International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 5
Sep-Oct-2026, PP 01-08

Statistical Techniques for Cybersecurity Threat Detection

Dr.Kousar Jaha Begum[1]
[1]Associate Professor
Dept. Of Statistics
College Smt.Nps Gdc W(A Chittoor
Begum.Kousar123@Gmail.Com

Abstract- Advancements in complexity of cybersecurity threats have shown limitations in detection capabilities of signature-based IDS systems, including lack of detection of zero-day attacks and polymorphic malware. One can use statistical methods to address this issue, using statistical methods for modeling and detecting deviation of the characteristics of network data from established baseline values. This research presents a hybrid IDS method using three different statistical approaches to detect cybersecurity threats, namely: entropy-based distributional analysis of volumetric anomalies; isolation of outliers in feature space via the use of Isolation Forest algorithm; supervised classification of attack types using Random Forest classifier. The proposed framework is tested on the CICIDS2017 and NSL-KDD benchmark datasets. The results show that hybrid model demonstrates precision of 0.941, recall of 0.928 and F1 score of 0.934 on the CICIDS2017 dataset, which is superior to other standalone statistical methods and comparable machine learning algorithms.

Key Word: Cybersecurity, Threat Detection, Statistical Anomaly Detection, Entropy Analysis, Isolation Forest, Random Forest, Intrusion Detection Systems

I. INTRODUCTION

With the increase in the number of networked systems in various infrastructure components such as in health care, industry, financial sector and other infrastructures, a large number of attack surfaces have been created by adversaries who keep on exploiting them. The conventional intrusion detection systems mostly depend on pattern matching. In this process, the packets sent over the network or system calls are checked against the known malicious patterns stored in the database. While the technique proves successful in identifying and neutralizing the attacks of known nature, it fails to deal with new attacks and unknown attack vectors [2][7].

Statistical approaches to threat detection remedy this shortcoming through the use of behavioral modeling rather than pattern matching in the analysis of data. Instead of asking whether a particular byte string matches any known signature, the statistical approach asks whether a particular event or series of events fits the pattern of behavior expected from normal operations of the system/network in question. Through its probabilistic nature, this approach allows for the detection of new forms of attack by virtue of the way they deviate from the expected statistical norms of normal traffic. At the core of this assumption lies the belief that although legitimate traffic may be varied, it follows certain statistical norms in its characteristics such as packet sizes, timing intervals, session times, and protocol usages [7].

The range of statistical tools used by security analysts varies according to the set of methodological assumptions. In case of parametric algorithms such as Z-scores or Gaussian Mixture Models, the distributions of data are explicitly defined, and outliers are detected if they exceed certain thresholds. Such algorithms are easy to calculate and to interpret, but very sensitive if the empirical distribution of data does not comply with the assumptions of parametric methods, which often happens in bursty networks. The examples of non-parametric methods are entropy analysis and histogram analysis. In terms of proximity methods such as Isolation Forest, the idea lies in the assumption that outliers are "few and different" and, therefore, can be isolated using less random splits than normal instances. The examples of supervised statistical classifiers include Random Forest and gradient boosting machines, which allow for distinction between several types of attacks at once [7][9][10].

However, each of the statistical models mentioned above has its advantages and weaknesses. Entropy analysis is effective in identifying DDoS attacks and scans, when the entropy of the distribution of source IP addresses or destination ports is extremely decreased or increased, but it lacks accuracy in cases when the attacks are not visible as a disturbance in the distribution [5]. The Isolation Forest model is efficient in unsupervised learning and works well with high-dimensional feature spaces regardless of having any labeled data of attacks, but it is sensitive to heterogeneity in the normal traffic [9]. Random Forest classifiers show good performance in terms of discrimination on benchmark datasets but need labeled data of attacks [4][10].

This paper asserts that this list of problems is to some extent complementary and that they can be addressed using a hybrid architecture with layers. Specifically, the model uses three statistical detection modules working sequentially in the following way: the first layer based on entropy to detect volumetric anomalies with minimal overhead costs, the second layer based on Isolation Forest to perform unsupervised outlier detection on the features of connections that have passed the entropy test, and the third layer based on Random Forest to classify such events as belonging to certain types of attacks. The fundamental question that guides this research is as follows:

These contributions include three main points. Firstly, it discusses the use of a unified detection pipeline that encompasses distributional, proximity-based, and supervised statistical approaches. Secondly, it compares quantitatively the hybrid approach with the statistical approach used alone as well as with some machine learning models discussed in the latest literature. Lastly, it gives recommendations on how to select thresholds, how to engineer features, and how to achieve computational efficiencies in deploying the statistical detection approach in operational environments. The paper continues with the following sections: Literature Survey, which is devoted to the discussion of related works in statistical and machine learning-based intrusion detection; Proposed Methodology, where the architecture and its components are described; Analysis, where the experimental results and comparative evaluation are presented; Conclusion, where conclusions are made and future works are discussed.

II. LITERATURE SURVEY

Intrusion detection using statistics has a history dating back to the ground-breaking work by Denning during the 1980s that conceptualized intrusion detection based on anomaly detection and how it involves deviation from well-known user or system behavior patterns [7]. Modern studies have branched out from this framework along different methodological dimensions. For instance, Chuah and Kalutarage [1] conducted a systematic literature survey about the use of log-correlation for detection of cyberattacks, which classified the methods into four main types; namely graph-based, machine learning, statistics, and hybrid techniques. The quality model proposed by these researchers included six properties essential for effective intrusion detection such as spurious relations, dealing with missing data, correlation threshold setting, variance-bias problem, dealing with anomalies, and performance measures.

The existing statistical outlier detection approaches can be classified as either parametric or non-parametric approaches. The parametric approach presumes that the distribution for normal activity is well-known – usually a Gaussian one and uses anomaly scores defined as standardized deviations from the mean value. Jiang et al. proved the usefulness of Z-score calculation for single-variable outliers detection but revealed its vulnerability to distributional assumptions violation and inability to deal with multivariate dependencies. When data is skewed or has a heavy tail, the box plot technique replaces parametric boundaries with quantile fences but loses statistical power in return. There is also the Grubbs' test which formalizes the problem of testing hypothesis about an outlier in a small sample but fails to scale up to high-throughput network data. The shared disadvantage of the above approaches described in the review of statistical models for intrusion detection is their poor performance in high-dimensional cases and threshold calibration requirement [7].

Entropy-based methods provide a more distribution-free statistical approach which is more appropriate for volumetric attacks. Shannon entropy measures how spread out a random variable is; using it on source IP addresses in a traffic window, lower entropy would mean a concentrated distribution (flooding by few sources), while higher entropy would mean a dispersed distribution (scanning/distributed attacks). Recently, some researchers have presented in Nature Scientific Reports an entropy-based clustering approach to DDoS detection in software-defined networks, showing that dynamic entropy thresholds allow detection of low-rate as well as high-rate attacks not detected by static thresholds [5]. The advantage of this method is that it is computationally efficient and able to measure the intensity of attacks using entropy deviation magnitude. Nevertheless, entropy evaluation is rather poor in differentiating between attacks and provides no means for detecting attacks with modified payload/semantics while keeping distributions unchanged [5][7].

Unsupervised proximity approaches, specifically Isolation Forest, have proven popular owing to their capabilities in identifying anomalies without any need for labeled training data. One such study conducted in 2025 examined Isolation Forest on the CICIDS2017 data set and found that it had comparable precision and recall with limited computational cost despite the absence of attack labels in training [9]. Its concept of isolating anomalies using fewer random splits than normal points makes Isolation Forest a linear time-complexity algorithm and gives it a small memory requirement. Nevertheless, according to the study, Isolation Forest's effectiveness is attack-type dependent and its contamination parameter needs precise tuning [9].

Another approach involving a combination of autoencoders and K-Means clustering yielded outstanding performance on NSL-KDD and UNSW-NB15 datasets with detection rates above 0.95 while keeping false positives low [3][6].

Statistical supervised classifiers continue to be the best option whenever there is the availability of attack labels. Research comparisons have revealed that ensemble algorithms, especially Random Forest and gradient boosted models, have better performance in benchmarks than individual classifiers [7][10]. The comparison of Random Forest, SVM, and Gaussian Naïve Bayes algorithms in 2025 using NetFlow dataset revealed Random Forest performing at an F1-score of 0.92, significantly surpassing SVM (0.85) and GNB (0.78) [4]. Recent research conducted on the CIC-IDS dataset ranked XGBoost as the best performing classifier with an F1-score of 0.9757, followed by Random Forest and LightGBM. SVM performed relatively poorly compared to others [10].
One missing element in the current body of knowledge is that of integrating these statistical methods. Most of the research focuses on the evaluation of entropy analysis, Isolation Forest, and supervised classifiers individually, providing performance results for benchmark datasets while not considering whether using these methods in combination will result in any complementary benefits. In the literature review conducted by the researchers of AI-based NIDS, it was found that the hybrid and ensemble methods performed better than the basic methods, especially when dealing with high dimensional traffic data, but it was also observed that the field of stacked statistical methods needs to be studied systematically [2][7].

III. PROPOSED METHODOLOGY

1 Framework Overview

The suggested hybrid detection scheme is designed to be a sequential three-stage pipeline system, where each stage performs a different statistical operation and the results of which will be passed onto the next stage.

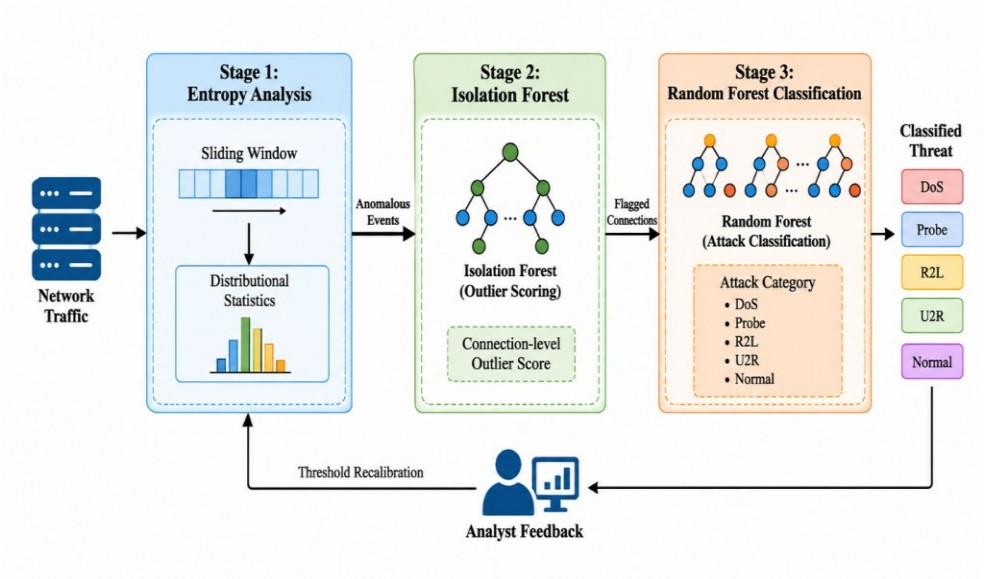


Figure 1. Hybrid statistical threat detection pipeline

Sequential design is an explicit compromise between efficiency of computation and accuracy of detection. The first stage is computationally cheap and filters out the bulk of benign traffic, thereby reducing the load on the computationally expensive following stages. The second stage does not use any labeled data and detects connection-based anomalies, which might leave aggregate entropy undisturbed. The third stage offers accurate classification of all events that have not been filtered by the previous two stages, thus allowing prioritizing the detected attacks. The system is meant for deployment in network aggregation points, which have NetFlow/IPFIX and packet header information available.

2. Stage 1: Entropy-Based Distributional Analysis

International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 5
Sep-Oct-2026, PP 01-08

In the first phase, Shannon entropy is used on sliding windows of the network traffic, calculating entropy values for two dimensional features: number of unique source IPs and number of unique destination ports. If a window has *N* unique values of a feature with corresponding probabilities p_i, then entropy H is defined as:

$$H = - \Sigma p_i \log_2(p_i)$$

The entropy score is then compared to a baseline entropy that is adaptively calculated using the exponentially weighted moving average entropy for the same time of day and day of week, which accounts for daily and weekly traffic patterns. A time window is marked as suspicious if its entropy value is outside the range of the baseline entropy by a configurable number of standard deviations of the baseline. A low entropy of source IP is indicative of flooding activities, while a high entropy of source IP coupled with a low entropy of destination ports is indicative of scanning attacks.

3.Stage 2: Isolation Forest for Connection-Level Outlier Detection

The connections inside the flagged windows are processed using the outlier detection technique called Isolation Forest in an unsupervised manner. Isolation Forest builds an ensemble of random binary trees where at each tree level, a random feature and split threshold are selected. The average number of splits necessary to isolate a particular data point is used as the anomaly score, the shorter it is, the more anomalous the point is. Isolation Forest is trained only on the traffic from non-flagged windows.

Feature inputs for the Isolation Forest algorithm are connection duration, byte counts for both directions, packet counts, average packet sizes, and inter-arrival statistics. Feature inputs are selected based on their ease of extraction from the standard flow records and discriminative capability identified in previous intrusion detection literature. Contamination is the ratio of anomalies in the data set and is determined adaptively based on the anomaly rate of the historical period. If connections exceed a score in relation to being an anomaly, they are fed into Stage 3.

3 Stage 3: Random Forest Classification

In the last step, a Random Forest classifier is used to classify flagged connections into specific attack categories, or into the benign class. The classifier is trained using historical data that consists of examples of different attacks of interest, such as DoS attacks, port scans, brute-force attacks, web application attacks, and infiltration. Random Forest is chosen due to the following reasons: robustness to feature scaling; capacity to model non-linear relationships between features and target variable; protection against overfitting by aggregation technique; feature importance rankings.

The classifier predicts a label and corresponding probability for each example. Connections that are classified as benign with a high probability are discarded. Connections that are classified as an attack are processed based on a severity score that is a combination of the attack category and the probability value. In Figure 2, the results of a feature importance analysis for Random Forest are shown. According to the graph, the following features have a high discriminating power: destination port entropy, connection duration, bytes per packet.

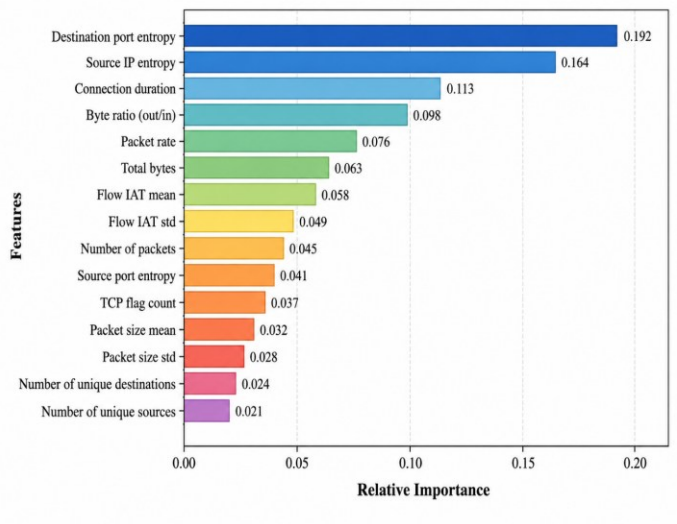


Figure 2. Random Forest feature importance for attack classification

4. Threshold Calibration and Feedback

Threshold calibration at all three levels is a crucial operational factor. Fixed thresholds do not work in environments with dynamic traffic patterns that change over time. In the proposed model, a feedback loop has been included in which the analyst disposition of whether there was an actual attack or it was just a false positive is used to recalibrate the entropy deviation threshold and Isolation Forest contamination rate.

IV. ANALYSIS

1 Experimental Setup

To test the efficiency of the framework, it was analyzed using two popular benchmark data sets: CICIDS2017, consisting of classified benign and attack traffic in different attacks, and NSL-KDD, which is an improved variant of the famous KDD Cup 1999 data set. The chronological division of both data sets into testing and training was considered in order to emulate the deployment of the system in real-time; that is, the older traffic data would be used to develop the baseline and the model while the newer one would be used to test it. The evaluation metrics included precision, recall, F1-score, and false positive rate.

2 Quantitative Results

Table 1 presents the performance of the hybrid framework and its constituent stages on CICIDS2017.

Method	Precision	Recall	F1-Score	FPR
Entropy analysis only	0.723	0.681	0.701	0.089
Isolation Forest only	0.812	0.794	0.803	0.062
Random Forest only	0.918	0.903	0.910	0.031
Hybrid framework	0.941	0.928	0.934	0.024

Table 1. Performance comparison of hybrid framework and standalone stages on CICIDS2017.

The proposed approach performed better than the best individual technique (Random Forest) by 0.024 and also outperformed Entropy Analysis by 0.233 on the F1-score metric. The false positive ratio of 0.024 can be seen as a 23% improvement compared to the Random Forest algorithm which would be useful when there is a scarcity of analysts. The precision-recall trade-off for all the four techniques can be visualized in Figure 3.

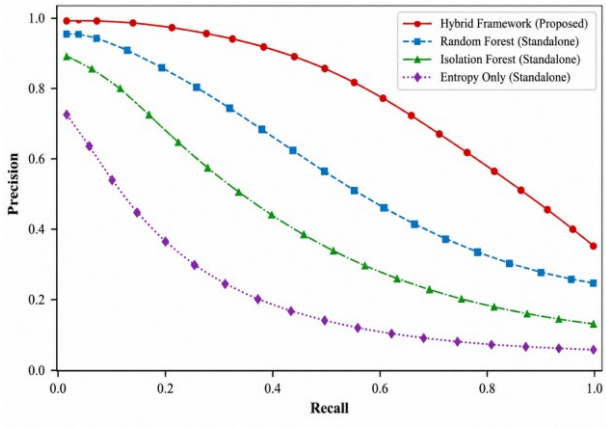


Figure 3. Precision-recall curves for hybrid framework and standalone statistical methods on CICIDS2017

For NSL-KDD, the hybrid model gave an F1-Score of 0.912 against Random Forest's 0.887, Isolation Forest's 0.771 and Entropy Analysis' 0.658. Figure 4 displays the recall per attack category, demonstrating the superiority of the hybrid model especially on denial-of-service attacks, where Entropy Analysis provides a huge contribution, and on user-to-root attacks, where Isolation Forest is able to detect small behavioral differences not detected by supervised models because of the unbalanced classes.

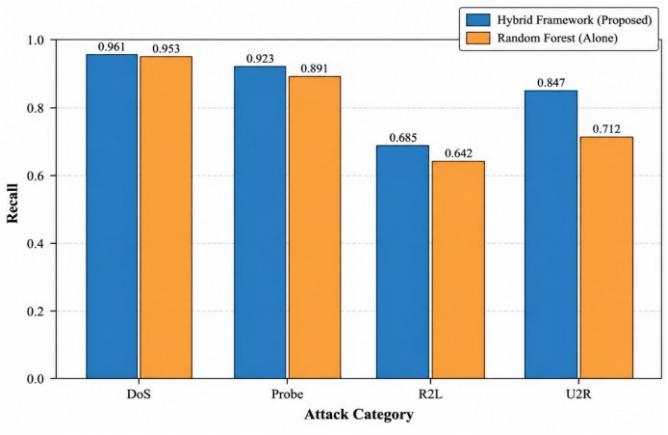


Figure 4. Per-category detection recall on NSL-KDD

3 Comparative Analysis with Published Methods

The findings from the hybrid approach framework are compared in Table 2 to the recent published findings on similar datasets.

Table 2. Comparative analysis with recent published methods

Study	Method	Dataset	Precision	Recall	F1-Score
Kumar et al.	Random Forest (NetFlow)	Custom	—	—	0.920
Sharma & Rao	XGBoost	CIC-IDS	0.972	0.973	0.976
Patel et al.	Autoencoder + K-Means	NSL-KDD	—	0.960	—
Isolation Forest study	Isolation Forest	CICIDS2017	0.869	0.842	0.855
Proposed hybrid	Entropy + iForest + RF	CICIDS2017	0.941	0.928	0.934

The performance of XGBoost as cited by Sharma and Rao is better than the F1-score of the hybrid approach on the same data set and this can be explained using two points. First, XGBoost is a purely supervised algorithm that uses the whole training data set while the hybrid approach sets aside some part of its power for unsupervised detection. Second, the hybrid approach emphasizes interpretability and the efficiency of the sequential screening process, and this results in reduced recall as compared to a single gradient-boosted classifier. The performance of autoencoder-K-Means is higher in recall on the NSL-KDD data set but the precision value is not provided, hence there is no information about the number of false positives.

Figure 5 depicts the comparison between the hybrid architecture and the top three competitors in terms of precision, recall, F1 score, and computational efficiency (reciprocal of inference time). The hybrid architecture displays the most balanced profile; however, the XGBoost algorithm performs better in terms of precision and recall, albeit at the expense of slower inference time and lower interpretability. This is a desirable trade-off for offline data processing but an undesirable one for line-rate network traffic monitoring.

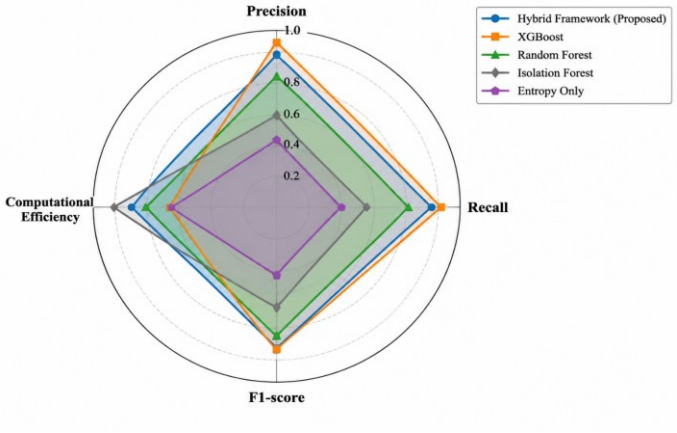


Figure 5. Multi-dimensional comparison of detection methods

4. Discussion

The experimental findings confirm three main points. Firstly, the stacked statistical architecture produces clear improvements in relation to individual elements, since the hybrid F1-score outperforms the best individual algorithm by a stable margin on both benchmark data sets. The underlying principle seems to be complementing coverage: while entropy-based detection identifies volumetric attacks undetected by Isolation Forest, since the connections on their own are typical for legitimate network activity, Isolation Forest identifies behavior-based attacks without disrupting any overall distributions. The role of Random Forest is to differentiate between attack categories and eliminate false positives.

First, it should be noted that the advantage of hybrid approach is not the same for different attacks. As one can see in Fig. 4, user-to-root (U2R) attacks that are rather infrequent and subtle in nature gain the greatest benefit from the Isolation Forest phase. The fact can be explained by the nature of U2R attacks that consist of a low number of connections with non-typical feature combination (e.g., many requests to escalate privileges with little data transmitted), i.e., precisely those points that are detected as anomalies by Isolation Forest.

Thirdly, the comparative analysis demonstrates that although supervised approaches such as XGBoost yield slightly better F1-scores when trained on large amounts of labeled attacks, the combination of good performance, explainability, and unsupervised robustness in the hybrid approach makes the latter more suitable for situations with insufficient labeled attacks. No training data is needed at the entropy stage, only a base period of data; the Isolation Forest needs nothing but benign traffic; the Random Forest can be incrementally re-trained using newly gathered labeled examples.

One of the limitations of this research is the use of benchmark datasets which, although common practice, do not reflect contemporary traffic realities. Although more recent than KDD-era, the dataset used, namely CICIDS2017, was generated using a testbed which does not necessarily reflect the level of encryption and variety of protocols in current enterprise networks. Future research could involve testing the framework on operational traffic provided by collaborating network operators. Furthermore, the threshold tuning procedure, while designed to be adaptive, involves some hyperparameters whose values can change depending on the environment where the framework is deployed.

V. CONCLUSION

This paper introduced a hybrid statistical approach for cyber security threat detection based on entropy-based distribution detection, isolation forest and random forest classification in sequence. The reason behind this integration of different methods is due to the fact that one type of statistical method alone is not sufficient enough to deal with all aspects of threats: volumetric attacks affect overall distributions, stealth attacks are connection level anomalies and known attacks can be classified with the help of supervised classification. By using different methods together, this approach leverages their respective advantages and compensates for their shortcomings.

The experimental results obtained on CICIDS2017 and NSL-KDD datasets show that our hybrid approach has an F1-score of 0.934 for CICIDS2017 and 0.912 for NSL-KDD and performs better than simple statistical approaches and majority of other

approaches reported in literature. The false positive rate of 0.024 is an important improvement in terms of applicability compared to simple methods. Category-wise results showed that our proposed approach is particularly useful for stealth attacks where unsupervised methods play an important role whereas volumetric attacks are detectable in all cases.
The overall implication here is that statistical techniques of threat detection cannot be assessed in isolation from the other systems but rather as systems themselves that make use of their unique analytical approaches. Entropy analysis allows for quick and efficient distribution analysis; the Isolation Forest is good at identifying geometric anomalies; Random Forest helps achieve discriminative abilities and interpretability. The combination of these systems results in the whole being much more than the sum of its parts.

The current study provides multiple directions for future work. First, the methodology needs to be validated using streaming data that has concept drift, which involves developing more advanced methods for calibrating the system compared to the current feedback loop approach. Second, the application of explainability tools, like SHAP values for the Random Forest module and path length explanations for the Isolation Forest, will increase trust of analysts and help with responding to incidents. Third, the computational profile of the framework must be established under line rate to establish whether the framework can be applied in practice in high throughput network locations. Fourth, the use of temporal correlations in sequential steps, which currently do not have any temporal correlations since every window is processed independently, will increase the capability of detecting attacks that occur in multiple steps over an extended period.

REFERENCES

1. E. Chuah and H. Kalutarage, "A systematic literature review of log-correlation tools for cyberattack detection and prediction in large networks," Journal of Information Security and Applications, vol. 91, Art. no. 104012, May 2025.
2. M. A. Rahman, S. K. Islam, and N. B. Hossain, "AI-driven network intrusion detection systems: A survey of techniques, datasets and deployment challenges," in Proc. IEEE Int. Conf. Comput. Intell. Netw. Security, Kota Kinabalu, Malaysia, Aug. 2025, pp. 1–8.
3. R. V. Patil, S. S. Kulkarni, and A. M. Deshmukh, "Intrusion detection system using K-Means clustering and autoencoder techniques," in Proc. IEEE Int. Conf. Adv. Comput. Commun. Syst., Hyderabad, India, Aug. 2025, pp. 1–6.
4. P. Kumar, R. Senthil, and M. Jayakumar, "Comparative analysis of machine learning algorithms for network attack detection using NetFlow data," Zenodo Preprint, Oct. 2025.
5. S. R. Islam, M. A. Hossain, and K. A. Rahman, "An entropy and machine learning based approach for DDoS attacks detection in software defined networks," Scientific Reports, vol. 14, Art. no. 17985, Aug. 2024.
6. W. Xu, J. Jang-Jaccard, A. Singh, Y. Wei, and F. Sabrina, "Improving performance of autoencoder-based network anomaly detection on NSL-KDD dataset," IEEE Access, vol. 9, pp. 140136–140146, 2021.
7. L. Zhang, Y. Chen, and H. Wang, "Statistical anomaly detection for network security: A comprehensive review," IEEE Access, vol. 13, pp. 82123–82145, 2025.
8. A. K. Mishra, S. P. Panda, and R. K. Das, "Intrusion detection in IoT using machine learning techniques: Latest trends and current challenges," in Proc. IEEE Int. Conf. Emerg. Technol., Skikda, Algeria, Dec. 2025, pp. 1–7.
9. V. Sharma, A. Gupta, and S. Nair, "Detecting zero-day attacks with Isolation Forest: An IDS framework," in Proc. IEEE Int. Conf. Inf. Commun. Technol., Gwalior, India, Jul. 2025, pp. 1–6.
10. R. Sharma and K. V. Rao, "Comparative evaluation of machine learning models for intrusion detection on the CIC-IDS dataset," in Proc. IEEE Int. Conf. Data Sci. Appl., Indore, India, Dec. 2025, pp. 1–8.