Article

Not peer-reviewed version

Predicting Consumer Purchase Intentions Using Machine Learning

Yaraib Arif and Rizwan Ayazuddin *

Posted Date: 11 November 2025

doi: 10.20944/preprints202511.0708.v1

Keywords: E-commerce prediction; machine learning models; shoppers behavior analysis

 doi:10.20944/preprints202511.0708.v1

Article

Predicting Consumer Purchase Intentions Using Machine Learning

Yaraib Arif [1] and Rizwan Ayazuddin [2,*]

[1] Software Engineering, University of Sialkot, Sialkot, Pakistan

[2] School of Computer Science, Taylor's University, Subang Jaya, Malaysia

* Correspondence: rizayazuddin@gmail.com

Abstract

This research investigates the application of machine learning methods in predicting the intention of online shoppers, a very important task in the fast-growing e- commerce industry. In this paper, the performance of some machine learning algorithms, namely K-Nearest Neighbors, Decision Tree, Naive Bayes, Random Forest, Random Tree, Gradient Boosting Tree, and Logistic Regression, is explored for forecasting online purchase behaviors. The features involved in the creation of the dataset for this study are enumerated as the traffic source, session time, amount of product pages visited, and finally, users' feedback. In such a way, the best models that were successful within the framework of the current paper were Gradient Boosting Tree with the rate 88.89% and Decision Tree with the following rate 88.89%, which helped predict their likelihood to buy. These models can help reduce predictive errors and mitigate the variations that exist in consumer behavior-for instance, on personal tastes or browsing behaviors-since they make the targeting of potential buyers better. It also pointed to how machine learning is powerful enough to enhance decision- making in e-commerce by offering more insight into the understanding of consumer intention, thus allowing marketers to shape a better strategy toward increasing sales.

Keywords: E-commerce prediction; machine learning models; shoppers behavior analysis

1. Introduction

E-commerce is growing hugely and rapidly in the retail sector, something which was not seen even a few years back[1]. More businesses are paying attention to understanding and predicting consumer behavior for improving user experience and increasing income[2]. This happens because more people shop online. Knowing whether or not the customer's interest is to buy or just browse will help in better customization of various campaigns to improve overall customer satisfaction and make the website more useful as well [3].

Sometimes the online shopping objectives are hard to determine because online shopping behaviors are robust as well as difficult to interpret[4-7]. Traditional ways of checking what customers want include surveys, interviews, or direct observation. These techniques are tedious, narrow in scope, and often fail to detect real-time patterns in large datasets, though they are informative[8-10].

The emergence of data-driven approaches has significantly improved the understanding of complex behavioral patterns with the help of machine learning (ML)[11-13]. ML is particularly well-suited for online shopping behavior analysis as it can process large volumes of data efficiently, identify trends, and offer actionable insights [14]. To comprehend online customer actions, it is essential to consider factors such as age, gender, browsing habits, time spent on pages, and clicks. Factors like advertisement exposure and seasonality are also important [15].

Traditional methods often underperform when handling the complexity and volume of such data. Machine learning, however, is capable of managing these challenges effectively[16-19]. ML

 doi:10.20944/preprints202511.0708.v1

systems can automatically extract features, detect intent, and forecast behaviors. Their usefulness in supporting data-driven decision-making is therefore enhanced [20].

A substantial number of research studies have highlighted the application of machine learning in e-commerce, particularly in understanding shoppers' intent for online purchases. Algorithms such as ensemble methods, Support Vector Machines (SVM), Random Forest (RF), and Logistic Regression (LR) are frequently employed to increase prediction precision, enabling firms to better anticipate and respond to customer needs [21]. Additionally, integrating diverse data sources has been shown to enhance model robustness and fairness across different user segments.

Using a strong dataset with rich parameters like Traffic Source, Session Time, Product Pages Visited, and User Interaction, we aim to build a comprehensive machine learning framework for predicting online buyers' intentions. The machine learning techniques[22-25] employed in the proposed framework include Decision Tree, Random Forest, K-NN, Naive Bayes, Random Tree, Gradient Boosting Tree, and Logistic Regression [26].

Our work focuses on optimal feature selection, parameter tuning, and model evaluation to correctly classify user intentions. Diversity in user behavior—driven by individual tastes, casual browsing, or external influences—poses a significant challenge in studying online consumer behavior. Our approach uses machine learning to reduce prediction errors and uncover key behavioral patterns, thereby addressing this unpredictability.

A high accuracy score reflects our framework's practical applicability in real-world e-commerce settings and confirms the effectiveness of our solution. This paper contributes to the growing literature on consumer behavior analysis by proposing an adaptive and scalable system to infer online consumer intentions[27-29]. Future work will focus on incorporating more contextual and behavioral data to improve prediction accuracy and scalability and explore the integration of advanced deep learning systems [30-33].

2. Literature Review

It has been observed that Shaifali Yadav (2023), using Decision Trees from PySpark's MLlib, achieved an accuracy of 81.8% in predicting whether an online shopper would complete a purchase. Further optimization and data scaling could enhance the model's utility, helping e-commerce companies boost sales and improve customer targeting by leveraging the model's predictive capabilities [1].

Topal (2019) identified that a decision tree method, by enhancing Page Value, could predict online purchases with up to 88% efficiency. His proposed technique emphasized low data consumption and addressed several privacy concerns. This method can help businesses improve user experience while also enhancing marketing outcomes [2].

Boosting algorithms such as XGBoost and Gradient Boosting have been applied to predict online purchase intentions based on browsing behavior. In a study by Köktürk, Güzel, and Ünay (2021), XGBoost delivered the best results, demonstrating its value for understanding customer behavior on e-commerce websites to enhance sales [3].

A study by Abdullah-All-Tanvir et al. (2023) reported that XGBoost outperformed other models, achieving an accuracy of 90.65%. Their work provides a real-time solution for addressing cart abandonment and increasing revenue by considering key factors like bounce rate [4].

According to Abdul Aziz et al. (2024), the prediction of online shopping behavior was conducted using decision tree and rule-based models. The Random Tree and PART models achieved accuracies of 87.56% and 89.34%, respectively. The insights from such models help e-commerce companies refine their marketing strategies [5].

Liu and Shi (2016) demonstrated that the C4.5 decision tree algorithm is more accurate and better structured than Naive Bayes, especially as dataset size increases. Their findings suggest that decision trees are more reliable for predicting customer behaviors in large-scale e-commerce environments [6].

 doi:10.20944/preprints202511.0708.v1

Raed A. Abd-Alhameed and Ahmad Aldelemy showed that machine learning models such as Naive Bayes, Decision Tree, and Random Forest can predict up to 92% of bank term deposit subscriptions when combined. This strategy aids banks in better understanding customer behavior [7].

[8] applied multiple models including Random Forest, XGBoost, and LightGBM, achieving an accuracy of 85% and an AUC of 0.928. Their approach supports inventory planning and sales prediction, and it could be adapted to estimate purchase quantities in similar studies.

[9] reported that the Random Tree model was the most effective at predicting consumer behavior for Malaysian e-commerce companies. They recommend integrating multiple models and using diverse datasets to improve marketing strategies.

[10] found that XGBoost and Random Forest algorithms are effective in predicting customer purchasing behavior by analyzing data attributes such as age, gender, and loyalty program participation. They advocate for advanced real-time prediction tools to enhance marketing performance and efficiency.

[11] developed a predictive model for JD.com using Random Forest, XGBoost, and LightGBM. Their hybrid XGBoost-LightGBM model outperformed others by accurately identifying potential customers and adapting marketing strategies accordingly.

[12-13] examined how predictive models and strategies can enhance targeted advertising. Similarly, Parihar and Yadav (2022) explored machine learning techniques for predicting online shopping behavior and found Gradient Boosting to be the most successful. Their work highlights how machine learning can significantly help e-commerce platforms better connect with and target their customers.

In 2022, [14] developed a system using MLP and LSTM models to simulate customer behavior. This system supports real-time customer engagement, improves targeting, and boosts sales.

Finally, [15] demonstrated that combining Random Forest models with LSTM networks greatly improves the accuracy of predicting online purchase behavior. Their method leads to more effective e-commerce marketing strategies and enhances sales forecasting capabilities.

3. Proposed Methodology

A robust methodology has been designed and implemented to forecast customer purchase intentions in the e-commerce sector. Initially, a comprehensive dataset was collected, incorporating key parameters such as traffic source, session duration, number of product pages visited, and user feedback. The dataset was then preprocessed to handle missing values, normalize relevant features, and encode categorical variables to ensure compatibility with machine learning algorithms.

Following preprocessing, various machine learning models were developed to classify user intent. Feature selection techniques were applied to identify the most relevant predictors, and hyperparameter tuning was conducted to optimize model performance. The models were trained and validated using standard practices, including data splitting and cross-validation.

Figure 1 illustrates the proposed methodology employed in this study, encompassing data preparation, model development, and evaluation procedures. Models are given below:

 doi:10.20944/preprints202511.0708.v1

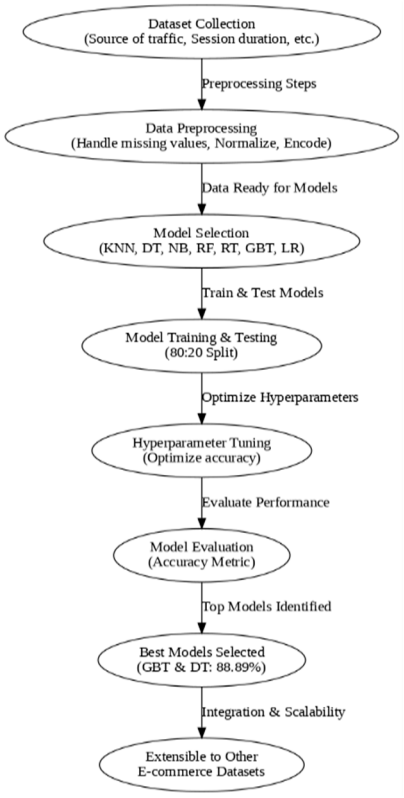

Figure 1. Proposed Methodology.

3.1. K-Nearest Neighbours (KNN)

K-Nearest Neighbors: K-Nearest Neighbor or KNN is a simple classification and regression technique that predicts the target values by relying on the nearest data points in the feature space. However, this technique suffers from two major drawbacks: it is computationally very expensive when it comes to big data, and sometimes it is sensitive to factors that have really no relevance with the target variable. We used a K-NN model, and my accuracy was 82.48% on that.

3.2. Decision Tree

A decision tree is a tree-like machine learning model which, for providing predictions, divides data into subsets based on feature values. It can solve problems of classification and regression; it's pretty simple to understand even if it could overfit complex data. Using the decision tree approach, it is possible to achieve an accuracy rate of 88.89% using the decision tree approach.

3.3. Naïve Bayes

The Naive Bayes algorithm is a probabilistic classification technique based on Bayes' theorem and assuming the feature independence of each other. Fast, easy, and efficient in doing tasks such as text categorization and so on, especially if one deals with big data. We used a Naive Bayes model where accuracy is 81.85%.

3.4. Random Forest

Contrary to single trees, the ensemble learning method random forest trains a multitude of decision trees-which avoid overfitting-considerably enhancing predictive power. By nature, random forests are always reliable and work really well with huge databases on either a classification or a regression task. Then, we took the Random Forest model, which gave me accuracy of 84.59%.

 doi:10.20944/preprints202511.0708.v1

3.5. Gradient Boosting Tree

It is an ensemble-based machine learning technique whereby the development of trees based on gradient boosting is designed one after the other, improving mistakes from the previous ones that had been developed. Though quite effective for both classification and regression tasks with generally very good predictive accuracy, determination of careful calibration cannot be afforded either to prevent overfitting problems.

3.6. Logistics Regression

Logistic regression is one of the techniques used for the prediction of an outcome using some input variables. It is simple, efficient, and effective to use in data that could easily be segregated into two categories. This model involves a special function-the sigmoid function-which shows how the input variables relate to the predicted results. This model has an accuracy of 87.59%.

3.7. Proposed Framework

The 12,330 entries and 18 attributes dataset represent a row of different instances or different sessions. Special Day, Bounce Rates, Exit, Rates, Page Values, Informational Duration. The attributes that capture certain metrics about the online behavior of customers are as follows Administrative, Administrative_ Duration, Informational, Informational Duration, Product Related, Product Relate Duration. It also has categorical variables that provide context surrounding the environment and features of the session such as Month, Operating Systems, Browser, Region, Traffic Type, Visitor Type, Weekend etc. This dataset is suitable for a classification task and prediction model as the target variable Revenue tells us if a purchase was made during the session. The proposed framework applies machine learning methods in predicting the intent of customers online. Data is collected on sources of traffic, session duration, product pages viewed, and user feedback; then preprocessed and selected for features[34-35]. It evaluates different models comprising KNN, Decision Tree, Naive Bayes, Random Forest, Gradient Boosting Tree, and Logistic Regression. Among these, the maximum of 88.89% accuracy was obtained by Gradient Boosting Tree and Decision Tree models. The proposed framework allows real-time prediction for personalized marketing to capture variable user behavior with a view to improving e- commerce experiences.

4. Result

The experimental results indicate that the Decision Tree and Gradient Boosting Tree models achieved the highest accuracy at 88.89% in predicting online shoppers' behavior. These results are consistent with existing literature, where decision tree-based and ensemble methods are known to deliver superior performance.

Other models showed moderate performance, including Random Forest and Random Tree, both achieving an accuracy of 84.59%, while KNN and Naive Bayes performance 82.48% and 81.85%, respectively. Logistic Regression, despite being a simpler model, attained 87.59%, placing it close to the top-performing models.

These relatively high accuracy values demonstrate that the selected models are effective in predicting customer behavior, with Decision Tree and GBT models standing out for their balance between efficiency and predictive power.

The Random Tree classifier, which randomly selects a subset of features at each split, introduces controlled randomness to reduce overfitting. Although it may not always yield the optimal result, it offers faster training times and increased simplicity, making it suitable for large-scale or real-time applications. As shown in Figure 2, the Decision Tree and Gradient Boosting Tree models outperform the others in terms of classification accuracy.

 doi:10.20944/preprints202511.0708.v1

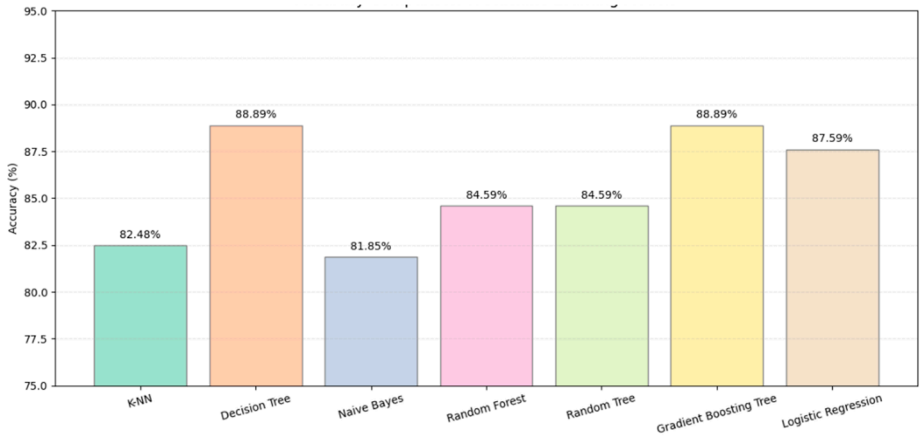

Figure 2. Accuracy Comparison of Machine Learning Models.

Figure 3 presents the confusion matrix of the binary classification model, including class-wise precision and recall. The results highlight that the model performs significantly better at predicting the negative class (FALSE), both in terms of precision (91.54%) and recall (95.92%). In contrast, predictions for the positive class (TRUE) show lower performance, with a precision of 68.91% and recall of 50.75%.

Figure 3. Confusion Matrix with Precision and Recall (Accuracy: 88.89%)

	true FALSE	true TRUE	class precision
pred FALSE	6966	654	91.54%
pred TRUE	306	676	68.91%
class recall	95.62%	50.75%	

Figure 3. Confusion matrix with precision and recall values for the binary classification model.

5.Conclusions

This study has successfully employed KNN, Decision Tree, Naive Bayes, Random Forest, Random Tree, Gradient Boosting Tree, and Logistic Regression to predict online shoppers' purchase intentions. The results show that Gradient Boosting Tree and Decision Tree performed best, each achieving an accuracy of 88.89%. These models can assist companies in enhancing their marketing strategies by focusing on key features such as traffic source, session duration, and user feedback. This approach minimizes prediction errors and improves the user experience. Future research may explore more advanced techniques and incorporate additional contextual data to better evaluate model performance and scalability in the evolving field of e-commerce.

References

1. S. Yadav and M. T. Student, "Prediction of Online Shopper's Buying Intention Using Algorithms of PySpark MLlib," 2023. [Online]. Available: www.ijcspub.org
2. İ. Topal, "Estimation of Online Purchasing Intention Using Decision Tree," Yönetim ve Ekonomi Araştırmaları Dergisi, vol. 17, no. 4, pp. 269–280, Dec. 2019. https://doi.org/10.11611/yead.542249.

 doi:10.20944/preprints202511.0708.v1

3. B. E. Köktürk Güzel and D. Ünay, "Predicting Purchase Interest of Online Shoppers Using Boosting Algorithms," Natural and Applied Sciences Journal, vol. 4, no. 2, pp. 1–15, Dec. 2021. https://doi.org/10.38061/idunas.848233.
4. Abdullah-All-Tanvir, I. Ali Khandokar, A. K. M. Muzahidul Islam, S. Islam, and S. Shatabda, "A Gradient Boosting Classifier for Purchase Intention Prediction of Online Shoppers," Heliyon, vol. 9, no. 4, Apr. 2023. https://doi.org/10.1016/j.heliyon.2023.e15163.
5. M. Abdul Aziz, A. Mustakim, S. A. Rahman, T. Mara, and S. Alam, "Decision Tree and Rule-Based Classification for Predicting Online Purchase Behavior in Malaysia," Malaysian Journal of Computing, vol. 9, no. 2, pp. 1905–1915, 2024. https://doi.org/10.24191/mjoc.v9i2.27130.
6. L. Bing and S. Yuliang, "Prediction of User's Purchase Intention Based on Machine Learning," in Proceedings of the 2016 3rd International Conference on Soft Computing and Machine Intelligence (ISCMI), Institute of Electrical and Electronics Engineers Inc., Oct. 2017, pp. 99–103. https://doi.org/10.1109/ISCMI.2016.21.
7. A. Aldelemy and R. A. Abd-Alhameed, "Binary Classification of Customer's Online Purchasing Behavior Using Machine Learning," Journal of Techniques, vol. 5, no. 2, pp. 163–186, Jun. 2023. https://doi.org/10.51173/jt.v5i2.1226.
8. Z. Liu and X. Ma, "Predictive Analysis of User Purchase Behavior Based on Machine Learning," International Journal of Smart Business and Technology, vol. 7, no. 1, pp. 45–56, May 2019. https://doi.org/10.21742/IJSBT.2019.7.1.05.
9. N. A. Mustakim, M. Abdul Aziz, and S. A. Rahman, "Predicting Consumer Behavior in E-Commerce Using Decision Tree: A Case Study in Malaysia," 2024.
10. E. Deniz and S. Ç. Bülbül, "Predicting Customer Purchase Behavior Using Machine Learning Models," Information Technology in Economics and Business, Jul. 2024. https://doi.org/10.69882/adba.iteb.2024071.
11. X. Zhai, P. Shi, L. Xu, Y. Wang, and X. Chen, "Prediction Model of User Purchase Behavior Based on Machine Learning," in Proceedings of the 2020 IEEE International Conference on Mechatronics and Automation (ICMA), Institute of Electrical and Electronics Engineers Inc., Oct. 2020, pp. 1483–1487. https://doi.org/10.1109/ICMA49215.2020.9233677.
12. "A Machine Learning Model for Prediction of Consumer Purchasing Behavior: A Review," [Online]. Available: www.ijater.com
13. O. J. Adaramola and J. R. Olasina, "Evaluation of Mobile ZigBee Technology Performance with Simulation Techniques," International Journal of Advanced Networking and Applications, vol. 13, no. 6, pp. 5159–5168, 2022. https://doi.org/10.35444/ijana.2022.13602.
14. N. Anastasiia, "Predictions of Customer Behaviour over E-Commerce Websites and Anticipating Their Intention," 2022. [Online]. Available: http://jesne.org/
15. W. Hu and Y. Shi, "Prediction of Online Consumers' Buying Behavior Based on LSTM-RF Model," in Proceedings of the 2020 5th International Conference on Communication, Image and Signal Processing (CCISP), Institute of Electrical and Electronics Engineers Inc., Nov. 2020, pp. 224–228. https://doi.org/10.1109/CCISP51026.2020.9273501.
16. A. U. Rehman et al., "A Machine Learning-Based Framework for Accurate and Early Diagnosis of Liver Diseases: A Comprehensive Study on Feature Selection, Data Imbalance, and Algorithmic Performance," International Journal of Intelligent Systems, vol. 2024, no. 1, Jan. 2024. https://doi.org/10.1155/2024/6111312.
17. 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.
18. Nugroho, D. A. (2025). The Role of AI in Predicting Consumer Behaviour and Effective Marketing Strategies. *Return: Study of Management, Economic and Bussines*, *4*(3), 302-312.
19. Hooshmand Pakdel, G., He, Y., & Chen, X. (2025). Predicting customer demand with deep learning: An LSTM-based approach incorporating customer information. *International Journal of Production Research*, 1-13.

 doi:10.20944/preprints202511.0708.v1

20. 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.
21. 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.
22. 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.
23. 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.
24. Khan, N. A., Jhanjhi, N. Z., Brohi, S. N., Almazroi, A. A., & Almazroi, A. A. (2022). A secure communication protocol for unmanned aerial vehicles. CMC-Computers Materials & Continua, 70(1), 601-618.
25. Muzafar, S., & Jhanjhi, N. Z. (2020). Success stories of ICT implementation in Saudi Arabia. In Employing Recent Technologies for Improved Digital Governance (pp. 151-163). IGI Global Scientific Publishing.
26. Jabeen, T., Jabeen, I., Ashraf, H., Jhanjhi, N. Z., Yassine, A., & Hossain, M. S. (2023). An intelligent healthcare system using IoT in wireless sensor network. Sensors, 23(11), 5055.
27. Parthasarathy, K., Ayyadurai, R., Panga, N. K. R., Bobba, J., Bolla, R. L., & Ogundokun, R. O. (2025). Deep learning for monitoring customer behavior in insurance industry. *Service Oriented Computing and Applications*, 1-16.
28. Gangineni, V. N., Penmetsa, M., Bhumireddy, J. R., Chalasani, R., Tyagadurgam, M. S. V., & Pabbineedi, S. (2025). Big Data and Predictive Analytics for Customer Retention: Exploring the Role of Machine Learning in E-Commerce. *Available at SSRN 5478047*.
29. Lakkimsetty, N. R. S. C. G. (2025). Role of AI in Business Analytics: Predictive Insights for Future Trends.
30. Shah, I. A., Jhanjhi, N. Z., & Laraib, A. (2023). Cybersecurity and blockchain usage in contemporary business. In Handbook of Research on Cybersecurity Issues and Challenges for Business and FinTech Applications (pp. 49-64). IGI Global.
31. Hanif, M., Ashraf, H., Jalil, Z., Jhanjhi, N. Z., Humayun, M., Saeed, S., & Almuhaideb, A. M. (2022). AI-based wormhole attack detection techniques in wireless sensor networks. Electronics, 11(15), 2324.
32. Suresh, B. S., & Suresh, M. (2024). Efficient customer behaviour prediction in I ndian metropolitan cities for E-commerce applications. *Expert Systems*, *41*(9), e13604.
33. Shah, I. A., Jhanjhi, N. Z., Amsaad, F., & Razaque, A. (2022). The role of cutting-edge technologies in industry 4.0. In Cyber Security Applications for Industry 4.0 (pp. 97-109). Chapman and Hall/CRC.
34. Humayun, M., Almufareh, M. F., & Jhanjhi, N. Z. (2022). Autonomous traffic system for emergency vehicles. Electronics, 11(4), 510.
35. Muzammal, S. M., Murugesan, R. K., Jhanjhi, N. Z., & Jung, L. T. (2020, October). SMTrust: Proposing trust-based secure routing protocol for RPL attacks for IoT applications. In 2020 International Conference on Computational Intelligence (ICCI) (pp. 305-310). IEEE.