JOURNAL OF COMMUNICATION AND INFORMATION SYSTEMS, VOL. 39, NO. 1, 2024.

Overcoming the Challenges of Data Lack, Leakage, and Dimensionality in Intrusion Detection Systems: A Comprehensive Review

Mohamed Aly Bouke, Azizol Abdullah, Nur Izura Udzir and Normalia Samian

***Abstract*—The Internet of Things (IoT) and cloud computing are rapidly gaining momentum as decentralized Internet-based technologies, leading to increased information in nearly every technical and commercial industry. However, ensuring the security of IoT systems is a pressing issue due to the complexities involved in connected and shared environments. Networks are guarded by Intrusion Detection Systems (IDS) against various cyber threats such as malware, viruses, and unauthorized access. IDS has recently adopted Machine Learning (ML) and Deep Learning (DL) techniques to identify and classify security risks. However, the effective utilization of these technologies depends on the availability, quality, and characteristics of the data used to train models. Moreover, data lack, data leak, and dimensionality (DLLD) are common problems in data science and ML. This paper surveys existing research and suggests solutions for overcoming DLLD-related issues to improve the IDS model.**

Index Terms—**Intrusion Detection Systems (IDS), Data Lack, Data Dimensionality, Data Leakage, Cybersecurity.**

INTRODUCTION

The rapid growth of decentralized and internet-based technologies, such as cloud computing and the IoT, has led to an information explosion [1]. However, the security of IoT systems has become one of the most complex issues in shared and connected environments, as cyber threats like hackers, viruses, and malicious software can compromise the security and stability of data [2]. Furthermore, data breaches can directly undermine the safety of the IoT, resulting in various harmful behaviors [3]. Therefore, IoT protection has become a hot topic [3].

Moreover, IDS is a software or hardware system that monitors network traffic and analyzes it for malicious activity or policy violations. If such activity is detected, the IDS can take various actions, such as logging the event, reporting it to an administrator, or blocking it [4]. In addition, IDS systems can be configured to monitor specific protocols, ports, or types of traffic. IDS systems are used to protect networks from a variety of cyber threats, such as malware, viruses, and unauthorized access.

Recently, ML and DL algorithms have increasingly been applied in Intrusion Detection Systems (IDS) to identify and classify security threats [5]–[8]. Their application enhances the accuracy and effectiveness of IDS by training systems on datasets that include both standard and abnormal network traffic, enabling these systems to identify patterns indicative of malicious activities [9]. ML can be utilized in IDS through various approaches [10]:

- **Anomaly Detection**: An IDS using ML algorithms learns to recognize regular network traffic patterns. It then identifies deviations from these established norms, which could indicate potential intrusions [11]. This method is particularly effective for detecting novel or previously unseen threats.
- **Signature-based Detection**: This approach involves ML identifying established patterns or 'signatures' of known malicious activities within network data [9]. Unlike anomaly detection, which focuses on deviations from normalcy, signature-based detection relies on a pre-defined database of known threat signatures. It's highly effective for detecting known threats but may not identify new, unrecorded types of attacks.
- **Behavior-based Detection**: ML analyzes and understands typical user or system behavior patterns. The IDS then detects activities that deviate from these patterns, such as repeated failed login attempts or unusual network scanning activities [12]. This method is useful for identifying threats that may not have a known signature but exhibit suspicious behavior patterns.

Mohamed Aly Bouke, Azizol Abdullah, Nur Izura Udzir and Normalia Samian, Universiti Putra Malaysia (Faculty of Computer Science and Information Technology), Serdang 43400, Malaysia, e-mail: bouke@ieee.org, azizol@upm.edu.my,izura@upm.edu.my,normalia@upm.edu.my,ORCID:0000-0003-3264-601X, 0000-0001-8321-9259,0000-0002-0543-3329,0000-0002-0905-8140.

Funding Disclosure: This work is not funded by any external or internal organization. DOI: 10.14209/jcis.2024.3

Each ML approach offers distinct advantages in detecting and mitigating security threats, making them integral to modern IDS solutions.

Moreover, using ML in an IDS can significantly improve its accuracy by allowing it to adapt to changing network traffic patterns and new types of intrusions. It can also reduce the number of false positives, as the IDS can learn to distinguish between normal and malicious activity more effectively [13].

However, the effective use of these technologies depends on the availability, quality, and characteristics of the data used to train models. **Data Lack, Data Leak, and Data Dimensionality (DLLD)** issues are all common problems that can arise in the field of data science and ML [14]–[16].

Data Lack, also known as "data scarcity," refers to insufficient data available to train a model or perform a specific analysis. This can occur for various reasons, such as difficulty in collecting data, high costs associated with data acquisition, or limitations on the data types that can be managed. Data lack can be a significant barrier to developing effective ML models, as a model's performance is directly related to the quality and quantity of the data it is trained on.

Data Leak, also known as "data leakage," occurs when a model is trained on data that includes information that would not be available when the model is used in production [17]. This can lead to models that perform well on the training data but poorly in practice [14]. Data leaks can occur for various reasons, such as poor data cleaning, lack of feature engineering, or inadequate data partitioning[15].

Data Dimensionality refers to the number of features or variables in a dataset. High dimensionality can make it challenging to model and analyze data, as the number of potential interactions between features increases exponentially with dimensionality [18]. High dimensionality can also lead to the "curse of dimensionality," a phenomenon in which models trained on high-dimensional data perform poorly due to a lack of data. In addition, high-dimensional data can make it more challenging to identify patterns and relationships in data, which can also lead to overfitting [16].

As a result, DLLD can significantly impact the performance of IDS systems. The research community has been working on different techniques to address these issues [19]–[24]. In this survey paper, we will be discussing the various methods and approaches used to overcome these problems in the field of IDSs. We will also highlight recent advances and ongoing challenges in this area of research. This paper will provide a comprehensive overview of the current state-of-the-art and help researchers and practitioners understand the latest trends in this field and design better IDS.

LITERATURE REVIEW

DLLD issues are significant challenges faced by researchers and practitioners in IDS. These issues can significantly impact the performance of IDS models, making it difficult to effectively identify and prevent unauthorized access, misuse, alteration, or destruction of information systems [25]–[32].

The literature in this field has been focusing on different approaches to address DLLD issues to improve the performance of IDS systems. Various techniques have been explored and applied in addressing the challenges of DLLD in IDS. Table 1 summarizes these techniques and categorizes them based on their specific challenge. Additionally, we have provided references for each technique where these methods have been discussed or applied in the context of IDS. This table serves as a comprehensive guide for researchers and practitioners in the field, offering a quick reference to the relevant literature on each technique.

Table 1 Techniques used to address DLLD.

Techniques	Data Leakage	Data Lack	Data Dimensionality	References
Data masking	X			[33]
Data encryption	X			[34]
Data tokenization	X			[35]
Data anonymization	X			[36]
Data sub-setting	X			[37]
Data augmentation		X		[38]
Data generation		X		[39]
Data synthesis		X		[40]
Data imputation		X		[41]
Transfer learning		X		[42]
Feature selection			X	[43]
Feature extraction			X	[44]
Reduction			X	[45]
Manifold learning			X	[46]
PCA			X	[47]

Data Dimensionality

IDS are designed to detect malicious activity or violations of security policies in computer networks. One of the significant challenges in designing and implementing IDSs is dealing with the high dimensionality of the data [22], [48]–[50].

Feature selection selects a subset of relevant features for model construction [51]. The goal is to choose a set of features that improves the accuracy and interpretability of the model while reducing dimensionality and minimizing overfitting [52]. Furthermore, the three primary feature selection methods are filter, wrapper, and embedding approaches. [53].

A study by Subba et al. [54] proposed a Principal Component Analysis (PCA) technique to lower the time complexity of anomaly-based IDS. The PCA method reduces large data sets

by organizing the input features based on their correlation while maintaining a suitable level of data accuracy. The study tested the Support Vector Machine (SVM), Multilayer Perceptron (MLP), C4.5, and Naive Bayes algorithms using multi- and binary classification. The results showed high accuracy, but it was noted that the testing was only conducted on the training data, which may not yield the same results when applied to test data.

Can et al. [55] introduced a neural network-based DDoS attacks classification model. The model takes advantage of a proposed automatic feature selection scheme, and the model's effectiveness was assessed using the CICDDoS2019 dataset and proved the efficiency of the proposed model. However, the model is significantly affected by the disparity between the test and training data distribution.

Di Mauro et al. [56] reviewed ML techniques for network intrusion detection, explicitly focusing on statistical features such as inter-arrival times and packet length distribution for classifying and recognizing data traffic. However, the authors note that dealing with the large number and diversity of features that typically characterize data traffic can be challenging and lead to issues such as lengthy training processes and introducing bias during classification. The authors propose that feature selection is a crucial preprocessing step in network management to address these issues, specifically for network intrusion detection.

In a study by Hassan et al. [57], an improved Binary Manta-Ray Foraging (BMRF) Optimization Algorithm is proposed. This algorithm is based on an adaptive S-shape function and a Random Forest (RF) classifier. The essential features are found in the intrusion detection datasets, and the redundant and unnecessary ones are removed using the BMRF algorithm. Conversely, the RF classifier is used to build the intrusion detection model and evaluate the feature set. The CIC-IDS2017 and NSL-KDD datasets are used as benchmarks to test the proposed method and compare it to other approaches. The findings show that the suggested model performs admirably in terms of precision, recall, F-measure, and accuracy. The proposed model and the compared methods differ significantly in terms of F-measure, according to a statistical significance test that was performed.

D'hooge et al. [58] suggest a hybrid feature selection mechanism for intrusion detection to increase the effectiveness and efficiency of using datasets. The proposed strategy is based on a first-pass filter method and a second-pass embedding method, with statistical testing playing a pivotal role in identifying hierarchies of dominating feature sets. The method is verified by creating feature hierarchies for current datasets supplied by the Canadian Institute for Cybersecurity. The findings demonstrate that attack classes with a distinct network component may be identified with reasonable accuracy, recall, and precision even when the classification model is constructed from a limited set of features.

Adnan et al. [1] introduce a novel approach for intrusion detection in IoT-based wireless sensor networks. The proposed Intelligent IDS system combines a rule-based feature selection algorithm with a multi-objective optimization (PSO) technique and a multiclass SVM classification algorithm with an enhanced rule-based approach. The system's performance was evaluated using the KDD'99 Cup and CIDD datasets. The results showed that the proposed IDS system reduced the false positive rate and improved detection accuracy.

Mushtaq et al. [59] offer a stacked ensemble-based intrusion detection system (SE-IDS) that uses optimum feature selection to increase detection accuracy and decrease false positives. As foundation learners, the system contains a Decision Tree (DT), XGBoost, bagging classifier, additional tree, RF, and an MLP as a meta-learner. The system was evaluated on the NSL-KDD dataset, and the findings revealed that the suggested method outperformed previous strategies regarding accuracy, detection rate, and false alarm rate.

Halim et al. [60] propose a new feature selection method called Genetic Algorithm-based Feature Selection (GbFS) that utilizes a genetic algorithm to improve the accuracy of classifiers. In addition, the proposed method includes a novel fitness function and parameters tuned for the genetic algorithm. The GbFS method was evaluated on three standard network security datasets: UNSW-NB15, Bot-IoT, and CIRA-CIC-DOHBrw-2020. The results were compared to traditional feature selection methods and showed that GbFS achieved a maximum accuracy of 99.80%.

Aksu and Aydin [61] propose a new intrusion detection framework for Controller Area Network (CAN) bus systems. The proposed framework involves feature selection and classifier techniques to improve IDS performance. The feature selection method uses a modified genetic algorithm (MGA) to reduce the dimensionality of the data and select the optimal feature subset. In contrast, the classifier uses five different linear and nonlinear methods to identify intrusions, including SVM, logistic regression, DT, K-Nearest Neighbor (KNN), and linear discriminant analysis. The proposed method is tested on three datasets (HCRL-car hacking, UNSW-NB15, and CIC-IDS2017), and results show that the MGA-DTC combination presents the best performance in several metrics.

Kiplinger et al. [62] propose a new dataset for IDS called Switch port anomaly-based IDS (SPA-IDS). They introduce an automated classification model that utilizes signals collected from the dataset, which includes the generation of features using vertical mode decomposition and statistics, as well as iterative feature selection and classification phases. The model employs ML methods such as decision tree, bagged tree, SVM, and KNN and is evaluated using ten-fold cross-validation.

Results indicate that the proposed method achieves high accuracy rates, with the KNN classifier achieving 96.65%, the SVM performing 98.52%, the DT reaching 98.39%, and the bagged tree achieving 99.11%. In addition, the study demonstrates the effectiveness and efficiency of the proposed vertical mode decomposition and iterative feature selection-based IDS.

Chopra et al. provide an algorithm for non-deterministic feature selection for IDS that blends swarm intelligence and ensemble approaches. Three datasets—NSL-KDD, UNSW-NB15, and IoT-Zeek—produced from Zeek network logs and malicious and benign threat intelligence are used to test the approach. The findings demonstrate that the suggested method surpasses current ML models regarding the f1 score, scoring 92.092% on NSL-KDD, 92.904% on UNSW-NB15, and 97.302% on IoT-Zeek.

Panigrahi et al. [63] present a multiclass IDS for a cyber-physical environment using the CICIDS 2017 dataset. The proposed IDS uses a multi-objective evolutionary feature selection (MOEFS) algorithm to select the most informative features from the dataset and a hybrid classification mechanism combining the efficiency of decision tree naive Bayes (DTNB) for detecting threats in network traffic. The system achieved 96.8% accuracy using only five features and successfully detected benign instances and 11 out of 13 attack classes. However, it struggled to detect Heartbleed and Web-SQL Injection attack instances due to low participation in the training module. The authors also suggest that the system could be improved with a feedback approach, class relabeling, and exploring other feature selection schemes.

Artur [64] examined the data using the Naive Bayes classifier and the Recursive Feature Elimination with Cross-Validation (RFECV) feature selection approach. The research indicated that the best number of features for binary classification using the Naive Bayes approach is 32 after stratified cross-validation with ten folds and five repeats on a dataset. Furthermore, according to the study's findings, the F-measure and Receiver operating characteristic (ROC) curve scores suggest that binary classification using the Bernoulli Naive Bayes method works effectively.

Herrera-Semenets et al. [65] feature selection algorithm, called Multi Measure Feature Selection Algorithm (MMFSA), combines three measures to estimate different qualitative information in the features. The algorithm was evaluated and compared with other feature selection algorithms on a dataset, and it was found that MMFSA outperforms the other algorithms regarding classifier efficacy. In addition, the algorithm does not require manually pre-defining several features, which is a limitation in most other algorithms. However, the paper also points out that the optimal parameter value for the algorithm could not be determined, which could lead to overfitting and a resulting classification model with less predictive power. The authors suggest that this limitation could be studied in depth in future work, and they also plan to evaluate MMFSA in other application domains.

Kasongo and Sun [66] used the XGBoost algorithm and several ML methods to propose a new feature selection. The performance of the suggested techniques was assessed using the UNSW-NB15 dataset.

The results show that using a reduced (optimal) feature vector generated by the XGBoost algorithm can reduce model complexity and increase detection accuracy on test data. However, the study also highlights that the XGBoost-ANN method performs poorly for minority classes in the UNSW-NB15 dataset. Therefore, to improve the occurrence of minority classes during the training phase in future work, the authors advise employing a synthetic oversampling technique.

Disha and Waheed [67] proposed a Gini Impurity-based Weighted RF (RF) feature selection technique for ML-based IDS using two imbalanced datasets: UNSW-NB 15 and Network TON_IoT. The feature selection reduced the number of features in both datasets. In addition, the accuracy, false positive rate, precision, recall, and F1 score of the following machine learning models: DT, AdaBoost, Gradient Boosting Tree (GBT), MLP, Long-Short Term Memory (LSTM), and Gated Recurrent Unit (GRU) were examined. The results showed that the feature selection strategy outperformed DT for both datasets. However, the study did not perform multiclass classification and time complexity analysis, which could be considered future work.

Kshirsagar and Kumar [68] proposed an ensemble feature selection method that combines four different feature selection techniques, ReliefF, Gain Ratio, and ReliefF-Weight Information gain, to obtain a subset of 24 reduced features from a more extensive set of features. The proposed framework was implemented and validated on the CICIDS 2017 dataset using a J48 classifier and produced an improved 99.9909% detection rate with a relatively short model built-up time of 11.08 seconds. The study also reported that the proposed ensemble method outperforms other state-of-the-art feature selection methods in model built-up time, detection rate, and accuracy. However, this study does not discuss the limitations or potential weaknesses of the proposed ensemble feature selection method and how it would perform on other datasets.

In summary, the literature review on feature selection for IDS highlights the importance of selecting relevant features to improve the model's accuracy and efficiency while reducing dimensionality and minimizing overfitting. Different categories of feature selection methods, such as wrapper, filter, and embedded forms, have been proposed in the literature.

Moreover, Table 2 comprehensively overviews IDS's diverse feature selection methods. Each method has distinct strengths and limitations, making them suitable for specific scenarios. For example, PCA and Neural Network-based methods are noted for their efficiency in large and complex datasets. Still, they also face challenges such as the risk of overfitting and variance in performance across data distributions.

Hybrid methods and ensemble approaches, represented by studies like D'hooge et al. [58] and Kshirsagar and Kumar [68], offer innovative solutions for creating feature hierarchies and improving detection rates but often come with increased complexity and computational demands. Advanced approaches like BMRF Optimization and Genetic Algorithm-based methods, as shown in the works of Hassan et al. [57] and Halim et al. [60], demonstrate high precision and accuracy but require careful consideration in parameter tuning and broader scenario applications.

Table 2 Comparison of the data dimensionality works.

Study	Method	Key Characteristics	Strengths	Limitations
Subba et al. [54]	PCA	Reduces data complexity	Efficient for large datasets	Risk of overfitting; Limited real-world testing
Can et al. [55]	Neural Network-based	Automatic feature selection	Adaptable to complex datasets	Performance varies with data distribution
Di Mauro et al. [56]	Statistical Feature Analysis	Focuses on statistical features	Manages feature diversity	Potential biases: Lengthy training
Hassan et al. [57]	BMRF Optimization	High precision and accuracy	Effective in feature reduction	Limited exploration in diverse scenarios
D'hooge et al. [58]	Hybrid Feature Selection	Combines filter and embedding methods	Efficient for creating feature hierarchies	Complexity in implementation
Adnan et al. [1]	Rule-based & Multi-Objective Optimization	Balances selection with optimization	Reduces false positives in IoT networks	Computationally intensive
Mushtaq et al. [59]	Stacked Ensemble-based IDS	Incorporates multiple algorithms	High detection accuracy	Management complexity of multiple algorithms
Halim et al. [60]	Genetic Algorithm-based (GbFS)	Utilizes a novel fitness function	High accuracy	Parameter tuning challenges
Aksu and Aydin [61]	Modified Genetic Algorithm	Feature selection for CAN Bus systems	Versatile in various methods	Unexplored in real-world scenarios
Kiplinger et al. [62]	SPA-IDS Dataset Model	Utilizes vertical mode decomposition	High accuracy rates	Specificity to the SPA-IDS dataset
Chopra et al.	Swarm Intelligence & Ensemble	Non-deterministic feature selection	High f1 score in multiple datasets	Requires large data for validation
Panigrahi et al. [63]	MOEFS	Selects informative features	High accuracy with few features	Struggles with certain attack classes
Artur [64]	RFECV with Naive Bayes	Optimal feature number determination	Effective binary classification	Limited to Bernoulli Naive Bayes method
Herrera-Semenets et al. [65]	MMFSA	Combines three measures for feature estimation	Outperforms other algorithms	Optimal parameter value undetermined
Kasongo and Sun [66]	XGBoost with ML Methods	Reduced feature vector	Increases detection accuracy	Poor performance for minority classes
Disha and Waheed [67]	Gini Impurity-based Weighted RF	Feature selection for imbalanced datasets	Improves various model metrics	No multiclass classification analysis
Kshirsagar and Kumar [68]	Ensemble Feature Selection	Combines multiple feature selection techniques	Improved detection rate and accuracy	Performance on different datasets unexplored

In conclusion, the choice of feature selection method in IDS should be carefully aligned with the data's specific characteristics and the model's intended goals. Future research in this area should aim at developing more versatile and robust feature selection methods that can effectively address the evolving challenges in network security.

Data lack

IDS have become crucial in securing networks and systems from cyber-attacks. ML-based IDSs have gained popularity as they can learn and adapt to changing patterns in network traffic. However, using ML in IDSs has also introduced a new challenge: data lack. Data lack refers to the lack of sufficient and representative data for training and testing ML-based IDSs. This can lead to poor performance of the IDSs in real-world scenarios, as the models cannot generalize well to unseen data. In this literature review, we will explore the impact of data lack on the performance of ML-based IDSs and various techniques that have been proposed to address this issue. We will examine the current state of research on this topic and identify potential avenues for future work.

Kenyon et al. [69] discuss the challenges of developing a reliable and representative dataset for IDS. The authors argue that existing datasets are often outdated and not suggestive of current threat landscapes and that there is a lack of

standardization and transparency in the design and availability of datasets. The paper also addresses de-identifying sensitive information to meet regulations such as the General Data Protection Act (GDPR) and the lack of standardized metrics to compare datasets equally. Finally, the authors attempt to classify the most widely used public intrusion datasets, highlighting their limitations and best practices in dataset design. The paper concludes by suggesting that these contributions will facilitate ongoing research and development in IDS.

Thakkar and Lohiya [70] review the datasets developed for use in the IDS field and find that they need to be updated to identify recent attacks better and improve performance. The CIC-IDS-2017 and CSE-CIC-IDS-2018 datasets have been introduced to address this need for more realistic network traffic and updated attacks. The paper also discusses the characteristics of these datasets and notes some shortcomings. The authors plan to further study these datasets' performance using ML and data mining techniques and incorporate feature engineering and data sampling to address the shortcomings.

Yang et al. [71] present an overview and analysis of intrusion detection research in network security. The authors examined 119 highly cited publications, emphasizing preprocessing and intrusion detection techniques, evaluation methods, and the research community. They discovered that network anomaly detection research is imbalanced across different target networks. In addition, the lack of available datasets is a significant obstacle to cybersecurity research in the ICN and SDN domains. The authors also found that supervised learning is currently the mainstream approach but that unsupervised and semi-supervised learning and automated data labeling may be more promising. Finally, they also suggest that more research is needed on anti-perturbation anomaly detection in adversarial environments.

C. Zhang et al. [12] provide a summary of the application and research of ML in IDS and compare different ML algorithms that have been used in the field in recent years. The study finds that there is more research on ensemble learning and DL models, with ensemble learning being more mature and applied in various fields, while deep understanding is still in the early stages of exploration. The paper conducts experimental research using KDD99 and NSL-KDD datasets and finds that ensemble learning generally has better results, but other algorithms also have their advantages. The article suggests that further research should explore new models that integrate multiple algorithms and use different data and preprocessing methods. Additionally, the paper states that future research should focus on detecting unbalanced and new data types and finding or building good algorithms for this purpose.

Bui et al. [72] describe a toolchain that automates the process of feature extraction, data labeling, and assessing the quality of created datasets from various sources, including network traffic, system logs, and monitoring software reports. It also allows for the validation and customization of datasets. The experiment findings show that the dataset obtained through this framework performs better regarding coverage and detection efficiency for ML-based IDS. However, the limitations of this work include a lack of diversity in attack behaviors, a lack of support for IoT device types, and the need for further studies to compress the inter-learning time.

Lawrence et al. [73] present the design of a framework called CUPID, a publicly available, labeled network traffic dataset that includes a human pentester activity. CUPID aims to aid in the investigation of anomaly-based IDS and to serve as a representative of modern enterprise networking. In addition, the authors provide a discussion on the challenges faced and how they were overcome during the design process, with the hope that this will assist others in creating similar datasets in the future.

Jatti and Kishor Sontif [74] present the UNSW-NB15 dataset as a modern representative of network traffic and attack scenarios for use in IDS research. The IXIA PerfectStorm tool created the dataset, and includes 49 features and nine prominent attack families. The authors claim that benchmark datasets, such as KDD98, KDDCUP99, and NSLKDD, are outdated and limited in their representation of attacks and packet information. The paper includes a comparison of UNSW-NB15 with the KDDCUP99 dataset to demonstrate the benefits of the new dataset. The article provides a straightforward methodology for creating the dataset and a clear comparison with existing benchmark datasets. However, it would be beneficial for the authors to provide more concrete examples and statistics to back up their claims about the limitations of existing benchmark datasets and the superiority of the UNSW-NB15 dataset. Additionally, more discussion on the potential limitations and drawbacks of the UNSW-NB15 dataset would have helped provide a more comprehensive understanding of its utility in IDS research.

Ferriyan et al. [75] present a new IDS dataset called HIKARI-2021, which addresses the lack of publicly available and up-to-date datasets for benchmarking and comparing IDS performance. The authors make two main contributions: first, they propose a set of requirements for building new datasets, such as anonymization, payload capture, ground-truth data, and encryption, which are lacking in existing datasets. Second, they generate the HIKARI-2021 dataset, which includes network traffic with encrypted traces and a mix of ground-truth data, making it publicly available. The dataset consists of more than 80 features from CICIDS-2017 and additional features such as source and destination IP and port addresses, and it is labeled as benign or an attack. The authors also provide scripts and guidelines for generating new data and evaluating the dataset.

Finally, they assess the performance of the dataset using four ML algorithms and suggest future research could include background traffic analysis and application identification.

Bhattacharya [76] presents a dataset, SSENet-2014, which has been analyzed for its suitability in detecting multiconnection attacks. The dataset has 28 attributes, similar to those in the 10% KDD Cup 99 dataset, but with the addition of real-world traffic trace files. The SSENet-2014 dataset is compared with the 10% KDD Cup 99 dataset, and it is found that SSENet-2014 is better in terms of the distribution of points and distinct attribute values for the 'normal' and 'attack' classes. Therefore, the authors suggest that SSENet-2014 can be used with the 10% KDD Cup 99 dataset to evaluate detection algorithms for multiconnection attacks. However, the paper does not provide a detailed description of the data collection process, and it is unclear how the real-world traffic trace files were obtained. Additionally, the article does not offer any results from experiments using the dataset.

Rajasinghe et al. [77] introduce INSecS-DCS, a software framework that can create a labeled network intrusion dataset. The framework can accept inputs from real-time packet capture streams or imported PCAP files and choose between a raw and processed dataset as the output. The paper's authors highlight the framework's flexibility, which allows researchers or IDS users to recreate datasets with attributes that meet their specific needs. The plans for INSecS-DCS include using the dataset creation capability to provide a Real-Time Network IDS with real-time datasets. The authors mention that this will significantly impact the commercial and private network security industry. One possible criticism of the work could be that the authors do not evaluate the INSecS-DCS dataset's performance compared to other existing datasets.

The KDD99 dataset, created in 1999 by Lippmann et al. [78], is widely used in IDS research but is considered inadequate for evaluation due to its age and the fact that it was captured in a simulated environment. In addition, it contains 24 attack types but is not updated with recent attack vectors. Nevertheless, despite some researchers recognizing its limitations, it is still widely used for benchmarking in IDS studies.

Sharafaldin et al. [79] discuss the challenges in developing a reliable and publicly available evaluation dataset for IDS. First, the authors analyze 11 publicly available datasets from 1998 to 2016 and find that they are limited in terms of traffic diversity and volume, anonymized packet information and payload, various attacks, and feature sets and metadata. In response, the authors propose a new IDS dataset, called CICIDS2017, that includes seven updated attack families and is publicly available. The authors then evaluate the dataset using 80 traffic features and seven ML algorithms and compare it to publicly available datasets using a proposed evaluation framework. Finally, the authors suggest that they will increase the number of PCs and conduct more up-to-date attacks in the future.

In this literature review, we have examined the impact of data lack on ML-based IDS performance and the various techniques proposed to address this issue. The studies reviewed indicate that the lack of sufficient and representative data for training and testing ML-based IDSs can lead to poor performance in real-world scenarios.

The authors of the reviewed studies have highlighted the challenges in developing reliable and representative datasets for IDSs and the need for standardization and transparency in the design and availability of datasets. They also suggested various techniques such as ensemble learning, DL models, unsupervised and semi-supervised learning, and automated data labeling to address the issue of data lack.

Furthermore, the comprehensive table (Table 3) presents a detailed overview of the diverse range of studies addressing the lack of data in ML-based IDS. A recurring theme across these studies is the need for updated, realistic, and representative datasets that reflect current threat landscapes. Kenyon et al. [69] and Thakkar and Lohiya [70] emphasize the necessity for modern datasets, while Yang et al. [71] and C. Zhang et al. [12] highlight the need for innovative preprocessing techniques and the exploration of new ML models.

The creation of new datasets, such as CUPID by Lawrence et al. [73], UNSW-NB15 by Jatti and Kishor Sontif [74], and HIKARI-2021 by Ferriyan et al. [75], represents a significant effort to address data lack. These datasets aim to offer more realistic and comprehensive data for IDS research. However, challenges such as dataset design, lack of diverse attack behaviors, and the need for further validation are common limitations.

Table 3 Comparison of the data dimensionality works.

Study Reference	Key Focus	Proposed Solutions	Limitations
Kenyon et al. [69]	Challenges in dataset development	Standardization, transparency, de-identification	Outdated datasets; Lack of standardized metrics
Thakkar and Lohiya [70]	Review of IDS datasets	Introduction of updated datasets (CIC-IDS-2017, CSE-CIC-IDS-2018)	Need for further study on dataset performance
Yang et al. [71]	Overview of Intrusion Detection Research	Emphasis on preprocessing, unsupervised and semi-supervised learning	Imbalance in research across networks; Lack of datasets in specific domains
C. Zhang et al. [12]	Comparison of ML algorithms in IDS	Exploration of new models integrating multiple algorithms	Challenges with unbalanced and new data types

Bui et al. [72]	Automated toolchain for dataset creation	Feature extraction, data labeling, and quality assessment tools	Limited attack behavior diversity; Lack of IoT device support
Lawrence et al. [73]	Design of the CUPID dataset	Inclusion of human pentester activity in the dataset	Challenges in the dataset design process
Jatti and Kishor Sontif [74]	Introduction of the UNSW-NB15 dataset	Modern representative of network traffic and attacks	Comparison with outdated datasets; Lack of concrete examples
Ferriyan et al. [75]	Creation of the HIKARI-2021 dataset	Anonymization, payload capture, ground-truth data, encryption	Need for background traffic analysis; Encryption challenges
Bhattacharya [76]	Analysis of the SSENet-2014 dataset	Real-world traffic trace files for detecting attacks	Lack of detailed data collection process description
Rajasinghe et al. [77]	INSecS-DCS software framework	Creation of labeled network intrusion datasets	Lack of performance evaluation compared to other datasets
Lippmann et al. [78]	Evaluation of the KDD99 dataset	Widely used benchmark in IDS studies	Outdated and simulated environment; Not updated with recent attacks
Sharafaldin et al. [79]	Development of CICIDS2017 dataset	Updated attack families; Evaluation using ML algorithms	Future plans for more up-to-date attacks and increased PCs

In conclusion, addressing data lack in ML-based IDS is critical for improving system performance in real-world scenarios. The studies reviewed underscore the importance of developing robust, representative, and up-to-date datasets. Future research should continue to focus on overcoming the challenges in dataset creation and exploring new techniques to mitigate the impact of data lack, including anti-perturbation anomaly detection in adversarial environments.

Data Leak

In recent times, IDS that employ ML techniques have gained prominence in detecting cyber-attacks. These systems typically depend on extensive labeled data, often sourced from real-world network environments. However, a common issue arises when the training and testing data do not accurately represent real-world scenarios, leading to what is known as pattern leakage.

Pattern leakage is a phenomenon where training and testing data are not independent and identically distributed. This situation often results in overly specific models to the training data, leading to suboptimal generalization when applied to new, unseen data [80]. Consequently, this affects the IDS's effectiveness in real-world deployment.

A study by Bouke and Abdullah [81]. Delves into the impact of pattern leakage during data preprocessing in ML-based IDS. Using datasets like NSL-KDD, UNSW-NB15, and KDDCUP99, they demonstrated that data leakage leads to inflated accuracy scores and unreliable models. Their findings indicate a heightened sensitivity of certain algorithms like Decision Trees and Gradient Boosting to data leakage, compared to others like Support Vector Machines. They emphasized the importance of proper data preprocessing and cautious model selection to prevent data leakage, ensuring the reliability and generalization capability of IDS models.

Further studies have explored various aspects of pattern leakage in ML-based IDS.

Several studies have delved into the impact of pattern leakage on the efficacy of ML-based IDS. Zheng and Casari [82] discuss how leakage in ML can occur when information intended for model training is inadvertently included in the test dataset. They recommend a strict division between the training and testing phases, with different data batches for each. However, this approach may lag behind current data trends, thus affecting the model's immediacy.

Dong [17] introduces a Bayesian inference-based method to detect data leakage in ML models. This innovative approach successfully predicts leakage in a dataset from sports wearables by estimating the marginal probability lower limit for observed variables. This Bayesian method provides a unique way to detect leakage in complex data distributions and can be coupled with other ML techniques for a more robust defense against attacks.

Farokhi and Kaafar [83] propose quantifying membership information leakage in ML models using mutual information and Kullback–Leibler divergence. Their findings indicate that this leakage decreases with larger training datasets, higher regularization weights, and increased model sensitivity. Adding noise to the data is also explored as a privacy-preserving measure against membership inference attacks. They suggest future exploration in applying their measures to more complex models like deep neural networks.

W. Zhang et al. [84] describe a technique in a centralized multi-party ML context for accessing sensitive data of other parties. This black-box attack method can extract information about sensitive attributes with limited queries, underscoring that traditional security measures are insufficient for data protection. They evaluated this method across various data types and correlations, highlighting the limitations of secure computation and differential privacy techniques.

Kuhn and Johnson [85] examine information leakage in resampling, particularly concerning input variable normalization. They assert that utilizing test set data during training can lead to overly optimistic results that do not replicate in future data instances. They also discuss the complexities in applying differential privacy and secure computation techniques to prevent this type of leakage.

Hannun et al. [86] explore using Fisher information loss (FIL) to measure the information an ML model leaks about its training data. They argue that FIL offers several advantages, including leakage assessment at different levels and the ability to design models with evenly distributed leakage. While FIL

helps tailor privacy to specific levels, the authors acknowledge its limitations and the need for further research.

To this end, data leak, or pattern leakage, is a critical issue in ML-based IDS where training and testing data are not independent and identically distributed, often leading to models that do not generalize well to unseen data. The comprehensive table (Table 4) presents an overview of the studies addressing data leak or pattern leakage in ML-based IDS. The studies explore various aspects of this issue, ranging from the impact of improper data preprocessing to advanced methods for detecting and quantifying data leakage.

Bouke and Abdullah [81] highlight the crucial role of data preprocessing and careful model selection to prevent data leakage, while Zheng and Casari [82] recommend a strict separation between training and testing data. Dong [17] introduces a Bayesian inference-based method for predicting leakage, providing a novel approach to this challenge.

Farokhi and Kaafar [83] propose using mutual information and Kullback–Leibler divergence to quantify membership information leakage, suggesting that larger datasets and certain model configurations can reduce leakage. W. Zhang et al. [84] and Kuhn and Johnson [85] discuss the limitations of current security measures and the complexities involved in implementing differential privacy and secure computation techniques.

Hannun et al. [86] introduce Fisher information loss (FIL) as a method for assessing information leakage, offering a way to design models with evenly distributed leakage. However, they acknowledge the need for further research in this area.

In conclusion, addressing data leaks in ML-based IDS is essential for ensuring the effectiveness and reliability of these systems in real-world scenarios. The reviewed studies indicate the need for robust data preprocessing, model selection, and innovative techniques to detect and mitigate data leakage. Future research should continue to develop methods to address these challenges and enhance the generalization capabilities of IDS models.

Table 4 Comparison of the data leak works.

Study	Key Focus	Proposed Solutions	Limitations
Bouke and Abdullah [81]	Impact of pattern leakage in data preprocessing	Proper data preprocessing and model selection	Heightened sensitivity of certain algorithms to data leakage
Zheng and Casari [82]	Leakage in ML due to training-test data overlap	Strict division between training and testing phases	Potential lag behind current data trends
Dong [17]	Detection of data leakage using Bayesian inference	Bayesian inference-based method	Specific to complex data distributions; Applicability in broader contexts
Farokhi and Kaafar [83]	Quantification of membership information leakage	Mutual information and Kullback–Leibler divergence measures	Applicability to more complex models like deep neural networks
W. Zhang et al. [84]	Black-box attack method in centralized multi-party ML	Highlighting limitations of traditional security measures	Evaluation across various data types and correlations
Kuhn and Johnson [85]	Information leakage in resampling	Discussion on differential privacy and secure computation	Complexities in applying privacy techniques
Hannun et al. [86]	Measurement of information leak using FIL	Fisher information loss (FIL) for leakage assessment	Limitations in tailoring privacy to specific levels

CONCLUSION

In conclusion, ML-based IDS have emerged as a crucial approach for detecting and preventing cyber-attacks. The effectiveness of these systems heavily relies on the quality, availability, and characteristics of the training data. Data Lack, Data Leak, and Data Dimensionality (DLLD) issues are significant challenges that can adversely affect the performance of IDS models. These challenges can hinder the systems' ability to effectively detect and respond to unauthorized access, misuse, alteration, or destruction of information systems.

Throughout this paper, we have reviewed and analyzed how DLLD issues impact IDS models. Based on our critical analysis of the literature, we have identified various techniques to address these challenges:

- **Data Lack**: Techniques such as data augmentation, synthetic data generation, and modern, realistic datasets are recommended. These approaches help compensate for data scarcity and enhance datasets' representativeness.
- **Data Leakage**: Preventive measures include rigorous cross-validation, strict separation between training and testing datasets, and feature engineering to ensure that the models do not inadvertently learn from test data.
- **Data Dimensionality**: Employing feature selection, dimensionality reduction, and data preprocessing techniques are essential. These techniques help manage high-dimensional data and improve the models' interpretability and accuracy.

Furthermore, DL-based intrusion detection methods and adversarial ML-based approaches have shown promise in improving the performance of IDS models, offering advanced capabilities in handling complex and evolving cyber threats.

However, addressing DLLD issues is just one facet of developing effective and efficient IDS models. Future research should focus on several key areas:

- **Development of Robust and Representative Datasets**: Creating datasets that accurately reflect current and emerging threat landscapes is crucial. These datasets should account for the latest attack vectors and network behaviors.
- **Exploration of Diverse ML Algorithms and Models**: Investigating the efficacy of different ML algorithms and models in IDS can provide insights into their suitability for various scenarios and challenges.

- **Advancement in Data Preprocessing and Feature Selection Methods**: Exploring new data preprocessing and feature selection methods can lead to more efficient and accurate models, especially in handling large and complex datasets.
- **Integration of Privacy and Security Concerns**: In the dataset creation, handling, and sharing process, integrating privacy and security considerations is essential. This includes compliance with data protection regulations and ensuring that datasets do not expose sensitive information.

Through addressing these challenges and exploring these research directions, the field of IDS can continue to evolve, offering more robust defenses against an increasingly sophisticated landscape of cyber threats.

REFERENCES

[1] A. Adnan, A. Muhammed, A. A. A. Ghani, A. Abdullah, and F. Hakim, "An intrusion detection system for the internet of things based on machine learning: Review and challenges," *Symmetry (Basel).*, vol. 13, no. 6, pp. 1–13, 2021, doi: 10.3390/sym13061011.

[2] M. J. Kang and J. W. Kang, "Intrusion detection system using deep neural network for in-vehicle network security," *PLoS One*, vol. 11, no. 6, pp. 1–17, 2016, doi: 10.1371/journal.pone.0155781.

[3] G. R. Gauthama, C. M. Ahmed, and A. Mathur, "Machine learning for intrusion detection in industrial control systems: challenges and lessons from experimental evaluation," *Cybersecurity*, vol. 4, no. 1, 2021, doi: 10.1186/s42400-021-00095-5.

[4] J. Verma, A. Bhandari, and G. Singh, "iNIDS: SWOT Analysis and TOWS Inferences of State-of-the-Art NIDS solutions for the development of Intelligent Network Intrusion Detection System," *Comput. Commun.*, vol. 195, no. August, pp. 227–247, 2022, doi: 10.1016/j.comcom.2022.08.022.

[5] D. Soni and N. Kumar, "Machine learning techniques in emerging cloud computing integrated paradigms: A survey and taxonomy," *J. Netw. Comput. Appl.*, vol. 205, no. May, p. 103419, 2022, doi: 10.1016/j.jnca.2022.103419.

[6] T. Saranya, S. Sridevi, C. Deisy, T. D. Chung, and M. K. A. A. Khan, "Performance Analysis of Machine Learning Algorithms in Intrusion Detection System: A Review," *Procedia Comput. Sci.*, vol. 171, no. 2019, pp. 1251–1260, 2020, doi: 10.1016/j.procs.2020.04.133.

[7] A. Balla, M. H. Habaebi, M. R. Islam, and S. Mubarak, "Applications of deep learning algorithms for Supervisory Control and Data Acquisition intrusion detection system," *Clean. Eng. Technol.*, vol. 9, no. June, p. 100532, 2022, doi: 10.1016/j.clet.2022.100532.

[8] M. A. Bouke, A. Abdullah, S. H. ALshatebi, and M. T. Abdullah, "E2IDS: An Enhanced Intelligent Intrusion Detection System Based On Decision Tree Algorithm," *J. Appl. Artif. Intell.*, vol. 3, no. 1, pp. 1–16, 2022, doi: 10.48185/jaai.v3i1.450.

[9] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: techniques, datasets and challenges," *Cybersecurity*, vol. 2, no. 1, 2019, doi: 10.1186/s42400-019-0038-7.

[10] C. A. de Souza, C. B. Westphall, R. B. Machado, L. Loffi, C. M. Westphall, and G. A. Geronimo, "Intrusion detection and prevention in fog based IoT environments: A systematic literature review," *Comput. Networks*, vol. 214, no. March, p. 109154, 2022, doi: 10.1016/j.comnet.2022.109154.

[11] A. Heidari and M. A. Jabraeil Jamali, "Internet of Things intrusion detection systems: a comprehensive review and future directions," *Cluster Comput.*, vol. 0123456789, 2022, doi: 10.1007/s10586-022-03776-z.

[12] C. Zhang, D. Jia, L. Wang, W. Wang, F. Liu, and A. Yang, "Comparative research on network intrusion detection methods based on machine learning," *Comput. Secur.*, vol. 121, p. 102861, 2022, doi: 10.1016/j.cose.2022.102861.

[13] E. M. Onyema, S. Dalal, C. A. T. Romero, B. Seth, P. Young, and M. A. Wajid, "Design of Intrusion Detection System based on Cyborg intelligence for security of Cloud Network Traffic of Smart Cities," *J. Cloud Comput.*, vol. 11, no. 1, 2022, doi: 10.1186/s13677-022-00305-6.

[14] X. Hu *et al.*, "A Systematic View of Model Leakage Risks in Deep Neural Network Systems," *IEEE Trans. Comput.*, vol. 71, no. 12, pp. 3254–3267, 2022, doi: 10.1109/TC.2022.3148235.

[15] D. Z. Abidin, S. Nurmaini, R. Firsandava Malik, Erwin, E. Rasywir, and Y. Pratama, "RSSI Data Preparation for Machine Learning," *Proc. - 2nd Int. Conf. Informatics, Multimedia, Cyber, Inf. Syst. ICIMCIS 2020*, pp. 284–289, 2020, doi: 10.1109/ICIMCIS51567.2020.9354273.

[16] I. Souiden, M. N. Omri, and Z. Brahmi, "A survey of outlier detection in high dimensional data streams," *Comput. Sci. Rev.*, vol. 44, p. 100463, 2022, doi: 10.1016/j.cosrev.2022.100463.

[17] Q. Dong, "Leakage Prediction in Machine Learning Models When Using Data from Sports Wearable Sensors," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/5314671.

[18] S. Yuan and X. Wu, "Deep learning for insider threat detection: Review, challenges and opportunities," *Comput. Secur.*, vol. 104, p. 102221, 2021, doi: 10.1016/j.cose.2021.102221.

[19] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[20] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[21] A. Kołcz and C. H. Teo, "Feature weighting for improved classifier robustness," *6th Conf. Email Anti-Spam, CEAS 2009*, no. 1, 2009.

[22] B. Zong *et al.*, "Deep autoencoding Gaussian mixture model for unsupervised anomaly detection," *6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc.*, pp. 1–19, 2018.

[23] H. Purohit, R. Tanabe, T. Endo, K. Suefusa, Y. Nikaido, and Y. Kawaguchi, "Deep Autoencoding GMM-based Unsupervised Anomaly Detection in Acoustic Signals and its Hyper-parameter Optimization," no. November, pp. 1–5, 2020, [Online]. Available: http://arxiv.org/abs/2009.12042

[24] A. Derhab, A. Aldweesh, A. Z. Emam, and F. A. Khan, "Intrusion Detection System for Internet of Things Based on Temporal Convolution Neural Network and Efficient Feature Engineering," *Wirel. Commun. Mob. Comput.*, vol. 2020, no. April, 2020, doi: 10.1155/2020/6689134.

[25] T. Kanamori, S. Hido, and M. Sugiyama, "Efficient direct density ratio estimation for non-stationarity adaptation and outlier detection," *Adv. Neural Inf. Process. Syst.*, vol. 21, 2008.

[26] G. C. Cawley and N. L. C. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *J. Mach. Learn. Res.*, vol. 11, pp. 2079–2107, 2010.

[27] K. M. Ting and I. H. Witten, "Issues in stacked generalization," *arXiv e-prints*, p. arXiv--1105, 2011.

[28] G. C. Cawley and N. L. C. Talbot, "Kernel learning at the first level of inference," *Neural networks*, vol. 53, pp. 69–80, 2014.

[29] Y. Liu, S. Liao, S. Jiang, L. Ding, H. Lin, and W. Wang, "Fast cross-validation for kernel-based algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1083–1096, 2019.

[30] R. Bellman, "On the approximation of curves by line segments using dynamic programming," *Commun. ACM*, vol. 4, no. 6, p. 284, 1961.

[31] W. H. R. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Trans. Inf. Theory*, vol. 37, no. 2, pp. 269–275, 1991.

[32] A. Kanlis and P. Narayan, "Error exponents for successive refinement by partitioning," *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 275–282, 1996.

[33] S. Gattani, *Reference models for network trace anonymization*. Iowa State University, 2008.

[34] J. Wang, Z. Xia, Y. Chen, C. Hu, and F. Yu, "Intrusion detection framework based on homomorphic encryption in AMI network," *Front. Phys.*, vol. 10, p. 1102892, 2022.

[35] J. Stapleton and R. S. Poore, "Tokenization and other methods of security for cardholder data," *Inf. Secur. J. A Glob. Perspect.*, vol. 20, no. 2, pp. 91–99, 2011.

[36] S. Murthy, A. A. Bakar, F. A. Rahim, and R. Ramli, "A comparative study of data anonymization techniques," in *2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC)*

and IEEE Intl Conference on Intelligent Data and Security (IDS), 2019, pp. 306–309.

[37] R. Bonifazi, J. Vandenplas, J. ten Napel, K. Matilainen, R. F. Veerkamp, and M. P. L. Calus, "Impact of sub-setting the data of the main Limousin beef cattle population on the estimates of across-country genetic correlations," *Genet. Sel. Evol.*, vol. 52, pp. 1–16, 2020.

[38] D. A. Van Dyk and X.-L. Meng, "The art of data augmentation," *J. Comput. Graph. Stat.*, vol. 10, no. 1, pp. 1–50, 2001.

[39] K. Houkjær, K. Torp, and R. Wind, "Simple and realistic data generation," in *Proceedings of the 32nd international conference on Very large data bases*, 2006, pp. 1243–1246.

[40] D. Evans, "Systematic reviews of interpretive research: interpretive data synthesis of processed data," *Aust. J. Adv. Nursing*, vol. 20, no. 2, 2002.

[41] Z. Zhang, "Missing data imputation: focusing on single imputation," *Ann. Transl. Med.*, vol. 4, no. 1, 2016.

[42] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *J. Big data*, vol. 3, no. 1, pp. 1–40, 2016.

[43] M. A. Bouke, A. Abdullah, J. Frnda, K. Cengiz, and B. Salah, "BukaGini: A Stability-Aware Gini Index Feature Selection Algorithm for Robust Model Performance," *IEEE Access*, vol. 11, pp. 59386–59396, 2023, doi: 10.1109/ACCESS.2023.3284975.

[44] M. Sarhan, S. Layeghy, N. Moustafa, M. Gallagher, and M. Portmann, "Feature extraction for machine learning-based intrusion detection in IoT networks," *Digit. Commun. Networks*, 2022, doi: 10.1016/j.dcan.2022.08.012.

[45] M. A. Bouke, A. Abdullah, S. H. ALshatebi, M. T. Abdullah, and H. El Atigh, "An intelligent DDoS attack detection tree-based model using Gini index feature selection method," *Microprocess. Microsyst.*, vol. 98, no. March, p. 104823, 2023, doi: 10.1016/j.micpro.2023.104823.

[46] A. J. Izenman, "Introduction to manifold learning," *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 4, no. 5, pp. 439–446, 2012.

[47] P. B. Udas, M. E. Karim, and K. S. Roy, "SPIDER: A shallow PCA based network intrusion detection system with enhanced recurrent neural networks," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 10, pp. 10246–10272, 2022, doi: 10.1016/j.jksuci.2022.10.019.

[48] A. Dahou *et al.*, "Intrusion Detection System for IoT Based on Deep Learning and Modified Reptile Search Algorithm," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–15, 2022, doi: 10.1155/2022/6473507.

[49] Y. Himeur, K. Ghanem, A. Alsalemi, F. Bensaali, and A. Amira, "Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives," *Appl. Energy*, vol. 287, no. November 2020, p. 116601, 2021, doi: 10.1016/j.apenergy.2021.116601.

[50] S. Rao, A. K. Verma, and T. Bhatia, "A review on social spam detection: Challenges, open issues, and future directions," *Expert Syst. Appl.*, vol. 186, no. August, p. 115742, 2021, doi: 10.1016/j.eswa.2021.115742.

[51] R. K. Deka, D. K. Bhattacharyya, and J. K. Kalita, "Active learning to detect DDoS attack using ranked features," *Comput. Commun.*, vol. 145, no. June, pp. 203–222, 2019, doi: 10.1016/j.comcom.2019.06.010.

[52] T. Hamed, R. Dara, and S. C. Kremer, "Network intrusion detection system based on recursive feature addition and bigram technique," *Comput. Secur.*, vol. 73, pp. 137–155, 2018, doi: 10.1016/j.cose.2017.10.011.

[53] N. Pilnenskiy and I. Smetannikov, "Modern Implementations of Feature Selection Algorithms and Their Perspectives," *Conf. Open Innov. Assoc. Fruct*, pp. 250–256, 2019, doi: 10.23919/FRUCT48121.2019.8981498.

[54] B. Subba, S. Biswas, and S. Karmakar, "Enhancing performance of anomaly based intrusion detection systems through dimensionality reduction using principal component analysis," in *2016 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*, 2016, pp. 1–6.

[55] D. C. Can, H. Q. Le, and Q. T. Ha, "Detection of Distributed Denial of Service Attacks Using Automatic Feature Selection with Enhancement for Imbalance Dataset," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2021, pp. 386–398. doi: 10.1007/978-3-030-73280-6_31.

[56] M. Di Mauro, G. Galatro, G. Fortino, and A. Liotta, "Supervised feature selection techniques in network intrusion detection: A critical review," *Eng. Appl. Artif. Intell.*, vol. 101, no. October 2020, p. 104216, 2021, doi: 10.1016/j.engappai.2021.104216.

[57] I. H. Hassan, M. Abdullahi, M. M. Aliyu, S. A. Yusuf, and A. Abdulrahim, "An improved binary manta ray foraging optimization algorithm based feature selection and random forest classifier for network intrusion detection," *Intell. Syst. with Appl.*, vol. 16, no. November 2021, p. 200114, 2022, doi: 10.1016/j.iswa.2022.200114.

[58] L. D'hooge, M. Verkerken, T. Wauters, B. Volckaert, and F. De Turck, "Hierarchical feature block ranking for data-efficient intrusion detection modeling," *Comput. Networks*, vol. 201, no. February, p. 108613, 2021, doi: 10.1016/j.comnet.2021.108613.

[59] E. Mushtaq, A. Zameer, and A. Khan, "A two-stage stacked ensemble intrusion detection system using five base classifiers and MLP with optimal feature selection," *Microprocess. Microsyst.*, vol. 94, no. December 2021, p. 104660, 2022, doi: 10.1016/j.micpro.2022.104660.

[60] Z. Halim *et al.*, "An effective genetic algorithm-based feature selection method for intrusion detection systems," *Comput. Secur.*, vol. 110, p. 102448, 2021, doi: 10.1016/j.cose.2021.102448.

[61] D. Aksu and M. A. Aydin, "MGA-IDS: Optimal feature subset selection for anomaly detection framework on in-vehicle networks-CAN bus based on genetic algorithm and intrusion detection approach," *Comput. Secur.*, vol. 118, p. 102717, 2022, doi: 10.1016/j.cose.2022.102717.

[62] I. F. Kilincer, T. Tuncer, F. Ertam, and A. Sengur, "SPA-IDS: An intelligent intrusion detection system based on vertical mode decomposition and iterative feature selection in computer networks," *Microprocess. Microsyst.*, vol. 96, no. December 2021, p. 104752, 2023, doi: 10.1016/j.micpro.2022.104752.

[63] R. Panigrahi *et al.*, "Intrusion detection in cyber–physical environment using hybrid Naïve Bayes—Decision table and multi-objective evolutionary feature selection," *Comput. Commun.*, vol. 188, no. September 2021, pp. 133–144, 2022, doi: 10.1016/j.comcom.2022.03.009.

[64] M. Artur, "Review the performance of the Bernoulli Naïve Bayes Classifier in Intrusion Detection Systems using Recursive Feature Elimination with Cross-validated selection of the best number of features," *Procedia Comput. Sci.*, vol. 190, no. 2019, pp. 564–570, 2021, doi: 10.1016/j.procs.2021.06.066.

[65] V. Herrera-Semenets, L. Bustio-Martínez, R. Hernández-León, and J. van den Berg, "A multi-measure feature selection algorithm for efficacious intrusion detection," *Knowledge-Based Syst.*, vol. 227, p. 107264, 2021, doi: 10.1016/j.knosys.2021.107264.

[66] S. M. Kasongo and Y. Sun, "Performance Analysis of Intrusion Detection Systems Using a Feature Selection Method on the UNSW-NB15 Dataset," *J. Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00379-6.

[67] R. A. Disha and S. Waheed, "Performance analysis of machine learning models for intrusion detection system using Gini Impurity-based Weighted Random Forest (GIWRF) feature selection technique," *Cybersecurity*, vol. 5, no. 1, pp. 1–22, 2022, doi: 10.1186/s42400-021-00103-8.

[68] D. Kshirsagar and S. Kumar, "Towards an intrusion detection system for detecting web attacks based on an ensemble of filter feature selection techniques," *Cyber-Physical Syst.*, vol. 00, no. 00, pp. 1–16, 2022, doi: 10.1080/23335777.2021.2023651.

[69] A. Kenyon, L. Deka, and D. Elizondo, "Are public intrusion datasets fit for purpose characterising the state of the art in intrusion event datasets," *Comput. Secur.*, vol. 99, p. 102022, 2020, doi: 10.1016/j.cose.2020.102022.

[70] A. Thakkar and R. Lohiya, "A Review of the Advancement in Intrusion Detection Datasets," *Procedia Comput. Sci.*, vol. 167, no. 2019, pp. 636–645, 2020, doi: 10.1016/j.procs.2020.03.330.

[71] Z. Yang *et al.*, "A systematic literature review of methods and datasets for anomaly-based network intrusion detection," *Comput. Secur.*, vol. 116, 2022, doi: 10.1016/j.cose.2022.102675.

[72] H. K. Bui, Y. D. Lin, R. H. Hwang, P. C. Lin, V. L. Nguyen, and Y. C. Lai, "CREME: A toolchain of automatic dataset collection for machine learning in intrusion detection," *J. Netw. Comput. Appl.*, vol. 193, no. March, p. 103212, 2021, doi: 10.1016/j.jnca.2021.103212.

[73] H. Lawrence *et al.*, "CUPID: A labeled dataset with Pentesting for evaluation of network intrusion detection," *J. Syst. Archit.*, vol. 129, no. May, p. 102621, 2022, doi: 10.1016/j.sysarc.2022.102621.

[74] S. A. V. Jatti and V. J. K. Kishor Sontif, "UNSW-NB15: A Comprehensive Data set for Network Intrusion Detection systems," *Int. J. Recent Technol. Eng.*, vol. 8, no. 2 Special Issue 11, pp. 3976–3983, 2019, doi: 10.35940/ijrte.B1540.0982S1119.

[75] A. Ferriyan, A. H. Thamrin, K. Takeda, and J. Murai, "Generating network intrusion detection dataset based on real and encrypted synthetic attack traffic," *Appl. Sci.*, vol. 11, no. 17, 2021, doi: 10.3390/app11177868.

[76] S. Bhattacharya and S. Selvakumar, "SSENet-2014 Dataset: A Dataset for Detection of Multiconnection Attacks," *Proc. - 2014 3rd Int. Conf. Eco-Friendly Comput. Commun. Syst. ICECCS 2014*, pp. 121–126, 2015, doi: 10.1109/Eco-friendly.2014.100.

[77] N. Rajasinghe, J. Samarabandu, and X. Wang, "INSECS-DCS: A Highly Customizable Network Intrusion Dataset Creation Framework," *Can. Conf. Electr. Comput. Eng.*, vol. 2018-May, 2018, doi: 10.1109/CCECE.2018.8447661.

[78] R. P. Lippmann *et al.*, "Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation," in *Proceedings DARPA Information Survivability Conference and Exposition. DISCEX'00*, 2000, pp. 12–26.

[79] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," *ICISSP 2018 - Proc. 4th Int. Conf. Inf. Syst. Secur. Priv.*, vol. 2018-Janua, no. Cic, pp. 108–116, 2018, doi: 10.5220/0006639801080116.

[80] J. Brownlee, "Machine Learning Mastery With Python: Data Cleaning, Feature Selection, and Data Transforms in Python," p. 500, 2020.

[81] M. A. Bouke and A. Abdullah, "An empirical study of pattern leakage impact during data preprocessing on machine learning-based intrusion detection models reliability," *Expert Syst. Appl.*, vol. 230, no. June, p. 120715, 2023, doi: 10.1016/j.eswa.2023.120715.

[82] A. Zheng and A. Casari, *Feature engineering for machine learning*, no. September. 2018.

[83] F. Farokhi and M. A. Kaafar, "Modelling and Quantifying Membership Information Leakage in Machine Learning," pp. 1–13, 2020, [Online]. Available: http://arxiv.org/abs/2001.10648

[84] W. Zhang, S. Tople, and O. Ohrimenko, "Leakage of dataset properties in multi-party machine learning," *Proc. 30th USENIX Secur. Symp.*, pp. 2687–2704, 2021.

[85] M. Kuhn and K. Johnson, *Feature Engineering and Selection*. 2019. doi: 10.1201/9781315108230.

[86] A. Hannun, C. Guo, and L. van der Maaten, "Measuring Data Leakage in Machine-Learning Models with Fisher Information (Extended Abstract)," *IJCAI Int. Jt. Conf. Artif. Intell.*, no. Uai, pp. 5284–5288, 2022, doi: 10.24963/ijcai.2022/736.

BIOGRAPHIES

MOHAMED ALY BOUKE holds a Master's and a Ph.D. in Information Security from the University of Putra Malaysia, specializing in cybersecurity, cyber warfare, and machine learning applications in information security. He is a member of the IEEE (Institute of Electrical and Electronics Engineers), reflecting his involvement in the technology and engineering community. In his role with the International Information System Security Certification Consortium (ISC2), Mohamed contributes to advancing cybersecurity practices. He is a certified trainer for various international organizations and engages in educating students worldwide through training programs. His expertise extends to authoring publications and participating as a manuscript reviewer for recognized journals, furthering his engagement in the cybersecurity field. As a public speaker and author, Mohamed shares his knowledge and insights, adding value to discussions and literature in information security.

AZIZOL ABDULLAH received the M.Sc. degree in engineering (telematics) from The University of Sheffield, U.K., in 1996, and the Ph.D. degree in parallel and distributed systems from Univer- siti Putra Malaysia, Malaysia, in 2010. He is an Associate Professor with the Department of Tech- nology and Communication Networking, Faculty of Computer Science and Information Technol- ogy, Universiti Putra Malaysia. He is the Head of the Network, Parallel, and Distributed Computing Research Group and a member of the Information Security Research Group at the Faculty of Computer Science and Information Technology, UPM. At the national level, he is a member of Cyber Security Academia Malaysia (CSAM). He was also appointed as a Fellow Researcher for ITU-UUM Asia Pacific Center of Excellence For Rural ICT Development (ITU-UUM). He has also been involved as a consultant for AnyCast@MyDNS Project, MyNIC and Ministry of Science and Innovation projects, Malaysia (MOSTI) and Integrated Sports Management System Project, Ministry of Youth and Sports, Malaysia. His main research areas include cloud and grid computing, network security, wireless and mobile computing and computer networks. He is engaged in Malware Detection research, SDN, SDWAN network research and SDWAN Security research.

NUR IZURA UDZIR is an Associate Professor at the Faculty of Computer Science and Information Technology, Universiti Putra Malaysia (UPM) since 1998. She received her Bachelor of Computer Science (1995) and Master of Science (1998) from UPM, and her PhD in Computer Science from the University of York, UK (2006). Her areas of specialization are computer security, intrusion detection systems, access control, secure operating systems, steganography, coordination models and languages, and distributed systems. She is a member of IEEE Computer Society, Malaysian Board of Technologists (MBOT), Information Security Professionals Association of Malaysia (ISPA.my), Society of Digital Information and Wireless Communications (SDIWC). Dr. Nur Izura has supervised and co-supervised over 50 PhD students and over 15 Master (by research) students. She has written a book on Introduction to C++ Programming (2001), edited 3 books in information security topics, and has published over 120 articles in journals and as book chapters, and over 100 international conference proceedings, thus earning a H-index of 17 with 1331 citations in Scopus (H-index 26 and 2679 citations in Google Scholar) as of December 2022. For her contributions in academic and research, she has won various awards, i.e. the MIMOS Prestigious Award 2015 for the supervision of her student's doctoral thesis, the Young Scientist Award 2021, and seven Best Paper Awards at international conferences. In addition to keynote speeches and invited expert talks, she has also been invited as a visiting lecturer/foreign scientist at the M. Auzeov South Kazakhstan State University, Kazakhstan in 2014, 2019 and 2020.

NORMALIA SAMIAN (Member, IEEE) is a Senior Lecturer at the Faculty of Computer Science and Information Technology, Universiti Putra Malaysia (UPM). She received her Ph.D. degree from Universiti Putra Malaysia (UPM) in 2017 in the area of cooperation in wireless multihop networks. During her Ph.D. candidature, she has been awarded the N2 Women Young Researcher Fellowship at IEEE LCN2016 in Dubai. Her main research interests include ad hoc networks security, cooperation, and trust management in wireless networks, the Internet of Things (IoT), and blockchain technology. She is now leading a grant project on securing IoT networks using blockchain technology. She has published several impact factors journals and tier-A conferences related to her fields and has served as a reviewer/technical program committee in international journals/conferences. Currently, she is the leader of the Wireless, Mobile, and Quantum Computing (WiMoQ) research group and also the head of the academic advisor in her department.