Statistical Feature Fusion Driven Enhanced Network Intrusion Detection

Ayush Verma[1]* and Manju Khari[1]

[1]Jawaharlal Nehru University, New Delhi110067, India
*ayush.k1998@gmail.com, manjukhari@yahoo.co.in

Abstract. This study presents a feature selection methodology designed to improve the efficacy of Deep Neural Network (DNN)-based intrusion detection systems (IDS). The suggested method uses a statistical fusion strategy that combines variance thresholding and pairwise correlation analysis to find a small but useful set of network traffic features. The method improves model interpretability and performance by concentrating on features that demonstrate significant variability and minimal redundancy. The proposed approach is examined using three established datasets: NSL-KDD, UNSW-NB15, and CIC-IDS2017. Before training the model, one-hot encoding is used to encode categorical features, and standard normalisation is used to make sure that all the features are scaled the same way. The statistical methods create smaller groups of features, which are then used as inputs for a multilayer DNN classifier. Experimental results show that accuracy, precision, recall, F1-score, and false positive rate (FPR) have all improved a lot compared to traditional feature selection methods like Recursive Feature Elimination, Chi-Square, and Random Forest. The suggested method also cuts down on execution time, even when it uses more features in some cases showing that it is more computationally efficient. The results show that the suggested feature selection strategy could greatly improve the performance and reliability of IDS in different network environments.

Keywords: Intrusion Detection, Statistical Fusion, Feature Selection, Deep Learning.

1 Introduction

A Network Intrusion Detection System (NIDS) is important for keeping networks safe because it watches traffic to find unauthorized access, bad behavior, or policy violations. Generally, NIDS are categorized in two central categories: signature-based, which are quick and work well for attacks that are already known, and anomaly-based, which can find new threats but often give false positives [1]. As networks become more complicated and cyber threats become more sophisticated, Deep Learning (DL), especially Deep Neural Networks (DNN), has been used to enhance the capabilities of IDS by automatically extracting features and making them more generalized to attacks that have never been seen before [2]. Despite these benefits, key challenges persist, including handling imbalanced datasets, achieving

S. Bhalerao et al. (eds.), *Proceedings of the 2nd International Conference on Recent Advancement and Modernization in Sustainable Intelligent Technologies & Applications (RAMSITA-2026)*, Advances in Intelligent Systems Research 207,
https://doi.org/10.2991/978-94-6239-678-4_22

real-time detection, coping with evolving threats, and defending against adversarial attacks.A major limitation is the scarcity of high-quality current datasets. To address these issues, research is advancing toward hybrid models using CNNs, RNNs, and attention mechanisms, along with techniques like ensemble learning, federated learning for data privacy, and explainable AI (XAI) to improve scalability, accuracy, interpretability, and adaptability of IDS [3].

This study proposes a DNN-based ID that incorporates a hybrid combination of feature selection methods. Since DNNs require large datasets for effective training, publicly available intrusion detection datasets are used, captured via tools like Wireshark and Nmap, and processed from pcap/tcpdump files to extract header and payload features. However, these datasets often include redundant or irrelevant features that may hinder classification performance. To address this, a filter-based feature selection technique is proposed that statistically interprets features, independent of the learning algorithm. The proposed approach focuses on identifying significant features through a hybrid method of Variance Threshold Feature Selection and Pairwise Correlation Feature Selection, aiming to enhance DNN performance.

Feature selection is defined as the process of recognizing and selecting the most relevant feature within the data set in order to improve the performance of the machine learning model. It is an important step in the machine learning process [4]. Feature selection is used to remove unnecessary information in the data set in order to improve the accuracy of the model, prevent overfitting, reduce training time, and improve the interpretability of the results. In the intrusion detection and classification problem domain, there are three main feature selection techniques used to identify and utilize the most relevant feature in the data set in order to improve the precision, effectiveness, and robustness of intrusion detection systems. As shown in Figure 1, feature selection techniques include filter methods, wrapper methods, and embedded methods.

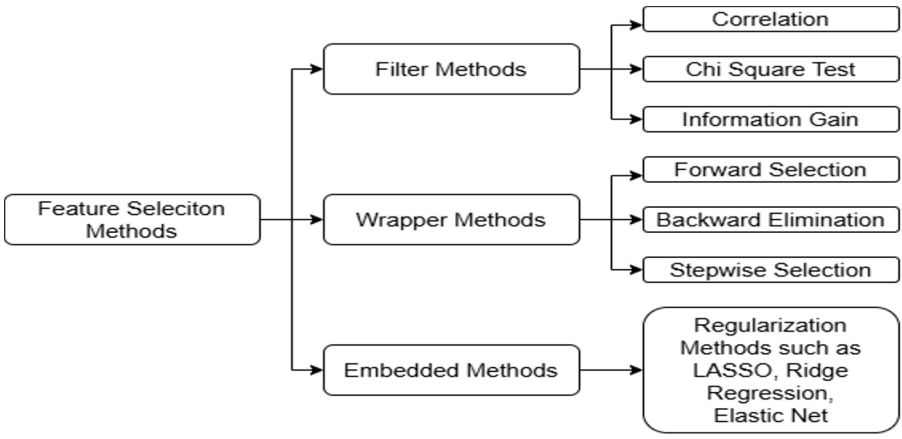

Fig. 1. Types of Feature Selection Methods

Filter Methods. These approaches do not use machine learning models; instead, they assess feature relevance using statistical methods. The correlation between each feature and the target variable is used to evaluate each one separately. ANOVA, mutual information, chi-square tests, and correlation coefficients are common methods. Although these techniques are quick and helpful for preprocessing, they might retain redundant features and overlook feature interactions.

Wrapper Methods. These techniques assess various feature subsets using machine learning models, choosing the best one based on model performance. This includes methods such as recursive feature elimination (RFE), backward elimination, and forward selection. They are computationally costly and may overfit, particularly on small datasets, even though they capture feature interactions and frequently produce better accuracy.

Embedded Methods. These techniques combine the benefits of filter and wrapper methods to perform feature selection during model training. LASSO, Ridge, and tree-based models like Random Forest or XGBoost are a few examples. These are usually model-specific but provide a good trade-off between efficiency and accuracy.

Outcomes of this work highlight:

- A statistical feature selection method for IDS based on fusion of Variance Threshold Feature Selection and Pairwise Correlation Feature Selection.
- Using the reduced feature subset, DNN is applied for classification.
- Evaluation on NSL-KDD, UNSW-NB15, and CIC-IDS2017 datasets.
- Evaluation of performance using metrics like recall, accuracy, precision, F1-score, and false positive rate (FPR).

The article's remaining sections are organized as follows: Literature Study of current feature selection techniques are examined in Section 2, both generally and in connection with IDS. The suggested hybrid approach used is described in Section 3. In Section 4, the experimental procedure is described and the data obtained from various evaluation criteria is examined. The paper concludes with conclusive observations and future directions in Section 5.

2 Literature Study

This section explores a deep analysis of the previous work performed by researchers in the domain of NIDS pertaining to statistical feature selection. Optimization of NIDS by statistical feature analysis has increasingly been researched upon in recent times. These methods have shown progressive research related to precision, accuracy and applicability.

A method by Kumar et al. [6] proposes an ML-based IDS using statistical ranking method. Their work focuses on network traffic data analysis for malware detection. Their method of IDS using statistical analysis identifies 9 traffic features for intrusion detection, ranked using the T-test for unequal variances. The top 7 features were more

effective than all 9 features. The Naive Bayes algorithm evaluates these features, achieving a maximum accuracy of 95.69% by removing the two features with the lowest‘t’ values.

Lan et al. [7] offer the 'quantitative logarithmic transformation-based intrusion detection system (QLT-IDS)', which examines network behavior through a statistical methodology, excluding machine learning or deep learning techniques. It can identify a broad spectrum of malicious attacks with high accuracy and is effective for both real-world and simulated network traffic without requiring extensive data gathering or training, hence improving real-time detection performance. Their method achieves detection accuracy of 93.1% for SSH intrusion attacks, 92.4% for RDP intrusion attacks, and 95.7% for port scanning assaults.

The use of statistical measures, particularly Information Gain, for feature selection in IDS is emphasized in a method by Adhao et al. [8]. This method lowers false alarms and computational complexity in flow-based IDS while increasing detection rates by identifying informative features. They employ ensemble techniques which improves detection rate, reduces false alarms and computation time. The proposed bio-inspired “Krill Herd Algorithm” with statistical measures enhances flow-based IDS performance.

Heryanto et al. [9] propose a correlation-based feature selection (CFS) to enhance IDS. Their correlation-based feature selection improves IDS performance with high accuracy. The proposed CFS-based IDS achieves high accuracy, recall, specificity, precision, F1-score, true positive rate, and true negative rate.It highlights high performance metrics such as accuracy and precision to improve detection efficiency and reduce false alarms.

A Novel IDS model using statistical processing techniques and machine learning algorithms is proposed by Veeranna et al. [10] which employs statistical features such as correlation and Joint Entropy, which are used to discover and remove duplicate connections and features in the data, enhancing the efficiency of the IDS model. The proposed method utilizes Fuzzy C-means and Support Vector Machine (SVM) for classification and obtains accuracy of 97.6% and false alarm rate of 0.7% which beats many other proposed methods.

Research by Sharma et al. [11] identifies key statistical features for IDS, utilizing ANOVA and Chi-Square tests. Notably, "Bytes Sent" ranks highest in ANOVA, while "Bytes Sent" and "Bytes Received" are top ranked in the Chi-Square test for normal and malware data. The study aims to distinguish normal traffic from intrusion network traffic using statistical techniques. ANOVA and Chi-Square tests rank features for effective intrusion detection.

Umair et al. [12] presents a statistical methodology for IDS, employing a multilayer convolutional neural network for feature extraction and selection, succeeded by a softmax classifier for classification, and attaining good accuracy in identifying network intrusions. Their approach utilizes a hybrid deep learning model that incorporates multilayer CNN for feature extraction, classified via softmax and DNN, achieving 99% accuracy on benchmark datasets.

A work by Imrana et al. [13] presents chi-square BiLSTM IDS which utilizes the chi-square statistical model for ranking and selecting features based on their $\chi 2$ test scores. It integrates the chi-square model with BiLSTM for intrusion detection for effective feature selection. The proposed method optimizes feature subsets to enhance

classification accuracy, reduce complexity in network intrusion detection and reduce false alarms.

3 Methodology

This section will highlight the methodology followed in detail. The section will throw light on the background of associativity between type of intrusion and selection of features. It will also discuss the conceptualization of the performed work incorporating statistical features, fusion of features etc.

3.1 Dataset Description

The efficacy of the proposed methodology is validated using three intrusion detection system datasets: UNSW-NB15, NSL-KDD, and CIC-IDS-2017. The datasets are diverse in terms of network characteristics and are created under various network environments. Moreover, the datasets include real and synthetic network traffic. Hence, the efficacy of the proposed approach can be justified through the analysis of various network traffic using these three different datasets. A brief description of each of the three datasets is provided in the subsections below.

NSL-KDD Dataset. This dataset [14] is a refined version of the KDD CUP 99 dataset, created by removing duplicates and missing values. It includes network features and samples from four attack types along with normal traffic. The dataset provides separate training and test sets with 125,973 and 22,544 samples, respectively.

UNSW_NB-15 Dataset. This dataset [15] was generated using the IXIA Perfect Storm tool to simulate network traffic, including both normal behavior and various attacks. Features were extracted using Argus and Bro-IDS tools. It includes 175,341 training samples and 82,332 test samples.

CIC-IDS-2017 Dataset. This dataset [16] is a recent and extensive intrusion detection dataset built by capturing real-time traffic over five days. It covers diverse network services, protocols, and modern attack types. Features were extracted using the CICFlowMeter tool.

3.2 Data Preprocessing

Preprocessing of the intrusion detection datasets is done to make it easier to experiment and train models [17]. There are two main steps in this process: feature encoding and normalization. As the datasets have attributes which are categorical in nature like protocol type, service, and flag status, these are first changed into numbers using one-hot encoding, which is a common way to show categorical variables [18] in binary format. After encoding, normalization is done to make sure that all of the feature scales are the same. This is important because the value ranges of the features

can be very different. Standard scaling is used to do this. It subtracts the mean and scales the values to unit variance, which lets the model treat all features equally important during training.

3.3 Feature Selection

Feature selection is employed to extract a compact set of informative features from each intrusion detection dataset, ensuring the elimination of redundant and irrelevant data. The proposed selection method, detailed earlier, is implemented on the NSL-KDD, UNSW_NB-15, and CIC-IDS-2017 datasets. As a result, the original feature sets are reduced to 21 features (from 41) for NSL-KDD, 21 features (from 42) for UNSW_NB-15, and 64 features (from 79) for CIC-IDS-2017. These refined subsets are then used as input for the deep learning model, enhancing both learning efficiency and detection accuracy.

Variance Threshold Feature Selection. It is a widely used filter-based feature selection technique. It deletes all features whose value of variance doesn't equal to a certain threshold. The core idea is that features with low variance are unlikely to be informative for machine learning models, as they don't provide much distinguishing power. Variance is calculated using Equation (1), where σ^2 denotes the, N is the quantity of data instances, x_i stands for the individual values of the feature, and μ represents the mean of that feature.Features with very low variance (i.e., almost constant values) across samples don't contribute much to the model's ability to discriminate between outputs

$$\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N} \tag{1}$$

Pairwise Correlation Feature Selection. It is a filter-based, unsupervised feature selection method that identifies and removes redundant features based on the correlation between them. The idea is to keep only one feature among a group of highly correlated features, as they carry similar information. When two features are highly correlated, they are linearly dependent. This implies that one feature can be approximated as a linear transformation of the other. Therefore, keeping both doesn't add meaningful new information to the feature space. This helps reduce dimensionality, potentially improving model performance and interpretability by removing features that offer little new information. Correlation Coefficient, denoted by 'r' is used to measure the correlation between two variables x and y, denoted in Equation (2).

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - \left(\Sigma x\right)^2][n\Sigma y^2 - (\Sigma y)^2]}} \tag{2}$$

The result of the feature selection procedure is a collection of pertinent features that are significantly associated with class output labels and enhance the extraction of patterns from data. Fig. 2 illustrates the overarching methodology employed for the statistical feature selection approach aimed at improving network intrusion detection.

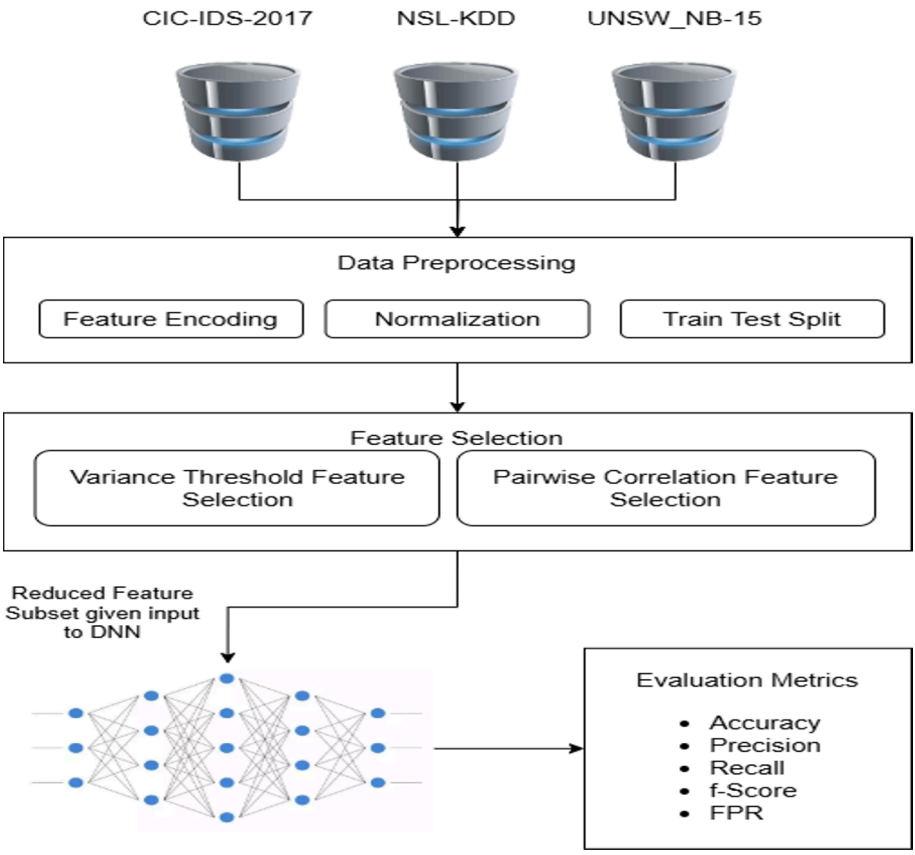

Fig. 2. Proposed Methodology

3.4 Intrusion Detection & Classification using Deep Neural Network

A multilayered Deep Neural Network (DNN) architecture is constructed to perform the intrusion detection and classification task [19]. The input layer's size is configured to match the number of selected features: 64 for CIC-IDS-2017 and 21 for both NSL-KDD and UNSW_NB-15. The network comprises three hidden dense layers with 1024, 768, and 512 neurons, respectively, to facilitate hierarchical feature learning. A dropout layer comes after each dense layer to lessen overfitting and encourage generalization. Non-linear transformations are made possible by applying the ReLU activation function to each hidden layer. Binary classification is made

possible by the output layer's use of the sigmoid activation function. The binary cross-entropy loss function is used to assess the model's performance after 300 epochs of training with a batch size of 1024.

4 Results and Discussion

We tested the suggested approach using Python on Google Colaboratory. The computer had 12.7 GB of RAM, 15 GB of GPU memory, and an NVIDIA T4 GPU. The experiments use pre-processed intrusion detection datasets and a reduced feature subset that is obtained from the suggested feature selection technique. The performance analysis entails a comparison of the proposed feature selection approach with current methodologies.

4.1 Evaluation Metrics

These evaluation criteria are used to evaluate the proposed approach. The criteria are the Confusion Matrix, which specifies true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Accuracy (Acc) is defined as the ratio of correct predictions to the total number of instances, as shown in Equation (3). Precision (P) is the proportion of true positives among all the true and false positive predictions and is given by Equation (4), which represents the proportion by which the proposed approach minimizes the inclusion of false data. Recall (R) is the proportion of true positives among all true and false positive cases and is shown in Equation (5), which represents the proportion by which the proposed approach is able to identify relevant cases. The F1 score (F-score) is the harmonic mean of precision and recall and is shown in Equation (6). The False Positive Rate (FPR) is the proportion of false positives and is used to measure the proportion of negative cases that are incorrectly identified as positive in the proposed approach and is shown in Equation (7).

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

$$P = \frac{TP}{TP+FP} \tag{4}$$

$$R = \frac{TP}{TP+FN} \tag{5}$$

$$F1\ Score = 2 \times \frac{P \times R}{P+R} \tag{6}$$

$$FPR = \frac{FP}{FP+TN} \tag{7}$$

4.2 Results

The results of the experiments on the CIC-IDS2017 NSL-KDD and UNSW_NB-15 datasets, which are shown in Tables 1, 2, and 3, show that the proposed feature

selection method works well. This method consistently outperformed conventional feature selection techniques when combined with a DNN-based Intrusion Detection System (IDS) across all three datasets. The optimized feature subset led to a peak accuracy of 99.93% in the CIC-IDS2017 dataset, which is an improvement over the previous best accuracy. The proposed method achieved an accuracy of 99.61% for the NSL-KDD dataset. The UNSW_NB_15 dataset also received an accuracy score of 89.50%.

We have also investigated the performance of the model based on precision and recall, which indicate the efficiency and sensitivity of the classifier model. The proposed technique performed extremely well on all the datasets. On the CIC-IDS2017, the proposed technique achieved 99.75% correctness and 98.81% recall values. Precision and recall values on the NSL-KDD dataset were 99.84% and 98.71%, respectively. The precision and recall values of the UNSW_NB-15 dataset were 95.10% and 98.90%, respectively.

The F-score is used for evaluating the performance of the imbalanced data. The F-score of the CIC-IDS2017 dataset was 99.93%, the F-score of the NSL-KDD dataset was 99.50%, and the F-score of the UNSW NB-15 dataset was 95.62%. The proposed methodology had a lower false positive rate compared to other approaches.

Also, the execution time of all datasets was considered in terms of preprocessing, feature selection, training, and classification. The proposed method of DNN-based IDS took smaller time for execution, even when it considered more features than other methods. This shows how efficient the derived feature subset is.

Table 1. Performance of proposed method on CIC-IDS-2017 dataset

Method	# Selected Features	Accuracy	Precision	Recall	f-Score	FPR
Recursive Feature Elimination	13	97.86	97.00	97.55	98.50	0.047
Chi-Square	13	96.80	97.65	98.50	97.90	0.061
Random Forest	37	98.44	98.13	98.10	97.68	0.060
Proposed Method	**64**	**99.93**	**99.75**	**99.81**	**99.93**	**0.011**

Table 2. Performance of proposed method on NSL-KDD dataset

Method	# Selected Features	Accuracy	Precision	Recall	f-Score	FPR
Recursive Feature Elimination	13	98.06	97.40	97.23	98.80	0.012
Chi-Square	13	97.54	97.79	97.63	98.21	0.012
Random Forest	16	98.61	97.29	97.68	98.10	0.021
Proposed Method	**21**	**99.61**	**99.84**	**98.71**	**99.50**	**0.011**

Table 3. Performance of proposed method on UNSW_NB-15 dataset

Method	# Selected Features	Accuracy	Precision	Recall	f-Score	FPR
Recursive Feature Elimination	13	83.36	79.77	95.60	86.44	0.013
Chi-Square	13	82.41	78.22	97.85	87.90	0.013
Random Forest	17	82.69	79.40	98.57	87.62	0.050
Proposed Method	**21**	**89.50**	**95.10**	**98.90**	**95.62**	**0.011**

The suggested method for choosing features is very good for use in real-time IDS that work with data streams that are always changing. Its practical strength comes from the fact that the training phase, which is very computationally intensive, is separate from the detection phase, which is very light and happens in real time. The statistical fusion approach is a filter-based method that combines variance thresholding and pairwise correlation. This means that the feature selection process happens offline, before the model is put into use, and is not part of the model training. The live IDS only needs to do a little bit of work once the best, smaller set of features has been found. It needs to extract the specific, pre-determined features from network packets, use pre-calculated normalization parameters, and send the small feature vector to the trained DNN for a prediction that happens almost instantly. This operational model avoids the big overhead of running complicated selection algorithms on live traffic. This makes sure that modern networks can be monitored effectively without becoming a bottleneck.

The main problem to solve in order to keep practical relevance in a changing threat landscape is "concept drift." This is when the statistical properties of network traffic and attack patterns change over time. A static model, no matter how accurate on its training data, will see its performance degrade as new threats emerge. So, a production-level implementation would need a strong MLOps (Machine Learning Operations) framework that includes regular retraining. This involves consistently gathering new network traffic data and labelling them, as well as repeating the whole process offline, from selecting statistical features to building the DNN model, in order to construct a new model. By employing these new models without any difficulties, it is ensured that the IDS is able to adapt to new threats, as well as maintain its high detection rate, ensuring it is operational in a successful manner for a long time in security-related tasks.

5 Conclusion and Future Scope

This paper proposes a feature selection method in conjunction with statistical significance tests based on standard deviation and absolute difference of mean and median, with the aim of improving intrusion detection and classification. The proposed method is based on selecting a small subset of features with high variance. Then, a Deep Neural Network (DNN) model is used to learn from these features. The

effectiveness of this model is evaluated by using three datasets for intrusion detection, NSL-KDD, UNSW_NB-15, and CIC-IDS-2017. The evaluation of these datasets demonstrates that the proposed model is better than other traditional feature selection methods. The evaluation also reveals that computation time is reduced. Therefore, it is concluded that the proposed feature selection method is effective in selecting a feature subset that improves the efficiency of intrusion detection systems based on a DNN model. In future, this feature selection method could be improved to be applicable to real-time data. The model could be extended to learn in real-time, enabling it to respond to changes in intrusion patterns. The model could be made robust against evasion attacks by integrating adversarial defense techniques. For applications like power grid management or financial transactions, it is of utmost importance to ensure robustness. At the same time, it is imperative to integrate techniques like Explainable AI (XAI) into this model to gain the trust of security analysts. XAI techniques could be used to provide justification for alerts beyond a simple black box model. The coupling of adversarial robustness with XAI could lead to the development of this model as a reliable tool for protecting valuable digital resources. The model could be extended to other domains like IoT, cloud computing, etc.

References

1. Satilmiş, H., Akleylek, S., & Tok, Z. Y. (2024).: A Systematic Literature Review on Host-Based Intrusion Detection Systems. *IEEE Access*, *12*, 27237–27266. https://doi.org/10.1109/access.2024.3367004
2. Dietz, K., Mühlhauser, M., Kögel, J., Schwinger, S., Sichermann, M., Seufert, M., Herrmann, D., & Hoßfeld, T. (2024).: The missing link in network intrusion detection: taking AI/ML research efforts to users. *IEEE Access*, *12*, 79815–79837. https://doi.org/10.1109/access.2024.3406939
3. Arreche, O., Guntur, T., & Abdallah, M. (2024).: XAI-IDS: toward proposing an explainable artificial intelligence framework for enhancing network intrusion detection Systems. *Applied Sciences*, *14*(10), 4170. https://doi.org/10.3390/app14104170
4. Saif, S., Das, P., Biswas, S., Khari, M., & Shanmuganathan, V. (2022).: HIIDS: Hybrid intelligent intrusion detection system empowered with machine learning and metaheuristic algorithms for application in IoT based healthcare. *Microprocessors and Microsystems*, 104622. https://doi.org/10.1016/j.micpro.2022.104622
5. Barbieri, M. C., Grisci, B. I., & Dorn, M. (2024).: Analysis and comparison of feature selection methods towards performance and stability. *Expert Systems With Applications*, *249*, 123667. https://doi.org/10.1016/j.eswa.2024.123667
6. Kumar, A., & Kumar, S. (2023).: Intrusion detection based on machine learning and statistical feature ranking techniques. In *2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 606-611). IEEE.
7. Lan, B., Lo, T., Wei, R., Tang, H., & Shieh, C. (2023).: A quantitative logarithmic Transformation-Based intrusion detection system. *IEEE Access*, *11*, 20351–20364. https://doi.org/10.1109/access.2023.3248261
8. Adhao, R., & Pachghare, V. (2022).: Ensemble of Bio-inspired Algorithm with Statistical Measures for Feature Selection to Design a Flow-Based Intrusion Detection System. *INTERNATIONAL JOURNAL OF NEXT-GENERATION COMPUTING*. https://doi.org/10.47164/ijngc.v13i4.455

9. Heryanto, A., Stiawan, D., Idris, M. Y. B., Bahari, M. R., Al Hafizin, A., & Budiarto, R. (2022).: Cyberattack feature selection using correlation-based feature selection method in an intrusion detection system. In *2022 9th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)* (pp. 79-85). IEEE.
10. Sharma, Y., Sharma, S., & Arora, A. (2022).: Feature Ranking using Statistical Techniques for Computer Networks Intrusion Detection. *2022 7th International Conference on Communication and Electronics Systems (ICCES)*, 761–765. https://doi.org/10.1109/icces54183.2022.9835831
11. Umair, M. B., Iqbal, Z., Faraz, M. A., Khan, M. A., Zhang, Y., Razmjooy, N., & Kadry, S. (2022).: A network intrusion detection system using hybrid multilayer deep learning model. *Big Data*, *12*(5), 367–376. https://doi.org/10.1089/big.2021.0268
12. Imrana, Y., Xiang, Y., Ali, L., Abdul-Rauf, Z., Hu, Y., Kadry, S., & Lim, S. (2022).: χ2-BidLSTM: A Feature Driven Intrusion Detection System Based on χ2 Statistical Model and Bidirectional LSTM. *Sensors*, *22*(5), 2018. https://doi.org/10.3390/s22052018
13. *NSL-KDD | Datasets | Research | Canadian Institute for Cybersecurity | UNB*, https://www.unb.ca/cic/datasets/nsl.html, last accessed 2025/06/28
14. Moustafa, N., & Slay, J. (2015).: UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In *2015 military communications and information systems conference (MilCIS)* (pp. 1-6). IEEE.
15. IDS 2017 | Datasets | Research | Canadian Institute for Cybersecurity | UNB, https://www.unb.ca/cic/datasets/ids-2017.html, last accessed 2025/06/28
16. Famili, A., Shen, W. M., Weber, R., & Simoudis, E. (1997).: Data preprocessing and intelligent data analysis. *Intelligent data analysis*, *1*(1-4), 3-23.
17. Chaudhary, P., Verma, A., & Khari, M.: Harnessing Language Models and Machine Learning for Rancorous URL Classification. In *Cybersecurity and Data Science Innovations for Sustainable Development of HEICC* (pp. 273-288). CRC Press.
18. S, P., & Kallimani, J. S. (2019).: Machine learning based predictive action on Categorical Non-Sequential data. *Recent Advances in Computer Science and Communications*, *13*(5), 1020–1030. https://doi.org/10.2174/2213275912666190417150421
19. Bhati, N. S., & Khari, M. (2022).: An ensemble model for network intrusion detection using AdABoost, random forest and logistic regression. In *Lecture notes in electrical engineering* (pp. 777–789). https://doi.org/10.1007/978-981-19-4831-2_64