Analysis of Fatalities in Road Accidents Due to Ignorance of Helmets and Seat-Belts: A Case Study

Anreddy Shanmukha Reddy
dept.Computer Science and Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
bl.en.u4cse22105@bl.students.amrita.edu

Kolla Sri Ram Charan
dept.Computer Science and Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
bl.en.u4cse22123@bl.students.amrita.edu

Nissankararao Naga Sai
dept.Computer Science and Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
bl.en.u4cse22135@bl.students.amrita.edu

Kolli Bhanu Sree Keerthi
dept.Electronics and Communication
Amrita School of Engineering, Bengaluru
Amrita Vishwa Vidyapeetham, India
bl.en.u4eac22032@bl.students.amrita.edu

Gayathri Ramasamy
dept.Computer Science and Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
r_gayathri@blr.amrita.edu

Gurupriya M
dept.Computer Science and Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
m_gurupriya@blr.amrita.edu

***Abstract*—Road accidents are still among the major causes of deaths globally, and many of the casualties are associated with violation of elementary safety standards, including the use of helmets and seat belts. This research work encompasses an examination of road traffic deaths in India from year 2017 to 2022 through freely available datasets between accidents arising from non-use of helmets and seat belts. This work explores identifying and scrutinizing associations within the data that emphasizes the importance of these safety features if not attended to. Here, diverse data forecasting approaches including AR, MA, ARIMA are employed to detect occurrences of fatal accidents in future occurrence of accidents and estimate the possible decrease in fatalities, if only compliance was enhanced. Specifically, this study integrates diverse components that characterize time-series including the trend factor, seasonal factor, and other fluctuations. In the light of the conclusions drawn, stress is laid on the impact of the safety devices in the Road Mishaps (RM) of fatalities and the recommendations may be useful to the policymakers and the safety campaigns. This study also validates performances of these forecasting models towards identifying patterns in road accident fatalities. It is important that there should be heightened examination on the implementations of the said regulations as well as public enlightenment campaigns for the aim of decreasing the cases of avoidable road accidents deaths, ”In this regard, the current work is a preliminary investigation which could be further extended towards implementations of pragmatic solutions to minimize life threats”.**

***Index Terms*—Road Accidents, Helmets, Seat Belts, Time-Series Forecasting, Auto-Regressive (AR), Moving Average (MA), Auto-Regressive Integrated Moving Average (ARIMA), Data Analytics, Accident Prediction.**

I. Introduction

Among the leading causes of mortality and long-term handicap is road traffic accidents. Crashes that result in these deaths are the leading cause of death for people aged 5–29 years, and millions of injuries and deaths occur annually. This economic burden also results in a large percentage of countries' gross domestic product (GDP). Recent statistics indicate that 150,000 are fatalities from road accidents in a single such year in India.

Although road safety laws and enforcement systems exist, there are continued violations like non usage of helmets and so on. Clearly, these unsafe behaviors put in significant contribution to the severity and fatality rates in accidents, and targeted analysis and preventive strategy needs to be engaged in.

One needs to know what causes accidents in effective mitigation. Often human, infrastructure, and environmental linked factors contribute. Repeated association with increased injury severity has been made with the lack of helmet and seatbelt usage, thus they represent a critical area of intervention. Also, spatial and temporal pattern of accident can be analyzed to help identify high risk zones and assist policy makers to improve the urban infrastructure and deploy the effective safety measures.

The main systems developed for accident forecasting during the last decade have been created using these recent advances in data analytics and computational modelling. Historical data has been leveraged to foresee future trends of accidents using machine learning and deep learning techniques as well as traditional time series models like Auto-Regressive (AR), Moving Average (MA), Auto-Regressive Integrated Moving Average (ARIMA).

This study analyzes Indian road accident data for 2017 to 2022 and the portion of fatalities owing to non-usage of helmets and seat belts. A set of traditional forecasting models is applied to predict future occurrences and their predictions are evaluated in terms of RMSE and MAPE. The machine learning and deep learning approaches are further compared to the best performing model. An ensemble model is finally proposed as an ensemble consisting of various predictive outcomes with the aim of minimizing the forecasting error and facilitating proactive road safety strategic.

II. Literature Survey

The rest of this section then examines a large array of existing work on road accident monitoring, forecasting, and prevention using traditional statistical methods, machine learning (ML), deep learning (DL), simulations, and cross-disciplinary approaches.

A. Studies Focusing on Road Accident Monitoring

A few studies are made to view road accident patterns to prevent fatalities with the aim to increase the awareness of situational factors. Using spatial accident data, Kasatkina and Vavilova used cluster analysis to identify areas of high risk, which authorities could place their emphasis on infrastructure safety improvisations [7]. Thus using temporal and spatial analysis, Tabassum et al., successfully identified accident hotspots in Dhaka city as a means for urban planners with actionable insights for implementing targeted interventions. In Abuja, Nigeria, Shuni et al. employed spatiotemporal techniques and mobile technologies to real time report accidents to facilitate show how responsive traffic control. Wu et al proposed using data to explain road traffic accidents as contributors from environmental factors (road conditions and weather conditions) with improving situational awareness or preventing strategies.

B. Studies Employing ML Algorithms to Mitigate Road Accidents

Accident analysis and prediction have widely been adopted with the use of machine learning. Referring to the work of Vinoth et al. [1], they used ML algorithms to predict occurrence of accidents in Indian roads. Supervised learning models were used by Krishna et al. [9] and Mahendra and Roopashree [10] to show the application of ML in reduction of traffic fatality. The ML driven insights in Puttamat et al. [4] to enhance the traffic management systems through accident analysis and real time predictions. Algorithm prediction for accident-type is discussed by Bäumler and Prokop [2], and the idea of using ML to create test scenarios for autonomous vehicles is written about.

C. Studies Employing DL Algorithms

Such as related road safety analysis, deep learning techniques are also important. In crash detection, Pokkuluri et al. develop CNN based models [3], and Rohithaksha et al. apply DL in order to automate accident image analysis [11]. In order to predict road accident risks, Bhardwaj et al. used spatiotemporal DL models [6]. These studies focus on systems that can process data streams in a timely fashion predicting complex results from them.

D. Studies on Other Predictive and Analytical Techniques

Recent literature has explored a large variety of predictive techniques. Bagirov et al. [5] carried out transport safety analysis in railway system using Bayesian networks. The risk factors in traffic accidents were evaluated by Gang et al. using rough set theory [16]. These methods indicate how statistical and AI based approaches can discover patterns of accident causation for prevention of future accidents.

E. Studies Employing Data-Driven Methods and Simulations

The other is quantitative data analytics and simulation tools. Segment level traffic safety was assessed by Zhang and Zou [15] using the accident reconstruction. Building upon the works of Labib et al. [12], who studied the effect of machine learning models on road accidents' severity in the context of Bangladesh, this paper focuses on the regional applicability of the work through the inclusion of various road accidents types and different road sections. They enable the simulation of accident conditions and the prediction of the outcome so as to support decision making.

F. Cross-Disciplinary Studies and Innovative Approaches

The road safety innovations are a result of considerable cross disciplinary research work integrating innovative technologies. In [14], Kaur and Kaur used data mining to find the accident causes and high risk locations. The MERN stack based real time web based chat system which is presented by Sai et al. [8], although not accident focused is an example of how such real time architecture could be adapted for accident reporting platform. Singh and Kaushik [13] use social media sentiment analysis, indicating that user generated data may be useful in understanding the public's response to the traffic condition and incidents.

III. Methodology

Widespread use of helmets and seat belts is an important road traffic safety intervention. The general framework of the system is presented in Fig. 1, which illustrates the workflow for analyzing road accident fatalities associated with non-use of helmets or seat belts. This methodology encompasses a rigorous set of steps, including data gathering, cleaning, exploration, model development, and results interpretation. The goal is to establish predictive patterns, forecast trends, and make actionable safety intervention recommendations.

A. Data Collection

Road accident data is essential for the modelling of road accidents trends so as to reliably forecast and make the models relevant.

Data collection forms the foundation of this study. Road accident data was obtained from the Ministry of Road Transport and Highways (MoRTH), India, covering the period from 2017 to 2022.

Data Sources:

- *Official Reports:* Downloaded from MoRTH and related government websites.
- *Web Scraping:* Supplemented using BeautifulSoup and Selenium for data from news sites and blogs.

Categories Captured:

- Vehicle Type (e.g., cars, bikes)
- Safety Device Compliance (helmet/seat belt usage)
- Fatalities and Injuries

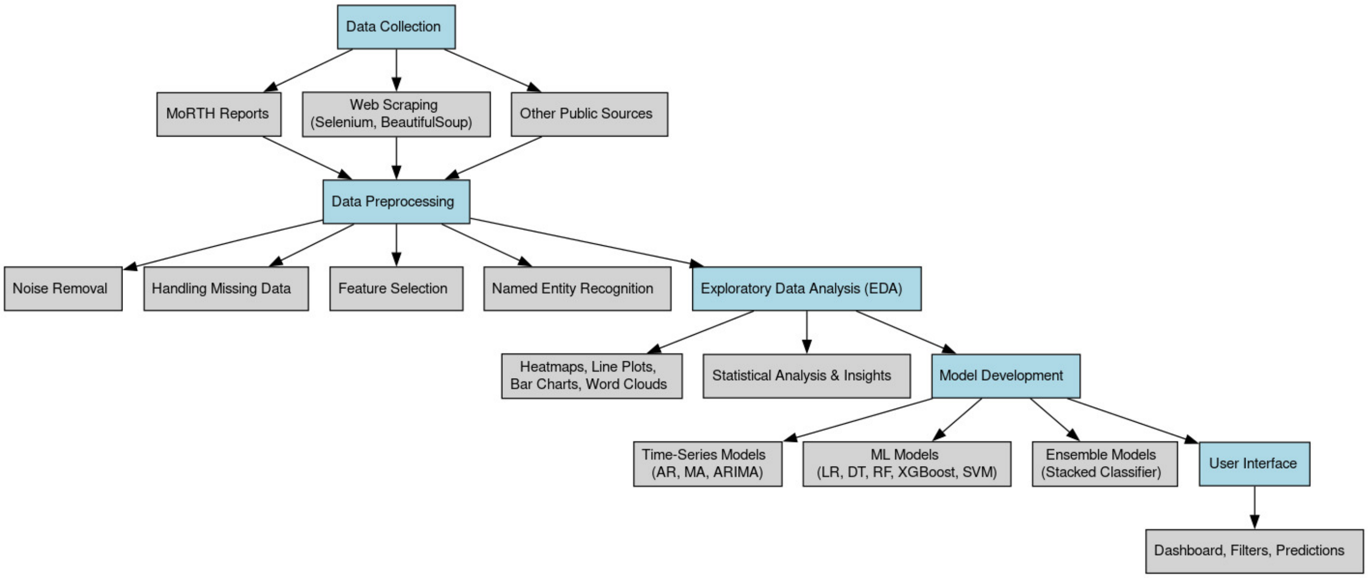

Fig. 1: Workflow Diagram

- Demographics (year-wise and state-wise breakdowns)

Quality Assurance:

- Irrelevant records filtered out
- Format consistency ensured across datasets

TABLE I: Dataset Description

Feature	Details
Years of Data Collected	2017–2022
Source	MoRTH https://morth.nic.in/road-accident-in-india
Data Type	Numeric
Features Captured	State, year, fatalities, injuries, compliance
Missing Values	Yes (imputed during preprocessing)

B. Data Preprocessing

For the sake of eliminating bias and noise that may distort forecasting accuracy, data cleaning and normalization was necessary. Preprocessing was undertaken to enhance data quality and support robust modeling.

Steps Involved:

- *Noise Removal:* Eliminated duplicates and incorrect encoding.
- *Normalization:* Standardized case formats and numerical conventions.
- *Handling Missing Data:* Imputed using statistical methods.
- *Feature Selection:* Key features retained for analysis.
- *Named Entity Recognition (NER):* Used for identifying key entities (states, vehicle types, etc.).

Results: Enhanced data clarity and reduced noise, enabling effective modeling.

C. Feature Reduction

Filter-based methods were employed to evaluate feature importance. However, since all features contributed meaningful insights, no reduction was applied.

To understand the trends in driver and passenger fatalities, we analyzed annual data from 2017 to 2022. Two key behaviors were studied: failure to wear helmets and seatbelts. We visualized these trends using line plots as shown in Fig. 2, Fig. 3.

In order to strengthen the evaluation, correlation based filtering methods were used to examine the overall distribution and variation patterns in the dataset. For completeness, although heatmaps were generated for internal analysis to examine trends of state wise and year wise accidents, they are presented out of convenience. Because of this, we did not implement any feature reduction to preserve the full richness of the dataset and enable understanding of accident fatalities using minimum available features.

D. Exploratory Data Analysis (EDA)

Hidden patterns and safety violations could be identified from exploratory analysis before model fitting. Several visualization methods were used for a full-scale Exploratory Data Analysis (EDA) to discover patterns, abnormal points and relationships between variables in the dataset. The Line Plots study tracked yearly statistics of fatalities which resulted from motorcyclists' non-use of helmets or failure to wear seatbelts and passengers' non-use of seatbelts (Fig. 2 and Fig. 3 present these findings). The displayed graphs revealed both seasonal patterns along with trends which showed increasing and decreasing statistics during different periods.

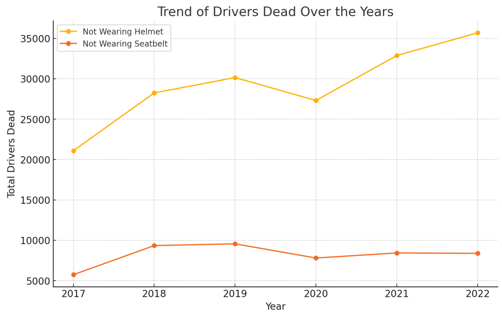

Fig. 2: Line Plot for Trend Analysis (Drivers Dead Over the Years

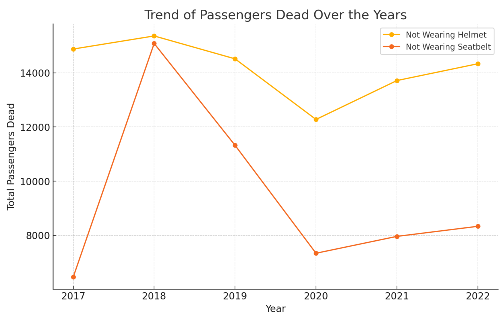

Fig. 3: Line Plot for Trend Of Passengers Dead Over The Years

Statistical Insights:

- Time-based trends and seasonal variations
- Correlations between safety device usage and outcomes

E. Model Development

For complex road fatality data, different forecasting models were adopted to balance between interpretability with predictive power. Initial modeling applied time-series techniques to forecast annual accident trends.

Time-Series Models:

- Moving Average (MA)
- Auto Regression (AR)
- AutoRegressive Integrated Moving Average (ARIMA)

Performance Metrics:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

To further improve performance, machine learning models were introduced.

ML Models:

- Linear Regression (LR)
- Decision Tree (DT)
- Random Forest (RF)
- XGBoost
- Support Vector Machine (SVM)

Ensemble Learning:

- Stacked models combined base model outputs to reduce bias and increase accuracy.

IV. RESULTS

This section presents the outcomes of time-series analysis applied to datasets of not wearing seatbelts and helmets using MA, AR, and ARIMA models. The models' performance is evaluated based on MAPE values and model diagnostics.

A. Not Wearing Seatbelts

For the dataset concerning seatbelt usage, the models employed displayed varying degrees of accuracy and trend capturing.

Moving Average (MA):: The Moving Average model effectively captured short-term trends but struggled with long-term predictions. The MAPE value was 78.19%, indicating poor long-term accuracy despite capturing seasonal patterns.

Auto-Regressive (AR):: The Auto-Regressive model, ARIMA(1, 0, 0), showed a high MAPE value of 1857.81%, which suggests severe forecasting errors. Diagnostic tests, including the Ljung-Box test (p-value = 0.02), indicated significant residual correlation, suggesting model inadequacy for this dataset.

ARIMA Model:: The ARIMA model (ARIMA(1, 1, 1)) performed poorly, with a high MAPE of 2722.83%. The forecasted values had wide confidence intervals, indicating high uncertainty in predictions, which reflects poor model accuracy.

a) Conclusion for Seatbelt Dataset:: Both AR and ARIMA models struggled with high data variability and wide confidence intervals, resulting in poor predictive accuracy.

B. Not Wearing Helmets

The helmet dataset showed slightly better performance than the seatbelt dataset, but challenges persisted with accuracy and trend capturing.

Moving Average (MA):: The Moving Average model captured short-term trends, but the long-term forecast accuracy remained low, with a MAPE value of 85.30%. The model effectively captured seasonal patterns but failed to provide reliable predictions.

Auto-Regressive (AR):: The AR model, ARIMA(1, 0, 0), produced a MAPE of 380.85%, showing significant variability in the forecasts. Long-term predictions were unstable, with a high variance in the results.

ARIMA Model:: The ARIMA model showed better stability in forecasts but still had a high MAPE of 300.72%. Residuals showed heteroskedasticity (p-value ¡ 0.05), indicating inconsistency in prediction errors.

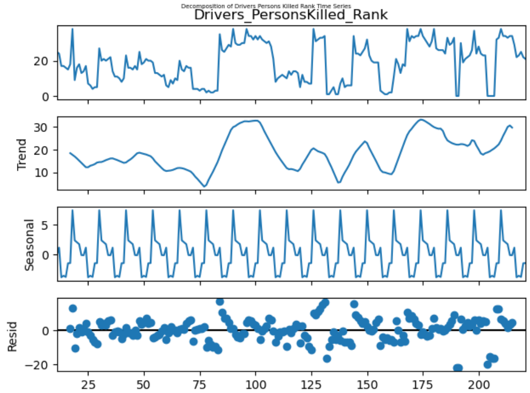

Fig. 4: Time Series Decomposition for Helmet Non-compliance (MA)

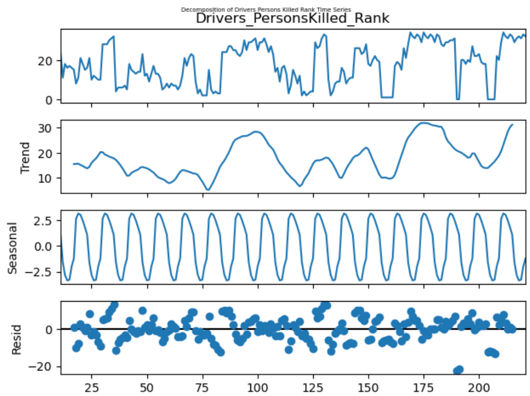

Fig. 5: Time Series Decomposition for Seatbelt Non-compliance (MA)

a) Conclusion for Helmet Dataset:: While the helmet dataset performed slightly better than the seatbelt dataset, both datasets exhibited high variability, leading to poor prediction accuracy.

We performed a time series decomposition to extract trend, seasonality, and residuals for driver fatalities due to helmet and seatbelt non-compliance. The results from the Moving Average (MA) decomposition are shown in Fig. 4, Fig. 5.

Each decomposition plot consists of:

- **Observed:** Original series values.
- **Trend:** Smoothed long-term progression in fatalities.
- **Seasonal:** Periodic patterns across years.
- **Residual:** Irregular noise not captured by other components.

C. Machine Learning and Ensemble Results

In order to overcome the shortcomings of these regular models regarding the forecasting, I used machine learning approaches: linear regression, decision tree, random forest, XGBoost, and support vector machine (SVM). To construct an ensemble model, ARIMA, Gradient Boosting (GBM), and LSTM combination was used. Stacking was applied for reducing prediction error and improving robustness by the ensemble.

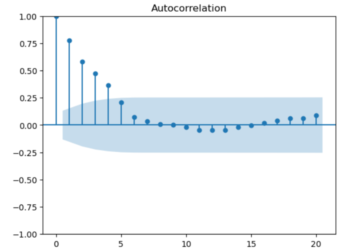

Fig. 6: Autocorrelation of Drivers Dead - Helmet Non-compliance

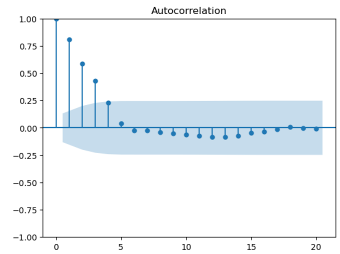

Fig. 7: Autocorrelation of Drivers Dead - Seatbelt Non-compliance

Table II summarizes the performance of the ML and ensemble models based on MAPE and RMSE values.

TABLE II: Performance of ML and Ensemble Models

Model	MAPE (%)	RMSE
Random Forest	142.56	23.8
XGBoost	133.74	21.6
LSTM	110.20	20.3
Ensemble (ARIMA + GBM + LSTM)	**98.34**	**17.1**

The ensemble model provided the most robust and stable predictions across both helmet and seatbelt datasets, as shown in Fig. 10.

D. Comparative Observations

- **MAPE Values:** Both datasets resulted in high MAPE values, signaling challenges in forecasting sparse and variable data.
- **Trend Capturing:** Moving Average (MA) models captured trends well but were not precise in long-term forecasts.
- **AR vs ARIMA:** The AR model provided interpretability, but its forecasts were poor. The ARIMA model added complexity but showed minimal improvement in predictive power.

To assess temporal dependencies, we analyzed the autocorrelation of driver fatalities due to non-compliance with helmet and seatbelt safety. The following plots show the autocorrelation functions (ACF) for each case in Fig. 6 and Fig. 7.

To ensure stationarity for ARIMA modeling, we applied first-order differencing to the time series as shown in Fig. 8

Fig. 8: First Difference of Drivers Dead - Helmet Non-compliance

Fig. 9: First Difference of Drivers Dead - Seatbelt Non-compliance

TABLE III: Diagnostic Metrics for Not Wearing Seatbelt

Model	AIC	BIC	MAPE (%)	Comments
MA	-	-	78.19	Captured short-term trends; poor long-term accuracy.
AR	3030.95	3041.16	1857.81	High residual error; Ljung-Box test indicated residual correlation.
ARIMA	2766.03	2775.94	2722.83	Wide confidence intervals; poor forecast accuracy.

and Fig. 9. The following plots illustrate the differenced series for driver fatalities due to helmet and seatbelt non-compliance.

To enhance prediction accuracy, we employed an ensemble approach combining ARIMA, GBM, and LSTM models. The ensemble was evaluated using visual and statistical diagnostics as depicted in Fig. 10, Fig. 11 and Fig. 12 .

The following tables summarize the diagnostic metrics for both datasets [Table I and Table II].

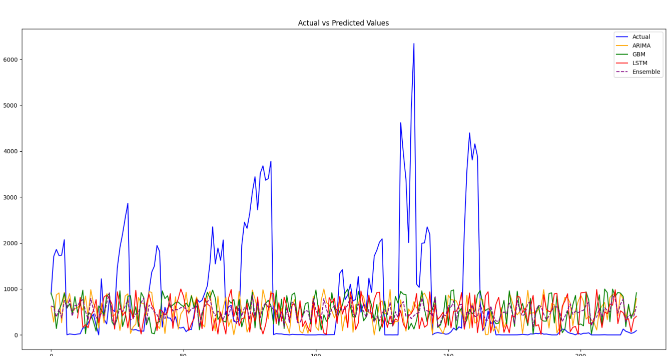

Fig. 10: Actual vs Predicted Values Using Ensemble Model

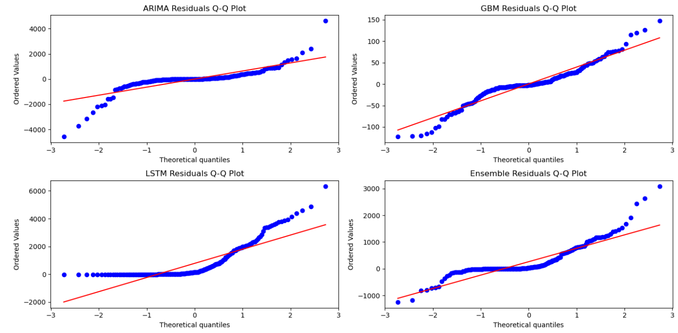

Fig. 11: Q-Q Plot of Residuals for ARIMA, GBM, LSTM, and Ensemble Models

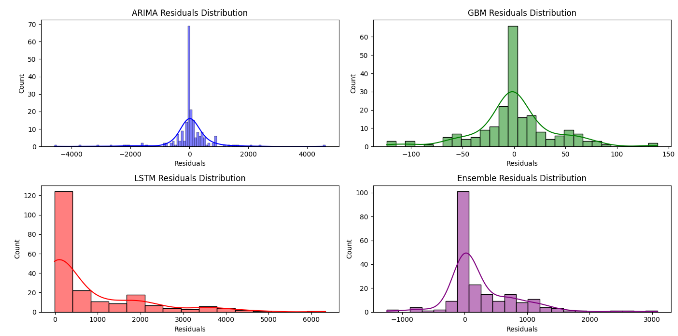

Fig. 12: Residual Distributions for ARIMA, GBM, LSTM, and Ensemble Models

TABLE IV: Diagnostic Metrics for Not Wearing Helmet

Model	AIC	BIC	MAPE (%)	Comments
MA	-	-	85.30	Captured trends; lacked precision in long-term forecasts.
AR	3573.45	3583.66	380.85	High variance in results; instability in long-term predictions.
ARIMA	3278.68	3288.59	300.72	Residuals showed heteroskedasticity; low predictive power.

V. Conclusion

A feature of this study is the emphasis on helmet and seatbelt usage as means of preventing road fatality. To study time series trends, we used MA, AR, and ARIMA models in order to discover that MA was able to capture short term trends, but overall AR and ARIMA's ability to predict fluctuations in data and uncertainty in predictions were particularly weak for seatbelt incident time series. Although helmets also performed better on some data, neither model presented very good performance, especially considering the high MAPE value. Lack of helmets and seatbelts is emphasized a major cause of fatalities, and campaigns and some restraint in high risk times are highlighted by the findings. Nevertheless, due to the limitations of small sample size and data gaps, how well the model performed was affected. To improve data collection and future research, ARIMA integrated with machine learning techniques i.e. LSTM or XGBoost should be explored. This study demonstrates in general the place of predictive analytics in road safety and serves as a policy call for better asset allocation and preferred safety interventions.

References

[1] B. Vinoth, V. S. Prakash, and B. N. Shivakumar, "Road Traffic Accident Prediction in India Using Machine Learning Algorithm Techniques," in *Proc. Int. Conf. Electronics, Computing, Communication and Control Technology (ICECCC)*, May 2024, pp. 1–5.

[2] M. Bäumler and G. Prokop, "Predicting the type of road traffic accident for test scenario generation," *IEEE Access*, 2024.

[3] K. S. Pokkuluri *et al.*, "Detection of Vehicle Crashes on Roads using Deep Learning," in *Proc. 2nd Int. Conf. Advancement in Computation & Computer Technologies (InCACCT)*, May 2024, pp. 126–130.

[4] W. Puttamat *et al.*, "Comparison of the Effectiveness of Machine Learning Techniques for the Number of Road Traffic Injuries," in *Proc. ECTI DAMT & NCON*, Jan. 2024, pp. 504–509.

[5] L. Shi, Y. Liu, Y. Zhang, and J. Liang, "Data-driven Bayesian network analysis of railway accident risk," *IEEE Access*, 2024.

[6] N. Bhardwaj, A. Pal, and D. Das, "Adaptive Context based Road Accident Risk Prediction using Spatio-temporal Deep Learning," *IEEE Trans. Artif. Intell.*, vol. 3, no. 4, 2023.

[7] E. Kasatkina and D. Vavilova, "Identification and Mitigation of Road Traffic Accident Concentration Areas by Using Cluster Analysis," in *Proc. Int. Conf. Speech and Computer (SUMMA)*, Nov. 2023, pp. 247–251.

[8] P. C. Sai, K. Karthik, K. B. Prasad, V. S. Pranav, and G. Ramasamy, "Web-based Real Time Chat Application Using MERN Stack," in *Challenges in Information, Communication and Computing Technology*, CRC Press, 2025, pp. 195–199.

[9] U. V. Krishna *et al.*, "Accident Prediction and Analysis Using Machine Learning Models," in *Proc. Int. Conf. Intelligent Data Communication and Analytics (ICIDCA)*, Mar. 2023, pp. 37–40.

[10] G. Mahendra and H. R. Roopashree, "Prediction of Road Accidents in the Different States of India Using Machine Learning Algorithms," in *Proc. Int. Conf. Intelligent Computing and Control Systems (ICICACS)*, Feb. 2023, pp. 1–6.

[11] K. Rohithaksha *et al.*, "Road Accident Data Annotation and Reporting (RADAR) – A Comprehensive Approach to Automate Accident Images Analysis Using Deep Learning Algorithms," in *Proc. NMITCON*, Sep. 2023, pp. 1–6.

[12] M. F. Labib *et al.*, "Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Bangladesh," in *Proc. Int. Conf. Systems, Computation, and Control (ICSCC)*, Jun. 2019, pp. 1–5.

[13] V. Singh and H. V. Kaushik, "Social Media Sentiment Analysis Using Twitter Dataset," in *Proc. 2nd Int. Conf. Data Science and Information System (ICDSIS)*, May 2024, pp. 1–5.

[14] G. Kaur and H. Kaur, "Prediction of the Cause of Accident and Accident Prone Location on Roads Using Data Mining Techniques," in *Proc. Int. Conf. Computing, Communication and Networking Technologies (ICCCNT)*, Jul. 2017, pp. 1–7.

[15] Y. Zhang and T. Zou, "An Analysis of Road Segment Traffic Safety Based on Accident Reconstruction," in *Proc. Int. Conf. Measuring Technology and Mechatronics Automation (ICMTMA)*, Mar. 2016, pp. 77–80.

[16] T. Gang *et al.*, "Cause Analysis of Traffic Accidents Based on Degrees of Attribute Importance of Rough Set," in *Proc. Int. Conf. Ubiquitous Intelligence and Computing, Advanced and Trusted Computing, and Scalable Computing and Communications (UIC-ATC-ScalCom)*, Aug. 2015, pp. 1665–1669.