Jestr

Journal of Engineering Science and Technology Review 18 (1) (2025) 199-209

Research Article

JOURNAL OF
Engineering Science and Technology Review

www.jestr.org

Evaluating Effectiveness: A Critical Review of Performance Metrics in Intrusion Detection System

Vandana Kadam* **and Rakesh Verma**

Analytics and Data Science, Indian Institute of Management, Mumbai, 400 087, India.

Received 27 August 2024; Accepted 24 December 2024

Abstract

This critical review analyzes the effectiveness of performance metrics for Intrusion Detection Systems (IDS) and applies the PRISMA methodology to conduct a systematic review. The investigation focuses on core metrics like accuracy, precision, recall, specificity, and F1 score, where each exerts varied information about IDS performance. On the PRISMA framework, we systematically conducted searching, screening, and analysing academic articles on performance-related IDS topics. The articles are searched for in reputed academic article indexing databases such as Scopus, IEEE Explore, Web of Science, and the ACM digital library. Based on the objective evaluation model for symptoms, the major findings of the study are that while accuracy is clear and relatively easy to conceptualise and compute, it often gives no real insight. Our findings give a comprehensive framework for further research and practical measures to be implemented so that increased effectiveness is ensured in threat detection strategies.

Keywords: Intrusion detection system, Accuracy, Precision, Recall, F1 score, PRISMA

1. Introduction

IDSs form critical components of information system security, ensuring system integrity. An IDS is a monitoring system for networks or systems that could involve possible malicious activities or system policy violations. It has early warnings that assist an organization in preventing the harmful effects of a potential security breach from causing significant destruction to the organization [1, 2]. The basic requirement of IDS is founded on the fact that the threat landscape is changing and common securities through firewall and antivirus applications are not enough against APTs, zero-day exploits, and insider threats [3]. They monitor the traffic on the network and identify any suspicious activity, indicating to the administrators a potential security breach. An IDS works mainly to identify attempts to gain unauthorized access inappropriate patterns of activity, and symptoms that may arise from a skilled attack. In a nutshell: it can detect such activities in real time and make alerts for administrators to act immediately to block or thwart the threat [4]. Because an attack can led to data theft, system downtime, a financial loss, and may tarnish an organization's image, this approach proves to be crucial for reducing such potential impacts.

The effectiveness of IDS is wholly dependent upon the performance metrics of detection and response to intrusions [5]. Because cyber threats are continuously changing both in sophistication and in vector numbers, the detection of intrusion has considerably gained importance. Cybersecurity Ventures predicts global cybercrime costs will balloon to $10.5 trillion annually by 2025 [6]. With organizations managing the sea of threats, from malware to advanced persistent threats, IDS have a pivotal role in making early detection and mitigation of such attacks successful [7]. The compelling need to protect sensitive information and provide assurance of the confidentiality, integrity, and availability of the systems has been the driving force behind the development and deployment of the IDS. The conventional security controls with firewalls and antivirus are outdated and stand deficient to probably catch anywhere near the sophisticated cyberattacks [8]. In this context, IDS is an integral part of security mechanisms. They provide an added layer of "net" to computer systems against cyber threats, designed to detect and react to attacks which other firewalls and other security applications may overlook or breach [9]. The importance of IDS is increasing due to the governmental laws, regulation needs, and industry standards demanding having adequate and appropriate complex security systems in place. As an illustration, the United States' Health Insurance Portability and Accountability Act (HIPAA) requires security mechanisms, including IDS, to be used in healthcare organizations for the security of electronic health information [10].

In this study the performance metrics of the IDS will be systematically reviewed, including their respective limitations. Knowing the metric limitations will guide the reader toward assessing the IDS performance evaluation status in a more informed manner and, in turn, help to find areas for improvement and future research.

Measuring the performance of the IDS truly is a task of many facets, with many metrics at a given time. For instance, metrics such as overall accuracy, precision, recall, and F1 score, variously regard the IDS, but an individual metric cannot realize the IDS performance genuinely on its own [11]. This review is intended to summarize what is known up to now, identify the gaps in knowledge, and point towards what future research might look like in order to develop effective IDS evaluation frameworks [12]. Even with the use of different performance measures, no specific standardized framework for IDS evaluation exists. The lack of standardization complicates the comparison of different IDS solutions and hampers, in general, the development of best

*E-mail address: vandana.kadam.2015@iimmumbai.ac.in
ISSN: 1791-2377
doi:10.25103/jestr.181.20

practices for IDS deployment and management [13-14]. Additionally, considering frequent changes in the cyber threat landscape, those updates to the evaluation frameworks should be continuous in most relevance and effectiveness assurance [15].

This review allows a reflection on practical issues that could be faced by security experts when implementing and interpreting performance metrics. This way, an organization will be prepared to make informed decisions about the deployment, configuration, and maintenance of an IDS. In this paper, a critical look at current practices is presented, hopefully highlighting areas for improvement; it is expected that such an exposition will set the stage for increasing research efforts into improving the state of IDSs.

2. Methodology

In this study, several journal articles indexing databases like Scopus, Web of Science, IEEE Explore, ACM digital library, and Google Scholar are used to find the articles. Scopus, Web of Science, and IEEE Xplore are frequently used by researchers to source academic articles [16]. Coverage in these databases is comprehensive, with peer-reviewed literature across disciplines. Scopus has extended tools for citation analysis and a broader range of indexed journals compared to Web of Science, which is recognized to have the most rigorous indexing and impact factor metrics [17]. Engineering and computer science are particularly favourable toward IEEE Xplore for high-quality access to conference proceedings and journals [18]. It enables researchers to check on credible and influential sources with which to reference their work. To access the latest articles, we restricted the publication year to 2001-2025 only, published in the English language. The following figure.1 shows the trend of publications that discussed the performance metrics in intrusion detection systems.

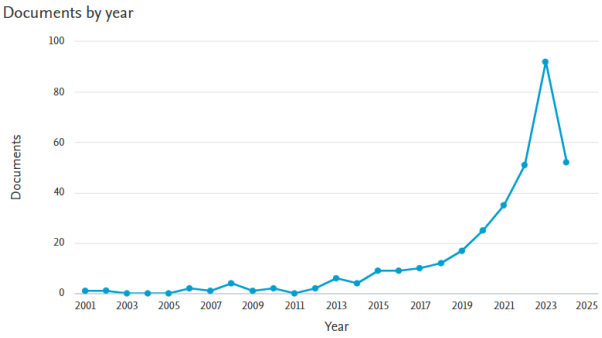

Fig. 1. Publication trends from 2001-2025

2.1 PRISMA Approach

This literature review adopts the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology to ensure a systematic and transparent approach to synthesizing existing research on performance metrics in intrusion detection systems [19-20]. PRISMA is widely recognized in academia for its structured framework, which helps minimize bias and enhance the rigour of literature reviews [21] (Mishra & Mishra, 2023). PRISMA consists of a four-phase flow diagram that guides the identification, screening, eligibility, and inclusion of studies in systematic reviews [19]. By following PRISMA guidelines, this review aims to provide a comprehensive and unbiased synthesis of the available literature on IDS performance metrics, including both quantitative and qualitative studies [22].

2.2 Selection Criteria

The selection of literature for this review was guided by specific inclusion and exclusion criteria. Relevant studies were identified through comprehensive searches of academic databases such as IEEE Xplore, ACM Digital Library, Scopus (Fig. 2). The search terms included variations of "intrusion detection systems", "performance metrics", "evaluation", "accuracy", "precision", "recall", and "F1 score", “completeness”. Only peer-reviewed journal articles, conference papers, and technical reports published in English from the past two decades were considered to ensure relevance and currency of information.

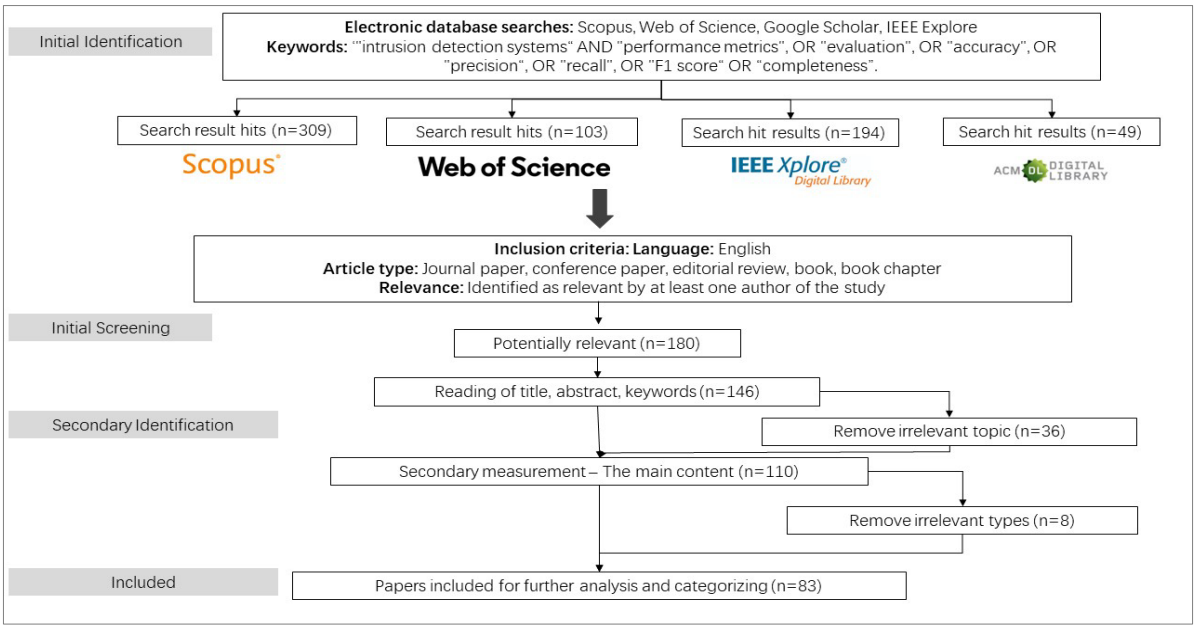

Fig. 2. PRISMA literature review procedure

The inclusion criteria focused on studies that evaluated performance metrics in the context of IDS, either through empirical testing, theoretical analysis, comprehensive review or comparative studies. Studies that provided insights into the strengths, limitations, and applicability of different metrics were prioritized for inclusion in the review. The researcher of this study manually went through the details of each article and filtered the relevant articles. The authors' institute has subscriptions to Scopus, Web of Science, IEEE Explore, and ACS digital libraries. The searching and filtering of articles as done in May 2024.

2.3 Data Extraction and Synthesis

Data extraction involved systematically collecting information from selected studies, including study objectives, methodologies, metrics evaluated, and main findings related to performance metrics in IDS. The extracted data were synthesized to identify common trends, gaps in research, and emerging themes regarding the use and limitations of performance metrics. Synthesis of the extracted data involved categorizing and summarizing findings from individual studies, comparing methodologies and outcomes, and identifying recurring themes or patterns in the literature. This process enabled a holistic understanding of the current state of research on IDS performance metrics and provided a basis for critical analysis and discussion in subsequent sections of the review.

2.4 Quality Assessment

The quality and reliability of each article were then sought to be established through its careful evaluation against predefined criteria: methodological rigour, relevance to the topic of the review, and clarity of findings reported. Those studies considered to have fallen below the predefined threshold of quality were excluded from synthesis to protect their integrity. Based on this, ensuring quality parameters are maintained, the researcher of the present study went through proper scrutiny as to why the article was held to be included or excluded for further analysis. Quality assessment criteria used were whether the research objectives were clearly stated, the appropriateness of the research design and methodology used, whether data collection and analysis were transparent, and the conclusions drawn from the findings were valid. This review applied stringent quality criteria to the rigours of academic standards and ensured the trustworthiness of the synthesis of its findings.

2.5 Limitations

Despite the adoption of a systematic approach in this review, we want to declare several of its limitations. Since this search is based on the published literature, as usual, publication bias may happen, as human studies that have statistically significant findings are more likely to get published. Also, the exclusion of publications in a language other than English and grey literature may further limit comprehensiveness. Such limitations could be overcome in future studies through the adoption of a wider source range with advanced search strategies, thereby seeking to capture a myriad of perspectives on performance metrics in IDS.

3. Performance Metrics in Intrusion Detection

3.1 Overview of Performance Metrics

Performance metrics in intrusion detection systems (IDS) are essential for evaluating their effectiveness in detecting and responding to security threats. These metrics provide quantitative measures of various aspects of IDS performance, such as detection accuracy, efficiency, and reliability. Commonly used metrics include accuracy, precision, recall (sensitivity), specificity, F1 score, ROC-AUC (Receiver Operating Characteristic - Area Under Curve), and others [23].

3.1.1 Accuracy

Accuracy measures the proportion of correctly classified instances (both true positives and true negatives) out of all instances examined. While accuracy provides an overall assessment of IDS performance, it may not be sufficient when dealing with imbalanced datasets, where the number of normal instances far exceeds the number of intrusion instances [24].

$$\text{Accuracy} = \frac{True\ Positives + True\ Negatives}{Total\ Instances} \quad (1)$$

One major advantage of accuracy is its simplicity. Computing and understanding it is very easy, so it makes for a good metric for one who just wants to make an initial evaluation of IDS performance. The simplicity of this metric makes it convenient to do assessments and quick comparisons. Furthermore, accuracy provides a broad view of how good or bad an IDS performance is. Probably one of the most interesting things about accuracy in this context is that it works very well for benchmarking the various IDS solutions out there. This allows drawing direct comparisons between different systems or configurations so that the choice of the most appropriate IDS for any specific environment will be easy to make. In the context of IDS, accuracy measures the system's ability to discriminate between various activities, mostly legitimate from malicious ones. High accuracy would mean, therefore, that it had been up to the mark in identifying intrusions alongside normal activities enough to be reliable over network security. Accuracy may be the first metric looked at while assessing the performance of IDS since it gives an instant feel about how good a system is doing.

3.1.2 Precision

Precision, also known as positive predictive value, measures the proportion of true positive alerts out of all alerts raised by the IDS. It indicates how precise the IDS is in identifying true intrusions without generating false alarms.

$$\text{Precision} = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (2)$$

This accuracy is, therefore, very appropriate in contexts where the cost of false alarms may be high. For example, running real-time monitoring systems would tend to get overwhelmed by many false positives, resulting in alert fatigue by security personnel and possibly missing real threats. High precision thus puts forward the expectation that the IDS is pretty good at filtering out so-called false alarm traffic, hence assuring that most of the alerts issued by the tool will be related to intrusion. This makes precision a rather good metric to convey the quality of an IDS's positive predictions. In practice, precision aids in establishing the reliability of a given IDS in detecting threats with low false positives. It will help security analysts judge how much credibility they should accord to the alerts raised by the IDS. With a focus on precision, organizations then pay more attention to the implementation of those IDS solutions which give more actionable and credible alerts in enhancing their security

infrastructure. This will also help in comparing different IDS models or configurations, showing which one is more balanced and would detect intrusions without too many false positives.

3.1.3 Recall (Sensitivity)

Recall measures the proportion of true positive alerts that are correctly identified by the IDS out of all actual intrusion instances. It indicates the ability of the IDS to detect all intrusions without missing any [25].

$$\text{Precision} = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (3)$$

It is particularly useful when security under concern needs to find out all the possible intrusions. A high recall value suggests that an IDS would be able to pick out most, if not all, of such intrusions and similarly be good at curtailing the threat of undetectable menace. This metric is basically for the security coverage needs, and missing out on intrusion would mean very significant vulnerabilities or even bigger breaches. Recall can guide the tuning and optimization of IDS systems by allowing the trade-offs between the thoroughness of detection and the incidence of false alarms to be quantified. For instance, raising the sensitivity of the IDS to enhance recall could at the same time compound false positives and hence burden security teams with many more alerts. In the analysis of recall alongside precision, which can measure accuracy in positive detection, security professionals can try to strike a balance that maximizes detection capability against efficiency.

More precisely, recall is very vital when comparing IDS models or configurations from various models against one another. An organization can identify the fact that which of its systems, will be more effective at detecting intrusion in different scenarios, by comparing recall values. This greatly helps in selecting and deploying IDS solutions toward gaining maximum security coverage concerning their own specific needs.

3.1.4 Specificity

Specificity measures the proportion of true negative alerts out of all instances that are actually non-intrusions. It indicates the IDS's ability to correctly identify non-intrusions and avoid false alarms.

$$\text{Specificity} = \frac{True\ Negatives}{True\ Negatives + False\ Positives} \quad (4)$$

One of the major advantages of the F1 score is that it provides only one metric, which encompasses the ability of the system to avoid false positives-precisions and to detect real intrusions-recall. This will be very useful, especially in the case of conflicting precision-recall performance, since the F1 score will balance them out and provide a comprehensive review on the performance of the IDS. A high F1 score indicates that the system does a great balance between precision and recall, which means it correctly classifies instances of intrusions without generating too many false alarms. For practical purposes, the F1 score is viewed as a way to compare and contrast the performance of different IDS models or configurations. This helps explain to security professionals exactly how an IDS is doing on not just detecting intrusions but minimizing false positives. This hence proves quite important in an environment where both false positives and missed intrusions cost dear. The F1 score will help ensure that, from the organizational point of view, there is an adequately reliable and exhaustive security cover from the given intrusion detection system.

3.1.5 F1 Score

The harmonic mean of recall and precision yields the F1 score, which is a balanced metric that incorporates both measures. It is particularly useful when there is an uneven class distribution in the dataset.

$$\text{F1 score} = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (5)$$

Another major advantage of the F1 score is that it gives just one metric, which prosecutors the understanding of what the system does to avoid false positives and its ability to detect true intrusions. This will be important in scenarios where precision comes at the cost of a recall or vice versa since the F1 score balances them; therefore, it can provide a comprehensive evaluation of the performance of the IDS. A high value in F1 means that the system has attained a good balance between precision and recall, directly interpreting the fact that it can precisely identify intrusions but has a foreground, hence avoiding excessive false alarms. The F1 score is practically useful in evaluating and comparing several performance measures for different IDS models or configurations. It helps security professionals understand how well an Intrusion Detection System is performing, either on its ability to detect intrusions or its ability to minimize false positives. This balanced view of performance is important when false positives and missed intrusions are expensive in equal portions in the working environment. By putting more emphasis on the F1 score, an organization is guaranteed to have its IDS provide reliable and comprehensive security coverage.

The F1 score can also guide tuning and optimization for IDS. In conjunction with the component metrics of precision and recall, looking at the F1 score will help security professionals find out where improvements in the system are needed and make an informed decision for optimized performance. For example, if an IDS has high recall along with poor precision, resources can then be focused on reducing false positives without impacting the system's intrusion detection capability. In such cases, targeted optimization may help to enhance the effectiveness of the IDS. In addition, the F1 score makes a comparative study and the selection of the IDS solutions possible. By comparing the F1 score of various systems, an organization can know which models better trade-off intrusion detection and reduction of false positives. This will facilitate the selection and fielding for the most effective and most efficient solutions, thus helping to obtain needed security coverage for the fielding of IDSs.

3.1.6 ROC-AUC

It simply plots the ROC curve of true positive rates against the false positive rate for different threshold settings. Already one measure which can be derived directly from the ROC is the so-called Area Under the Curve (AUC) as a single number giving the measure of the receiver operating characteristic of the IDS concerning the performance of being able to discriminate between normal behavior and intrusions. The higher the value of the ROC-AUC, the better the discrimination performance. In practice, because it is robust to imbalanced datasets often the case in intrusion detection, the ROC-AUC score historically dominated IDS evaluations.

Focusing on the efficiency of a classifier to draw a boundary between normal and intrusive activities, the ROC-AUC score gives insight into how well the models are at detecting attacks with very low false alarms. This is an important capability in real-world applications of timely and accurate anomaly detection, critical to system security. Moreover, the ROC-AUC score is very useful during model selection and optimization. It quantifies the discriminative power of classifiers under different operational conditions. It provides a dependable benchmark for measuring improvements in the performance of IDSs compared to baseline methods or when comparing different detection algorithms. It will also help researchers and practitioners find the optimal configurations in comparative analysis, some of which can balance detection sensitivity with specificity to improve general reliability and efficiency of intrusion detection systems.

For IDS evaluation, this very high value of the ROC-AUC score would mean very good model performance quality, signifying a great ability to distinguish normal activities from malicious ones at very minimal false alarm rates. This not only raises confidence in an IDS deployment; the results also support continuous refinement and adaption against ever-evolving cyber security threats. By having that insight from the ROC-AUC score, organizations can perform a better risk priority resource allocation and deftly manage risk while maintaining robust defenses in an increasingly sophisticated set of intrusion attempts within dynamic digital environments.

3.1.7 Completeness

The "completeness" metric in an intrusion detection system gives information about the system's ability to detect any possible intrusions into a network. This shows how an IDS will detect different styles of attacks under different conditions so that no attack or harmful situation will go undetected. Completeness is therefore considered critical in ensuring comprehensive security by indicating the ability of the intrusion detection system to cover the entire spectrum of threats. However, high completeness may be hard to achieve due to the very nature of cyber threats, and mechanisms for detecting them have to be incessantly updated since they are ever-evolving in nature. Thus, an IDS of higher completeness reduces the risk of undetected intrusion and strengthens the overall security of a network.

This makes completeness a key metric for measuring IDS performance, which otherwise has direct implications for the system's effectiveness in terms of threat detection and response. Measuring the percentage of detected attacks against the total known number of threats, completeness presents a clear metric for the operational readiness and reliability of the IDS. It is a metric that helps security analysts and system administrators estimate if the principal aim of the IDS, which is the identification and mitigation of possible security incidents in real-time, can be achieved. Besides, completeness enables informed decisions on IDS deployment and configuration tuning. Knowing exactly where there are gaps in detection coverage, organizations can prioritize enhancements in sensor placement, rule refinement, or algorithmic improvements to improve overall threat visibility and response capabilities. Finally, a high completeness score represents that a strong IDS implementation will protect against most threats of known classes and thus provide power to assist in making an organization resilient; it reduces the risks that are likely to be involved in case of a breach of cybersecurity.

3.2 Few important articles in the literature

Table 1. Most (rated) significant studies in literature

Authors	Citations	Objectives of the study	Limitations
[26]	96	A comprehensive survey of various performance metrics used in IDS evaluation, including accuracy, detection rate, false positive rate, precision, recall, and F1-score.	Does not provide empirical validation; lacks detailed analysis of trade-offs between different metrics.
[27]	27	Analysis of commonly used metrics like True Positive Rate, False Positive Rate, ROC curves, and AUC. Proposes new metrics such as Cost per Detection.	Limited focus on real-world applicability and operational challenges of implementing new metrics.
[28]	260	Proposes a standardized benchmarking framework using metrics such as detection latency, throughput, and resource utilization.	Benchmarking framework requires extensive computational resources; not easily replicable in smaller settings.
[29]	49	Evaluation of machine learning models for IDS using metrics like precision, recall, F1-score, and confusion matrix analysis.	Focuses primarily on machine learning approaches; less emphasis on traditional IDS techniques.
[30]		Introduces a comprehensive metric framework including security effectiveness, operational performance, and cost metrics. Provides a case study application.	Framework complexity can be a barrier to practical implementation; requires extensive data collection and analysis.
[31]	39	Focuses on metrics specific to anomaly-based IDS, including detection accuracy, time to detection, and anomaly score thresholds.	Limited generalizability to signature-based IDS; lacks discussion on metric standardization across different IDS types.
[32]	114	Discusses challenges in evaluating network-based IDS and introduces metrics such as detection rate under load, and resilience to evasion techniques.	The article discusses the challenges in evaluating machine learning-based NIDS, such as the variability in attack patterns and the dynamic nature of network environments. It provides recommendations for researchers and

			practitioners to improve the reliability and robustness of their evaluations.
[33]	59	Reviews various performance metrics and proposes strategies for improving IDS performance through metric optimization.	Lack of detailed empirical results to support proposed strategies; theoretical focus.
[34]	116	Comparative analysis of classification techniques using different metrics and their applicability in various IDS scenarios.	Comparative study limited to a few metrics; does not cover recent advancements in IDS evaluation.
[35]	121	This study summarises the metrics of IDS and discusses all metrics carefully.	This study highlights that there is no benchmark metric exists to date for intrusion detection and finalizing it is still under process.
[36]	338	This study measures the effectiveness of multiple machine-learning models using several performance metrics	To save the availability and the confidentiality of the network resources, single performance metric is not sufficient to detect the intrusion. Multiple performance metrics should be taken into consideration for measuring the efficiency or effectiveness of IDS.

3.3 Metrics Evaluation in IDS

Evaluation of performance metrics in IDS involves empirical testing, theoretical analysis, and comparative studies to assess their strengths, limitations, and applicability in different scenarios. Researchers use benchmark datasets, simulation environments, and real-world deployments to evaluate IDS performance metrics under varying conditions [12]. Studies by [23] highlight the importance of selecting appropriate metrics based on the specific goals and operational requirements of the IDS. For example, in high-security environments such as financial institutions or government agencies, precision may be prioritized to minimize false alarms and ensure operational continuity. In contrast, in environments where detecting all intrusions is critical, such as military or defence sectors, recall may be emphasized to avoid missing any potential threats [37].

3.4 Challenges and Considerations

While performance metrics in the IDS can be beneficial, they have also presented numerous challenges. In fact, trade-offs within the basic metrics, such as that between precision and recall, are one of the foremost. That is, in most situations, maximizing one metric will tend to work to the diminution of another, and a balanced approach calls for a need to take different operational goals and threat environments [38] into consideration.

Another challenge arises from imbalanced datasets where the amount of normal instances far outweighs the amount of intrusion instances. In this kind of situation, the traditional metrics, including accuracy, will be deceiving because high accuracy can be obtained when the classifier correctly classifies the majority class (normal instances) while falsely treating the minority class (intrusions). Techniques including oversampling, undersampling, and synthetic data usage have been applied to treat such imbalanced data and aid in the improvement of approximations for the performance evaluation of IDS [39].

3.5 Emerging Trends and Future Directions

Conversely, the ongoing trend in IDS research is the further development of more sophisticated metrics capable of accounting for more dynamic and evolving threats. Such metrics have previously been referred to in other contexts but continue to gain increased active attention, with research aimed at new attack vectors for existing IDS gaining research attention [40-41]. Such metrics will generally enhance the resilience and effectiveness among IDS toward the detection and mitigation of advanced cyber threats. Additionally, with these IDS frameworks, the integration of machine learning, artificial intelligence, and big data analytics holds promise in developing predictive and proactive metrics to foresee and, to a degree, curtail possible breach occurrences in advance. These two aspects together create promise for development [13, 42]. Future research will explore the technologies and their impact on IDS performance.

3.6 Usability or integration issues for organizations deploying IDS.

IDS generates a lot of information including alarms and logs, without the help of automation or proper tools it becomes very difficult for the security or IT support people to use these complex performance metrics. This can lead to overlooking the frequency of alarms, ignoring the alerts after a period of time, leading to fatigue and stress for the people handling the cyber security and finally missing nuanced attacks. IDS have often been subjected to high false alarm rates, and without the expertise and proper systems to handle noise, and misinformation the reliance on the system decreases over time. The user interface of IDS should be simple and transaction time should be quick so that the cyber security personnel can quickly log into the system and check the alarm metrics quickly and helping them to make decisions rapidly. A normal IDS may simply give the alarms without providing any actionable insights and recommendations for corrective actions.

Apart from the usability issues, there are several integration-related issues in effectively utilising IDS. Organisations often install equipment and software purchased from multiple vendors having different operating systems and software, the integration of IDS with existing systems is a complex and time-consuming process. Scalability and maintenance of IDS is also one of the biggest challenges because adjusting the IDS capacity as per the growing IT infrastructure without causing any bottlenecks in the systems is a costly procedure. Also, organisations handling personal data, need to ensure that all policy and regulatory laws have been complied with the performance of IDS.

4 Drawbacks of Performance Metrics

4.1 Limitations of Traditional Metrics

Performance metrics in an intrusion detection system play a critical role in determining the effectiveness in which security

threats are detected and mitigated, as reported by [43]. Nevertheless, they bear inherent limitations and, therefore, they need careful addressing to ensure that their evaluation is done accurately and meaningfully.

4.1.1 Imbalanced Datasets and Misleading Metrics

A prominent issue faced by evaluation methods for IDSs occurs during the occurrence of imbalanced datasets, where a huge amount of normal instances, representing benign traffic, are compared to intrusion instances, representing malicious activities [44-45]. Common metrics such as accuracy, precision, and recall are usually unreliable in the case of imbalanced datasets because they are mainly dominated by the majority class (normal instances) and almost completely ignore the minority class (intrusions) [46-47].

For instance, accuracy shows which proportion of correctly classified cases—both true positive and true negative is within the overall amount of cases researched. In other words, high accuracy can be found in a classifier that predicts the majority class when the datasets are imbalanced, inferring that such high accuracy is not an effective intrusion detection mechanism [48, 39]. Precision and recall are two measures whose trade off is characteristic of imbalanced datasets. While it may be possible to achieve such an objective by having a maximal recall, the number of false negatives is high, making maximization of precision the goal. On the other hand, maximizing either recall or precision induces tremendous amounts of false positives and false negatives, respectively, because it increases the correct instances classifying as normal or intrusion [49, 12].

4.1.2 Contextual Challenges and Operational Realities

Performance measures related to IDS effectiveness probably depend upon context-specific issues that involve network topologies, traffic patterns, and sophistication in attack techniques [1, 50]. This only serves to illustrate the point more strongly: metrics performing well in a controlled lab environment might not generalize in reality to large, complex, dynamic network environments with a diversity of traffic types and volumes [23]. Operational challenges include high false rates, resource constraints, and timely response to the threats detected, with which IDS deployment in real-world conditions commonly contends [51]. This means that many operational realities could be missed out from the traditional metrics used to evaluate them, thereby resulting in huge discrepancies between the laboratory evaluations and the actual field performance [52, 7].

4.2 Limitations of Single Metric Assessments

This is because, in many instances, for a single performance measure, the estimation of the performance of an intrusion detection system could provide a very narrow, if not misleading view, of the overall performance of the system. Using only accuracy, precision, and recall as performance metrics is wrong when evaluating the performance of an IDS since they bring forth different characteristics of the latter. For example, accuracy is an overall measurement of classification correctness and therefore does not identify the type of errors committed [53]. Precision and recall reflect how well the IDS can reduce false alarms and detect intrusions, respectively; however, when trying to optimize one such metric, it could be at the cost of the other [54]. Multi-metric, integrated evaluation frameworks should consider their trade-offs to give a balanced view of the performance of an IDS [55, 37].

4.3 *Ethical* and Legal Implications, Algorithmic biases

This is where the use of performance metrics in IDS raises the highest ethical and legal considerations. Metrics that overemphasize detection accuracy can turn into a deal-breaker, spotting sensitive information from users in an event dated back without consent or proper regulator supervision [56]. This could further promote discriminative effects in the event of biases in training data or simply in algorithmic decisions, generally increasing already existing gaps in the digital access to resources and digital opportunities. In that light, there is a greater need for regulatory frameworks and industry standards for the responsible use of IDS performance metrics to ensure compliance with ethical considerations and legal obligations [57]. When the models are trained on the imbalance or incomplete or inappropriate information, this may induce algorithmic biases in the IDS leading to skewed results. For example, if the historical data has already flagged an event as malicious, the events similar to that will be tagged as malicious events and may give false alarms. To overcome these limitations organisations should use balanced datasets, and data from diverse settings, and implement context-aware machine learning models to reduce biases.

IDS also have to deal with the ethical issues which may arise from the data. For example, based on the historical data or frequency of attacks arising from certain geographical locations or communities or user profiles, if these data have been stored as imbalanced data and provided for further training the IDS system, in the near future it may flag the request arising from that particular geographical location or group as a malicious activity. Such issues, certainly damage the reputation of the companies in this emerging world, to avoid these issues organisations may use balanced and rich datasets for training and also maintaining transparency in designing and deploying the algorithms is crucial.

4.4 Mitigation Strategies and Future Directions

To address the drawbacks of performance metrics in IDS, researchers and practitioners are exploring several mitigation strategies:

- **Advanced Machine Learning Techniques:** Employ advanced machine learning algorithms with ensemble methods and deep learning to enhance the robustness and adaptability of IDS performance metrics [38, 58].
- **Context-Aware Evaluation Frameworks:** Develop context-aware evaluation frameworks concerning exact operational environments, threat profiles, and organizational priorities [59, 12].
- **Human-in-the-Loop Approaches:** Methods that incorporate human expertise and domain knowledge in the process of making evaluations about IDS to improve the interpretability and relevance of performance metrics [60, 7].
- **Regulatory Frameworks:** They involve the development of regulatory frameworks and industry standards that determine the responsible use of IDS performance metrics to ensure conformity to ethical principles and legal requirements [61, 23].

5. Critical Analysis of Literature

5.1 Overview of Existing Research

The literature on performance metrics in intrusion detection systems (IDS) spans various methodologies, metrics, and evaluation frameworks. This section critically analyzes the findings and contributions of existing studies to identify trends, gaps, and areas for further exploration.

5.2 Methodological Approaches

Studies have employed diverse methodological approaches to evaluate IDS performance metrics, including empirical testing, simulation-based experiments, and comparative analyses using benchmark datasets. Each approach offers unique insights into the strengths and limitations of different metrics under controlled and real-world conditions [12].

5.3 Key Findings and Trends

5.3.1 Effectiveness of Traditional Metrics

Precision, sensitivity/recall, and the F1 score also remain in such mass use for the evaluation of IDS effectiveness. There is one piece of consistent research showing the trade-offs between these and thus giving implications to the performance of the IDS in detecting different sorts of intrusion points [62, 23]. For example, accuracy universally quantifies the correct classification but may obscure the capability of the IDS to detect rare or emerging threats in imbalanced tabular data. Precision and recall suggest the ability of an IDS to cut back on the case of false positives and the ability then to detect intrusions, respectively, but improving one metric has an associated tradeoff against the other [63, 39].

5.3.2 Challenges in Real-World Deployments

The research underscores the challenges of translating laboratory-based performance characteristics into actual performance by an IDS in the wild. Variability within the network, the dynamics of the traffic, and evolving attack techniques are typical sources of topological changes in operational settings, which could significantly impact the IDS performance metrics [64, 7]. Research has documented differences in performance metrics calculated in controlled or lab-simulated environments from those measured in live network environments. High false alarm rates, resource constraints, and the need for timely threat response are some additional challenges posed by the timely response to threats that traditional metrics may not be able to fully capture [65, 66].

5.3.3 Emerging Trends in Metric Development

Recent literature has focused on developing adaptive, context-aware performance metrics with the necessary flexibility to adapt to the dynamism of the threat landscape and operational conditions, as demonstrated in [67]. An example is that research on adversarial robustness, sensitivity to abnormal detection, and responsiveness to incidents are some of the metrics that are gaining more research emphasis in IDS [41]. These trends indicate the creation of increasingly sophisticated frameworks for evaluating system performance, not relying on very basic measures but ensuring multiple ways of proving that an IDS is resilient against advanced cyber threats [68].

5.3.4 Addressing the metrics trade-offs in resource Constrained Environment

Balancing the metric trade-offs where IDS works in complex dynamic environments is very critical. While the accuracy is most commonly metric for evaluating performance, using accuracy in isolation will lead to inefficient resource utilisation, especially the data is imbalanced. Since accuracy is sensitive to true positives and false negatives, other metrics like precision, recall, F1 score can help organisations to get more insights about the performance of IDS. Financial institutes like banks and E-commerce companies like Amazon, eBay, Flipkart etc deal with thousands of transactions every day, these organisations' presence of false positives, and false negatives can give wrong indications about the activities of bank operations. In these cases, the organisation should prioritise precision over other performance metrics to flag genuine threats avoiding the wrong indicators.

5.4 Gaps and limitations in current research

While existing literature provides valuable insights into performance metrics in IDS, several gaps and limitations warrant further investigation:

- **Limited Diversity in Evaluation Datasets**: Many studies rely on standard benchmark datasets, which may not fully represent the diversity of real-world network environments and threat scenarios.
- **Over-reliance on Technical Metrics:** There is usually too much focus on technical metrics that measure aspects of technical performance—such as accuracy and precision—at the expense of qualitative attributes, such as usability, scalability, and adaptability to organizational contexts.
- **Comparative Studies:** The comparative analysis of different IDS measurements over a very broad set of working conditions (for instance, different types of network traffic and intensities of attack) is not sufficient, making it very difficult to generalize the results over different working operational environments.

5.5 Future Research Directions

To address these gaps and advance the field of IDS performance evaluation, future research directions include:

- **Enhanced Evaluation Frameworks**: Developing comprehensive evaluation frameworks that integrate technical metrics with qualitative assessments of IDS usability, scalability, and operational impact.
- **Context-Aware Metrics**: Designing adaptive performance metrics that can dynamically adjust to changing threat landscapes and operational priorities.
- **Real-World Validation**: Conducting extensive field trials and case studies to validate performance metrics in diverse organizational settings and under varying threat scenarios.
- **Ethical and Legal Considerations**: Addressing ethical and legal implications of IDS performance metrics, including privacy concerns, data protection, and algorithmic bias.

6. Conclusion

In many ways, this critical literature on which performance metrics for IDSs are based shows quite a rich landscape of methodologies, findings, and new trends. Although traditional metrics offer rather basic insight into the efficacy of an IDS, future research should aim to develop more adaptive, context-aware metrics and full evaluation frameworks that naturally align with the changing nature of cyber threats and operation realities [69, 70]. Second, a review of the published academic literature identifies the

challenges for IDS evaluation. In general, common themes are the complexity of real-world network environments, the need for standardised datasets, and the evolving nature of cyber threats. For instance, the works of the authors [71] and [72] propose a comprehensive metric framework but often cause a lack of implementation because of the complexity and the many requirements for data necessary to make the framework work. Generally, the evaluation of IDS is very challenging and does entail multipart processes that require a mixture of the diverse performance metrics. From the literature, the critical elements that will ensure IDS performance with robustness and reliability are standardized IDS evaluation frameworks, comprehensive metric integration, and applicability in real-world scenarios. At the same time, it requires continuous innovation and adaptation to answer the fast-changing cyberspace threat landscape. IDS has to deal with broader levels of organisational and operational issues, and balancing the trade-off in resource resource-constrained environment is very crucial. Since this study is a review of published academic literature, readers can study how reputed organisations can manage algorithmic bias, data imbalance, and ethical issues in the real world will provide more insight to improve the performance of IDS.

Future research in performance assessment for IDS will focus on the design of adaptive, context-aware metrics, which are capable of dynamic accommodation toward the dynamic nature of cyber threats and operational settings. The metrics seek to improve the resilience and strength of IDS under the issues of network variance, variation of traffic, and attack techniques of varying sophistication [41]. ML techniques applied to intrusion detection, deep learning, and reinforcement learning will boost accuracy and efficiency, respectively [73, 74]. It is these techniques that call for large-scale data analytics and automated decision processes that enable new technologies to be more effective detect threats. Machine learning and artificial intelligence techniques will significantly advance performance metrics. Other studies have also shown that ML-based technologies enhance the security performance criteria of IDS systems toward major radical advancements. The techniques are deep learning and reinforcement learning; they enhance both accuracy and performance efficiency during the threat detection process. These technologies are. The future IDS frameworks would include providing real-time and proactive threat-detecting capabilities [75, 76]. The detection latency, response time to incidents, and integration of adaptive threat intelligence are some of the important key metrics. These shall aid in minimizing the threats of cyber events prior to their escalation, thus effectively elevating the overall security of the system [77, 65, 42].

It is also necessary for research done on IDS performance metrics to embrace the ethical and legal questions at stake [78]. Among such critical issues, the leading questions are privacy, data protection, and algorithmic bias. The future should involve the development of transparent and accountable frameworks for evaluation based on ethics and regulations [23]. This will only be fully realized if future studies bring up an interdisciplinary collaboration for the betterment of IDS performance evaluation as a field of study. Collaboration among cybersecurity professionals, data scientists, legal professionals, and policymakers can help in innovation, addressing complex problems, and ensuring that the metrics of IDS are practically relevant under various organizational contexts [79, 7].

References

[1] A. Heidari, and M. A. J. Jamali, "Internet of Things intrusion detection systems: a comprehensive review and future directions," *Cluster Comput.,* vol. 26, no. 6, pp. 3753–3780, Dec. 2023, doi: 10.1007/s10586-022-03776-z.

[2] A. Patel, Q. Qassim, and C. Wills, "A survey of intrusion detection and prevention systems," *Inf. Comput. Secur.*, vol. 18, no. 4, pp. 277–290, Dec. 2010, doi: https://doi.org/10.1108/09685221011079199

[3] N. Sfetcu, *Advanced Persistent Threats in Cybersecurity – Cyber Warfare*. MultiMedia Publishing, 2024.

[4] D. P. Möller, "Intrusion detection and prevention," in *Guide to Cybersecurity in Digital Transformation: Trends, Methods, Technologies, Applications and Best Practices*, Cham: Springer Nature Switzerland, pp. 131–179, 2023.

[5] I. Corona, G. Giacinto, and F. Roli, "Adversarial attacks against intrusion detection systems: Taxonomy, solutions and open issues," *J. Inf. Sci.*, vol. 239, pp. 201–225, Aug. 2013, doi: https://doi.org/10.1016/j.ins.2013.03.022.

[6] S. Morgan, "Cybercrime to cost the world $10.5 trillion annually by 2025," *Cybersecurity Ventures*, 2020.

[7] K. Scarfone and P. Mell, "Guide to Intrusion Detection and Prevention Systems (IDPS)," *NIST Special Publication*, vol. 800, no. 94, pp. 1–127, Feb. 2007.

[8] Ö. Aslan, S. S. Aktuğ, M. Ozkan-Okay, A. A. Yilmaz, and E. Akin, "A comprehensive review of cyber security vulnerabilities, threats, attacks, and solutions," *J. Electron.,* vol. 12, no. 6, Mar. 2023, Art. no. 1333, doi: https://doi.org/10.3390/electronics12061333

[9] R. Bace and P. Mell, "Intrusion Detection Systems." Special Publication (NIST SP), National Institute of Standards and Technology, Gaithersburg, MD, Nov. 01, 2001.

[10] HHS, "Health Insurance Portability and Accountability Act (HIPAA)," *U.S. Department of Health & Human Services*, 2013.

[11] A. Aldweesh, A. Derhab, and A. Z. Emam, "Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues," *Knowl. Based Sys.,* vol. 189, Feb. 2020, Art. no. 105124, doi: 10.1016/j.knosys.2019.105124.

[12] P. Garcia-Teodoro, J. Diaz-Verdejo, G. Macia-Fernandez, and E. Vazquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," *Comput. Secur*, vol. 28, no. 1–2, pp. 18–28, Mar. 2009, doi: https://doi.org/10.1016/j.cose.2008.08.003.

[13] A. Thakkar and R. Lohiya, "A review on machine learning and deep learning perspectives of IDS for IoT: Recent updates, security issues, and challenges," *Arch. Comput. Methods. Eng.,* vol. 28, no. 4, pp. 3211–3243, Oct. 2021, doi: https://doi.org/10.1007/s11831-020-09496-0.

[14] H. Debar, M. Dacier, and A. Wespi, "Towards a taxonomy of intrusion-detection systems," *J. Comput. Netw.*, vol. 31, no. 8, pp. 805–822, Apr. 1999, doi: https://doi.org/10.1016/S1389-1286(98)00017-6

[15] K. Al-Dosari and N. Fetais, "Risk-management framework and information-security systems for small and medium enterprises (SMES): A meta-analysis approach," *Electronics*, vol. 12, no. 17, Aug. 2023, Art. no. 3629 doi: https://doi.org/10.3390/electronics12173629.

[16] M. E. Falagas, G. A. Pitsouni, G. A. Malietzis, and G. Pappas, "Comparison of PubMed, Scopus, web of science, and Google scholar: Strengths and weaknesses," *FASEB Journal*, vol. 22, no. 2, pp. 338–342, Sep. 2008, doi: 10.1096/fj.07-9492LSF.

[17] P. Mongeon and A. Paul-Hus, "The journal coverage of Web of Science and Scopus: A comparative analysis," *Scientometrics*, vol. 106, pp. 213–228, Oct. 2016, doi: https://doi.org/10.1007/s11192-015-1765-5.

[18] R. Tomaszewski, "A study of citations to STEM databases: ACM Digital Library, Engineering Village, IEEE Xplore, and MathSciNet," *Scientometrics*, vol. 126, no. 2, pp. 1797–1811, Jan. 2021, doi: https://doi.org/10.1007/s11192-020-03795-w.

[19] O. H. Abdulganiyu, T. Ait Tchakoucht, and Y. K. Saheed, "A systematic literature review for network intrusion detection system (IDS)," *Int. J. Inf. Secur.,* vol. 22, no. 5, pp. 1125–1162, Mar. 2023, doi: https://doi.org/10.1007/s10207-023-00682-2.

[20] A. Liberati et al., "The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration," *Ann. Intern. Med.*, vol. 151, no. 4, Jun. 2009, Art. no. 65, doi: 10.1136/bmj.b2700.

[21] V. Mishra and M. P. Mishra, "PRISMA for Review of Management Literature – Method, Merits, and Limitations – An Academic Review," in *Review of Management Literature*, S. Rana, J. Singh, and S. Kathuria, Eds., Emerald Publishing Limited, 2023, pp. 125–136. doi: 10.1108/S2754-586520230000002007.

[22] M. J. Page et al., "Updating guidance for reporting systematic reviews: Development of the PRISMA 2020 statement," *J. Clin. Epidemiol.,* vol. 134, pp. 103–112, Feb. 2021. doi: 10.1016/j.jclinepi.2021.02.003.

[23] A. Patcha and J. M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," *Comput. Netw.*, vol. 51, no. 12, pp. 3448–3470, Aug. 2007, doi: https://doi.org/10.1016/j.comnet.2007.02.001

[24] A. Divekar, M. Parekh, V. Savla, R. Mishra, and M. Shirole, "Benchmarking datasets for Anomaly-based Network Intrusion Detection: KDD CUP 99 alternatives," in *2018 IEEE 3rd Int. Conf. Comp., Communic. Sec. (ICCCS)*, Kathmandu: IEEE, Oct. 2018, pp. 1–8. doi: 10.1109/CCCS.2018.8586840.

[25] R. A. Hubbard et al., "Cumulative probability of false-positive recall or biopsy recommendation after 10 years of screening mammography: A cohort study," *Ann. Intern. Med.*, vol. 155, no. 8, pp. 481–492, May. 2011, doi: 10.7326/0003-4819-155-8-201110180-00004.

[26] A. Alhomoud et al., "Performance evaluation study of intrusion detection systems," *Procedia Comput. Sci.,* vol. 5, pp. 173–180, Apr. 2011, doi: https://doi.org/10.1016/j.procs.2011.07.024

[27] W. P and L. H, "A survey of Intrusion Detection System," *Int. J. Infor. Comput.*, vol. 1, no. 1, pp. 1–10, Jan. 2020, doi:10.35842/ijicom.v1i1.7

[28] Z. K. Maseer et al., "Benchmarking of machine learning for anomaly-based intrusion detection systems in the CICIDS2017 dataset", *IEEE Access*, vol. 9, pp. 22351–22370, Feb. 2021, doi: 10.1109/ACCESS.2021.3056614

[29] M. Alrowaily, F. Alenezi, and Z. Lu, "Effectiveness of machine learning-based intrusion detection systems," in *SpaCCS 2019*, Atlanta, GA, USA, Jul. 2019, pp. 277–288, doi: https://doi.org/10.1007/978-3-030-24907-6_21

[30] T. Mages, M. Almgren, and C. Rohner, "Towards an information-theoretic framework of intrusion detection for composed systems and robustness analyses," *Comput. Secur.*, vol. 116, p. 102633, May. 2022, doi: https://doi.org/10.1016/j.cose.2022.102633

[31] V. Jyothsna and K. Munivara Prasad, "Anomaly-Based Intrusion Detection System," in *Computer and Network Security*, J. Sen, Ed., IntechOpen, 2020. doi: 10.5772/intechopen.82287.

[32] R. Magán-Carrión, D. Urda, I. Díaz-Cano, and B. Dorronsoro, "Towards a reliable comparison and evaluation of network intrusion detection systems based on machine learning approaches," *Appl. Sci.*, vol. 10, no. 5, p. 1775, Mar. 2020, doi: https://doi.org/10.3390/app10051775

[33] A. H. Azizan et al., "A machine learning approach for improving the performance of network intrusion detection systems," *Ann. Emerg. Technol. Comput*., vol. 5, no. 5, pp. 201–208, Mar. 2021, doi: 10.33166/AETiC.2021.05.025

[34] H. Chauhan, V. Kumar, S. Pundir, and E. S. Pilli, "A comparative study of classification techniques for intrusion detection," in *2013 ISCBI*, IEEE, Aug. 2013, pp. 40–43, doi: 10.1109/ISCBI.2013.16.

[35] G. Kumar, "Evaluation metrics for intrusion detection systems – A study," *Evaluation*, vol. 2, no. 11, pp. 11–17, Oct. 2014.

[36] M. Almseidin, J. Al-Sawwa, and M. Alkasassbeh, "Generating a benchmark cyber multi-step attacks dataset for intrusion detection," *Int. J. Fuzzy Log.*, vol. 43, no. 3, pp. 3679–3694, 2022, doi: 10.3233/JIFS-213247

[37] G. Loukas et al., "A taxonomy and survey of cyber-physical intrusion detection approaches for vehicles," *Ad Hoc Netw.*, vol. 84, pp. 124–147, Mar. 2019, doi: https://doi.org/10.1016/j.adhoc.2018.10.002

[38] A. Nassar and M. Kamal, "Machine learning and big data analytics for cybersecurity threat detection: A holistic review of techniques and case studies," *J. Artif. Intellig. Mach. Learn. Manag.*, vol. 5, no. 1, pp. 51–63, Feb. 2021.

[39] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell.,* vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[40] A. McCarthy et al., "Functionality-preserving adversarial machine learning for robust classification in cybersecurity and intrusion detection domains: A survey," *J. Cybersecur. Priv.,* vol. 2, no. 1, pp. 154–190, Mar. 2022, doi: https://doi.org/10.3390/jcp2010010

[41] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," arXiv preprint, arXiv:1412.6572, pp. 1-11, Mar. 2014, doi: https://doi.org/10.48550/arXiv.1412.6572

[42] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cybersecurity intrusion detection," *IEEE Commun. Surv. Tutor*, vol. 18, no. 2, pp. 1153–1176, Oct. 2016, doi: 10.1109/COMST.2015.2494502

[43] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: Techniques, datasets, and challenges," *Cybersecurity*, vol. 2, no. 1, pp. 1–22, Jul. 2019, doi: https://doi.org/10.1186/s42400-019-0038-7

[44] P. Bedi, N. Gupta, and V. Jindal, "I-SiamIDS: An improved Siam-IDS for handling class imbalance in network-based intrusion detection systems," *Appl. Intell.*, vol. 51, no. 2, pp. 1133–1151, Oct. 2021, doi: https://doi.org/10.1016/j.procs.2020.04.085

[45] V. Bulavas, V. Marcinkevičius, and J. Rumiński, "Study of multi-class classification algorithms' performance on highly imbalanced network intrusion datasets," *Informatica*, vol. 32, no. 3, pp. 441–475, Jul. 2021, doi: 10.15388/21-INFOR457

[46] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*. Cham: Springer International Publishing, 2018. doi: 10.1007/978-3-319-98074-4.

[47] J. Attenberg and Ş. Ertekin, "Class Imbalance and Active Learning," in *Imbalanced Learning*, 1st ed., H. He and Y. Ma, Eds., Wiley, 2013, pp. 101–149. doi: 10.1002/9781118646106.ch6.

[48] R. Panigrahi et al., "A consolidated decision tree-based intrusion detection system for binary and multiclass imbalanced datasets," *Mathematics*, vol. 9, no. 7, Mar. 2021, Art. no. 751, doi: https://doi.org/10.3390/math9070751

[49] H. Ding et al., "Imbalanced data classification: A KNN and generative adversarial networks-based hybrid approach for intrusion detection," *Future Gener. Comput. Syst.,* vol. 131, pp. 240–254, Jun. 2022, doi: https://doi.org/10.1016/j.future.2022.01.026

[50] M. Ozkan-Okay et al., "A comprehensive systematic literature review on intrusion detection systems," *IEEE Access*, vol. 9, pp. 157727–157760, Nov. 2021, doi: 10.1109/ACCESS.2021.3129336

[51] I. Martins, J. S. Resende, P. R. Sousa, S. Silva, L. Antunes, and J. Gama, "Host-based IDS: A review and open issues of an anomaly detection system in IoT," *Fut. Gener. Comput. Syst.,* vol. 133, pp. 95-113, Aug. 2022, doi: https://doi.org/10.1016/j.future.2022.03.001

[52] Youping Zhao, Shiwen Mao, J. O. Neel, and J. H. Reed, "Performance Evaluation of Cognitive Radios: Metrics, Utility Functions, and Methodology," *Proc. IEEE*, vol. 97, no. 4, pp. 642–659, Apr. 2009, doi: 10.1109/JPROC.2009.2013017.

[53] C. Liu, P. Frazier, and L. Kumar, "Comparative assessment of the measures of thematic classification accuracy," *Remote Sens. Environ.,* vol. 107, no. 4, pp. 606-616, Apr. 2007, doi: https://doi.org/10.1016/j.rse.2006.10.010

[54] A. A. Cardenas, J. S. Baras, and K. Seamon, "A framework for the evaluation of intrusion detection systems," in *2006 IEEE Symp. Sec. Priv. (S&P'06)*, Berkeley/Oakland, CA: IEEE, 2006, pp. 15 – 77. doi: 10.1109/SP.2006.2.

[55] D. Amo-Filva, D. Fonseca, F. J. García-Peñalvo, M. A. Forment, M. J. Casany Guerrero, and G. Godoy, "Exploring the landscape of learning analytics privacy in fog and edge computing: A systematic literature review," *Comput. Human Behav.*, vol. 158, Sep. 2024, Art. no. 108303, doi: 10.1016/j.chb.2024.108303.

[56] T. du Preez, *DECIDE-The Art and Science of Choosing Wisely*. SG: Marshall Cavendish International (Asia) Private Limited, 2020.

[57] D. Dauda Wisdom, O. Rebecca Vincent, A.-A. Adebayo, F. Olusegun, I. Oniovosah Ayetuoma, and G. Alpha Baba, "Security Measures in Computational Modeling and Simulations," in *Computational Modeling and Simulation of Advanced Wireless Communication Systems*, 1st ed., Boca Raton: CRC Press, 2024, pp. 112–150. doi: 10.1201/9781003457428-6.

[58] H. Liu and B. Lang, "Machine learning and deep learning methods for intrusion detection systems: A survey," *Appl. Sci.*, vol. 9, no. 20, p. 4396, Oct. 2019, doi: https://doi.org/10.3390/app9204396

[59] V. Alagar, M. Mohammad, K. Wan, and S. A. Hnaide, "A framework for developing context-aware systems", CASA, vol. 1, no. 1, pp. 1-26, Mar. 2014, doi: https://doi.org/10.4108/casa.1.1.e2

[60] S. Neupane, J. Ables, W. Anderson, S. Mittal, S. Rahimi, I. Banicescu, and M. Seale, "Explainable intrusion detection systems (X-IDS): A survey of current methods, challenges, and opportunities," *IEEE Access*, vol. 10, pp. 112392-112415, Jul. 2022, doi: https://doi.org/10.48550/arXiv.2207.06236

[61]I. Brown and C. T. Marsden, *Regulating Code: Good Governance and Better Regulation in the Information Age*. The MIT Press, 2013. doi: 10.7551/mitpress/8879.001.0001.

[62] N. Munaiah, A. Meneely, R. Wilson, and B. Short, "Are intrusion detection studies evaluated consistently? A systematic literature review," *RIT Scholar Works*, Sep. 2016.

[63] P. K. Keserwani, M. C. Govil, and E. S. Pilli, "An effective NIDS framework based on a comprehensive survey of feature optimization and classification techniques," *Neural Comput. Appl.*, vol. 35, no. 7, pp. 4993-5013, May. 2023, doi: https://doi.org/10.1007/s00521-021-06093-5

[64] N. Chaabouni, M. Mosbah, A. Zemmari, C. Sauvignac, and P. Faruki, "Network intrusion detection for IoT security based on learning techniques," *IEEE Commun. Surv. Tutor.*, vol. 21, no. 3, pp. 2671-2701, Jan. 2019, doi: 10.1109/COMST.2019.2896380

[65] L. Gudala, M. Shaik, S. Venkataramanan, and A. K. R. Sadhu, "Leveraging artificial intelligence for enhanced threat detection, response, and anomaly identification in resource-constrained IoT networks," *Distrib. Learn. Broad Applic. Sci. Res.*, vol. 5, pp. 23-54, Jul. 2019.

[66] A. K. Jones and R. S. Sielken, "Computer system intrusion detection: A survey," *Computer Science Department, University of Virginia*, Tech. Rep. 22, 2000.

[67] B. R. Siqueira, F. C. Ferrari, K. E. Souza, V. V. Camargo, and R. de Lemos, "Testing of adaptive and context-aware systems: Approaches and challenges," *Softw. Test. Verif. Reliab,* vol. 31, no. 7, May 2021, Art. no. e1772, doi: https://doi.org/10.1002/stvr.1772

[68] W. Lim, K. S. C. Yong, B. T. Lau, and C. C. L. Tan, "Future of generative adversarial networks (GAN) for anomaly detection in network security: A review," *Comp. Secur.*, vol. 139, Apr. 2024, Art. no. 103733, doi: 10.1016/j.cose.2024.103733.

[69]A. Khanan, Y. Abdelgadir Mohamed, A. H. H. M. Mohamed, and M. Bashir, "From Bytes to Insights: A Systematic Literature Review on Unraveling IDS Datasets for Enhanced Cybersecurity Understanding," *IEEE Access*, vol. 12, pp. 59289–59317, 2024, doi: 10.1109/ACCESS.2024.3392338.

[70] I. H. Sarker, *AI-Driven Cybersecurity and Threat Intelligence: Cyber Automation, Intelligent Decision-Making and Explainability*. Cham: Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-54497-2.

[71] W. Jiang, K. Huang, J. Geng, and X. Deng, "Multi-scale metric learning for few-shot learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 3, pp. 1091-1102, May 2020, doi: 10.1109/TCSVT.2020.2995754.

[72] K. Siddique, Z. Akhtar, M. A. Khan, Y. H. Jung, and Y. Kim, "Developing an intrusion detection framework for high-speed big data networks: A comprehensive approach," *KSII Trans. Internet Inf. Sys.*, vol. 12, no. 8, pp. 4021-4037, Oct. 2018.

[73] S. Muneer, U. Farooq, A. Athar, M. Ahsan Raza, T. M. Ghazal, and S. Sakib, "A critical review of artificial intelligence-based approaches in intrusion detection: A comprehensive analysis," *J. Eng.*, vol. 2024, no. 1, Aug. 2024, Art. no. 3909173, doi: 10.3837/tiis.2018.08.026

[74] Z. Ahmad, A. Shahid Khan, C. Wai Shiang, J. Abdullah, and F. Ahmad, "Network intrusion detection system: A systematic study of machine learning and deep learning approaches," *Trans. Emerg. Telecommun. Technol.*, vol. 32, no. 1, Jan. 2021, Art. no. e4150, doi: 10.1002/ett.4150

[75] M. A. Rahman, A. T. Asyhari, O. W. Wen, H. Ajra, Y. Ahmed, and F. Anwar, "Effective combining of feature selection techniques for machine learning-enabled IoT intrusion detection," *Multimedia Tools Appl.*, vol. 80, no. 20, pp. 31381-31399, Sep. 2020, doi: https://doi.org/10.1007/s11042-021-10567-y

[76] M. Lopez-Martin, B. Carro, and A. Sanchez-Esguevillas, "Application of deep reinforcement learning to intrusion detection for supervised problems," *Expert Syst. Appl.*, vol. 141, Mar. 2020, Art. no. 112963, doi: https://doi.org/10.1016/j.eswa.2019.112963

[77] M. Al-Hawawreh, N. Moustafa, and J. Slay, "A threat intelligence framework for protecting smart satellite-based healthcare networks," *Neural Comput. Appl.*, vol. 36, no. 1, pp. 15-35, Apr. 2024. Doi: https://doi.org/10.1007/s00521-021-06441-5

[78] B. R. Maddireddy, and B. R. Maddireddy, "Real-time data analytics with AI: Improving security event monitoring and management," *Unique Endeavor Busin. Soc. Sci.*, vol. 1, no. 2, pp. 47-62, Mar. 2022.

[79] Computer Science and Telecommunications Board, Division on Engineering and Physical Sciences, and National Academies of Sciences, Engineering, and Medicine, *Foundational Cybersecurity Research: Improving Science, Engineering, and Institutions*. Washington, D.C.: National Academies Press, 2017, Art. no. 24676. doi: 10.17226/24676.