

Article

Not peer-reviewed version

Predicting Traffic Accidents Using Artificial Intelligence: A Machine Learning Approach

Sania Afzal and [Manzoor Hussain](#) *

Posted Date: 11 November 2025

doi: 10.20944/preprints202511.0767.v1

Keywords: traffic analysis; prediction model; KNN model; decision tree; Naive Bayes; logistic regression; random forest



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Predicting Traffic Accidents Using Artificial Intelligence: A Machine Learning Approach

Sania Afzal ¹ and Manzoor Hussain ^{2,*}

¹ Department of Software Engineering, University of Sialkot, Sialkot, Pakistan

² Department of CS, Indus University, Karachi, Pakistan

* Correspondence: manzoor.hussain@indus.edu.pk

Abstract

Traffic accidents are a major global concern, resulting in significant human and economic losses. This study aims to explore the predictive capabilities of machine learning algorithms in identifying the factors that contribute to such accidents. Utilizing a comprehensive dataset that includes weather, road type, time of day, traffic density, speed limits, and driver attributes (e.g., age and alcohol consumption), we evaluated various classifiers including logistic regression, decision trees, naïve Bayes, KNN, and random forest. Experimental results show that KNN achieved the highest prediction accuracy. This work demonstrates the potential of ensemble and classification techniques for improving traffic safety through proactive accident prediction.

Keywords: traffic analysis; prediction model; KNN model; decision tree; Naive Bayes; logistic regression ; random forest

1. Introduction

Traffic accidents represent a critical global issue, exacting a high toll in terms of human lives, injuries, and economic costs. According to the WHO (World Health Organization), road traffic accidents are the leading cause of preventable deaths among young people aged 15–29 [1]. The steady growth in vehicular traffic and the diversity of driving environments have contributed to both the frequency and severity of accidents worldwide.

The rise in population and urbanization makes it very important to have improved traffic management and protective strategies. Historically, traffic accident research has concentrated on identifying the causes of accidents, focusing on environmental factors, driver behavior, and vehicle speed. For instance, adverse weather conditions[19], such as rain, snow, and fog, have been found to significantly increase accident risks [4]. Additionally, factors like driving age, experience, and alcohol consumption have been correlated with crash severity [3]. In order to develop accurate predictive models, it is essential to integrate diverse and multidimensional data, including road infrastructure, weather conditions, vehicle telemetry, and driver demographics. These heterogeneous data sources enable the construction of robust machine learning models[20–22] that can capture the complex dynamics leading to traffic incidents. Figure 1 illustrates the key categories of data needed for effective traffic accident prediction, highlighting how various environmental, infrastructural, and human-related variables contribute to the forecasting process.

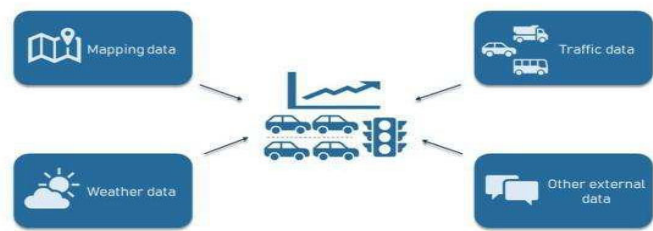


Figure 1. Data Needed for Traffic Prediction.

Hence there is a compelling need to integrate these variables into more advanced predictive models. The advent of machine learning and artificial intelligence offers promising new avenues to predict and manage traffic incidents [5,6]. With the development of connected vehicles and intelligent transportation systems, vast volumes of real-time data on traffic conditions, driver behavior, and environmental settings are now available. These datasets can be used to build complex probabilistic models that not only identify high-risk scenarios but also assist policy makers and traffic safety agencies in making informed decisions [7]. This study explores how modern data science and machine learning approaches can be applied to traffic accident prediction, particularly focusing on factors related to crash severity. The primary objective is to identify the most accurate predictors of accident occurrence and severity using a comprehensive dataset that includes weather conditions, road types, traffic density, and driver characteristics.

By comparing the effectiveness of advanced machine learning techniques such as neural networks and ensemble methods with traditional methods like logistic regression and decision trees, this research seeks to establish a universal analytical framework. This framework can support traffic safety planning, inform policy development, and enable proactive risk mitigation strategies.

2. Literature Review

The global burden of road traffic incidents remains severe. WHO reports over 1.35 million deaths annually due to traffic accidents [1], highlighting the urgent need for innovative traffic management and accident prevention strategies. Rapid urbanization, increased vehicle ownership, and inadequate road infrastructure, especially in developing countries, have exacerbated the situation. In India alone, over 151,000 deaths were recorded due to road traffic injuries in 2019 [2].

As the global car population is expected to reach 2 billion by 2025 [2], traffic congestion and the risk of accidents will rise accordingly. Analyzing and predicting traffic accidents is challenging due to the complex interplay of variables such as driver behavior, vehicle conditions, road environments, and geographic and climatic factors [4].

Traditional methods mainly based on statistical and physical models have limitations in capturing these interactions accurately. AI and ML have shown significant promise in addressing these gaps[13]. For instance, AI algorithms can process real-time data from cameras, sensors, and even social media to detect potential accidents before they happen [5,7]. But looking at the literature still lacks comprehensive models that incorporate diverse data sources[14–17]. Most studies have focused on accident frequency rather than severity, and many have been geographically limited. Variables such as driver behavior, socio-demographic factors, and heterogeneous data (weather, road types, and traffic intensity) are often underutilized [6,9].

The typical methodology for AI-based traffic prediction includes:

- 1. Data collection (historical crash records, weather, road geometry, driver traits).
- 2. Data integration (using GIS for geographic insights).
- 3. Model training using algorithms such as random forests, support vector machines, and artificial neural networks [6,10].

Random forests, for instance, perform well in scenarios with large datasets containing multiple parameters. Ensemble methods, which combine several models, further improve prediction accuracy by reducing overfitting [8]. Logistic regression also has been used to analyze the impact of infrastructure features such as road type, lighting conditions, and speed limits on crash severity [9]. Studies show that rural roads, major arteries, and poor lighting are associated with higher crash risks[11]. Despite these advancements, a lack of attention to pre-crash human factors persists. A landmark study categorizing 13,568 accidents found human error to be the primary cause in over 70% of cases [3]. Similarly, the integration of technologies like UAV has been proposed to enhance accident scene reconstruction and provide more accurate data under varied environmental conditions [6]. On the economic front, limited research has explored the valuation of road safety improvements. A study in Austria estimated the marginal value of safety (MVS) between AS 36–47 million per statistical life saved, suggesting a need for more efficient allocation of traffic safety resources [2].

3. Proposed Methodology

This study adopts a machine learning-based approach for traffic accident prediction by utilizing a diverse dataset that includes historical accident records, weather conditions, road types, traffic volume, and driver characteristics[12]. The data underwent preprocessing and feature selection, followed by the application of models such as Random Forest and Neural Networks[18]. Model performance was evaluated using cross-validation and standard metrics to ensure reliability and accuracy. This study adopts a machine learning-based approach for traffic accident prediction by utilizing a diverse dataset that includes historical accident records, weather conditions, road types, traffic volume, and driver characteristics. Figure 2 illustrates the overall methodological pipeline used in this study, starting from data preparation to model evaluation and result analysis.

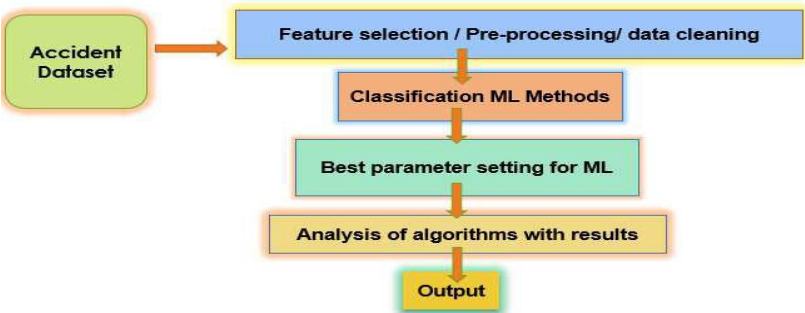


Figure 2. Machine Learning Pipeline for Traffic Accident Prediction.

3.1. Model-Specific Implementation for Accident Prediction

Machine learning techniques including KNN, decision tree, logistic regression, and random forest were applied for traffic accident prediction. The process began with data preprocessing such as handling missing values and feature scaling. The dataset was divided into training and testing sets, followed by model training and evaluation. Each model followed a distinct decision strategy, and their performance was assessed using accuracy scores and confusion matrices. Figure 3 provides a model-specific extension of the earlier pipeline, detailing the flow from data preprocessing to performance evaluation across various classifiers.

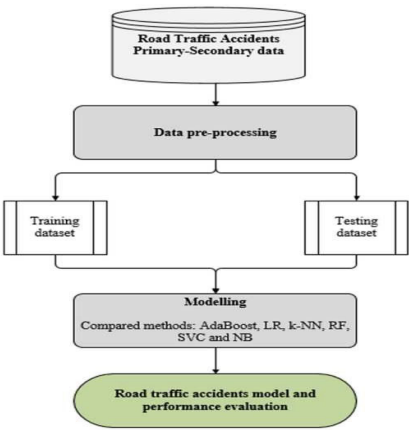


Figure 3. Model-specific process for traffic accident prediction using classification algorithms and evaluation metrics.

3.2. Dataset

The dataset used in this study was obtained from Kaggle and comprises detailed records of traffic accident cases. It includes 14 key attributes capturing a wide range of factors that influence road safety and traffic behavior. These features span environmental, temporal, and human-related variables, enabling comprehensive modeling and analysis of traffic incidents. Key features include weather conditions (e.g., Rainy, Clear, Foggy), road type (e.g., City Road, Highway), time of day (Morning, Afternoon, Evening, Night), and traffic density (on a scale from 0 to 2). Additional variables include speed limits, the number of vehicles involved, alcohol consumption by drivers, accident severity, and road conditions (e.g., Wet, Dry, Icy). The dataset as shown in Table 1 and Figure 3 also captures vehicle types and driver demographics such as age and experience.

Table 1. Featured Attributes of the Traffic Accident Dataset.

Attribute	description
WEATHER	The weather conditions at the time of the incident (e.g., Rainy, Clear, Foggy, Snowy, Stormy).
ROAD TYPE	The type of road where the incident occurred (e.g., City Road, Highway, Rural Road, Mountain Road).
TIME OF DAY	The time of day when the incident took place (e.g., Morning, Afternoon, Evening, Night).
TRAFFIC DENSITY	A numerical representation of traffic density, typically on a scale from 0 to 2.
SPEED LIMIT	The maximum speed limit on the road where the incident occurred, measured in kilometers per hour.
NUMBER OF VEHICLE	The total number of vehicles involved in the incident
DRIVER ALCOHOL	Indicates whether the driver was under the influence of alcohol (0 = No, 1 = Yes).
ACCIEDENT SAVERITY	The severity of the accident (e.g., Low, Moderate, High)
ROAD CONDITION	The condition of the road at the time of the incident (e.g., Wet, Dry, Icy, Under Construction).
TYPE OF VEHICLE	The type of vehicle involved in the incident (e.g., Car, Truck, Bus, Motorcycle).
DRIVER AGE	The age of the driver involved in the incident.

ROADLIGHT
CONDITION

The lighting conditions on the road (e.g., Daylight, Artificial Light, No Light)

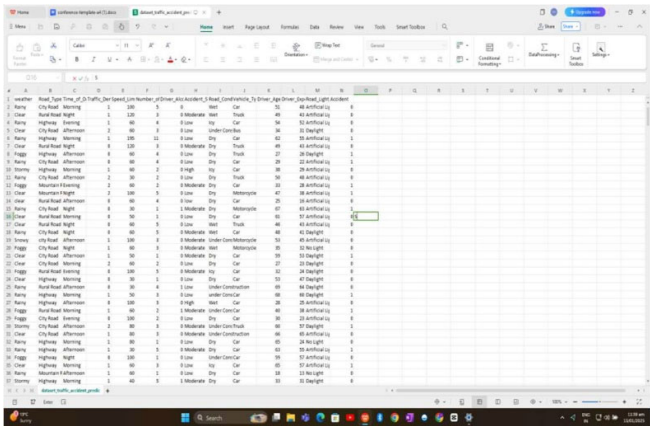


Figure 4. Dataset.

Model Selection

To evaluate classification performance for traffic accident prediction, several machine learning models were tested using RapidMiner, including k-Nearest Neighbors (kNN), decision tree, logistic regression, naive Bayes, and random forest. Among these, kNN demonstrated the highest accuracy and was selected as the final model for deployment. The process began with dataset importation into RapidMiner, followed by preprocessing to address missing values, which can significantly degrade model performance. A *Replace Missing Values* operator was employed to handle incomplete entries, as illustrated in Figure 5.

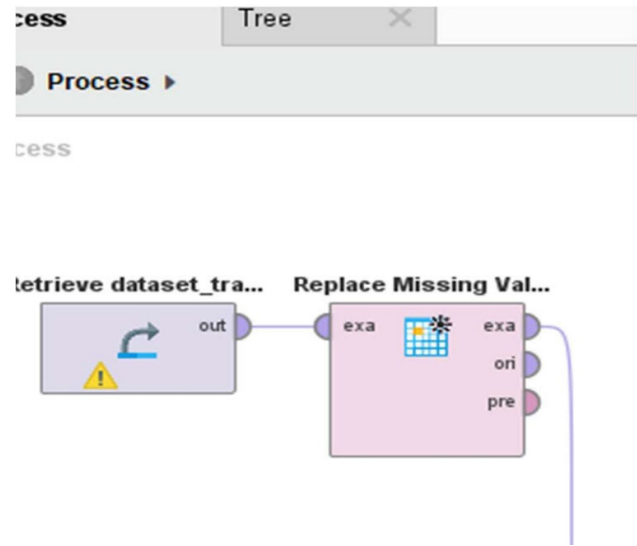


Figure 5. Replacing missing values using RapidMiner operator.

After data cleaning, the next step involved role assignment, where the target attribute was designated as the label. The data was then divided into training and testing sets using a 70:30 ratio through the Split Data operator (Figure 6).

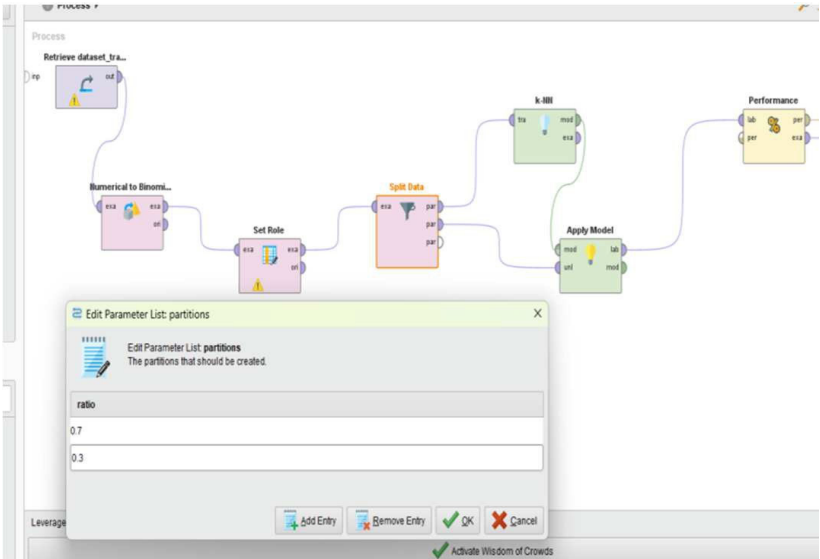


Figure 6. Data splitting into training and testing sets in RapidMiner.

Following the split, the kNN model was applied using the dedicated classification operator in RapidMiner. Finally, a *Performance* operator was included to evaluate model metrics such as accuracy and confusion matrix results. This workflow is shown in Figure 7.

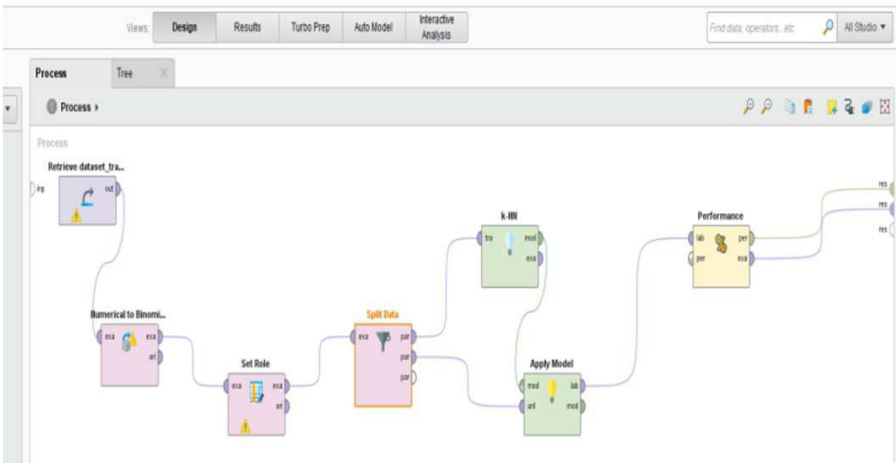


Figure 7. Applying the kNN model and evaluating performance in RapidMiner.

4. Results

This section presents the performance evaluation of the proposed system based on experimental results. To optimize the accuracy of the traffic accident prediction model, multiple classifiers were tested, including Decision Tree, Logistic Regression, Naive Bayes, and k-Nearest Neighbors (KNN). The accuracy of each model was recorded to identify the most effective approach. The comparative results are shown in **Table 2**.

Table 2. Accuracy of Different Classifiers.

Algorithm	Accuracy
KNN (N = 5)	81.85%
Decision Tree	42.23%
Naive Bayes	42.71%
Logistic Regression	60.71%

As shown in Table 2, the KNN algorithm outperformed all other classifiers with an accuracy of 81.85%. This clearly demonstrates its effectiveness in predicting traffic accidents using the given dataset. To further validate model performance, a confusion matrix was used to analyze true positives, true negatives, false positives, and false negatives. The evaluation results for the KNN model are visualized in **Figure 8**.

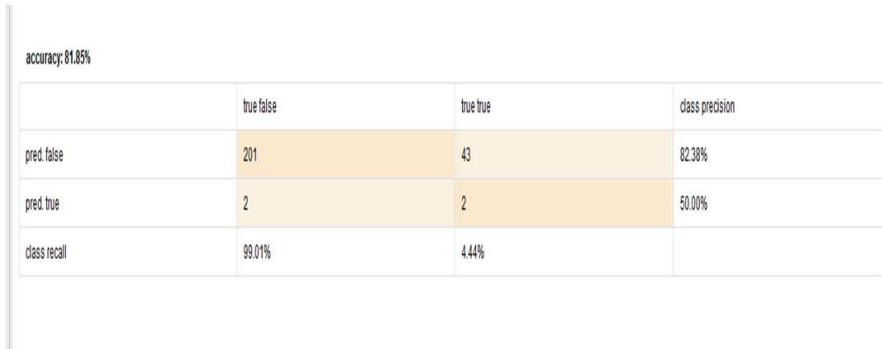


Figure 8. Performance metrics of the KNN model.

Additionally, a comparison with previous studies using the same dataset was conducted to highlight improvements in prediction accuracy. Table 3 summarizes accuracy levels achieved by other authors with various classifiers.

Table 3. Comparison with Previous Studies.

Author	Year	Classifier	Accuracy
[12]	2016	Decision Tree	42.23%
[12]	2016	Naive Bayes	42.71%
[10]	2023	Logistic Regression	60.71%
[11]	2020	Random Forest	78.78%
[11]	2020	KNN	78.53%

Compared to these prior results, the proposed KNN model demonstrated superior performance, achieving the highest accuracy of 81.85%. This improvement reflects the impact of proper preprocessing, feature selection, and parameter tuning applied in the current study.

5. Conclusions

In this research we examine the possibility of using machine learning for the prediction of traffic accidents together with a view of identifying the best method for the prediction of traffic accidents. Among all the algorithms employed in the current study, the KNN proved most effective in prediction mode since the highest accuracy of predictions was obtained. Not only the testing accuracy of KNN shows the best result but also indicated how well the algorithm successfully mapped the patterns and relationship of the data set for the traffic accident prediction problems. These results

support earlier studies in the field of this type of techniques, where such algorithms have shown certain efficiency. The meeting with prior studies also declare the viability of the utilized methodology and suggests the applicability of machine learning in the resolution of realistic issues concerning traffic safety. Data cleaning and transformation, as well as training and testing various models, were performed in RapidMiner which well meets the purpose of such tools. The advantage of using RapidMiner was the lack of wastage of lots of time developing the model and testing because of RapidMiner it was easy to complete both processes in a short duration and get a model with high accuracy. The study clearly shows how an creative accurate predictive model such as KNN can help improve traffic control and road safety solutions. The presented models may help decision-makers, traffic departments, and city administrators enhance preventive measures to decrease accident chances. Some work can be done in future for increase the accuracy, live data can also be employed to enhance further the effectiveness of those systems on the basis of predicting.

References

1. W. D. Glauz, K. M. Bauer, and D. J. Migletz, "Expected traffic conflict rates and their use in predicting accidents."
2. G. Maier, S. D. Gerking, and P. Weiss, "The economics of traffic accidents on Austrian roads."
3. J. R. Treat, N. S. Tumbas, S. T. McDonald, D. Shinar, R. D. Hume, R. E. Mayer, R. L. Stansifer, and N. J. Castellan, Institute for Research in Public Safety, Indiana University, Bloomington, Indiana, 47401.
4. J. R. Treat, "A study of precrash factors involved in traffic accidents," HSRI Research Review, vol. 10, no. 6, pp. 35, 1980.
5. Y. Wang and W. Zhang, "Analysis of roadway and environmental factors affecting traffic crash severities," Transportation Research Procedia, vol. 25, pp. 2119–2125, 2017, doi: 10.1016/j.trpro.2017.05.407.
6. A. M. Almeshal, M. R. Alenezi, and A. K. Alshatti, "Accuracy assessment of small unmanned aerial vehicle for traffic accident photogrammetry in the extreme operating conditions of Kuwait," Information (Switzerland), vol. 11, no. 9, 2020, doi: 10.3390/info11090442.
7. H. L. Gururaj et al., "Predicting traffic accidents and their injury severities using machine learning techniques," International Journal of Transport Development and Integration, vol. 6, no. 4, pp. 363–377, 2022, doi: 10.2495/TDI-V6-N4-363-377.
8. S. P. Ardakani et al., "Road car accident prediction using a machine-learning-enabled data analysis," Sustainability (Switzerland), vol. 15, no. 7, 2023, doi: 10.3390/su15075939.
9. A. Venkat, G. K. Vijey, and I. S. Thomas, "Machine learning based analysis for road accident prediction," 2020. [Online]. Available: <https://ssrn.com/abstract=3550462>
10. G. Prajapati, L. Kumar, and S. S. Rekha Patil, "Road accident prediction using machine learning," [Online]. Available: www.jsrtjournal.com
11. N. Fiorentini and M. Losa, "Handling imbalanced data in road crash severity prediction by machine learning algorithms," Infrastructures, vol. 5, no. 7, 2020, doi: 10.3390/infrastructures5070061.
12. S. Dhruvit and S. Pranay, "Road accident analysis and identification of black spot location on State Highway-5 (Halol-Godhra Section)," 2016. [Online]. Available: www.ijedr.org
13. D. Santos, J. Saias, P. Quaresma, and V. B. Nogueira, "Machine learning approaches to traffic accident analysis and hotspot prediction," Computers, vol. 10, no. 12, 2021, doi: 10.3390/computers10120157.
14. Das, S. R., Jhanjhi, N. Z., Asirvatham, D., Rizwan, F., & Javed, D. (2025). Securing AI-based healthcare systems using blockchain technology. In AI techniques for securing medical and business practices (pp. 333-356). IGI Global.
15. Khan, A., Jhanjhi, N., Hamid, D. H. H., Omar, H. A. H. B. H., Amsaad, F., & Wassan, S. (2025). Future Trends and Challenges in Cybersecurity and Generative AI. Reshaping CyberSecurity With Generative AI Techniques, 491-522.
16. Khan, M. R., Khan, N. R., & Jhanjhi, N. Z. (Eds.). (2024). Digital transformation for improved industry and supply chain performance. IGI Global.

17. Ashfaq, F., Jhanjhi, N. Z., Khan, N. A., Muzafar, S., & Das, S. R. (2024, March). CrimeScene2Graph: Generating Scene Graphs from Crime Scene Descriptions Using BERT NER. In International Conference on Computational Intelligence in Pattern Recognition (pp. 183-201). Singapore: Springer Nature Singapore.
18. Gill, S. H., Razzaq, M. A., Ahmad, M., Almansour, F. M., Haq, I. U., Jhanjhi, N. Z., ... & Masud, M. (2022). Security and privacy aspects of cloud computing: a smart campus case study. *Intelligent Automation & Soft Computing*, 31(1), 117-128.
19. Almulhim, M., Islam, N., & Zaman, N. (2019). A lightweight and secure authentication scheme for IoT based e-health applications. *International Journal of Computer Science and Network Security*, 19(1), 107-120.
20. Zaman, N., Low, T. J., & Alghamdi, T. (2014, February). Energy efficient routing protocol for wireless sensor network. In 16th international conference on advanced communication technology (pp. 808-814). IEEE.
21. Azeem, M., Ullah, A., Ashraf, H., Jhanjhi, N. Z., Humayun, M., Aljahdali, S., & Tabbakh, T. A. (2021). Fog-oriented secure and lightweight data aggregation in iomt. *IEEE Access*, 9, 111072-111082.
22. Ahmed, Q. W., Garg, S., Rai, A., Ramachandran, M., Jhanjhi, N. Z., Masud, M., & Baz, M. (2022). Ai-based resource allocation techniques in wireless sensor internet of things networks in energy efficiency with data optimization. *Electronics*, 11(13), 2071.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.