*Journal of Discrete Mathematical Sciences & Cryptography*
ISSN 0972-0529 (Print), ISSN 2169-0065 (Online)
Vol. 29 (2026), No. 10-C, pp. 4775–4786
DOI : 10.47974/JDMSC-3091

# A hybrid cryptographic framework with machine learning-based anomaly detection for secure network traffic using UNSW-NB15

Swami Nisha Bhagirath [@]
*Department of Computer Applications*
*Manipal University Jaipur*
*Jaipur 303007*
*Rajasthan*
*India*

Bindiya Jain [†]
*Faculty of Computer Science and Engineering*
*Poornima University*
*Jaipur 303905*
*Rajasthan*
*India*

Pragya Singh [§]
*Symbiosis Centre for Management Studies (Noida)*
*Symbiosis International (Deemed University)*
*Noida 201301*
*Uttar Pradesh*
*India*

Vaibhav Bhatnagar [‡]
*Department of Computer Applications*
*Manipal University Jaipur*
*Jaipur 303007*
*Rajasthan*
*India*

[@] *E-mail:* swami.bhagirath@jaipur.manipal.edu
[†] *E-mail:* Bindiya.jain@poornima.edu.in
[§] *E-mail:* spragyasingh@gmail.com
[‡] *E-mail:* Vaibhav.bhatnagar@jaipur.manipal.edu

© Taru Publications

Rajesh Handa *
*Centre for Distance and Online Education (CDOE)*
*Manipal University Jaipur*
*Jaipur 303007*
*Rajasthan*
*India*

**Abstract**

The advancement of technology is changing digital communication systems rapidly. The development of network topologies is equally growing. Both trends have led to an increase in the number of cyber threats. Intelligent and adaptive Network Intrusion Detection Systems (NIDS) have become more necessary to counter the growing threat of cyber-attacks that include zero-day exploits, polymorphic malware, and other malicious traffic patterns. Sophisticated attacks could go undetected in large-scale network environments. This paper proposed the possibility of using structured machine learning in intrusion detection by utilizing the UNSW-NB15 Dataset. The analysis involved comparison of five machine learning algorithms: Random Forest, Naive Bayes, Neural Network, Gradient Boosting, and Stochastic Gradient Descent. The Random Forest model achieved high accuracy.

***Subject Classification:*** *68M25, 68M12.*

***Keywords:*** *Network intrusion detection system, Cybersecurity, Machine learning, Network traffic analysis.*

## 1. Introduction

The rise in network traffic due to the rise in internet technology, cloud computing, and interconnected systems has reached unprecedented levels. Even though there have been many developments in internet technology, cloud computing, and interconnected systems that have facilitated easier communication and data exchange, these technological improvements have posed some very complex cybersecurity challenges [1]. Some examples of cyber attacks such as distributed denial-of-service attacks, phishing, malware injections, ransomware, reconnaissance, and unauthorized access have become complex [2]. Consequently, the protection of network communication and the real-time detection of intrusions become fundamental components of modern computer networks. Conventional security mechanisms, such as firewalls and signature-based intrusion detection systems are insufficient against modern and emerging cyber threats because they depend on recognizing predefined patterns of known attacks [3]. While cryptographic protocols such as Advanced Encryption Standard (AES) [4] and Rivest–Shamir–Adleman (RSA) [5] ensure confidentiality, authenticity, and secure communication. Cryptography

* *E-mail:* **rajesh.handa@jaipur.manipal.edu (Corresponding Author)**

alone cannot identify malicious activities occurring within encrypted connections or underlying network traffic. Because of increasing demand, there is need for intelligent security frameworks using the combination of secure communications with intelligent mechanisms for the detection of attacks [6]. Machine Learning (ML) provides a powerful means of detecting unusual network activity because ML algorithms can "learn" typical patterns of traffic in addition to recognizing unusual behaviours [7]. ML-based intrusion detection systems can automatically analyze massive amounts of network traffic data and identify both known and unknown threats [8]. Algorithms capable of successfully detecting anomalies in networks in cybersecurity are Random Forests [9], Neural Networks [10], Naive Bayes [11], Stochastic Gradient Descent [12], and Gradient Boosting [13]. Each of these algorithms has provided statistically significant performance improvements for the accurate detection of intrusions [14]. This work proposes a hybrid cryptographic security framework for external network traffic and uses the UNSW-NB15 Datasets for testing purposes [15]. This framework provides two levels of security for external network traffic combining cryptographic techniques with intelligent anomaly detection methodologies. The UNSW-NB15 dataset has gained immense popularity within the field of cybersecurity because of its accuracy in providing simulations of actual modern day network traffic including many categories of cyberattacks exploit, fuzzer, reconnaissance, worm, shellcode, denial of service. This study relates to the Sustainable Development Goals (SDG) 9: (Industry, Innovation and Infrastructure), since it will use intelligent ML based cyber security technologies to help improve secure digital infrastructure.

### 1.1 *Objective*

The primary objectives of proposed model are as follows:

- To develop an efficient machine learning-based intrusion detection system.
- To analyze and classify network traffic using multiple ML algorithms.
- To compare the performance of Random Forest (RF), Neural Network (NN), Gradient Boosting (GB), Stochastic Gradient Descent (SGD), and Naïve Bayes (NB) models.
- To improve network security using preprocessing, feature selection, and hybrid cryptographic techniques.

## 2. Methodology

The proposed methodology follows a structured machine learning approach for intrusion detection and secure network traffic analysis. The system performs data preprocessing, feature selection, normalization, and classification to improve detection accuracy.

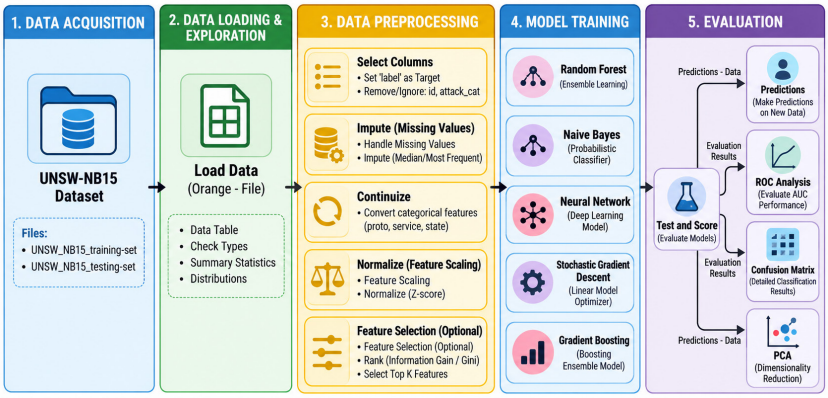



**Figure 1**

**Proposed Hybrid Cryptographic Framework**

The proposed methodology is shown in Figure 1. The overall framework combines intelligent anomaly detection techniques with security mechanisms to enhance cybersecurity performance in modern network environments.

2.1 *Dataset*

The approach relies on the UNSW-NB15 Dataset, which is a benchmarking dataset that has been constructed for modern network intrusion detection studies [15]. This dataset comprises 175,341 records of network traffic data without any missing values, including 33 predictive attributes derived from network flows. Binary classification has been conducted by label attribute, distinguishing normal traffic from malicious one.

2.2 *Data Preprocessing*

Preprocessing will result in transformation of the raw network traffic data set into an efficient feature space. Consider that the whole data set is defined in equation 1.

$$D = \{X, Y\} \tag{1}$$

Where $X = \{x_1, x_2, \ldots, x_n\}$represents input feature vectors, $Y \in \{0,1\}$ represents the target label, $Y = 0$: Normal traffic, $Y = 1$: Malicious (attack) traffic.

Irrelevant attributes such as attack_cat are removed from the training dataset to eliminate redundancy and reduce classification bias as in equation 2.

$$X' = X - \{attack_cat\} \tag{2}$$

Where $X'$= refined feature set used for training. The target feature was defined for classifying traffic as normal or malicious. Categorical attributes such as protocol, service, and connection state were converted into numerical form, while continuous features were normalized using Z-score standardization to ensure uniform scaling and improve model performance, as shown in Equation 3 [17].

$$Z_i = \frac{X_i - \mu}{\sigma} \tag{3}$$

where $Xi$= original feature value, $\mu$= mean of feature, $\sigma$= standard deviation, $Z_i$= standardized feature value. Further, refinement of the set of features was carried out using evaluation approaches like Information Gain and Gini Index ranking, which made it possible to identify the most significant characteristics of the traffic for intrusion detection. Feature importance is evaluated using Information Gain, defined as equation 4.

$$IG(Y, X) = H(Y) - H(Y \mid X) \tag{4}$$

Where, $H(Y)$= entropy of class distribution, $H(Y \mid X)$= conditional entropy after feature split, $IG$= reduction in uncertainty after observing feature $X$. Similarly, the Gini Index used for feature ranking is computed as equation 5 [18].

$$Gini(D) = 1 - \sum_{i=1}^{c} p_i^2 \tag{5}$$

where $p_i$ represents the probability of class $i$. Based on these ranking measures, the optimal feature subset is selected as equation 6.

$$F_{best} = \{f_1, f_2, f_3, \ldots, f_{10}\} \tag{6}$$

The final preprocessed feature space is then provided to the machine learning classifier for intrusion detection as equation 7.

$$f(X') \rightarrow Y \tag{7}$$

where $f(X')$represents the trained classification model used to predict normal and malicious network traffic. The process of Exploratory Data Analysis (EDA) and feature selection was conducted to determine important patterns in the data set and to increase training model efficiency. To filter out redundant features, a "Best Ranked:10" technique was employed, selecting the top ten features, which contribute the most to intrusion detection performance, as shown in Table 1.

**Table 1**

**Top Ranked Network Traffic Features Based on Information Gain and Gini Index Analysis**

| Rank | Attribute | Feature Type | Information Gain | Gini Score |
|---|---|---|---|---|
| 1 | dinpkt | Numeric | 0.377 | 0.210 |
| 2 | dload | Numeric | 0.332 | 0.189 |
| 3 | synack | Numeric | 0.295 | 0.165 |

*Contd...*

| | | | | |
|---|---|---|---|---|
| 4 | tcprtt | Numeric | 0.294 | 0.165 |
| 5 | ackdat | Numeric | 0.293 | 0.164 |
| 6 | dpkts | Numeric | 0.255 | 0.147 |
| 7 | state | Categorical | 0.248 | 0.136 |
| 8 | dmean | Numeric | 0.241 | 0.138 |
| 9 | dbytes | Numeric | 0.227 | 0.128 |
| 10 | sjit | Numeric | 0.191 | 0.118 |

The feature ranking analysis indicates that malicious network behavior is strongly associated with traffic flow characteristics and TCP handshake timing features.

2.3 *Random Forest*

RF is an ensemble learning technique that combines multiple decision trees to improve classification accuracy and reduce overfitting. The final prediction is obtained using majority voting among all generated trees as equation 8 [19].

$$\hat{Y} = \text{mode}(T_1(X), T_2(X), \ldots, T_M(X)) \tag{8}$$

Where, $\hat{Y}$ = final prediction, $T_m(X)$ = prediction of $m^{th}$ decision tree, M = number of trees. Each tree is trained on bootstrapped samples and random feature subsets, reducing overfitting and improving generalization.

2.4 *Naïve Bayes*

NB assumes that all input features are conditionally independent given the target class. The posterior probability for classification is computed using Bayes' theorem as equation 9 [20].

$$P(C|X) = \frac{\{P(X|C)P(C)\}}{\{P(X)\}} \tag{9}$$

Where $P(C \mid X)$= posterior probability of class $C$given feature vector $X$, $P(X \mid C)$= likelihood of observing feature vector $X$under class $C$, $P(C)$= prior probability of class $C$, $P(X)$= evidence probability of feature vector $X$. Under the conditional independence assumption, the probability model becomes as equation 10.

$$P(C \mid X) \propto P(C) \prod_{i=1}^{n} P(x_i \mid C) \tag{10}$$

The final classification rule is equation 11.

$$\hat{Y} = \arg\max_C \; P(C) \prod_{i=1}^{n} P(x_i \mid C) \tag{11}$$

Where, $x_i$= individual feature value, $n$= total number of features, $\hat{Y}$= predicted output class. NB provides fast classification with low computational complexity; however, its independence assumption may reduce performance for highly correlated network traffic features.

2.5 *Neural Network*

The NN model computes the weighted combination of input features followed by an activation function as equation 12 [21].

$$y = f(\sum_{i=1}^{n} w_i\, x_i + b)) \tag{12}$$

where $x_i$represents the input feature, $w_i$denotes the corresponding weight, $b$is the bias term, and $f(\cdot)$represents the activation function. The weighted input to a neuron is computed as equation 13.

$$z^{(l)} = W^{(l)} a^{(l-1)} + b^{(l)} \tag{13}$$

where: $z^{(l)}$ = weighted input at layer $l$, $W^{(l)}$= weight matrix of layer $l$, $a^{(l-1)}$= activation output from previous layer, $b^{(l)}$= bias vector

The activation output is then obtained using an activation function as equation 14.

$$a^{(l)} = f(z^{(l)}) \tag{14}$$

Where $f(\cdot)$= nonlinear activation function such as Sigmoid or ReLU.

The final prediction is calculated as equation 15.

$$\hat{Y} = \sigma(WX + b) \tag{15}$$

Where $\sigma(\cdot)$= output activation function, $W$= output weight matrix, $b$= output bias vector. NN can learn highly complex nonlinear relationships between network traffic attributes and intrusion patterns.

2.6 *Gradient Boosting*

In GB, each new learner attempts to correct the residual errors generated by the previous model. The boosting process is represented as equation 16 [22].

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{16}$$

Where $F_m(x)$ = updated prediction model at iteration $m$, $F_{m-1}(x)$= previous prediction model, $h_m(x)$= weak learner trained on residual errors, $\gamma_m$= learning rate controlling model updates.

2.7 *Stochastic Gradient Descent*

SGD updates parameters iteratively using individual training samples or small batches. The optimization objective is defined as equation 17 [23].

$$J(\theta) = \frac{1}{N} \sum_{i=1}^{N} L(Y_i, f(X_i; \theta)) \tag{17}$$

Where $J(\theta)$= total loss function, $N$= total number of training samples, $L(\cdot)$= loss function, $f(X_i; \theta)$= predicted output for sample $X_i$, $\theta$= model parameter vector

Model parameters are updated iteratively using equation 18.

$$\theta_{t+1} = \theta_t - \eta \nabla J(\theta_t; x_i, y_i) \tag{18}$$

Where $\theta_t$ = parameter vector at iteration t, $\theta_{t+1}$= updated parameter vector, $\eta$ = learning rate, $\nabla J(\theta_t; x_i, y_i)$= gradient of loss function for training

sample ($x_i$, $y_i$). SGD significantly reduces computational complexity and training time, making it suitable for large-scale and real-time intrusion detection systems.

## 3. Results and Discussion

In this section, the comparative performance analysis of the five machine learning algorithms used for network intrusion detection is provided. The prediction function of the classification system is represented as equation 19.

$$f(X') \rightarrow \hat{Y} \tag{19}$$

Where $X'$= preprocessed feature vector, $f(\cdot)$= trained machine learning classifier, $\hat{Y} \in \{0,1\}$= predicted output class, 0: Normal traffic, 1: Malicious traffic. The predicted output $\hat{Y}$is compared with the actual target value Yto compute classification performance. MCC is a robust evaluation metric for binary classification under imbalanced datasets as equation 24 [16].

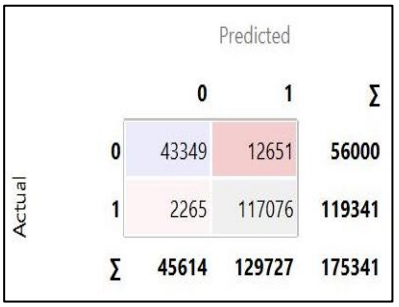
$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{20}$$

Where $MCC = +1$: perfect prediction, $MCC = 0$: random prediction, $MCC = -1$: completely incorrect prediction. The Classification Accuracy of RF is 95.2%, where F1-Score is 0.952 meaning that RF can efficiently classify malicious traffic vs benign traffic. Moreover, RF exhibited the highest MCC score of 0.889, showing that RF is one of the most robust and reliable classifiers. Also, NN and GB are among efficient classifiers, with a CA of 93.9% (NN) and 93.8% (GB). Both algorithms have an AUC of 0.987 meaning that they also have a great capacity to classify attack vs benign packets. The various performance measures used to validate each algorithm include AUC, CA, Precision, Recall, F1-Score, and MCC as shown in Table 2.

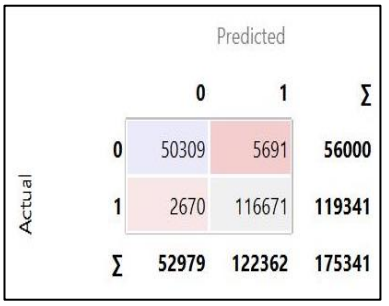
**Table 2**
**Comparison of Different Models**

| Model | AUC | CA | F1 | Precision | Recall | MCC |
|---|---|---|---|---|---|---|
| Random Forest | 99.1 | 95.2 | 95.2 | 95.2 | 95.2 | 88.9 |
| Naïve Bayes | 88.0 | 72.5 | 73.4 | 80.0 | 72.5 | 49.6 |
| Neural Network | 98.7 | 93.9 | 93.8 | 93.9 | 93.9 | 85.9 |
| Gradient Boosting | 98.7 | 93.8 | 93.7 | 94.1 | 93.8 | 85.8 |
| Stochastic Gradient Descent | 87.2 | 91.3 | 91.0 | 91.7 | 91.3 | 79.8 |

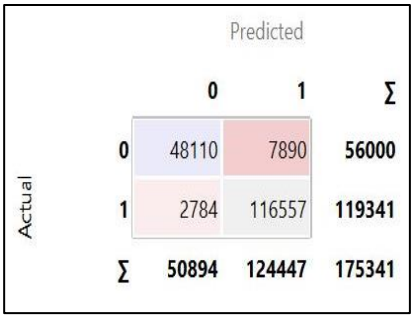
SGD classifier achieved model accuracy of 91.3% with stable performance and earned an AUC of 0.872. This is due primarily to the independent assumptions about the correlations among features in the model since network traffic features demonstrate significant correlation. Figure 2 (a), (b), (c), (d) and (e) shows all five different confusion matrix.



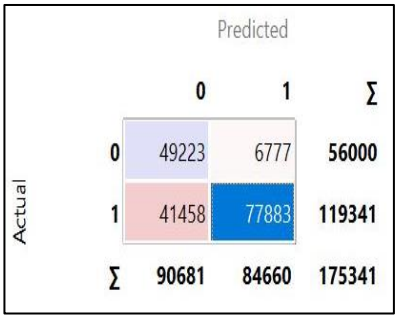

(a) SGD Confusion Matrix



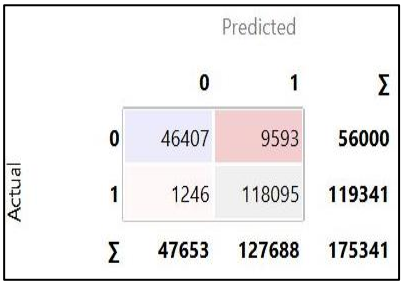

(b) Random Forest




(c) Neural Network




(d) Naïve Bayes




(e) Gradient Boosting

**Figure 2**

**Confusion Matrix of Different Machine Learning Models**

Figure 3 (a) represents the ROC curve that considers the performance of classifiers in detecting the attack traffic based on the trade-off between the True Positive Rate (TPR) and False Positive Rate (FPR). The maximum value of area under the curve is seen with random forest, equal to 0.991. Figure 3 (b) represents the Principal Component Analysis diagram. As per the findings from Principal Component Analysis, the contribution from the first principal component towards explaining the total variance is 16.5%, whereas the second principal component accounts for an extra 6.5%.

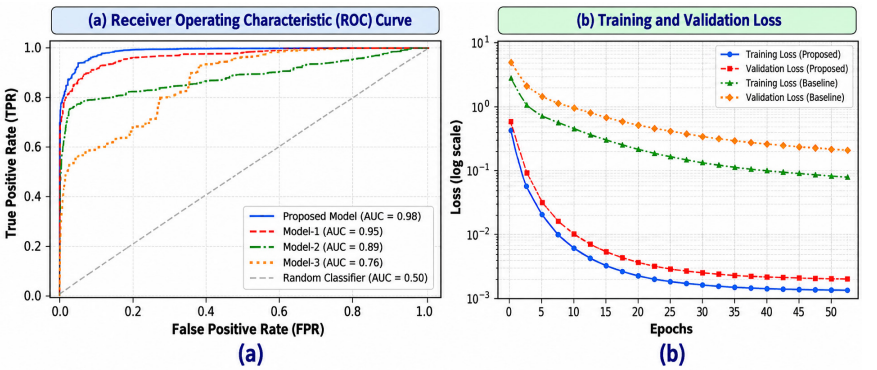



**Figure 3**
**ROC Curve and Principal Component Analysis**

## 4. Conclusion

The rapid increase in cyber threats and the growing complexity of network attacks have highlighted the limitations of traditional intrusion detection techniques. This study investigated the effectiveness of machine learning algorithms for network intrusion detection using the UNSW-NB15 dataset and the Orange data mining framework. Five machine learning algorithms, namely Random Forest, Neural Network, Gradient Boosting, Naïve Bayes, and Stochastic Gradient Descent (SGD), were implemented and evaluated to assess their classification performance. The experimental results demonstrated that Random Forest achieved the highest accuracy of 95.2%, outperforming the other evaluated algorithms. These findings indicate that machine learning techniques, particularly Random Forest, offer a promising approach to improving the accuracy and reliability of network intrusion detection systems.

## References

[1] A. D. Vibhute, M. Khan, C. H. Patil, S. V. Gaikwad, A. V. Mane, and K. K. Patel, "Network anomaly detection and performance evaluation of convolutional neural networks on UNSW-NB15 dataset," *Procedia Comput. Sci.*, vol. 235, pp. 2227–2236 (2024).

[2] A. Fathima, A. Khan, M. F. Uddin, M. M. Waris, S. Ahmad, C. Sanin, and E. Szczerbicki, "Performance evaluation and comparative analysis of machine learning models on the UNSW-NB15 dataset: A contemporary approach to cyber threat detection," *Cybern. Syst.*, vol. 56, no. 8, pp. 1160–1176 (2025).

[3] Y. F. Sallam, S. Abd El-Nabi, W. El-Shafai, H. E. D. H. Ahmed, A. Saleeb, N. A. El-Bahnasawy, and F. E. Abd El-Samie, "Efficient implementation

of image representation, visual geometry group with 19 layers and residual network with 152 layers for intrusion detection from UNSW-NB15 dataset," *Security Privacy*, vol. 6, no. 5, Art. no. e300 (2023).

[4] K. K. Ponniah and B. Retnaswamy, "A novel deep learning based intrusion detection system for the IoT-Cloud platform with blockchain and data encryption mechanisms," *J. Intell. Fuzzy Syst.*, vol. 45, no. 6, pp. 11707–11724 (2023).

[5] J. Jagannathan, S. G. Kumara, T. L. Narayana, and M. S. Alam, "Securing wireless communication using novel transfer learning for encryption in wireless networks," *ICTACT J. Commun. Technol.*, vol. 14, no. 3, p. 3005 (2023).

[6] M. Ali, S. Pervez, S. E. Hosseini, and M. K. Siddhu, "Evaluation and detection of cyberattack in IoT-based smart city networks using machine learning on the UNSW-NB15 dataset," *Int. J. Online Biomed. Eng.*, vol. 21, no. 2 (2025).

[7] A. Kumar, K. Guleria, R. Chauhan, and D. Upadhyay, "Advancing intrusion detection with machine learning: Insights from the UNSW-NB15 dataset," in *Proc. IEEE Int. Conf. Inf. Technol., Electron. Intell. Commun. Syst. (ICITEICS)*, pp. 1–5 (Jun. 2024).

[8] M. Shushlevska, D. Efnusheva, G. Jakimovski, and Z. Todorov, "Anomaly detection with various machine learning classification techniques over UNSW-NB15 dataset," in *Proc. 10th Int. Conf. Appl. Innov. IT (IC-AIIT),* Köthen, Germany, vol. 10, no. 1, pp. 21–27 (Mar. 2022), doi: 10.25673/76928.

[9] M. Manneh, P. Ansah, S. K. Tetarave, M. R. Mishra, and E. Kalaimannan, "A comparative analysis of random forest and support vector machine techniques on the UNSW-NB15 dataset," in *Proc. Int. Conf. Innovations Comput. Res.*, Cham, Switzerland: Springer Nature Switzerland, pp. 194–203 (2024).

[10] R. Sharma and S. Sobti, "A deep neural networks model for intrusion detection in UNSW-NB15 dataset," in *Proc. 4th Asian Conf. Innovation Technol. (ASIANCON)*, pp. 1–5 (Aug. 2024).

[11] R. Tahri, A. Jarrar, A. Lasbahani, and Y. Balouki, "A comparative study of machine learning algorithms on the UNSW-NB15 dataset," *ITM Web Conf.*, vol. 48, Art. no. 03002 (2022).

[12] N. Hassouneh and S. Al-Sharaeh, "Intrusion detection in IoT networks using LSTM deep learning models with the UNSW-NB15 dataset," in *Proc. Int. Conf. New Trends Comput. Sci. (ICTCS)*, pp. 263–269 (Apr. 2025).

[13] A. Al-Obaidi, A. A. Ibrahim, and A. M. Khaleel, "The effectiveness of deploying machine learning techniques in information security to detect nine attacks: UNSW-NB15 dataset as a case study," *Math. Model. Eng. Probl.*, vol. 10, no. 5 (2023).

[14] A. Thaljaoui, "Intelligent network intrusion detection system using optimized deep CNN-LSTM with UNSW-NB15," *Int. J. Inf. Technol.*, pp. 1–17 (2025).

[15] UNSW Canberra, "The UNSW-NB15 Dataset" (2021). [Online]. Available: https://research.unsw.edu.au/projects/unsw-nb15-dataset. [Accessed: Apr. 20, 2026].

[16] A. E. Curtis, T. A. Smith, B. A. Ziganshin, and J. A. Elefteriades, "The mystery of the Z-score," *Aorta*, vol. 4, no. 4, pp. 124–130 (2016).

[17] M. W. Ahmad, J. Reynolds, and Y. Rezgui, "Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees," *J. Cleaner Prod.*, vol. 203, pp. 810–821 (2018).

[18] Y. Ma, C. Su, J. Liu, Z. Ren, C. Xu, and Y. Gao, "Predictions of nuclear charge radii and physical interpretations based on the naive Bayesian probability classifier," *Phys. Rev. C*, vol. 101, no. 1, Art. no. 014304 (2020).

[19] A. D. Jagtap, K. Kawaguchi, and G. E. Karniadakis, "Adaptive activation functions accelerate convergence in deep and physics-informed neural networks," *J. Comput. Phys.*, vol. 404, Art. no. 109136 (2020).

[20] J. Guo, S. Yun, Y. Meng, N. He, D. Ye, Z. Zhao, L. Jia, and L. Yang, "Prediction of heating and cooling loads based on light gradient boosting machine algorithms," *Build. Environ.*, vol. 236, Art. no. 110252 (2023).

[21] R. C. Poonia, T. Singh, C. K. Rajashri, D. Mondal, and S. M. Patil, "Optimizing communication system security through comprehensive testing and intelligent network configurations," *J. Inf. Optim. Sci.*, vol. 47, no. 5-A, pp. 1709–1717 (2026), doi: 10.47974/JIOS-2261.

[22] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: An overview," *arXiv preprint arXiv:2008.05756* (2020).

[23] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, Art. no. 6 (2020).

*Received April, 2026*