ENHANCING CYBERSECURITY THROUGH MACHINE LEARNING: A CATBOOST-BASED FRAMEWORK FOR NETWORK INTRUSION DETECTION

Tahniyat Fatima[1], Dr. I. Samuel Peter James[2]
[1]PG Scholar, Department of AI&DS, Shadan Women's College of Engineering and Technology, Hyderabad, tahniyatfatima828@gmail.com
[2]Assoc. Professor, Department of AI&DS, Shadan Women's College of Engineering and Technology drisamuelpeterjames@gmail.com

To Cite this Article
Tahniyat Fatima, Dr. I. Samuel Peter James, "Enhancing Cybersecurity Through Machine Learning: A Catboost-Based Framework For Network Intrusion Detection", Journal of Science Engineering Technology and Management Science, Vol. 03, Issue 08, August 2026, pp: 708-713, DOI: http://doi.org/10.64771/jsetms.2026.v03.i08.pp708-713
Submitted: 01-07-2026 *Accepted: 07-08-2026* *Published: 14-08-2026*

ABSTRACT
The increasing frequency and sophistication of cyber-attacks have highlighted the need for accurate and efficient network intrusion detection systems. Conventional signature-based detection techniques often struggle to identify evolving and previously unseen attacks, making machine learning a promising solution for modern cybersecurity applications. This paper presents a CatBoost-based network intrusion detection framework for classifying network traffic into normal and malicious categories. The proposed approach utilizes the KDD Cup 1999 dataset, where data preprocessing and feature selection techniques are applied to improve model efficiency and reduce computational complexity. The performance of the proposed model is evaluated against widely used machine learning algorithms, including Decision Tree and Random Forest. Experimental results demonstrate that the CatBoost classifier provides superior classification performance due to its ordered boosting strategy, efficient handling of categorical features, and improved generalization capability. To demonstrate the practical applicability of the proposed framework, a Flask-based web application is developed for user authentication, network traffic prediction, and performance visualization. The proposed framework offers an efficient, scalable, and reliable solution for network intrusion detection and can contribute to strengthening cybersecurity through intelligent machine learning techniques.
Keywords: Cybersecurity, Network Intrusion Detection, Machine Learning, CatBoost, Network Traffic Analysis, Intrusion Detection System, Feature Selection, KDD Cup 1999 Dataset.

1. INTRODUCTION

The rapid growth of the Internet, cloud computing, and digital communication has significantly increased network traffic, making cybersecurity a critical challenge in modern computing environments. As organizations increasingly depend on interconnected systems and online services, cyber-attacks such as Denial-of-Service (DoS), Probe, Remote-to-Local (R2L), and User-to-Root (U2R) continue to threaten data confidentiality, system availability, and network integrity. Consequently, developing accurate and efficient Network Intrusion Detection Systems (NIDS) has become essential for protecting network infrastructures.

Traditional intrusion detection systems rely primarily on signature-based and rule-based techniques to detect malicious activities. Although effective against known attacks, these approaches often fail to identify emerging threats and require extensive preprocessing when combined with conventional machine learning algorithms. Furthermore, traditional models may struggle to efficiently handle high-dimensional network traffic data, reducing their effectiveness in dynamic cybersecurity environments.

Recent advances in machine learning have significantly improved intrusion detection by enabling models to learn complex patterns from network traffic. Among these approaches, the CatBoost algorithm offers several advantages, including efficient handling of categorical features, reduced overfitting through ordered boosting, and strong classification performance on structured datasets, making it well suited for network intrusion detection.

This paper proposes a CatBoost-based Network Intrusion Detection Framework using the KDD Cup 1999 benchmark dataset. The framework incorporates data preprocessing, feature selection, and supervised learning techniques to improve detection accuracy while reducing computational complexity. Its performance is evaluated against Decision Tree and Random Forest classifiers. A Flask-based web application is also developed to support user authentication, network traffic prediction, and performance visualization. Experimental results demonstrate that the proposed framework provides an efficient, scalable, and reliable solution for enhancing cybersecurity through machine learning-based network intrusion detection.

2. EXISTING SYSTEM

Existing Network Intrusion Detection Systems (NIDS) commonly employ traditional machine learning algorithms such as Decision Tree, Random Forest, and Logistic Regression to classify network traffic. Although these approaches have demonstrated satisfactory performance in detecting known attacks, they exhibit several limitations when handling high-dimensional data, complex attack patterns, and evolving cyber threats. A comparative summary of these existing approaches is presented in Table 1.

Table 1. Comparison of Existing Machine Learning Algorithms for Network Intrusion Detection

Algorithm	Strengths	Limitations
Decision Tree	Simple, interpretable, and fast for small datasets.	Prone to overfitting and reduced performance on complex network traffic.
Random Forest	Higher accuracy than a single decision tree and improved robustness.	Requires feature preprocessing and has higher computational complexity.
Logistic Regression	Computationally efficient and suitable for binary classification.	Limited ability to model complex non-linear relationships in network traffic.
Support Vector Machine (SVM)	Effective in high-dimensional feature spaces and small datasets.	High training time and limited scalability for large network datasets.

Limitations Of Existing Systems

The existing machine learning approaches require extensive preprocessing and often struggle to detect sophisticated or previously unseen cyber-attacks while maintaining high scalability. These limitations motivate the adoption of advanced gradient boosting techniques such as CatBoost, which offer improved classification accuracy, efficient handling of categorical features, and better generalization for modern network intrusion detection systems.

3. RELATED WORK

Machine learning has become one of the most widely adopted approaches for network intrusion detection due to its ability to identify complex attack patterns and improve detection accuracy. Several studies have investigated supervised learning algorithms, including Decision Tree, Random Forest, Support Vector Machine (SVM), and Naïve Bayes, for classifying network traffic into normal and malicious categories. These methods have demonstrated promising performance in detecting known attacks; however, their effectiveness often depends on extensive feature engineering and careful parameter tuning.

Recent research has focused on ensemble learning and gradient boosting techniques to improve classification performance and model generalization. Ensemble models combine the predictions of multiple classifiers to enhance robustness and reduce overfitting, making them suitable for intrusion detection applications. In particular, gradient boosting algorithms have shown improved capability in learning complex, non-linear relationships within network traffic data while maintaining high predictive performance.

Several benchmark datasets, including KDD Cup 1999, NSL-KDD, UNSW-NB15, and CICIDS2017, have been extensively used to evaluate intrusion detection models. Although these studies have reported significant improvements in detection accuracy, many existing approaches require substantial preprocessing, feature transformation, and computational resources, which may limit their applicability in real-world environments. Motivated by these challenges, this work proposes a CatBoost-based Network Intrusion Detection Framework that efficiently handles structured network traffic data with minimal preprocessing. By leveraging CatBoost's ordered boosting strategy and native support for categorical features, the proposed framework aims to provide a scalable, reliable, and effective solution for modern cybersecurity applications.

4. PROPOSED SYSTEM

In this study, the CatBoost classifier is adopted because it efficiently processes structured network traffic data while reducing preprocessing requirements through native support for categorical features. The model is trained and evaluated using the KDD Cup 1999 dataset, which contains both normal and malicious network traffic records. Data preprocessing and feature selection techniques are applied to remove redundant attributes and identify the most significant network features, improving model efficiency and classification performance. CatBoost effectively learns complex patterns in network traffic while reducing overfitting through its ordered boosting mechanism, enabling accurate detection of various cyber-attacks such as DoS, Probe, R2L, and U2R.

4.1 Advantages of The Proposed System

- Provides improved intrusion detection accuracy through an advanced gradient boosting method.
- Efficiently handles categorical and numerical features with minimal preprocessing.
- Reduces overfitting using ordered boosting, resulting in better model generalization.

- Offers scalable and reliable performance for network intrusion detection in dynamic cybersecurity environments.

5. METHODOLOGIES

1. Data Collection and Preprocessing

The KDD Cup 1999 dataset, which includes labelled network traffic records, was used for this project. In preprocessing, categorical features (such as protocol type, service, and flag) are handled by label encoding, redundant or unnecessary features are eliminated, and numerical properties are scaled as needed. The top 15 most crucial characteristics that greatly aid in anomaly identification are extracted using feature selection.

2. Feature Selection and Engineering

To determine which features, have the greatest influence, the CatBoost model is used to assess feature importance. By choosing the top 15 attributes, model accuracy is increased and computational overhead is decreased. Critical network characteristics including connection counts, error rates, and byte values that strongly suggest unusual activity are among these features.

3. Model Training and Evaluation

On the pre-processed dataset, 80% of the records are used for training and 20% are used for testing the CatBoost Classifier. In order to avoid overfitting and attain high generalisation accuracy, the model has been optimised. Accuracy, precision, recall, F1-score, and confusion matrix are used to gauge the model's performance. Over 99% of network anomalies are detected by the trained model.

4. Model Saving and Deployment

For later use and application integration, the trained CatBoost model is saved in the.cbm format. Predictions can be produced without retraining by storing the model, guaranteeing quick and effective real-time network intrusion detection.

5. Web Application Development (Flask Integration

A web application built with Flask is designed to offer an intuitive user interface. The modules that make up this system are Home, Temporary Register, Login, Predict, and Charts. Users can enter network traffic characteristics, log in, and get predictions right away. The Charts page improves interpretability by offering visual insights such as feature significance plots and confusion matrices**.**

6.Visualization and Results

Effective results presentation makes use of visualisation. Bar graphs are used to show feature relevance, while confusion matrices are used to show how well normal and attack classes are classified. Transparency in performance evaluation is ensured by reporting accuracy ratings for both testing and training.

6. TECHNIQUES USED

6.1 Existing Technique: Random Forest

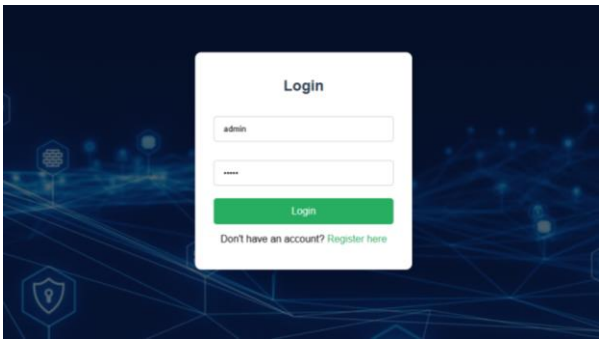
For classification tasks, Random Forest is an ensemble learning technique that generates the majority vote after building several decision trees during training. By merging the outcomes of multiple decision trees, it increases classification accuracy while lowering the possibility of overfitting in comparison to a single decision tree. However, Random Forests handle categorical data less optimally, frequently necessitating preprocessing, and their effectiveness may deteriorate when dealing with extremely big datasets or highly imbalanced classes.

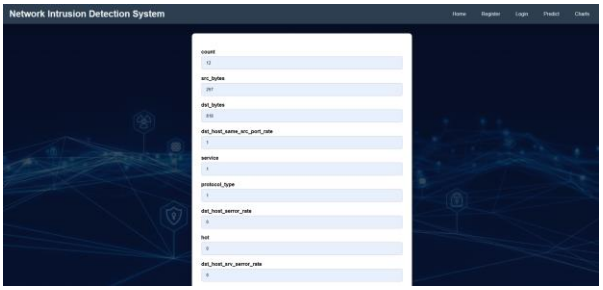
6.2 Proposed Technique: CatBoost Classifier

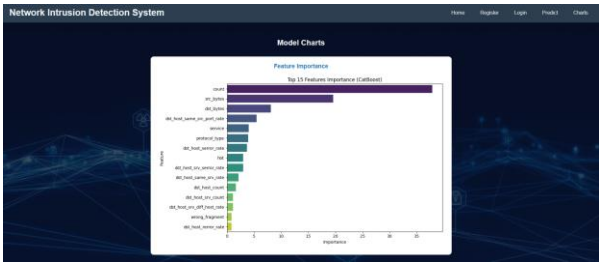
An advanced gradient boosting algorithm developed to efficiently handle structured datasets containing both numerical and categorical features, making it well suited for network intrusion detection. The algorithm builds an ensemble of symmetric decision trees using an ordered boosting technique, which minimizes prediction bias and reduces overfitting by preventing target leakage during training. Unlike traditional gradient boosting methods, CatBoost processes categorical network attributes directly without requiring extensive preprocessing such as one-hot encoding, thereby reducing dimensionality and computational overhead. In network traffic analysis, CatBoost effectively captures complex, non-linear relationships among traffic features, enabling accurate classification of normal and anomalous connections. Its robustness, high predictive accuracy, and ability to handle imbalanced data make CatBoost a reliable choice for intrusion detection systems operating in dynamic and large-scale networks.

Technique	Advantages	Limitations
Random Forest	Robust ensemble classifier with good prediction accuracy.	Requires preprocessing of categorical features and has higher computational complexity for large datasets.
CatBoost (Proposed)	Native support for categorical data, reduced overfitting, efficient learning, and high classification accuracy.	Slightly higher training time than simpler machine learning models.

7. RESULTS







8. CONCLUSION

This paper presented a CatBoost-based Network Intrusion Detection Framework for classifying network traffic as normal or malicious using the KDD Cup 1999 dataset. The proposed model integrates data preprocessing, feature selection, and supervised machine learning techniques to improve the efficiency and reliability of intrusion detection. Compared with conventional machine learning algorithms, the CatBoost classifier demonstrates enhanced capability in handling structured network traffic data while reducing preprocessing requirements and improving model generalization. To validate the practical applicability of the proposed approach, a Flask-based web application was developed to support user authentication, network traffic prediction, and performance visualization. The experimental evaluation demonstrates that the proposed approach provides an effective, scalable, and reliable solution for network intrusion detection, contributing to improved cybersecurity in modern network environments.

9. FUTURE ENHANCEMENT

There are a number of areas for further improvement even though the suggested system shows excellent accuracy in identifying abnormalities using the CatBoost algorithm. The system can be expanded from binary classification (normal vs. attack) to multi-class classification, which allows for the identification of distinct attack types including DoS, Probe, U2R, and R2L. By integrating with real-time network monitoring tools, anomalies might be detected in real time, increasing the system's viability for implementation at the corporate level. To further enhance the identification of intricate attack patterns, deep learning models like CNNs, RNNs, or hybrid architectures may be incorporated. Future research may also concentrate on utilising sophisticated sampling strategies to handle unbalanced datasets and enhancing the system's scalability for implementation in distributed or cloud systems.

REFERENCES

1] U. Fiore, F. Palmieri, A. Castiglione, and A. De Santis, ‘‘Network anomaly detection with the restricted Boltzmann machine,’’ Neurocomputing, vol. 122, pp. 13–23, Dec. 2013, Doi: 10.1016/j.neucom.2012.11.050.

[2] V. Chandola, A. Banerjee, and V. Kumar, ‘‘Anomaly detection,’’ ACM Comput. Surveys, vol. 41, no. 3, pp. 1–58, Jul. 2009, doi:10.1145/1541880.1541882.
[3] D. E. Difallah, P. Cudré-Mauroux, and S. A. McKenna, ‘‘Scalable anomaly detection for smart city infrastructure networks,’’ IEEE Internet Comput., vol. 17, no. 6, pp. 39–47, Nov. 2013, Doi: 10.1109/MIC.2013.84.
[4] E. Anceaume, Y. Busnel, E. L. Merrer, R. Ludinard, J. L. Marchand, and B. Sericola, ‘‘Anomaly characterization in large scale networks,’’ inProc. 44th Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw., Jun. 2014, pp. 68–79, Doi: 10.1109/DSN.2014.23.
[5] M. Ahmed, A. N. Mahmood, and J. Hu, ‘‘A survey of network anomaly detection techniques,’’ J. Netw. Comput. Appl., vol. 60, pp. 19–31, Jan. 2016, Doi: 10.1016/j.jnca.2015.11.016.
[6] M. Usama, J. Qadir, A. Raza, H. Arif, K. A. Yau, Y. Elkhatib, A. Hussain, and A. Al-Fuqaha, ‘‘Unsupervised machine learning for networking: Techniques, applications and research challenges,’’ IEEE Access, vol. 7, pp. 65579–65615, 2019, Doi: 10.1109/ACCESS.2019.2916648.
[7] K. Fotiadou, T.-H. Velivassaki, A. Voulkidis, D. Skias, S. Tsekeridou, and T. Zahariadis, ‘‘Network traffic anomaly detection via deep learning, ’Information, vol. 12, no. 5, p. 215, May 2021, Doi: 10.3390/info12050215.
[8] M. A. Umer, K. N. Junejo, M. T. Jilani, and A. P. Mathur, ‘‘Machine learning for intrusion detection in industrial control systems: Applications, challenges, and recommendations,’’ Int. J. Crit. Infrastruct. Protection, vol. 38, Sep. 2022, Art. no. 100516, Doi: 10.1016/j.ijcip.2022.
100516.
[9] A. A. Jihado and A. S. Girsang, ‘‘Hybrid deep learning network intrusion detection system based on convolutional neural network and bidirectional long short-term memory,’’ J. Adv. Inf. Technol., vol. 15, no. 2, pp. 219–232, 2024, Doi: 10.12720/jait.15.2.219-232.
[10] S. A. Jebur, K. A. Hussein, H. K. Hoomod, L. Alzubaidi, and J. Santamaría, ‘Review on deep learning approaches for anomaly event detection in video surveillance,’’ Electronics, vol. 12, no. 1, p. 29, Dec. 2022, doi:10.3390/electronics12010029.
[11] S. Naseer, Y. Saleem, S. Khalid, M. K. Bashir, J. Han, M. M. Iqbal, and K. Han, ‘‘Enhanced network anomaly detection based on deep neural networks,’’ IEEE Access, vol. 6, pp. 48231–48246, 2018, doi:10.1109/ACCESS.2018.2863036.
[12] N. Kumar and S. Sharma, ‘‘A hybrid modified deep learning architecture for intrusion detection system with optimal feature selection,’’ Electronics, Vol. 12, no. 19, p. 4050, Sep. 2023, Doi: 10.3390/electronics12194050.
[13] S. Hajj, R. El Sibai, J. B. Abdo, J. Demerjian, A. Makhoul, and C. Guyeux,‘‘Anomaly-based intrusion detection systems: The requirements, methods, measurements, and datasets,’’ Trans. Emerg. Telecommun. Technol., vol. 32, no. 4, p. e4240, Apr. 2021, Doi: 10.1002/ett.4240.
[14] L. I. Khalaf, B. Alhamadani, O. A. Ismael, A. A. Radhi, S. R. Ahmed, and S. Algburi, ‘‘Deep learning-based anomaly detection in network traffic for cyber threat identification,’’ in Proc. Cognit. Models Artif. Intell. Conf., May 2024, pp. 303–309, Doi: 10.1145/3660853.3660932.
[15] S. Gunupusala and S. C. Kaila, ‘‘Multi-class network anomaly detection using machine learning techniques,’’ Contemp. Math., vol. 5, no. 2, pp. 5–22, Jun. 2024, Doi: 10.37256/cm.5220243723.
[16] K. Lu, ‘‘Network anomaly traffic analysis,’’ Academic J. Sci. Technol., vol. 10, no. 3, pp. 65–68, Apr. 2024, Doi: 10.54097/8as0rg31.
[17] A. Alfardus and D. B. Rawat, ‘‘Machine learning-based anomaly detection for securing in-vehicle networks,’’ Electronics, vol. 13, no. 10, p. 1962, May 2024, Doi: 10.3390/electronics13101962.
[18] V. Takale, A. Patil, A. Lonikar, A. Shinde, and P. V. Rupnar, ‘‘SecureNet: Network intrusion detection using machine learning and deep learning techniques,’’ Int. J. Res. Appl. Sci. Eng. Technol., vol. 12, no. 3, pp. 439–443, Apr. 2024, Doi: 10.22214/ijraset.2024.59791.
[19] S. H. Rafique, A. Abdallah, N. S. Musa, and T. Murugan, ‘‘Machine learning and deep learning techniques for Internet of Things network anomaly detection—Current research trends,’’ Sensors, vol. 24, no. 6, p. 1968, Mar. 2024, Doi: 10.3390/s24061968.
[20] M. Mynuddin, S. U. Khan, Z. U. Chowdhury, F. Islam, M. J. Islam, M. I. Hossain, and D. M. A. Ahad, ‘‘Automatic network intrusion detection system using machine learning and deep learning, ’in IEEE MTT-S Int. Microw. Symp. Dig., Feb. 2024, pp. 1–9, doi:10.1109/aims61812.2024.10512607.
[21] S. Marappan and H. Marappan, ‘‘Deep learning-based intelligent algorithms for effective transmission authentication and anomaly identification in vehicular networks,’’ in Proc. Int. Conf. Innov. Comput., Intell. Commun. Smart Electr. Syst. (ICSES), Dec. 2023, pp. 1–7, Doi:
10.1109/icses60034.2023.10465346.
[22] S. Ola-Obaado and M. A. Suleiman, ‘‘Anomaly-based network intrusion detection using transfer learning,’’ in Proc. 2nd Int. Conf. Multidisciplinary Eng. Appl. Sci. (ICMEAS), Nov. 2023, pp. 1–5, doi:10.1109/icmeas58693.2023.10379384.
[23] L. Lahesoo, U. Do, R. Carnier, and K. Fukuda, ‘‘SIURU: A framework for machine learning based anomaly detection in IoT network traffic, ’in Proc. 18th Asian Internet Eng. Conf., Dec. 2023, pp. 87–95, doi:10.1145/3630590.3630601.
[24] M. Rele and D. Patil, ‘‘Intrusive detection techniques utilizing machine learning, deep learning,

and anomaly-based approaches,'' in Proc. IEEE Int. Conf. Cryptography, Informat., Cybersecurity (ICoCICs), Aug. 2023, pp. 88–93, Doi: 10.1109/icocics58778.2023.10276955.
[25] T. Ahmad and D. Truscan, ''Efficient early anomaly detection of network security attacks using deep learning,'' in Proc. IEEE Int.Conf. Cyber Secure. Resilience (CSR), Jul. 2023, pp. 154–159, doi:10.1109/csr57506.2023.10224923.
[26] M. A. Siddiqui, M. Kalra, and C. Rama Krishna, ''Anomaly detection for IoT-enabled kitchen area network using machine learning,'' inProc. Int. Conf. Machine Telligence Res. Innov., 2024, pp. 195–209, doi:10.1007/978-981-99-8129-8_17.
[27] Y. Akhiat, K. Touchanti, A. Zinedine, and M. Chahhou, ''IDS-EFS: Ensemble feature selection-based method for intrusion detection system, 'Multimedia Tools Appl., vol. 83, no. 5, pp. 12917–12937, Jul. 2023, doi:10.1007/s11042-023-15977-8.
[28] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, ''A detailed analysis of the KDD CUP 99 data set,'' in Proc. IEEE Symp. Comput.Intell. Secur. Defense Appl., Jul. 2009, pp. 1–6, Doi: 10.1109/CISDA.2009.5356528.
[29] G. Fernandes, J. J. P. C. Rodrigues, L. F. Carvalho, J. F. Al-Muhtadi, and M. L. Proença, ''A comprehensive survey on network anomaly detection,'' Telecommun. Syst., vol. 70, no. 3, pp. 447–489, Mar. 2019, doi:10.1007/s11235-018-0475-8.
[30] F. T. Liu, K. M. Ting, and Z.-H. Zhou, ''Isolation Forest,'' in Proc.8th IEEE Int. Conf. Data Mining, Dec. 2008, pp. 413–422, doi:10.1109/ICDM.2008.17.
[31] H. Chen, S. Hu, R. Hua, and X. Zhao, ''Improved naive Bayes classification algorithm for traffic risk management,'' EURASIP J. Adv. Signal Process., Vol. 2021, no. 1, p. 30, Dec. 2021, Doi: 10.1186/s13634-021-00742-6.16148 VOLUME 13, 2025S. Ness et al.: Anomaly Detection in Network Traffic Using Advanced ML Techniques
[32] M. Al-kasassbeh, M. A. Abbadi, and A. M. Al-Bustami, ''Light GBM algorithm for malware detection,'' in Proc. Sci. Inf. Conf., 2020, pp. 391–403, Doi: 10.1007/978-3-030-52243-8_28.
[33] D. M. Abdullah and A. M. Abdulazeez, ''Machine learning applications based on SVM classification a review,'' Qubahan Academic J., vol. 1, no. 2, pp. 81–90, Apr. 2021, Doi: 10.48161/qaj. v1n2a50.
[34] M. Nemeth, D. Borkin, and G. Michalconok, ''The comparison of machine-learning methods XG Boost and Light GBM to predict energy development,'' in Proc. Comput. Methods Syst. Softw., 2019, pp. 208–215, Doi: 10.1007/978-3-030-31362-3_21.
[35] O. Rainio, J. Teuho, and R. Klén, ''Evaluation metrics and statistical tests for machine learning,'' Sci. Rep., vol. 14, no. 1, p. 6086, Mar. 2024, doi:10.1038/s41598-024-56706-x.
[36] M. Hossin and M. N. Sulaiman, ''A review on evaluation metrics for data classification evaluations,'' Int. J. Data Mining Knowl. Manage. Process, Vol. 5, no. 2, pp. 1–11, Mar. 2015, Doi: 10.5121/ijdkp.2015.5201.
[37] G. Varoquaux and O. Colliot, ''Evaluating machine learning models and their diagnostic value,'' in Machine Learning for Brain Disorders, O. Colliot, Ed., New York, NY, USA: Springer, 2023, pp. 601–630, doi:10.1007/978-1-0716-3195-9_20.
[38] M. A. Aslam, F. Murtaza, M. E. U. Haq, A. Yasin, and M. A. Azam, 'A human-centered approach to academic performance prediction using personality factors in educational AI,'' Information, vol. 15, no. 12, p. 777, Dec. 2024, Doi: 10.3390/info15120777.
[39] A. Halbouni, T. S. Gunawan, M. H. Habaebi, M. Halbouni, M. Kartiwi, and R. Ahmad, ''CNN-LSTM: Hybrid deep neural network for network intrusion detection system,'' IEEE Access, vol. 10, pp. 99837–99849, 2022, doi:10.1109/ACCESS.2022.3206425.
[40] C. Xu, J. Shen, X. Du, and F. Zhang, ''An intrusion detection system using a deep neural network with gated recurrent units,'' IEEE Access, vol. 6, pp. 48697–48707, 2018, Doi: 10.1109/ACCESS.2018.2867564.