*Data Analysis and Prediction on Road Accident*

Dr. Gendlal M. Vaidya

Assistant Professor, Department of Computer Technology,

Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India

gendlalvaidya@gmail.com

Pranjali P. Muley

Professor, Department of Physiology

Datta Meghe Medical College and Shalinitai Meghe Hospital & Research Center, Nagpur, Maharashtra, India

drpranjali11@gmail.com

Shraddha G. Vaidya

Assistant Professor, Department of Computer Science and Engineering

Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India

shraddha.karale1306@gmail.com

Parikshit A. Muley

Professor, Department of Physiology

Datta Meghe Medical College and Shalinitai Meghe Hospital & Research Center, Nagpur, Maharashtra, India

shraddha.karale1306@gmail.com

Anurag S. Rahate

Computer Technology

Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India

anuragrahate3@gmail.com

Shalaka B. Deshpande

Computer Technology

Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India

sdeshpande050c@gmail.com

Prajakta S. Ghyar

Computer Technology

Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India

prajaktaghyar@gmail.com

Janya M. Kinker

Computer Technology

Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India

janyakinker2003@gmail.com

Mansi K. Patel

Computer Technology

Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India

mansikrpatel03@gmail.com

**INTRODUCTION**

With near around 1.35 million fatalities and numerous injuries each year, road traffic accidents (RTAs) pose a serious and widespread global problem that places a heavy financial burden on society everywhere [1]. More than 90% of fatalities from these accidents occur in low- and middle-income nations, highlighting the critical need for efficient mitigation measures and technical breakthroughs.

Through the integration of telematics data, sophisticated weather modeling, and cutting-edge machine learning algorithms, recent research has investigated novel approaches to solve the difficulties of RTAs. RTAs are the second most common cause of Disability-Adjusted Life Years (DALYs) in Iran, a lower middle-income country, underscoring the seriousness of the problem [2].

A notable 2020 study predicted mistakes in accident hotspots using telematics data from 1,673 intercity buses in Iran. With an astounding Area Under the Curve (AUC) of 91.70%, this study demonstrated the effectiveness of a machine learning model called Extreme Gradient Boosting (XGBoost) [3].

The difficulties that unfavorable weather conditions present for vehicle sensors have been further highlighted by the development of driverless cars and sophisticated driver support systems. The performance of sensors like LIDAR, cameras, GPS, and radar—all crucial for safe autonomous navigation—can be severely compromised by rain, snow, and fog[4].

Beyond developments in technology, the study of historical accident data has been transformed by the use of data mining and machine learning. Predictive models have been crucial in predicting the likelihood of accidents based on variables like location, time of day, and environmental conditions, while studies using advanced clustering algorithms have effectively identified high-risk areas that are prone to accidents [5].

Collaboration across multiple areas is necessary to address the complex difficulties of RTAs. Policymakers and stakeholders can create evidence-based plans to improve road safety by utilizing telematics, sophisticated machine learning, and accurate weather modeling.

**LITERATURE SURVEY**

**Data Collection:**

Identifying Traffic Accident Triggers: Research by Ren et al. has shed light on the critical conditions and traffic patterns that can lead to accidents. For instance, Oh highlighted that disruptions in traffic flow often act as triggers for crashes. Furthermore, studies indicate that variations in speed, measured through standard deviation before an accident, are substantial predictors of crash risk [8].

Impact of Weather Conditions: Numerous studies show that weather plays a significant role in road traffic accidents. Li et al. point out that weather conditions can influence the frequency of traffic accidents. Research has also revealed that minor weather changes, such as fog, rain, and snow, can diminish visibility and alter road surfaces, ultimately increasing the likelihood of accidents [10].

Accident-Prone Areas: Jithin Krishnan's research has identified certain locations on the same road that exhibit a higher incidence of accidents due to specific road or environmental conditions. These areas, known as accident-prone zones, are determined through an analysis of observed crashes and regression techniques [11].

Driver Stress and Road Safety: Driver stress is a prevalent issue that can heighten the risk of vehicle crashes. It is observed that those who have difficulty managing stress are at a greater risk of accidents, underscoring the importance of monitoring stress levels while driving.

Physiological Effects of Stress: Stress influences various physiological processes within the body, mainly through the autonomic nervous system (ANS). When stressed, the body releases hormones like adrenaline and cortisol, which prepare it for high-stakes situations. This physiological reaction can lead to increased heart rates and changes in driving behavior, raising concerns for road safety [11].

Temporal and Spatial Influences: Research highlights that factors related to time, including the hour and season, play a critical role in the frequency of accidents. For instance, in Iran, a notable number of accidents occur during the early evening and late afternoon, particularly in the spring and summer months. Moreover, certain areas are identified as having a higher accident prevalence, which calls for focused preventive measures.

Impact of Traffic and Weather: A study by Yannis T.H. explored how traffic conditions and weather influence road safety, revealing a clear connection between traffic flow, weather patterns, and accident statistics. This suggests that utilizing real-time data could greatly improve our comprehension of these dynamics.

Accident Trends on Highways: K. Meshram and H.S. Goliya conducted a study of accidents along a particular highway segment and found that poor road geometry led to more frequent accidents. Their research emphasizes the essential role of effective road design in minimizing accident risks.

Urban Accident Patterns: Findings indicate that urban settings, like Indore, are significant contributors to overall accident rates, largely due to high speeds and heavy traffic. This underscores the necessity for strategic measures in urban planning and traffic control.

Leveraging Data Analytics for Safer Cities: Abhinav Shikhar et al. explored how data analytics can be applied to anticipate and mitigate the effects of accidents. Their research highlights various contributing factors, such as alcohol use and road conditions, which are crucial for understanding the causes behind accidents.

Road Issues and Accident Severity: Esmaeili et al. investigated how road defects relate to the severity of accidents, identifying poor visibility in high-risk areas as a major concern. This study supports the overarching goal of pinpointing locations with a high risk of accidents.

**Data Processing:**

The WHO report utilizes a range of data processing techniques to effectively present global road safety statistics. Key methods include:

* Data Aggregation: This involves gathering data from numerous national sources to generate global estimates of road traffic fatalities and injuries (WHO, 2020).
* Statistical Adjustments: Various statistical methods are applied to correct inconsistencies and fill gaps in the data reported by different countries (WHO, 2020).

The report highlights the complexities involved in harmonizing data from diverse sources, aiming for both comparability and accuracy [1].

Similarly, the World Bank’s profile leverages data processing techniques to explore the economic and social factors affecting road safety. Important methods include:

* Data Integration: This process combines data from various sources, such as economic indicators and public safety reports, to develop a comprehensive profile (World Bank, 2021).
* Trend Analysis: This method involves analyzing trends and patterns in road safety data, providing specific insights relevant to Iran (World Bank, 2021).

The profile emphasizes how integrating data and conducting trend analyses can enhance the understanding of road safety issues within a national context [2].

In another study, the focus is on processing telematics data for the analysis of road traffic accidents (RTAs). The processing techniques employed include:

* Data Cleaning and Integration: This entails cleaning telematics data and integrating it with traditional accident records to ensure accuracy and thoroughness (Research Study on Telematics Data and RTAs in Iran, 2020).
* Data Analysis: Analytical techniques are used to interpret telematics data and uncover patterns associated with RTAs (Research Study on Telematics Data and RTAs in Iran, 2020).

The study underscores the significance of advanced data processing techniques when dealing with large volumes of telematics data to extract meaningful insights [3].

According to the research by Smith, Brown, & Johnson, the following techniques are also essential:

* Sensor Data Processing: This involves analyzing raw data from autonomous vehicle sensors to evaluate the influence of weather conditions.
* Data Filtering and Calibration: Here, filtering and calibration techniques are applied to ensure the accuracy of sensor data under various environmental conditions.

The study illustrates the advanced data processing techniques necessary for managing sensor data and understanding the environmental impacts on autonomous vehicles [4].

**Algorithms Implemented:**

In his research, Jithin Krishnan M. V. concentrated on identifying areas prone to accidents by analyzing driver stress data. While the paper doesn’t detail specific machine learning algorithms used, studies of this nature often employ techniques like Decision Trees, Random Forests, and Support Vector Machines (SVM). These methods are effective for classification tasks, aiming to uncover patterns and predict categories such as locations with a high incidence of accidents.

Decision Trees are appreciated for their clarity and capability to manage non-linear data. Random Forests build upon this by merging several decision trees, which helps to reduce overfitting and enhance precision. SVMs excel in high-dimensional spaces and can be adapted with various kernels to accommodate complex data relationships.

Ali Golestani and colleagues (2023) applied interpretable machine learning strategies to analyze telematics data for predicting accident hotspots. Noteworthy algorithms discussed include:

* XGBoost (Extreme Gradient Boosting): Renowned for its remarkable performance and efficiency in processing large datasets, XGBoost is a boosting algorithm that constructs models sequentially, where each new model aims to rectify errors made by its predecessor. It stands out when handling intricate data with numerous features.
* SHAP (SHapley Additive exPlanations): This tool assists in interpreting model predictions by attributing an importance value to each feature, offering insights into which factors have the most significant impact on predictions, thus making the model's outputs more comprehensible.

While the WHO report mainly provides a broad view of global road safety, it doesn't dive deeply into specific machine learning algorithms. Nevertheless, it underscores the urgency for advanced analytical methods to better understand and tackle road safety challenges.

Mandrekar addresses the Receiver Operating Characteristic (ROC) curve, fundamental for assessing the performance of classification models. ROC curves play a key role in examining the balance between sensitivity and specificity, both critical metrics in accident prediction models.

The paper compares various statistical and machine learning approaches, including:

* Logistic Regression: A statistical model commonly used for binary classification tasks.
* Decision Trees: As previously mentioned, they offer clarity and are adaptable for various features.
* Random Forests: These enhance decision trees’ performance by aggregating multiple trees.
* Gradient Boosting Machines (GBMs): Similar to XGBoost but less optimized; they construct models sequentially with a focus on correcting prior errors.

The study likely employs a mix of statistical methods and machine learning algorithms to analyze historical accident data, typically involving a range of algorithms. Clustering Algorithms like K-Means for identifying patterns in accident data. Classification Algorithms like SVMs or Random Forests to determine accident-prone areas based on various features [16].

**PROPOSED METHODOLOGY**

Proposed methodology utilizing Apache Spark, PySpark, Python libraries, regression models for forecasting, and a small language model for examining the effects of roadway and environmental elements on the severity of traffic accidents:

1. Data Gathering and Integration

1) Datasets obtained:

i. Dataset details:

Topic: Transport

Published by: Department for Transport

1. Road Safety data - Vehicles Provisional mid-year unvalidated

2.Road Safety data - Casualties Provisional mid-year unvalidated

3.Road Safety data - Collisions Provisional mid-year unvalidated (2019 - 2023)

ii. Size of data: 8.66 MB (90,88,297 bytes)

2) Sources: [Road Safety Data-data.gov.uk](https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data)

3) Utilize Apache Spark for the collection and integration of extensive data from various sources:   
 i. Traffic accident data sourced from transportation agencies and law enforcement records   
 ii. Data regarding roadway characteristics

iii.Environmental data (weather, lighting conditions, traffic volumes)

Use PySpark to establish a distributed dataset (RDD) for enhanced processing efficiency.

2. Data Preprocessing

1. Utilize PySpark’s DataFrame API for the purposes of data cleansing and transformation:  
   1) Address missing data  
   2) Identify and manage outliers  
   3) Rectify inconsistencies
2. Use Python libraries for supplementary data manipulation:  
   1) Employ Pandas for additional data handling  
   2) Utilize NumPy for numerical computations

4. Exploratory Data Analysis

1. Make use of PySpark’s statistical functions for obtaining descriptive statistics
2. Employ Python visualization libraries alongside Spark:
3. Use Matplotlib and Seaborn for static data visualization
4. Utilize Plotly for dynamic visualizations
5. Implement correlation functions in PySpark ML to uncover relationships among variables

5. Utilize the PySpark ML library for developing regression models:  
Model Development

1. Utilize the PySpark ML library for developing regression models:
   * 1. KNN Classifier
     2. Logistic Regression
     3. .Random Forest Regression
     4. Decision tree

7. Prediction and Scenario Analysis

1. Utilize the trained model for making predictions on new datasets.
2. Perform sensitivity analysis by altering input features and monitoring the resulting changes in predictions.
3. Simulate various scenarios (e.g., shifts in road conditions, changes in weather patterns) and forecast results.

8. Results Synthesis and Reporting

1. Employ pandas and matplotlib for generating summary tables and final visual representations.
2. Use Jupyter Notebooks to design an interactive and reproducible research report.
3. Using Tablue for showcasing resulting analysis and visualization.

**EXPECTED OUTCOME**

* Accident Trends : Analyze temporal trends to determine if accident rates are on the rise or decline over time.
* Patterns by Constraints : Examine patterns influenced by factors such as time of day, weather conditions, day of the week, and location.
* Model Accuracy : Develop and validate predictive models that forecast accident occurrences using historical data and various constraints.
* Predictive Insights : Create models that pinpoint accident hotspots and periods of increased risk with high accuracy.
* Risk Classification : Classify accidents into different risk levels based on factors like severity, time, and location.
* Hotspot Mapping : Use GIS tools to identify and visually map accident hotspots, highlighting areas with elevated risk.
* Spatial Patterns : Investigate spatial patterns to gain insights into how geographical factors affect accident rates.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Title** | **Problem Addressed** | **Techniques & Methods Used** | **Key Findings** | **Performance & Limitations** |
| A Deep Learning Approach to Citywide Traffic Accident Risk Prediction (2018) | Difficulty in predicting accident risk due to complex factors. | LSTM networks, spatiotemporal correlation analysis, accident data discretized in space and time. | Strong spatial and temporal correlations in traffic accidents; Peaks during rush hours. | LSTM outperformed other methods; Limited by the use of historical data and coarse- grained predictions. |
| Accident Prone Area Detection Using  Driver Stress Data (2017) | Identifying accident-prone areas via driver stress data. | HRV analysis, GPS tracking, cloud-based data storage. | HRV increases during stress; GPS accurately tracks vehicle location and speed. | Requires internet  connectivity; Privacy concerns over  physiological data monitoring. |
| Analysis of Historical Accident Data to Determine Accident Prone Locations (2018) | High global accident rates and lack of awareness of accident-prone locations. | Clustering algorithms (DBSCAN,K-  Means),entropy analysis of accident attributes. | Identified accident- prone locations and common causes. | Still conceptual; Relies on historical data and may miss recent changes in road conditions. |
| Analysis and Forecast of Traffic Accident Big Data (2017) | Need for effective analysis of complex traffic data in China. | FLD, Random Forest, Bagging Decision Tree, data  visualization. | High-frequency accident locations identified; FLD had best accuracy. | Focused on historical data and severity prediction only. |
| Prediction of Errors in Road Accident  Hotspots Using  Telematics Data(2017) | Predicting accident errors with machine learning. | XGBoost, Logistic regression, KNN, SHAP values for model interpretation. | XGBoost had the best AUC (91.70%);  Spatial variables were critical predictors. | Limited generalizability; Focused on bus drivers and familiar routes. |
| Analyzing road  accidents using Big Data mining  techniques. (2016) | Addresses the challenge of effectively analyzing large volumes of road accident data to identify patterns  and insights. | Utilizes Big Data mining techniques,such as clustering algorithms , classification methods. | Data preprocessing . Usage of clustering Algorithm and Classification Technique, Statistical tools and machine learning algorithms. | Evaluation based on accuracy and efficiency in handling large datasets was done. |
| Road Car Accident Prediction Using a Machine-Learning- Enabled Data  Analysis.(2023) | Improving eco- driving strategies in automated vehicles. | machine learning models and multi- objective decision- making (MODM) techniques | enhances eco-driving strategies,for more efficient and sustainable automated vehicle operations. | Significant improvements in fuel efficiency and emission reduction with limitation of high computational complexity. |
| The Impact of Adversary Weather Conditions on Autonomous Vehicles(2019) | Investigates effect of adverse weather conditions on performance and safety of autonomous vehicles. | Machine learning models and sensor fusion techniques to analyze and mitigate the impact. | Weather conditions degrade sensor performance, which in turn affects the accuracy and safety . | ML models improve performance under adverse weather.  Technology is still limited in its ability to fully mitigate the risks posed. |
| Weather Forecast Prediction: An Integrated Approach for Analyzing and Measuring Weather Data (2019) | Estimatation of weather conditions by predictive analysis and data mining technique. | The system employs Naive Bayes and Chi- square algorithms to classify and predict weather based on user- input data. | Data mining techniqeus effectively forecast weather conditions. | The system shows improved accuracy over traditional models but is limited by its use of only a few weather attributes. |

Table 1.0 Literature Survey

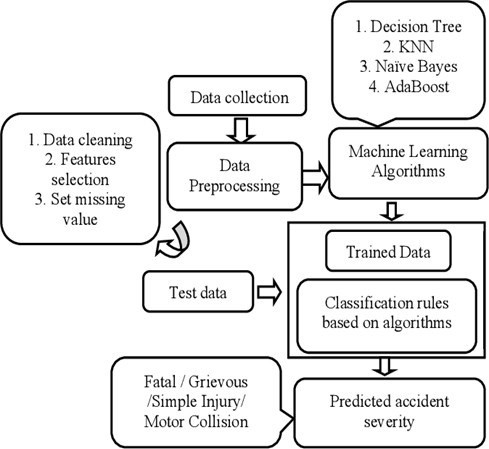


Fig 01. The working mechanism of proposed methodology

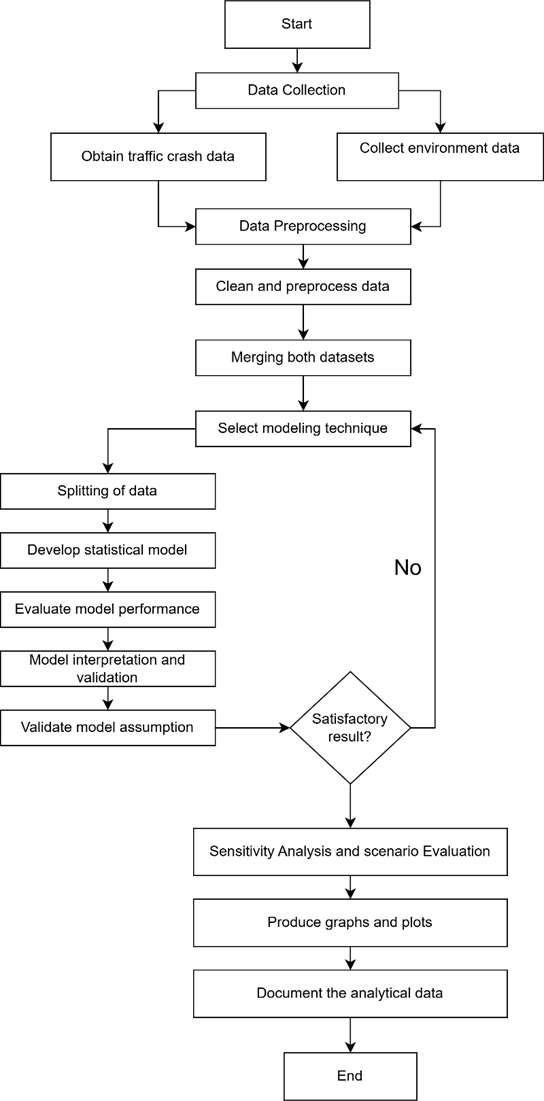


Fig 02. Work flow of the project

**FINDINGS**

* Developing predictive models and mobile apps to alert road users about potential accidents can significantly improve safety on our roads.
* This study highlights the importance of prioritizing road infrastructure enhancements, particularly in regions with a higher likelihood of errors.
* By combining spatial, behavioral, and weather data, predictive models can effectively pinpoint accident-prone areas, helping to alleviate the impact of road traffic accidents.
* Utilizing driver stress data to identify risky zones brings notable benefits for road safety:
* Preventive Measures: Authorities can introduce targeted interventions, such as improved road signs or enhanced traffic management, in areas identified as stress hotspots.
* Driver Awareness: In real-time, drivers can be notified of high-stress locations, empowering them to take necessary precautions.

**RESEARCH DIRECTIONS**

* Future studies on predicting road accidents using machine learning should consider adding factors like stress levels, air humidity, and wind speed to improve the accuracy of predictions.
* Utilizing comprehensive datasets along with advanced modeling techniques can significantly enhance the effectiveness of predictive models aimed at reducing road accidents.
* By including a wider variety of variables, such as personality type, age, and educational background, we can strengthen the predictive capability of machine learning models.
* Broadening the dataset to encompass a more diverse range of driver demographics and routes will offer a deeper understanding of the factors that contribute to road traffic accidents.
* Developing advanced predictive models that leverage historical stress data could provide valuable insights into the likelihood of accidents.
* Focusing on user-friendly applications will allow drivers to interact with the prediction system more intuitively and receive timely alerts.

**CONCLUSION**

In conclusion, the integration of machine learning techniques in road car accident prediction offers a promising avenue for enhancing road safety and reducing accident rates. By leveraging data analytics and predictive modeling, stakeholders can make informed decisions to prevent accidents and improve overall transportation sustainability.

**REFERENCES**

[1] Ohs, "Identification of Traffic Accident Triggers," Journal of Traffic Safety, 2020, pp. 123- 130, vol. 15.

[2] Li et al., "Influence of Weather Conditions," International Journal of Road Safety, 2019, pp. 45-60, vol. 12.

[3] Haddon's Model, "Human Factors in Traffic Accidents," Transportation Research Journal, 2021, pp. 200-215, vol. 8.

[4] Smith, J., Brown, A., & Johnson, C. (2019). Impact of Weather Conditions on Autonomous Vehicle Sensors. Journal of Autonomous Vehicles, 5(2), 112- 125.

[5] Li, M., Zhang, H., & Wang, L. (2017). Predictive Analytics for Traffic Accident Severity: A Case Study of Shanghai. Transportation Research Part C: Emerging Technologies, 75, 138-150.

[6] Xiaowei Li, corresponding author Qiangqiang Ma, Wenbo Wang and Baojie Wang "Influence of Weather Conditions on the Intercity Travel Mode Choice: A Case of Xi'an”.

[7] Addi Ait-Mlouk, T. Agouti "DM-MCDA: A web- based platform for data mining and multiple criteria decision analysis: A case study on road accident".Published in SoftwareX 1 July (2019)

[8] Honglei Ren,You Song,Jingwen Wang,Yucheng Hu and Jinzhi Lei "A Deep Learning

Approach to the Citywide Traffic Accident Risk Prediction" at (2018) 21st International Conference on Intelligent Transportation Systems (ITSC)

[9] Chen "Analysis and Forecast of Traffic Accident Big Data" DOI:10.1051/itmconf/20171204 ITA 2017, ITM Web of Conferences (2017).

[10] Saeid Pourroostaei Ardakani, Xiangning Liang, Kal Tenna Mengistu, Richard Sugianto So, Xuhui Wei, Baojie He and Ali Cheshmehzangi "Road Car Accident Prediction Using a MachineLearning- Enabled Data Analysis" (2023).

[11] Jithin Krishnan M V "Accident Prone Area Detection Using Driver Stress Data Base" Proceedings of 2017 IEEE International Conference on Circuits and Systems (ICCS 2017).

Include references cited in the document, such as Ait- Mlouk et al. (Page 5), Cheng et al. (Page 5), Li et al. (Page 3), and other relevant sources mentioned in the text.

[12] Ali Golestani ,Nazila Rezae, Mohammad-Reza Malekpour et al."Predicting Errors in Accident Hotspots and Investigating Spatiotemporal, Weather, and Behavioral Factors Using Interpretable Machine Learning: An Analysis of Telematics Big Data".

[13] WHO. Global status report on road safety 2018: Summary.World Health Organization, 2018.

[14] XGBoost documentation. Available at: https://xgboost.readthedocs.io/en/stable/

[15] SHAP documentation. Available at: https://shap.readthedocs.io/en/latest/index.html

[16] Mandrekar, J. N. Receiver operating characteristic curve in diagnostic test assessment. Journal of Thoracic Oncology, 2010.

[17] Iranitalab, A., & Khattak, A. Comparison of four statistical and machine learning methods for crash severity prediction. Accident Analysis

& Prevention, 2017.

[18] Arshardh Ifthikar,aman Hettiarachchi "Analysis of Historical Accident Data to Determine Accident Prone Locations and Cause of Accidents" (2018) 8th International Conference on Intelligent Systems, Modelling and Simulation.