles above each chart facilitate easy navigation and comparison, allowing researchers and stakeholders to quickly identify any recurring patterns or anomalies. This visualization approach is invaluable for gaining insights into the relationship between weekends, weekdays, and accident occurrences, which can inform targeted safety measures and interventions aimed at reducing accidents during critical periods. Overall, this code snippet exemplifies how data visualization techniques can unveil valuable trends within complex datasets, contributing to informed decision-making and accident prevention efforts. For monthly cases, the below figure (Fig:5) shows the pie chart visualization percentages in different months:

Figure 5 Monthly cases from pie slices

2.7.4. Scatter Plot analysis

Top ten street vs accident count in this experiment. In this section, we provide a code snippet for data analysis and visualization using Python libraries such as Pandas, Seaborn, and Matplotlib. This code is part of our research study, which aims to analyze and visualize accident data to gain insights into accident patterns in different cities. We import essential libraries, including Pandas for data manipulation, Seaborn for data visualization, and Matplotlib for customizing and displaying plots. From below figure (Fig :6), we can see the cities that have most accidents and the top ten streets for the most accidents:

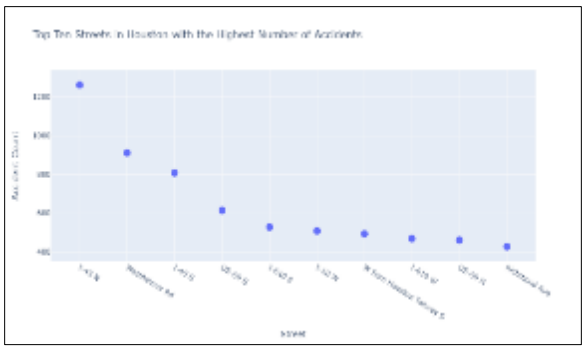

Figure 6 Top Ten Street that has most accidents

If we've already loaded our accident data into a DataFrame named 'df,' we proceed to group the data by the 'City' column and count the number of accidents in each city. The results are sorted in descending order to identify the cities with the highest accident counts. To focus on the top 10 cities with the highest accident counts, we create a new DataFrame, 'df6,' which selects the first 10 rows from the sorted 'df5' DataFrame. We created a bar plot to visualize the number of accidents in these top 10 cities. The figure size is adjusted to ensure clarity in the visualization. The x-axis represents the cities, while the y-axis shows the number of accidents. We add labels to the axes and a title to the plot for better interpretation. Finally, we print the number of accidents in the top 10 cities to provide a clear numerical representation alongside the visual plot. In summary, these visualizations provide valuable insights for improving road safety. They help identify factors like weather conditions, time of day, and geographic regions that are associated with higher

accident rates or severity. The findings can inform strategies and interventions to reduce accidents and enhance road safety.

3. Results and Discussion

Traffic accidents stand out as a pressing and ongoing public safety issue, necessitating continuous research, analysis, and proactive measures. The findings of this study highlight alarming statistics, underscoring the crucial need for a deep understanding of accident patterns and targeted interventions. Notably, California recorded an astonishing 254,987 traffic accidents, while Houston, Texas, experienced a significant 43,344 accidents, emphasizing the severity of the problem. Of particular concern is the concentration of accidents, exemplified by the 1,263 reported incidents on I-45 N street in Houston. These figures stress the vital importance of identifying high-risk zones within cities and states, offering valuable insights for authorities to allocate resources effectively and implement tailored safety measures. Given that Houston is one of the largest and busiest metropolitan areas in the United States, immediate attention from authorities and traffic management agencies is imperative. The remarkably high accident count in this city suggests that current safety measures may require reevaluation and enhancement in specific areas.

4. Conclusion

In summary, the data disclosed in this investigation underscore the alarming magnitude of the traffic accident predicament in both California and Houston, Texas. The clustering of accidents in specific locales, notably I-45 N street, underscores the immediate necessity for decisive action. It becomes imperative for authorities to adopt proactive measures, refining safety protocols, discerningly distributing resources, and sustaining endeavors to curtail accidents and protect the lives of residents and commuters. The pursuit of safer roads mandates a collective dedication to accident prevention and the alleviation of their profound repercussions. The substantial figure of 1,263 accidents on I-45 N street raises considerable alarm, warranting an in-depth examination and targeted interventions for this area. To mitigate accident risks, it is critical for authorities to implement a comprehensive strategy, encompassing intensified law enforcement, more stringent traffic regulations, and substantial investments in infrastructure enhancements. This involves improvements in road design, upgrades to signage, and the integration of advanced traffic management systems. Additionally, prioritizing public awareness initiatives and educational programs is essential to cultivate responsible driving behavior and disseminate information about accident-prone zones. Effective collaboration among government entities, law enforcement, urban planners, and transportation authorities is indispensable for tackling this public safety concern. Through concerted efforts, these stakeholders can devise holistic strategies and allocate resources strategically to address the root causes of traffic accidents.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Ciro Caliendo, Maurizio Guida, and Alessandra Parisi. 2007. A crash-prediction model for multilane roads. Accident Analysis and Prevention 39, 4 (2007), 657–670.

[2] Li-Yen Chang and Wen-Chieh Chen. 2005. Data mining of tree-based models to analyze freeway accident frequency. Journal of safety research 36, 4 (2005), 365–375

[3] Alameen Najjar, ShunâĂŹichi Kaneko, and Yoshikazu Miyanaga. 2017. Combining satellite imagery and open data to map road safety. In Thirty-First AAAI Conference on Artificial Intelligence. AAAI, Palo Alto, CA, USA

[4] Honglei Ren, You Song, Jingwen Wang, Yucheng Hu, and Jinzhi Lei. 2018. A Deep Learning Approach to the Citywide Traffic Accident Risk Prediction. In 2018 21st International Conference on Intelligent Transportation Systems (ITSC). IEEE, 3346–3351.

[5] Zhuoning Yuan, Xun Zhou, and Tianbao Yang. 2018. Hetero-ConvLSTM: A deep learning approach to traffic accident prediction on heterogeneous spatiotemporal data. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York, NY, USA, 984–992.

[6] Daniel Eisenberg. 2004. The mixed effects of precipitation on traffic crashes. Accident analysis and prevention 36, 4 (2004), 637–647.

[7] David Jaroszweski and Tom McNamara. 2014. The influence of rainfall on road accidents in urban areas: A weather radar approach. Travel behaviour and society 1, 1 (2014), 15–21.

[8] JD Tamerius, X Zhou, R Mantilla, and T Green"eld-Huitt. 2016. Precipitation effects on motor vehicle crashes vary by space, time, and environmental conditions. Weather, Climate, and Society 8, 4 (2016), 399–407

[9] Quanjun Chen, Xuan Song, Harutoshi Yamada, and Ryosuke Shibasaki. 2016. Learning deep representation from big and heterogeneous data for traffic accident inference. In Thirtieth AAAI Conference on Arti#cial Intelligence. AAAI, Palo Alto, CA, USA

[10] Lei Lin, Qian Wang, and Adel W Sadek. 2015. A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction. Transportation Research Part C: Emerging Technologies 55 (2015), 444–459

[11] Lu Wenqi, Luo Dongyu, and Yan Menghua. 2017. A model of traffic accident prediction based on convolutional neural network. In 2017 2nd IEEE International Conference on Intelligent Transportation Engineering (ICITE). IEEE, 198–202.

[12] Zhuoning Yuan, Xun Zhou, Tianbao Yang, James Tamerius, and Ricardo Mantilla. 2017. Predicting tra!c accidents through heterogeneous urban data: A case study. In Proceedings of the 6th International Workshop on Urban Computing (UrbComp 2017), Halifax, NS, Canada, Vol. 14. ACM, New York, NY, USA

[13] Chukwutoo C Ihueze and Uchendu O Onwurah. 2018. Road traffic accidents prediction modelling: An analysis of Anambra State, Nigeria. Accident Analysis and Prevention 112 (2018), 21–29.

[14] Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.

[15] Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. "A Countrywide Traffic Accident Dataset.", arXiv preprint arXiv:1906.05409 (2019).

[16] World Health publisher. 2015. Global status report on road safety 2015. World Health publisher.

[17] Rahman, S. M., Ibtisum, S., Podder, P., and Hossain, S. M. (2023). Progression and challenges of IoT in healthcare: A short review. *arXiv preprint arXiv:2311.12869*.

[18] Rahman, S. M., Ibtisum, S., Bazgir, E., and Barai, T. (2023). The significance of machine learning in clinical disease diagnosis: A review. arXiv preprint arXiv:2310.16978.

[19] Ibtisum, S., Bazgir, E., Rahman, S. A., and Saokat, S. M. H., A comparative analysis of big data processing paradigms: Mapreduce vs. apache spark. World Journal of Advanced Research and Reviews, 2023, 20(01), 1089–1098.

[20] Ullah, M. R., Molla, S., Siddique, I. M., Siddique, A. A., and Abedin, M. M. (2023). Utilization of Johnson's Algorithm for Enhancing Scheduling Efficiency and Identifying the Best Operation Sequence: An Illustrative Scenario. Journal of recent activities in production, 8(3), 11-29.

[21] Shakil, M., Ullah, M. R., and Lutfi, M. (2013). Process flow chart and factor analysis in production of jute mills. Journal of Industrial and Intelligent Information Vol, 1(4).

[22] Ullah, M. R., Molla, S., Siddique, I. M., Siddique, A. A., and Abedin, M. M (2023). Manufacturing Excellence Using Line Balancing and Optimization Tools: A Simulation-based Deep Framework., Journal of Modern Thermodynamics in Mechanical System, 5(3), 8-22.

[23] Fayshal, M. A., Jarin, T. T., Ullah, M. R., Rahman, S. A., Siddque, A. A., and Siddique, I. M. A Comprehensive Review of Drain Water Pollution Potential and Environmental Control Strategies in Khulna, Bangladesh. Journal of Water Resources and Pollution Studies, 8(3), 41-54.

[24] Fayshal, M. A., Ullah, M. R., Adnan, H. F., Rahman, S. A., and Siddique, I. M. Evaluating Multidisciplinary Approaches within an Integrated Framework for Human Health Risk Assessment., Journal of Environmental Engineering and Studies, 8(3), 30-41.

[25] Md Rahamat Ullah, et al. (2023). Optimizing Performance: A Deep Dive into Overall Equipment Effectiveness (OEE) for Operational Excellence, Journal of Industrial Mechanics, 8(3), 26-40.

[26] Hossain, M. Z., Rahman, S. A., Hasan, M. I., Ullah, M. R., and Siddique, I. M. (2023). Evaluating the effectiveness of a portable wind generator that produces electricity using wind flow from moving vehicles. Journal of Industrial Mechanics, 8(2), 44-53.,

[27] Iqtiar Md Siddique, Anamika Ahmed Siddique, Eric D Smith, Selim Molla. (2023). Assessing the Sustainability of Bitcoin Mining: Comparative Review of Renewable Energy Sources. Journal of Alternative and Renewable Energy Sources,10(1),1-12.

[28] Ahmmed, Syed, Prajoy Podder, M. Rubaiyat Hossain Mondal, S M Atikur Rahman, Somasundar Kannan, Md Junayed Hasan, Ali Rohan, and Alexander E. Prosvirin. 2023. "Enhancing Brain Tumor Classification with Transfer Learning across Multiple Classes: An In-Depth Analysis" BioMedInformatics 3, no. 4: 1124-1144. https://doi.org/10.3390/biomedinformatics3040068.

[29] Sifat Ibtisum, S M Atikur Rahman and S. M. Saokat Hossain, 2023. "Comparative analysis of MapReduce and Apache Tez Performance in Multinodeclusters with data compression". World Journal of Advanced Research and Reviews, 2023, 20(03), 519–526. 10.30574/wjarr.2023.20.3.2486.

[30] Kamal, T., Islam, F., and Zaman, M. (2019). Designing a Warehouse with RFID and Firebase Based Android Application. Journal of Industrial Mechanics, 4(1), 11-19.

[31] Parvez, M. S., Talapatra, S., Tasnim, N., Kamal, T., and Murshed, M. (2022). Anthropomorphic investigation into improved furniture fabrication and fitting for students in a Bangladeshi university. Journal of The Institution of Engineers (India): Series C, 103(4), 613-622.

[32] Parvez, M. S., Tasnim, N., Talapatra, S., Kamal, T., and Murshed, M. (2022). Are library furniture dimensions appropriate for anthropometric measurements of university students? Journal of Industrial and Production Engineering, 39(5), 365-380.

[33] Rahman, M. A., Bazgir, E., Hossain, S. S., and Maniruzzaman, M. (2024). Skin cancer classification using NASNet. International Journal of Science and Research Archive, 11(1), 775-785.

[34] Rahman, S. A., and S Shohan (2015). Supplier selection using fuzzy-topsis method: A case study in a cement industry, IASET: Journal of MechanicalEngineering.,4(1).31-42.

[35] Molla, S., Bazgir, E., Mustaquim, S. M., Siddique, I. M., and Siddique, A. A. (2024). Uncovering COVID-19 conversations: Twitter insights and trends. *World Journal of Advanced Research and Reviews*, *21*(1), 836-842.

[36] S M Mustaquim. (2024). "Utilizing Remote Sensing Data and ArcGIS for Advanced Computational Analysis in Land Surface Temperature Modeling and Land Use Property Characterization". World Journal of Advanced Research and Reviews, 2024, 21(01), 1496–1507.

[37] Noman, A. H. M., Das, K., and Andrei, S. (2020). A Modified Approach for Data Retrieval for Identifying Primary Causes of Deaths. ACET Journal of Computer Education and Research, 14(1), 1-13.

[38] Noman, A. H. M. (2018). WHO Data: A Modified Approach for Retrieval (Doctoral dissertation, Lamar University-Beaumont).

[39] Mohammad Fokhrul Islam Buian, Nafis Ahmed Pantho, Iqtiar Md Siddique, and Anamika Ahmed Siddique. (2024). Industrial Process Optimization for the Effective Removal of Per- and Polyfluoroalkyl Substances (PFAS) from Water Treatment Systems. European Journal of Advances in Engineering and Technology, 11(2), 1–12. https://doi.org/10.5281/zenodo.10671774.

[40] Mohammad Fokhrul Islam Buian, Iqtiar Md Siddique, and Anamika Ahmed Siddique. (2024). Efficient Parking Management through QR Technology. Journal of Scientific and Engineering Research, 11(2), 1–9. https://doi.org/10.5281/zenodo.10671733.

[41] Hasan, M. R., Hossain, M. S., and Rahman, K. P. (2017). Design and construction of a portable charger by using solar cap. Global Journal of Researches in Engineering: A Mechanical and Mechanics Engineering, 17(5), 14-18.

[42] Noman, A.H.M., Mustaquim S.M. Molla, S., and Siqqique, M.I., (2024). Enhancing Operations Quality Improvement through Advanced Data Analytics. Journal of Computer Science Engineering and Software Testing. Vol. 10, Issue 1 (January – April, 2024) pp: (1-14). https://doi.org/10.46610/JOCSES.2024.v10i01.001.

[43] Mustaquim, S. M., Noman, A. H. M., Molla, S., Siddique, A. A., and Siddique, A. A. (2024). Enhancing Accident Risk Prediction with Novel Data and Findings from Heterogeneous Sparse Sources. Journal of Data Mining and Management, 9(1), 1-16.

[44] Ibtisum, S., Rahman, S. A., and Hossain, S. S. (2023). Comparative analysis of MapReduce and Apache Tez Performance in Multinode clusters with data compression. World Journal of Advanced Research and Reviews, 20(3), 519-526.

[45] Jamil, M. A., Mustofa, R., Hossain, N. U. I., Rahman, S. A., & Chowdhury, S. (2024). A Theoretical Framework for Exploring the Industry 5.0 and Sustainable Supply Chain Determinants. Supply Chain Analytics, 100060.

[46] Mustofa, R. (2020) "Bullwhip Effect Minimization Strategy Formulation: Keys to Enhancing Competitiveness and Performance". International Conference on Mechanical, Industrial and Energy Engineering, December 19th-21st, Khulna, Bangladesh, 20-079.