Multi-Stage Enhanced Zero Trust Intrusion Detection System for Unknown Attack Detection in Internet of Things and Traditional Networks

MALEK AL-ZEWAIRI*, Department of Computer Science, Princess Sumaya University for Technology, Amman, Jordan
SUFYAN ALMAJALI*, Department of Computer Science, Princess Sumaya University for Technology, Amman, Jordan
MOUSSA AYYASH†, Department of Computing, Information, and Mathematical Sciences and Technology, Chicago State University, Chicago, United States
MOHAMED RAHOUTI‡, Computer and Information Science, Fordham University, New York, United States
FERNANDO MARTINEZ‡, Computer and Information Science, Fordham University, New York, United States
NORDINE QUADAR§, Royal Military College of Canada, Kingston, Canada

Detecting unknown cyberattacks remains an open research problem and a significant challenge for the research community and the security industry. This paper tackles the detection of unknown cybersecurity attacks in the Internet of Things (IoT) and traditional networks by categorizing them into two types: entirely new classes of unknown attacks (type-A) and unknown attacks within already known classes (type-B). To address this, we propose a novel multi-stage, multi-layer zero trust architecture for an intrusion detection system (IDS), uniquely designed to handle these attack types. The architecture employs a hybrid methodology that combines two supervised and one unsupervised learning stages in a funnel-like design, significantly advancing current detection capabilities. A key innovation is the layered filtering mechanism, leveraging type-A and type-B attack concepts to systematically classify traffic as malicious unless proven otherwise. Using four benchmark datasets, the proposed system demonstrates significant improvements in accuracy, recall, and error classification rates for unknown attacks, achieving an average accuracy and recall ranging between 88% and 95%. This work offers a robust, scalable framework for enhancing cybersecurity in diverse network environments.

CCS Concepts: • **Networks** → **Network monitoring**; • **Security and privacy** → **Network security**; **Intrusion detection systems**; • **Computing methodologies** → *Machine learning*.

Additional Key Words and Phrases: Unknown attacks, network anomaly, intrusion detection system, deep learning, IoT.

Authors' Contact Information: Malek Al-Zewairi, Department of Computer Science, Princess Sumaya University for Technology, Amman, Jordan; e-mail: m.alzewairi@jisdf.org; Sufyan Almajali, Department of Computer Science, Princess Sumaya University for Technology, Amman, Jordan; e-mail: s.almajali@psut.edu.jo; Moussa Ayyash, Department of Computing, Information, and Mathematical Sciences and Technology, Chicago State University, Chicago, Illinois, United States; e-mail: msma@ieee.org; Mohamed Rahouti, Computer and Information Science, Fordham University, New York, New York, United States; e-mail: mrahouti@fordham.edu; Fernando Martinez, Computer and Information Science, Fordham University, New York, New York, United States; e-mail: fmartinezlopez@fordham.edu; Nordine Quadar, Royal Military College of Canada, Kingston, Ontario, Canada; e-mail: quadar@rmc.ca.

ACM 2471-2574/2025/3-ART
https://doi.org/10.1145/3725216

1 Introduction

The tremendous advancement in information technology had significant effects on the current and future of cybersecurity. The immense number of Internet of Things (IoT) devices with unpatched security vulnerabilities makes them susceptible to massive security attacks. Moreover, advancements in machine learning and artificial intelligence have been widely utilized in the cybersecurity domain, including but not limited to intrusion detection system (IDS) solutions. With the large training datasets of modern traffic, intelligent algorithms, and powerful machine learning tools, security researchers have greatly improved the intrusion detection models and enhanced their ability to detect malicious traffic more accurately [17].

Over the past year, there has been a significant surge in cyber threats that has raised alarm bells across the cybersecurity community. Recent data from cyber security experts, Sonicware, reveals that in 2022 alone, there were a staggering 2.8 billion malware attacks, averaging around 8,240 attempts per customer [22]. This surge represents a significant increase from the previous year, where malware attacks hit a seven-year low in 2021. This surge in malware activity can largely be attributed to the proliferation of cryptojacking and IoT attacks, which saw increases of 30% and 77%, respectively, throughout the year. In fact, there were a total of 57 million IoT malware attacks only in 2022. What's even more concerning is that there were over 270,228 malware variants detected in 2022 that had never been seen before. This surge in newly developed attacks poses an eccentric challenge for intrusion detection systems, including next-generation solutions. Further, the cybersecurity landscape of 2023 was profoundly impacted by a notable escalation in covert operations and the advanced capabilities of cyber adversaries, as reported in the CrowdStrike 2024 Global Threat Report [1]. Remarkably, 75% of cyber attacks were executed without the use of malware, indicating a substantial shift in tactics compared to the previous two years. This evolution in cyber threat methodologies is attributed primarily to geopolitical tensions, underscoring a strategic pivot towards more stealthy and sophisticated techniques of compromise.

Unknown cyberattacks are becoming more prevalent than ever, consistently ranking as the second cause of attacks for the past two years and the number one source of attacks in this decade [31]. Detecting unknown attacks against computer networks and systems has been identified as a challenging research problem by the security research community [13]. Furthermore, unknown attacks constitute a significant percentage of the total cybersecurity attacks ranking as the second top ten attacks. During 2022, the number of unknown attacks has increased to 22.2% compared to 18.7% in 2021 [30], being second to malwares as the cause of cybersecurity attacks. The lack of a standard scientific definition for unknown attacks and the inability of modern machine learning models to detect all types of unknown attacks were identified as two major issues.

The research community has been relying on inconsistent definitions for unknown attacks where unseen instances of known classes (in the data science terminology) are often incorrectly referred to as unknown attacks when used in testing models. The researchers addressed this issue in a previous study [5], where a proper definition for unknown attacks was proposed, evaluated, and validated. The proposed definition distinguished between "completely never seen before" attacks and referred to them as type-A unknown attacks and new subtypes of "previously known" attacks and referred to those as type-B unknown attacks.

Nonetheless, under the proposed definition, researchers studied the ability of shallow and deep supervised classifiers in detecting unknown attacks against traditional networks and IoT devices. It was concluded that the classification error rate of deep and shallow machine learning models is too high to be acceptable in detecting various types of unknown attacks.

In this study, the authors build upon previous studies [4–6, 34] to solve the problem of detecting unknown security attacks and propose a multi-stage multi-layer zero trust architecture. The performance of the proposed architecture is then evaluated using four benchmarking datasets (i.e., CIC-IDS-2017 [38], CIC-IDS-2018 [39], Bot-IoT [23], and IoT-23 [29]). The datasets were selected as two IoT datasets and two traditional network datasets.

The proposed architecture distinguishes itself from prior works through its innovative multi-stage, multi-layer zero trust approach tailored to address the challenges of unknown cyberattacks, categorized into type-A (entirely new) and type-B (new within known classes). This architecture employs a hybrid methodology combining both supervised and unsupervised learning techniques across multiple layers, thereby enhancing the detection capabilities beyond what has been achieved in existing models. Specifically, it introduces a novel funneling process where traffic is subjected to multiple layers of scrutiny, significantly improving upon the precision, recall, and error classification rates for detecting unknown attacks, with results showing overall average accuracy and recall between 88% and 95% across four benchmark datasets.

This research not only advances the theoretical framework for intrusion detection systems (IDS) but also demonstrates practical applicability across a range of network environments, including IoT devices and traditional networks, offering a more robust solution to the ever-evolving threat landscape. Four benchmarking datasets (i.e., CIC-IDS-2017, CIC-IDS-2018, Bot-IoT, and IoT-23) are used to evaluate the performance of the proposed architecture. The key contributions of this work are summarized as follows:

- Multi-stage multi-layer zero trust architecture: The paper introduces a novel architecture for an IDS. This architecture assumes all network traffic is malicious by default unless proven otherwise, aligning with the zero-trust security model. It incorporates a multi-layer filtering funnel, allowing traffic to pass through multiple filters optimized to detect unknown attacks effectively.
- Significant performance enhancement: Demonstrates significant accuracy improvements in detecting unknown attacks using four benchmark datasets.
- Application across diverse network environments: The research extends the applicability of the proposed architecture to a wide range of network environments, including both traditional networks and IoT devices.

The rest of this article is structured as follows. Section 2 discusses the related work on unknown attacks. Section 3 presents the proposed architecture. This section first presents the multi-stage multi-layer zero trust architecture, then describes the datasets selection and processing steps. Section 4 highlights the evaluation process and provides the performance analysis results. Next, Section 5 discusses the complexity analysis of the proposed architecture, focusing on its space and time requirements, scalability, and performance in real-world scenarios. Finally, the concluding remarks of the paper and the future work are outlined in Section 6.

2 Related Work

This section highlights the studies that aimed to address the issue of detecting modern unknown network intrusions in the past few years, as well as several recently proposed novel anomaly-based intrusion detection models.

2.1 Known Attacks

Ajjouri et al. [3] proposed a distributed hierarchical agent-based IDS architecture, utilizing case-based reasoning to learn new patterns of security attacks from similarities with known attacks. However, without empirical or formal evaluation, the effectiveness of this architecture remains unverified. Sellami et al. [37] introduced a cloud, agent, and anomaly-based IDS, pushing an agent to devices attempting to connect to a Cloud Service Provider. This model, which triggers an anomaly event based on a predefined threshold, closely resembles traditional agent-anomaly-based IDS but was presented without empirical validation. Khraisat et al. [20] developed a hybrid two-layer model that combines signature-based IDS with anomaly-based IDS aimed at detecting both known and unknown attacks, blending established and novel detection methodologies.

Further, efforts to secure networks against known attacks are also a critical component of contemporary cybersecurity strategies. The intelligent zero trust framework proposed by Guo et al. [16] for software defined networking (SDN) integrates zero Ttust principles with the programmability of SDN to efficiently manage and

mitigate known vulnerabilities and threats. This framework, along with the other discussed works, leverages the flexibility and dynamic control offered by modern network technologies to implement robust defenses against identified threats. By combining proactive measures against unknown attacks with strong protections against known vulnerabilities, these research efforts contribute to the development of more secure and resilient network environments.

2.2 Unknown Attacks

Kukiełka and Kotulski [24] presented a supervised learning-based IDS classifier to detect new attack types; however, it requires retraining to recognize manually added new attack variants. Bao et al. [11] proposed a general reasoning methodology to infer new attacks, lacking practical evaluation and a clear definition of "new attacks." Ahmad et al. [2] explored the efficiency of using ML algorithms (SVM, RF, ELM) on large datasets for training IDS classifiers, suggesting that ELM (Extreme Learning Machine) is preferable for large data volumes. Santikellur et al. [34] proposed a multi-layer network-based IDS, using machine learning models optimized with evolutionary computing algorithms, yet did not evaluate the model's performance against truly unknown attacks. Qureshi et al. introduced both a Random Neural Network (RNN) classifier [43] and a hybrid model combining RNN with the Approximate Bayesian Computation (ABC) algorithm [33] for anomaly detection, showing promising results in precision and accuracy.

Meira et al. [27] conducted an experimental study into the detection of unknown attacks using unsupervised learning techniques, finding that these algorithms often misclassified attacks due to similarities with normal traffic. Amato et al. [9] explored the use of an Multi-Layer Perceptron (MLP) classifier for identifying new attack types without substantial experimental validation. Aljawarneh et al. [7] proposed a hybrid IDS using various classifiers in a voting-based ensemble model, showing improved detection accuracy. Khare et al. [19] introduced a hybrid model combining SMO algorithm with DNN, achieving significant improvements in detection rates. Tang et al. [42] developed a deep learning IDS for SDN environments, demonstrating superior performance compared to traditional classifiers.

Kim et al. [21] and Jo et al. [18] explored image-based deep learning models for IDS, converting packet-based datasets into image formats for evaluation. Bhavsar et al. [12] developed the PCC-CNN model, combining linear-based feature extraction with CNNs, showing high detection accuracy. Sharma et al. [40] introduced a DNN model for IoT networks, using GANs to address class imbalance, resulting in high accuracy. Xing et al. [45] proposed DUA-IDS, a dynamic IDS for detecting unknown attacks, incorporating a feature extraction module that combines a transformer encoder and CNN; while facing challenges with integrating detected unknown attack categories.

Moreover, the cutting-edge research in zero trust architecture (ZTA) and network security demonstrates a significant emphasis on addressing unknown attacks; particularly through the use of advanced technologies such as AI and ML. Sedjelmaci and Ansari's framework [36] for 6G edge computing and Sharma et al.'s distributed intrusion detection system for zero trust multi-access edge computing [41] highlight the move towards adaptive security measures that dynamically respond to emerging threats. The use of AI and ML, as discussed by Yang et al. [47] in the context of AutoML for zero-touch network security, further illustrates a focus on developing systems capable of detecting and mitigating previously unseen attacks. These approaches embody the principle of continuous verification inherent to zero trust, aiming to protect networks against the constantly evolving landscape of cyber threats.

As discussed earlier, recent advancements in IDS highlight a shift towards methodologies that are not only rigorously evaluated but also broad in their approach to detecting unknown threats, addressing past criticisms of insufficient validation. The introduction of AI and ML-based solutions by Yang et al. [47] marks a significant step forward in identifying new threats through adaptive learning. Similarly, the focus on empirical validation, as

seen in Sharma et al.'s [41] work on distributed intrusion detection for zero trust multi-access edge computing, and the emphasis on adaptability in IDS, exemplified by Sedjelmaci and Ansari [36] in their study on ZTA for 6G edge computing, underscore the importance of flexible and dynamic IDS solutions. This evolution towards more validated, adaptable, and comprehensive security methodologies reflects our motivation for creating more secure and resilient network environments amidst the continuously evolving landscape of cyber threats.

In conclusion, as evidenced in Table 1, a comprehensive review of the literature reveals that remarkable efforts have been made to develop methodologies for detecting unknown attacks [9–11, 18, 20, 26, 27, 34, 35, 42, 44, 46]. However, these approaches frequently suffer from shortcomings such as inadequate analytical rigor and evaluation [9, 11] or an imprecise conceptualization of what constitutes an unknown attack [10, 26, 34, 44]. The persistent absence of a definitive solution underscores the critical need for continued research in this field. To bridge this gap, our study introduces a novel multi-stage Zero Trust IDS that innovatively integrates supervised and unsupervised learning techniques within a funnel-like framework, specifically tailored to identify and mitigate both known and unknown types of attacks.

Table 1. Summary of the related work on unknown attacks.

Ref.	Proposed Solution	Testing Attacks Type	Datasets	Validation Techniques	Limitations
[44]	DNN-based IDS	20 types of attacks	Developed benchmark NetFlow-based dataset	Batch normalization and dropout layers	Scalability
[11]	Forward deduction method to infer new attacks	-	-	-	Lacks rigours evaluation
[3]	Case-based reasoning technique	-	-	-	Lacks rigours evaluation
[37]	Cloud agent-based NADS	DoS	-	Manual simulation	Not truly new attacks
[27]	One-class classification model: Autoencoder ANN K-means Nearest neighbour Isolation forest	Several	NSL-KDD ISCX-IDS	ML testing	Poor precision (high FAR)
[9]	MLP classifier	DoS Privilege escalation Probing	KDD99	ML testing	Not truly new attacks
[34]	Multi-layer IDS: AdaBoost ANN Naive Bayes Decision tree	Several	CIC-IDS-2017	ML testing	Not truly new attacks
[20]	Hybrid two-layer IDS: C5 decision tree and SVM	Several	NSL-KDD ADFA	ML testing	The results did not clearly highlight the model performance in detecting unknown attacks
[42]	DNN and GRU-RNN classifiers	Several/SDN	NSL-KDD	ML testing	The dataset used is not built for SDN

3 Proposed Architecture

Focusing on the distinct characteristics of IoT and traditional network traffic necessitates a deep understanding of the inherent differences between these environments, which directly impacts the development and performance of intrusion detection systems. IoT environments are typically defined by a vast number of devices with limited processing power and energy constraints, often communicating over wireless networks using specific protocols

like MQTT or CoAP, tailored for low power and bandwidth usage. This results in unique traffic patterns, such as periodic data transmissions from sensors or devices, and necessitates lightweight security solutions that are efficient in processing and energy usage [8].

On the other hand, traditional network environments, comprised of desktops, servers, and networking hardware, handle more complex and higher-volume traffic patterns, utilizing a broader range of protocols (e.g., HTTP, FTP, SMTP). These environments can support more computationally intensive security measures and face a different set of vulnerabilities and attack vectors.

Thus, the proposed architecture in this paper is meticulously designed to address these variations, implementing multi-stage, multi-layer detection strategies that adapt to the specific requirements and constraints of each environment. By leveraging advanced machine learning algorithms and adopting a zero-trust security model, the system effectively distinguishes between benign and malicious traffic in both IoT and traditional networks, ensuring high accuracy and recall rates. The proposed multi-stage multi-layer zero trust architecture for detecting unknown security attacks is defined and detailed next.

3.1 Multi-stage Multi-layer Zero Trust Architecture

This subsection describes the proposed multi-stage multi-layer zero trust architecture for detecting unknown attacks. Essentially, the proposed architecture assumes that all traffic is malicious unless proven otherwise (named zero trust). It acts as a multi-layer filtering funnel, where traffic passes through multiple filters designed to root out all malicious traffic using a combination of supervised and unsupervised machine learning algorithms logically organised to take advantage of the concept of type-A and type-B unknown attacks.

In the heart of our proposed IDS architecture lies a sophisticated multi-stage, multi-layer zero trust model, meticulously crafted to address the nuanced and evolving landscape of cyber threats in both IoT and traditional network environments. Fig. 1 provides a visual representation of this architecture, illustrating its foundational principles. The three stages of the proposed architecture are outlined as follows:

- The first stage consists of two layers of shallow deep learning classifiers designed to identify the primary type of attacks (i.e., type-A).
- The second stage includes two layers of deep learning models designed to identify the subtype of attacks (i.e., type-B).
- The third stage utilises an unsupervised clustering algorithm, namely DBSCAN [14], to distinguish benign from unknown traffic.

Two essential security principles were utilized in designing the proposed architecture: the Defense-in-Depth principle and the Closed Security model. The Defense-in-Depth principle is a cornerstone of information security, employing a multilayered approach to mitigate risks. Each layer is specifically tailored to address distinct threats. Consequently, the proposed architecture is structured into three aforementioned stages, each incorporating multiple ML layers, thereby demonstrating the defense-in-depth principle. The Closed Security model, on the other hand, is based on the Deny-by-Default principle, where actions are denied implicitly unless explicitly permitted. This model is implemented in the proposed architecture by assuming all network traffic is malicious until it successfully passes through all three stages. This approach embodies the zero-trust philosophy.

At each stage, if all ML layers classify an instance as malicious, the processing halts, and the instance is definitively labeled as malicious. Conversely, if the instance is not unanimously classified as malicious, it proceeds to the next stage for further analysis. This staged approach is designed to optimize overall processing time while simultaneously minimizing the false-positive rate.

Further, the paradigm shift provided by the proposed architecture is operationalized through a funnel-like process, where traffic undergoes scrutiny across multiple layers, each designed with specific filters that leverage a blend of supervised and unsupervised learning techniques. The initial layers employ shallow and deep learning

Algorithm 1 Generate testing, training, and validation sets for unknown attacks [5].

Input: D, *dataset of size* $(m \times n)$
Input: $C \leftarrow$ *[], set of size l; Type-A or Type-B*
Output: $l \times$*Testing subsets of size* $(m - 1 \times n_1)$
Output: $l \times$*Training subsets of size* $(m - 1 \times n_2)$
Ensure: $n_2 = (n - n_1) * 0.9$ **Output:** $l \times$*Validation subsets of size* $(m - 1 \times n_3)$
Ensure: $n_3 = (n - n_1) * 0.1$

for all $c \in C$ **do**
  $D_{temp} \leftarrow D$
  $D_{test^c} \leftarrow []$
  $D_{train^{C-c}} \leftarrow []$
  $D_{val^{C-c}} \leftarrow []$
  for all $d \in D$ **do**
    if Type-A **and** d.class == c **then**
      D_{test^c} += d
      D_{temp} -= d
    end if
    if Type-B **and** d.subclass == c **then**
      D_{test^c} += d
      D_{temp} -= d
    end if
    if BinaryClass **then**
      del *d.class*
      del *d.subclass*
    end if
    if MultiClass **and** Type-A **then**
      del *d.binary*
      del *d.subclass*
    end if
    if MultiClass **and** Type-B **then**
      del *d.binary*
      del *d.class*
    end if
  end for
  $D_{train^{C-c}}, D_{val^{C-c}}$ = *StratifiedRandomSplit*(D_{temp}, *0.9, 1586512076128*)
end for
Function StratifiedRandomSplit(l, r, s):
  return r, $(1 - r)$ *from l using stratified random sampling with seed s*
End Function

classifiers to sift through traffic, segregating benign from malicious activities based on pre-defined characteristics of known threats. Subsequently, an unsupervised learning stage employs clustering techniques to identify patterns that deviate from the norm, flagging them as potential unknown threats.

The hybrid combination between the supervised and unsupervised machine learning algorithms and the hierarchical multi-layer design allows the architecture to identify benign, known, and unknown attacks with notable recall and meagre classification error rate. Theoretically, all known attacks should be captured in the first two stages. While the first two stages are expected to catch some unknown attacks, the rest should be recognised as noises by the unsupervised clustering algorithm.

The first step in testing the proposed architecture is to create the testing sets of the different unknown attacks, represented by DS_t, using the same procedure defined in our previous work [5] and is shown in Algorithm 1.

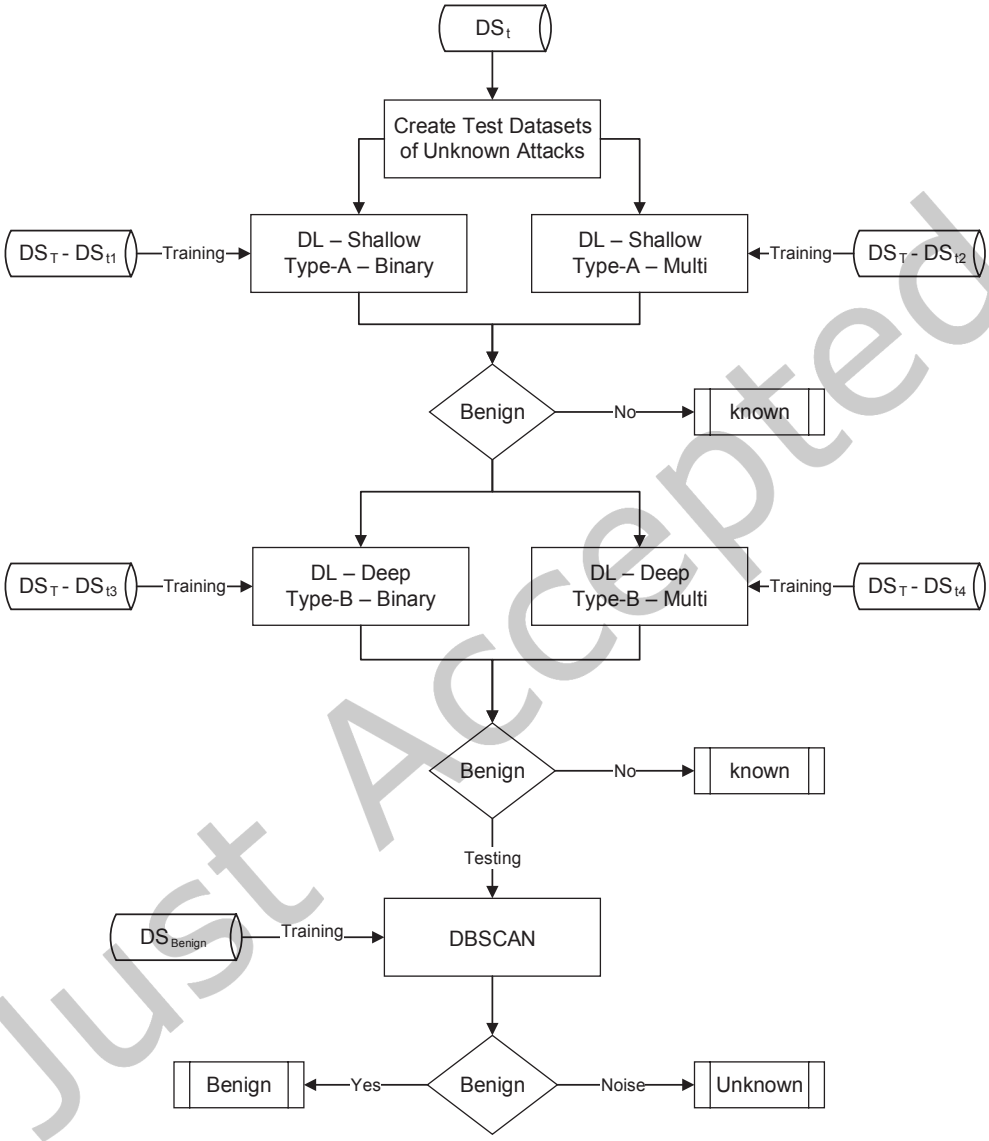

Fig. 1. The proposed multi-stage, multi-layer zero trust architecture.

An integral part of our study involves the rigorous evaluation of our IDS architecture against unknown attack types. This necessitates the creation of testing sets that accurately represent these unknown attacks; a process for which we have drawn upon methodologies proposed in prior works, notably [5]. The creation of these testing sets is pivotal to our research, as it allows for the simulation of real-world scenarios where an IDS must detect

novel attack patterns not present in the training data. The methodology involves categorizing attacks into two types: type-A (entirely new attack classes) and type-B (new instances within known classes). For type-A attacks, we remove all instances of a given attack category from the training dataset, ensuring that the system has no prior knowledge of these attacks. These instances are then reintroduced in the testing phase to simulate the scenario of encountering completely new attack vectors. Conversely, for type-B attacks, we exclude all instances of a specific attack subtype, following a similar rationale to ensure that the IDS faces genuinely unknown patterns during testing.

The dataset used for this procedure, delineated by Algorithm 1 in [5], undergoes a stratified split to ensure a balanced representation of various attack types, maintaining the integrity and challenge of the testing environment. This methodological approach, adopted from [5], has been modified in this study to adapt to the nuances of our proposed IDS architecture and the specific datasets employed. It underscores our commitment to evaluating our system's robustness in identifying and mitigating unknown cyber threats effectively.

All instances of an attack category are removed from the training dataset and reintroduced in the testing dataset to simulate type-A unknown attacks. Similarly, to simulate type-B unknown attacks, all instances of an attack subcategory are deleted in a similar approach. DS_T denotes the entire dataset, a specific type of an unknown attack is denoted by DS_{tn}, which is a subset of DS_t, and the combination of training/validation datasets are both denoted by $DS_T - DS_{tn}$. DS_{Benign} denotes a dataset which only contains benign instances.

To clarify the process of simulating an unknown attack, the first step involves identifying the class labels in the dataset that represent attack categories (e.g., DoS). These class labels are referred to as type-A attacks. To simulate a new category of unknown attack, all instances with the class label corresponding to DoS are removed from the training and validation datasets and then reintroduced in the testing dataset. Therefore, the ML model will have no prior knowledge of DoS attacks. Similarly, unknown attacks that are subtypes of known attacks are simulated using a similar approach. In this case, the target is the class label representing a subcategory of attacks (e.g., a UDP DoS attack). All instances in the dataset with the UDP DoS subclass are removed from the training and validation datasets and subsequently reintroduced in the testing dataset, thereby simulating type-B unknown attacks. Therefore, the ML model will have prior knowledge of other types of DoS attacks (e.g., TCP) but not the UDP type. By iteratively applying this process to all classes and subclasses in the dataset, we successfully simulated and evaluated the performance of the proposed architecture in detecting various types and subtypes of unknown attacks, as defined by type-A and type-B unknown attacks. This approach enables the detection of unknown attacks with varying levels of complexity. Unknown attacks that share common features with known attacks are identified in the first two stages by being classified as the most similar known attack. In contrast, attacks that do not resemble any known attacks are addressed in the third stage. This process of simulating unknown attacks establishes the foundation for evaluating the performance of deep learning models. Following dataset preparation, the next step involves configuring and optimizing these models for detailed analysis.

Once the datasets are created, the deep learning models are configured. In the literature, a **shallow model** is an artificial neural network (ANN) model with one or two hidden layers, whereas a **deep model** consists of more than two hidden layers [28, 32]. Several studies have examined the effectiveness of deep learning models in detecting cyberattacks within IDS [4, 13, 25, 32, 34]. In our implementation, a deep model consists of five hidden layers with ten neurons per layer, while a shallow model includes a single hidden layer with the same number of neurons. In this context, the number of hidden layers is the only distinction between shallow and deep models. Several experiments were conducted to evaluate the effectiveness of shallow and deep models in detecting unknown attacks under the proposed categorization [4–6, 34]. The experiments were divided into two main parts. The first part focused on assessing the ability of shallow and deep multilayer feedforward ANN models to detect type-A and type-B unknown attacks using binary and multiclass classification. The second part extended this evaluation to optimize the values for each hyperparameter. The findings generally showed that ANN models outperform other ML models (i.e., SVM, NB and RF) in detecting cyberattacks and deep ANN

models have been shown to outperform shallow ANN models in detecting unknown attacks but with notable computational complexity. Therefore, they have been used for multiclass classification, while shallow models were reserved for binary classification.

After configuring the deep learning models, clustering techniques are employed to enhance the architecture's ability to detect and classify patterns within the datasets. Among these techniques, the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) stands out as a critical component of the proposed framework.

DBSCAN [15] is a density-based clustering algorithm that groups n points into k clusters based on the change of density between regions from high to low. DBSCAN requires two parameters, *epsilon* and the $number of minimum points$. Epsilon defines the radius around a point; while, minimum points define the minimum number of points within the epsilon. Unlike centre-based clustering algorithms such as K-means, DBSCAN does not require the number of clusters to be known ahead of time and can distinguish noise points that do not belong to any cluster making it suitable for the proposed architecture.

A sensitivity analysis test is performed next to determine the best value for *epsilon*, given that the minimum number of points is set to twice the number of features in the dataset. The results of the sensitivity analysis test are presented in Table 3 and Fig. 2, which indicated that the best value for *epsilon* is between 14 and 15 (it was chosen to be 14.5) and that the minimum number of points is twice the number of features. The test was run for more than 200 iterations, and with each iteration, the hyperparameters were increased until the value of completeness stabilized. This was used to tune the hyperparameters of the DBSCAN model as shown in Table 3, which shows the *epsilon* value, number of clusters, number of noise points, completeness value, and the running time in seconds.

At the first stage, the proposed architecture works by classifying each instance using two shallow deep learning models, a binary classifier and a multiclass classifier. The binary classifier determines whether an instance is benign or malicious, and the multiclass classifier identifies the known attack category. If both classifiers predict the class as malicious, then the instance will be labeled as an attack of the predicted category. Otherwise, the instance is passed to the next stage.

At the second stage, the remaining instances from stage one are inspected by two deep ANN classifiers, a binary classifier and a multiclass classifier. However, the multiclass model is trained to detect subtypes of known attacks. Similar to the first stage, when both models identify an instance as malicious, it will be labeled as an attack with the predicted subtype. However, if either classifier identifies the instance as benign, then it will be passed to the final stage.

At the third stage, the remaining instances are fed to the DBSCAN model, an unsupervised clustering algorithm that does not require the number of clusters to be specified. The model is trained using only benign traffic along with the data passed from the previous stages. At this stage, the instances which are clustered as noise points are considered unknown attacks; otherwise, they are labeled as benign.

The hyperparameter configurations of the four deep learning models utilized in the proposed architecture are detailed in Table 2. The rationale behind these configurations, as well as the selection of specific algorithms, is thoroughly explored through experimentation and sensitivity analysis in [4–6, 34]. In the proposed architecture, an instance is classified as benign only if at least one classifier at each stage identifies it as benign, thereby adhering to the principles of zero trust. This means that if there is any doubt regarding the benign nature of an instance, it will be treated as malicious and forwarded to the next inspection stage.

3.2 Dataset Selection and Preprocessing

In order to evaluate the performance of the proposed architecture, four benchmarking datasets containing modern realistic traffic for IoT networks and traditional computer networks were selected (i.e., CIC-IDS-2017 [38], CIC-IDS-2018 [39], Bot-IoT [23], and IoT-23 [29]). The datasets were primarily selected as two IoT datasets and

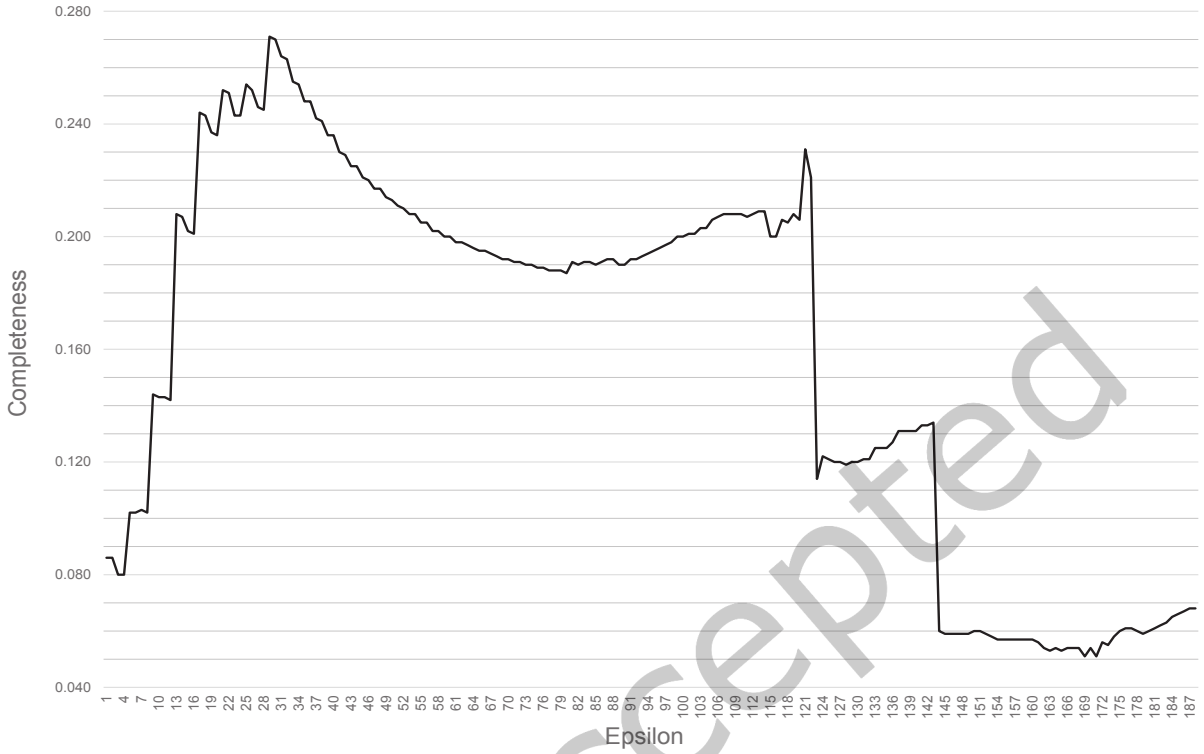

Fig. 2. Sensitivity analysis of the Completeness value of the DBSCAN.

Table 2. Proposed model hyperparameters settings.

Hyperparameter	**Value**
Activation function	Rectifier
Batch size	32
Dropout	Without dropout
Epochs	10
Fold assignment	Stratified assignment
K-fold	10
Learning rate	0.01
Prediction threshold	0.5078
Runs	5
Seed	1,586,512,076,128

two traditional network datasets. These datasets used for evaluation—CIC-IDS-2017, CIC-IDS-2018, Bot-IoT, and IoT-23—are independently developed and widely recognized in the cybersecurity domain. Each dataset captures unique network environments, attack scenarios, and traffic patterns, ensuring a robust and unbiased evaluation of the proposed architecture across diverse contexts.

Table 3. Sensitivity analysis of the DBSCAN hyperparameters; NP and Comp denote noise points and completeness, respectively.

Epsilon	Clusters	NP	Comp	Time (S)	Epsilon	Clusters	NP	Comp	Time (S)	Epsilon	Clusters	NP	Comp	Time (S)
0.5	11,013	54,077	0.086	15.832	35.5	3,422	13,373	0.191	16.671	270.0	299	2,334	0.133	21.500
1.0	11,014	54,051	0.086	18.561	36.0	3,426	13,326	0.191	16.808	280.0	297	2,305	0.133	21.275
1.5	10,855	51,823	0.080	16.466	36.5	3,409	13,050	0.190	16.986	290.0	286	2,263	0.134	28.690
2.0	10,867	51,758	0.080	15.928	37.0	3,419	12,985	0.190	19.432	300.0	283	2,221	0.060	28.813
2.5	2,774	34,159	0.102	18.601	37.5	3,401	12,763	0.189	19.863	310.0	274	2,192	0.059	28.700
3.0	2,808	34,049	0.102	18.544	38.0	3,406	12,709	0.189	16.940	320.0	267	2,159	0.059	21.866
3.5	2,847	33,726	0.103	18.306	38.5	3,352	12,515	0.188	20.836	330.0	261	2,137	0.059	21.889
4.0	2,893	33,575	0.102	20.152	39.0	3,347	12,457	0.188	21.124	340.0	259	2,104	0.059	22.150
4.5	1,198	32,069	0.144	16.528	39.5	3,297	12,237	0.188	20.252	350.0	253	2,071	0.059	29.785
5.0	1,238	31,935	0.143	18.828	40.0	3,301	12,185	0.187	16.704	360.0	247	2,037	0.060	22.157
5.5	1,370	31,475	0.143	18.887	40.5	1,707	9,695	0.191	19.701	370.0	246	2,010	0.060	22.413
6.0	1,407	31,350	0.142	18.960	41.0	1,565	9,217	0.190	20.350	380.0	243	1,999	0.059	30.464
6.5	1,094	30,552	0.208	19.040	41.5	1,423	8,832	0.191	20.040	390.0	234	1,984	0.058	31.071
7.0	1,129	30,431	0.207	18.972	42.0	1,360	8,571	0.191	19.757	400.0	231	1,963	0.057	25.055
7.5	1,295	29,834	0.202	17.143	42.5	1,237	8,254	0.190	17.717	410.0	232	1,927	0.057	29.347
8.0	1,335	29,693	0.201	18.986	43.0	1,217	8,092	0.191	20.415	420.0	236	1,891	0.057	42.486
8.5	1,414	28,896	0.244	18.945	43.5	1,144	7,917	0.192	21.775	430.0	234	1,869	0.057	39.286
9.0	1,441	28,793	0.243	19.061	44.0	1,128	7,762	0.192	21.652	440.0	233	1,844	0.057	41.598
9.5	1,583	28,230	0.237	17.180	44.5	1,083	7,612	0.190	26.566	450.0	232	1,831	0.057	50.597
10.0	1,606	28,131	0.236	19.289	45.0	1,064	7,523	0.190	22.163	460.0	231	1,790	0.057	50.519
10.5	1,743	27,403	0.252	19.229	45.5	1,023	7,444	0.192	19.346	470.0	221	1,782	0.056	51.115
11.0	1,765	27,316	0.251	17.917	46.0	1,009	7,382	0.192	20.803	480.0	220	1,755	0.054	52.063
11.5	1,936	26,624	0.243	19.591	46.5	978	7,295	0.193	20.944	490.0	215	1,735	0.053	51.924
12.0	1,950	26,557	0.243	16.467	47.0	963	7,233	0.194	20.497	500.0	213	1,713	0.054	52.660
12.5	2,093	25,819	0.254	19.187	47.5	944	7,146	0.195	20.476	525.0	206	1,660	0.053	50.859
13.0	2,124	25,699	0.252	19.380	48.0	928	7,068	0.196	20.326	550.0	194	1,613	0.054	51.589
13.5	2,248	25,111	0.246	19.427	48.5	901	7,002	0.197	20.854	575.0	192	1,579	0.054	52.700
14.0	2,278	24,994	0.245	19.495	49.0	887	6,959	0.198	21.092	600.0	183	1,547	0.054	53.012
14.5	2,414	24,271	0.271	20.187	49.5	868	6,890	0.200	20.971	625.0	168	1,480	0.051	57.014
15.0	2,436	24,186	0.270	19.324	50.0	877	6,842	0.200	23.165	650.0	183	1,547	0.054	47.085
15.5	2,555	23,608	0.264	19.487	50.5	849	6,784	0.201	19.087	675.0	168	1,480	0.051	49.060
16.0	2,575	23,534	0.263	19.558	51.0	841	6,733	0.201	19.202	700.0	155	1,437	0.056	50.487
16.5	2,697	22,804	0.255	20.128	51.5	827	6,683	0.203	17.870	750.0	136	1,384	0.055	48.367
17.0	2,715	22,731	0.254	19.468	52.0	818	6,622	0.203	18.130	800.0	129	1,336	0.058	50.999
17.5	2,807	22,169	0.248	19.764	52.5	805	6,593	0.206	21.372	850.0	128	1,276	0.060	50.787
18.0	2,826	22,095	0.248	19.492	53.0	801	6,545	0.207	21.683	900.0	121	1,262	0.061	51.073
18.5	2,914	21,488	0.242	16.911	53.5	789	6,502	0.208	18.601	950.0	116	1,240	0.061	51.091
19.0	2,939	21,401	0.241	16.417	54.0	779	6,452	0.208	21.143	1000.0	116	1,218	0.060	54.537
19.5	3,020	20,860	0.236	16.661	54.5	772	6,428	0.208	20.499	1100.0	101	1,185	0.059	58.202
20.0	3,032	20,795	0.236	16.362	55.0	769	6,396	0.208	20.607	1200.0	87	1,144	0.060	59.328
20.5	3,108	20,099	0.230	16.508	55.5	760	6,359	0.207	17.335	1300.0	83	1,126	0.061	59.556
21.0	3,129	19,993	0.229	16.625	56.0	762	6,301	0.208	20.499	1400.0	87	1,092	0.062	64.947
21.5	3,174	19,492	0.225	17.048	56.5	755	6,297	0.209	18.112	1500.0	86	1,059	0.063	68.294
22.0	3,185	19,427	0.225	16.534	57.0	758	6,257	0.209	21.416	1600.0	84	1,034	0.065	72.914
22.5	3,227	18,900	0.221	17.490	57.5	743	6,220	0.200	21.573	1700.0	83	1,013	0.066	69.710
23.0	3,242	18,832	0.220	16.771	58.0	744	6,184	0.200	21.264	1800.0	86	984	0.067	71.522
23.5	3,300	18,370	0.217	16.400	58.5	645	6,092	0.206	18.419	1900.0	83	974	0.068	65.172
24.0	3,307	18,292	0.217	16.463	59.0	620	6,005	0.205	18.154	2000.0	85	956	0.068	74.577
24.5	3,325	17,863	0.214	16.599	59.5	602	5,972	0.208	19.114	5000.0	26	824	0.091	103.156
25.0	3,340	17,792	0.213	16.734	60.0	601	5,934	0.206	21.757	5100.0	26	824	0.091	105.428
25.5	3,378	17,361	0.211	16.959	70.0	621	5,379	0.231	21.380	5200.0	26	822	0.091	107.709
26.0	3,398	17,270	0.210	16.601	80.0	651	4,948	0.221	17.189	5300.0	27	819	0.091	108.826
26.5	3,392	16,862	0.208	16.455	90.0	649	4,495	0.114	17.369	5400.0	27	819	0.091	110.364
27.0	3,405	16,797	0.208	17.138	100.0	660	4,186	0.122	17.486	5500.0	27	819	0.091	111.851
27.5	3,432	16,375	0.205	17.981	110.0	651	3,983	0.121	22.847	5600.0	27	819	0.091	112.494
28.0	3,445	16,312	0.205	17.339	120.0	645	3,784	0.120	22.373	5700.0	27	816	0.091	114.337
28.5	3,443	15,896	0.202	16.885	130.0	583	3,356	0.120	22.435	5800.0	26	814	0.093	111.099
29.0	3,456	15,816	0.202	16.901	140.0	575	3,146	0.119	22.769	5900.0	26	814	0.093	116.418
29.5	3,466	15,451	0.200	16.663	150.0	546	3,025	0.120	22.830	6000.0	26	814	0.093	120.242
30.0	3,479	15,374	0.200	16.798	160.0	532	2,908	0.120	22.946	7000.0	25	807	0.095	117.712
30.5	3,447	15,069	0.198	16.756	170.0	491	2,813	0.121	22.950	7100.0	25	807	0.095	125.098
31.0	3,460	15,005	0.198	16.792	180.0	473	2,763	0.121	23.322	7200.0	25	807	0.095	128.916
31.5	3,455	14,714	0.197	16.621	190.0	420	2,680	0.125	26.521	7300.0	25	807	0.095	129.226
32.0	3,471	14,650	0.196	17.192	200.0	400	2,623	0.125	26.631	7400.0	24	807	0.096	136.483
32.5	3,438	14,380	0.195	16.883	210.0	395	2,571	0.125	26.811	10000.0	20	789	0.086	142.496
33.0	3,449	14,310	0.195	16.701	220.0	378	2,523	0.127	26.928	15000.0	18	777	0.088	191.126
33.5	3,443	14,018	0.194	16.812	230.0	347	2,496	0.131	26.967	16000.0	18	777	0.088	212.326
34.0	3,456	13,934	0.193	16.865	240.0	337	2,455	0.131	27.366	17000.0	18	777	0.088	222.156
34.5	3,425	13,642	0.192	16.910	250.0	326	2,406	0.131	20.563 18000.0	18	777	0.088	237.136	
35.0	3,437	13,579	0.192	16.814	260.0	313	2,368	0.131	21.784	-	-	-	-	-

The dataset selection process and the preprocessing steps for the three datasets CIC-IDS-2017, Bot-IoT, IoT-23 have already been explained in previous research works in [4], [5] and [6], respectively. Similarly, the fourth

dataset, CIC-IDS-2018, underwent the same preprocessing steps. The CIC-IDS-2018 is a publicly available dataset consisting of ten different CSV files [39]. The dataset comprises 16, 232, 943 instances, including 11, 989, 837 unique instances, representing a significant enhancement over the CIC-IDS-2017 dataset.

The dataset contains 79 standard features extracted from data corresponding to ten days of network traffic captured in PCAP format using CICFlowMeter. Two features contain missing values (i.e., Flow Byts/s and Flow Pkts/s), where each feature has 95,760 missing values. Moreover, eight features were found empty (i.e., Bwd PSH Flags, Bwd URG Flags, Fwd Byts/b Avg, Fwd Pkts/b Avg, Fwd Blk Rate Avg, Bwd Byts/b Avg, Bwd Pkts/b Avg, and Bwd Blk Rate Avg). The 79 features were divided into 78 numerical features, one of which is a long number, 37 floating-point, 40 integers, and the remaining feature is a timestamp. Table 4 displays the statistical analysis of the dataset.

Table 4. Statistical analysis of the CIC-IDS-2018 dataset features.

#	Features	Type	μ	σ	σ^2	$\widetilde{\mu}_3$	#	Features	Type	μ	σ	σ^2	$\widetilde{\mu}_3$
1	Dst Port	Integer	9.16E+03	1.89E+04	3.58E+08	1.86E+00	40	Bwd Pkts/s	Integer	1.53E+04	9.24E+04	8.54E+09	9.07E+00
2	Protocol	Integer	8.75E+00	4.92E+00	2.42E+01	1.01E+00	41	Pkt Len Min	Integer	1.09E+01	2.26E+01	5.12E+02	6.98E+00
3	Timestamp	String	-	-	-	-	42	Pkt Len Max	Integer	3.89E+02	5.14E+02	2.64E+05	2.02E+00
4	Flow Duration	Long	1.18E+07	4.94E+08	2.44E+17	-1.43E+03	43	Pkt Len Mean	Double	7.76E+01	1.04E+02	1.08E+04	4.35E+00
5	Tot Fwd Pkts	Integer	2.35E+01	1.52E+03	2.31E+06	8.86E+01	44	Pkt Len Std	Double	1.21E+02	1.63E+02	2.64E+04	1.73E+00
6	Tot Bwd Pkts	Integer	6.31E+00	1.64E+02	2.69E+04	1.71E+02	45	Pkt Len Var	Double	4.12E+04	2.13E+05	4.54E+10	1.86E+03
7	TotLen Fwd Pkts	Integer	9.73E+02	6.22E+04	3.86E+09	8.37E+02	46	FIN Flag Cnt	Integer	4.80E-03	6.91E-02	4.77E-03	1.43E+01
8	TotLen Bwd Pkts	Double	4.73E+03	2.34E+05	5.50E+10	1.60E+02	47	SYN Flag Cnt	Integer	4.39E-02	2.05E-01	4.20E-02	4.45E+00
9	Fwd Pkt Len Max	Integer	2.01E+02	3.04E+02	9.21E+04	4.60E+00	48	RST Flag Cnt	Integer	1.87E-01	3.90E-01	1.52E-01	1.60E+00
10	Fwd Pkt Len Min	Integer	1.11E+01	2.42E+01	5.87E+02	1.14E+01	49	PSH Flag Cnt	Integer	3.92E-01	4.88E-01	2.38E-01	4.42E-01
11	Fwd Pkt Len Mean	Double	5.03E+01	6.05E+01	3.66E+03	7.06E+00	50	ACK Flag Cnt	Integer	3.32E-01	4.71E-01	2.22E-01	7.15E-01
12	Fwd Pkt Len Std	Double	7.08E+01	1.16E+02	1.36E+04	2.45E+00	51	URG Flag Cnt	Integer	4.17E-02	2.00E-01	4.00E-02	4.58E+00
13	Bwd Pkt Len Max	Integer	3.50E+02	4.97E+02	2.47E+05	1.67E+00	52	CWE Flag Count	Integer	1.64E-04	1.28E-02	1.64E-04	7.81E+01
14	Bwd Pkt Len Min	Integer	2.65E+01	5.10E+01	2.60E+03	2.11E+00	53	ECE Flag Cnt	Integer	1.87E-01	3.90E-01	1.52E+01	1.60E+00
15	Bwd Pkt Len Mean	Double	1.13E+02	1.64E+02	2.69E+04	4.32E+00	54	Down/Up Ratio	Integer	4.96E-01	9.97E-01	9.93E-01	1.28E+02
16	Bwd Pkt Len Std	Double	1.32E+02	2.04E+02	4.17E+04	1.50E+00	55	Pkt Size Avg	Double	8.95E+01	1.08E+02	1.17E+04	3.87E+00
17	Flow Byts/s	Double	2.57E+05	3.67E+06	1.35E+13	8.06E+01	56	Fwd Seg Size Avg	Double	5.03E+01	6.05E+01	3.66E+03	7.06E+00
18	Flow Pkts/s	Double	5.23E+04	2.65E+05	7.01E+10	6.80E+00	57	Bwd Seg Size Avg	Double	1.13E+02	1.64E+02	2.69E+04	4.32E+00
19	Flow IAT Mean	Double	3.34E+06	2.23E+08	4.97E+16	-3.38E+03	58	Fwd Byts/b Avg	Integer	-	-	-	-
20	Flow IAT Std	Double	1.28E+06	3.38E+08	1.14E+17	1.21E+03	59	Fwd Pkts/b Avg	Integer	-	-	-	-
21	Flow IAT Max	Double	6.60E+06	6.60E+08	4.35E+17	1.01E+03	60	Fwd Blk Rate Avg	Integer	-	-	-	-
22	Flow IAT Min	Double	2.32E+06	7.48E+08	5.59E+17	-1.17E+03	61	Bwd Byts/b Avg	Integer	-	-	-	-
23	Fwd IAT Tot	Double	1.15E+07	4.94E+08	2.44E+17	-1.43E+03	62	Bwd Pkts/b Avg	Integer	-	-	-	-
24	Fwd IAT Mean	Double	3.66E+06	2.23E+08	4.97E+16	-3.38E+03	63	Bwd Blk Rate Avg	Integer	-	-	-	-
25	Fwd IAT Std	Double	1.40E+06	3.38E+08	1.14E+17	1.21E+03	64	Subflow Fwd Pkts	Integer	2.35E+01	1.52E+03	2.31E+06	8.86E+01
26	Fwd IAT Max	Double	6.41E+06	6.60E+08	4.35E+17	1.01E+03	65	Subflow Fwd Byts	Integer	9.73E+02	6.22E+04	3.86E+09	8.37E+02
27	Fwd IAT Min	Double	2.41E+06	7.48E+08	5.59E+17	-1.17E+03	66	Subflow Bwd Pkts	Integer	6.31E+00	1.64E+02	2.69E+04	1.71E+02
28	Bwd IAT Tot	Double	7.60E+06	2.59E+07	6.69E+14	3.62E+00	67	Subflow Bwd Byts	Integer	4.73E+03	2.34E+05	5.50E+10	1.60E+02
29	Bwd IAT Mean	Double	8.24E+05	4.34E+06	1.88E+13	1.24E+01	68	Init Fwd Win Byts	Integer	8.79E+03	1.62E+04	2.64E+08	2.59E+00
30	Bwd IAT Std	Double	8.51E+05	3.38E+06	1.14E+13	6.22E+00	69	Init Bwd Win Byts	Integer	8.69E+03	2.06E+04	4.25E+08	2.13E+00
31	Bwd IAT Max	Double	2.61E+06	1.02E+07	1.05E+14	5.16E+00	70	Fwd Act Data Pkts	Integer	1.99E+01	1.52E+03	2.31E+06	8.88E+01
32	Bwd IAT Min	Double	2.91E+05	3.83E+06	1.46E+13	1.69E+01	71	Fwd Seg Size Min	Integer	1.80E+01	7.69E+00	5.92E+01	2.56E-01
33	Fwd PSH Flags	Integer	4.39E-02	2.05E-01	4.20E-02	4.45E+00	72	Active Mean	Double	1.73E+05	2.51E+06	6.28E+12	2.49E+01
34	Bwd PSH Flags	Integer	-	-	-	-	73	Active Std	Double	8.64E+04	1.51E+06	2.29E+12	2.59E+01
35	Fwd URG Flags	Integer	1.64E-04	1.28E-02	1.64E-04	7.81E+01	74	Active Max	Double	2.62E+05	3.32E+06	1.10E+13	2.03E+01
36	Bwd URG Flags	Integer	-	-	-	-	75	Active Min	Double	1.15E+05	2.11E+06	4.47E+12	3.43E+01
37	Fwd Header Len	Integer	2.58E+02	1.23E+04	1.50E+08	8.76E+01	76	Idle Mean	Double	5.02E+06	2.63E+08	6.93E+16	1.26E+03
38	Bwd Header Len	Integer	1.33E+02	3.27E+03	1.07E+07	1.70E+02	77	Idle Std	Double	2.87E+05	1.69E+08	2.86E+16	1.31E+03
39	Fwd Pkts/s	Double	3.67E+04	2.13E+05	4.55E+10	8.70E+00	78	Idle Max	Double	5.43E+06	6.25E+08	3.91E+17	1.33E+03
79	Idle Min	Double	4.69E+06	6.37E+07	4.05E+15	3.34E+03							

Originally, the dataset had a single multiclass label with six different attack categories constituting of 16.93% of the total samples and one class for benign traffic. Several preprocessing steps were taken to prepare the dataset. First, two new labels were generated to indicate the binary class label and the attack category, which each CSV file represents. The existed multiclass label was used as the subtype attack label. After that, nine features were eliminated: Timestamp, Bwd PSH Flags, Bwd URG Flags, Fwd Byts/b Avg, Fwd Pkts/b Avg, Fwd Blk Rate Avg, Bwd Byts/b Avg, Bwd Pkts/b Avg, and Bwd Blk Rate Avg. Then, all the ten CSV files were merged into a single file, and all repeated instances were removed, which reduced the dataset size by 26.14 times. Thereafter, the numerical

values were normalised to [0-1] interval using the Min-Max method. Tables 5, 6 and 7 show the different classes of the CIC-IDS-2018 dataset after the preprocessing stage.

Table 5. CIC-IDS-2018 dataset binary class distribution before & after preprocessing.

Class Label	Before		After	
	Instances	Ratio	Instances	Ratio
Benign	13,484,708	83.07%	10,637,200	88.72%
Malicious	2,748,235	16.93%	1,352,637	11.28%

Table 6. CIC-IDS-2018 dataset type-A class distribution before & after preprocessing.

Class Label	Before		After	
	Instances	Ratio	Instances	Ratio
Botnet	286,191	1.76%	144,535	1.21%
Brute Force	380,949	2.35%	94,102	0.78%
DoS	654,300	4.03%	196,568	1.64%
DDoS	1,263,933	7.79%	775,955	6.47%
Infiltration	161,934	1.00%	140,610	1.17%
Web	928	0.01%	867	0.01%

Table 7. CIC-IDS-2018 dataset type-B class distribution before & after preprocessing.

Class Label	Subclass	Before		After	
		Instances	Ratio	Instances	Ratio
Botnet		286,191	1.76%	144,535	1.205%
Brute Force	FTP	193,360	1.19%	54	0.0005%
	SSH	187,589	1.16%	94,048	0.784%
DDoS	HOIC	686,012	4.23%	198,861	1.659%
	LOIC-HTTP	576,191	3.55%	575,364	4.799%
	LOIC-UDP	1730	0.01%	1730	0.014%
DoS	GoldenEye	41,508	0.26%	41,406	0.345%
	Hulk	461,912	2.85%	145,199	1.211%
	SlowHTTPTest	139,890	0.86%	55	0.0005%
	Slowloris	10,990	0.07%	9908	0.083%
Infiltration		161,934	1.00%	140,610	1.173%
Web	Brute Force	611	0.004%	555	0.005%
	SQLi	87	0.001%	84	0.001%
	XSS	230	0.001%	228	0.002%

4 Performance Analysis

In this section, the results for evaluating the proposed architecture are presented and discussed for each dataset. The results are presented per the general class type (i.e., benign, known, and unknown) and the overall performance of the proposed architecture. Specifically, five generalisation error metrics are used to evaluate the performance of the proposed architecture, namely, the accuracy, precision, recall, F1-score, and classification error rate. The results are split into four subsections. Subsection 4.2 discusses the results of the unknown attacks, subsection 4.3 presents the results of the known instances, and subsection 4.4 presents the results for benign instances. Lastly in subsection 4.5, the overall results of the proposed architecture are discussed.

4.1 Evaluation Metrics

In this paper, the following evaluation metrics are utilized to assess the performance of the proposed architecture:

- **Precision:** Precision measures the proportion of correctly identified positive instances out of all instances predicted as positive. It is calculated as:

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}} \tag{1}$$

- **Recall (Sensitivity):** Recall assesses the proportion of actual positive instances that are correctly identified by the model. It is given by:

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \tag{2}$$

- **F1-Score:** The F1-score provides a harmonic mean of precision and recall, balancing their trade-offs. It is defined as:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

- **Accuracy:** Accuracy reflects the overall correctness of the model by measuring the proportion of correctly classified instances (both positive and negative) to the total instances. The formula is:

$$\text{Accuracy} = \frac{\text{True Positives (TP)} + \text{True Negatives (TN)}}{\text{Total Instances}} \tag{4}$$

- **Classification Error Rate:** This metric calculates the proportion of incorrect predictions to the total number of predictions. It is expressed as:

$$\text{Classification Error Rate} = 1 - \text{Accuracy} \tag{5}$$

4.2 Unknown Type

In this subsection, the results of testing the proposed architecture on unknown data are presented and discussed per dataset.

When examining the proposed architecture performance in detecting unknown attacks using the IoT-23 dataset shown in Table 8, the precision ranged between 0.00001 and 0.97 with a 0.15 average and a 0.28 standard deviation with six outliers between 0.39 and 0.97. The recall ranged between 0.5 and 1.0 with an average of 0.86, a 0.11 standard deviation and no outliers. The accuracy ranged between 0.79 and 0.97 with an average of 0.88 and a standard deviation of 0.06 with no outliers. The F1-score had a low average of 0.18, where the lowest value was near zero (i.e., 0.0001), and the maximum value 0.97 with a 0.3 standard deviation and five outliers ranged between 0.64 and 0.97.

The classification error rate was interestingly low, with an average of 0.12 and a standard deviation of 0.06, where the lowest value was 0.03 and the highest 0.21.

The results of detecting unknown attacks using the Bot-IoT dataset were significantly more optimistic, as shown in Table 9. The precision ranged between 0.08 and 0.96 with a 0.79 average and a 0.26 standard deviation with a single outlier (i.e., 0.08). The recall ranged between 0.78 and 0.98 with an average of 0.93, a 0.04 standard deviation and no outliers. The accuracy ranged between 0.86 and 0.98 with an average of 0.93 and a standard deviation of 0.04 with no outliers. The F1-score had a single outlier at 0.15 with an average of 0.83, ranging between 0.15 and 0.96. The classification error rate had an average of 0.07 and a standard deviation of 0.04, where the lowest value was 0.02 and the highest 0.14 without outliers.

On the other hand, the model performance in detecting unknown attacks using the CIC-IDS-2017 dataset shown in Table 10, the precision and F1-score were significantly poor when predicting rare attacks, that is,

Table 8. Results of detecting unknown attacks using the IoT-23 dataset.

Class	Subclass	Precision	Recall	Accuracy	F1	Error
Gagfyt	CC-HeartBeat	0.001	0.95	0.80	0.001	0.20
	DDoS	0.42	0.97	0.91	0.59	0.09
Hakai	CC	0.03	0.95	0.87	0.06	0.13
HideAndSeek	CC	0.00001	1.00	0.80	0.00001	0.20
	PartOfAHorizontalPortScan	0.74	0.87	0.90	0.80	0.10
IRCBot	Attack	0.004	0.97	0.82	0.01	0.18
	CC	0.02	0.94	0.92	0.04	0.08
	PartOfAHorizontalPortScan	0.14	0.96	0.81	0.24	0.19
Kenjiro	Attack	0.00004	0.75	0.92	0.0001	0.08
	CC-HeartBeat	0.06	0.94	0.84	0.12	0.16
	DDoS	0.01	0.80	0.89	0.02	0.11
	Okiru	0.04	0.82	0.94	0.08	0.06
	PartOfAHorizontalPortScan	0.03	0.96	0.83	0.05	0.17
	PartOfAHorizontalPortScan-Attack	0.0001	0.80	0.96	0.0002	0.04
Linux.Hajime	PartOfAHorizontalPortScan	0.49	0.93	0.86	0.64	0.14
Linux.Mirai	CC-HeartBeat	0.02	0.80	0.79	0.05	0.21
	DDoS	0.17	0.86	0.81	0.28	0.19
	Okiru	0.30	0.97	0.97	0.46	0.03
Mirai	Attack	0.01	0.85	0.82	0.03	0.18
	CC	0.07	0.91	0.84	0.14	0.16
	CC-FileDownload	0.0003	0.80	0.84	0.001	0.16
	CC-HeartBeat-Attack	0.005	0.90	0.82	0.01	0.18
	CC-HeartBeat-FileDownload	0.0001	0.91	0.83	0.0001	0.17
	CC-Mirai	0.00001	0.50	0.93	0.00003	0.07
	CC-PartOfAHorizontalPortScan	0.01	0.95	0.87	0.01	0.13
	DDoS	0.97	0.89	0.90	0.93	0.10
	FileDownload	0.0001	0.80	0.89	0.0002	0.11
	Okiru	0.001	0.91	0.80	0.002	0.20
	PartOfAHorizontalPortScan	0.97	0.96	0.95	0.97	0.05
Muhstik	Attack	0.03	0.79	0.83	0.06	0.17
	CC	0.0002	0.75	0.96	0.0004	0.04
	PartOfAHorizontalPortScan	0.72	0.85	0.95	0.78	0.05
Okiru	CC-HeartBeat	0.39	0.88	0.97	0.54	0.03
	Okiru	0.01	0.86	0.90	0.01	0.10
	Okiru-Attack	0.0001	0.67	0.97	0.0001	0.03
Torii	CC-Torii	0.0001	0.83	0.81	0.0002	0.19
Trojan	CC-FileDownload	0.00005	0.67	0.95	0.0001	0.05
	FileDownload	0.00002	0.67	0.91	0.00005	0.09

attacks with a low number of instances such as the Heartbleed attack and Infiltration attacks. However, the recall was extremely high in detecting such scarce type of attacks with a meagre classification error rate. The results showed that the precision was noticeably poor, ranging between 0.0004 and 0.91, with an average of 0.26 and a standard deviation of 0.26 with no outliers. Subsequently, the F1-score was poor as well, with a 0.31 average, a 0.001 minimum value and a 0.91 maximum value, with a standard deviation of 0.34 and no outliers across all metrics. Conversely, the recall was interestingly high, where it ranged between 0.86 and 0.97, with an average of 0.91 and a standard deviation of 0.04. Similarly, the classification error rate results were good as well, where the average was 0.10, and the lowest value was 0.04, and the highest value was 0.15 with a standard deviation of 0.03.

Table 9. Results of detecting unknown attacks using the Bot-IoT dataset.

Class	Subclass	Precision	Recall	Accuracy	F1	Error
DoS	HTTP	0.91	0.92	0.94	0.91	0.06
	TCP	0.77	0.88	0.86	0.82	0.14
	UDP	0.88	0.97	0.94	0.92	0.06
DDoS	HTTP	0.94	0.98	0.97	0.96	0.03
	TCP	0.81	0.92	0.89	0.86	0.11
	UDP	0.92	0.96	0.96	0.94	0.04
Reconnaissance	Data Exfiltration	0.08	0.87	0.93	0.15	0.07
	Keylogging	0.80	0.98	0.98	0.88	0.02
Theft	OS Fingerprint	0.96	0.92	0.95	0.94	0.05
	Service Scan	0.84	0.96	0.92	0.89	0.08

Table 10. Results of detecting unknown attacks using the CIC-IDS-2017 dataset.

Class	Subclass	Precision	Recall	Accuracy	F1	Error
Botnet		0.05	0.90	0.89	0.10	0.11
Brute Force	FTP	0.35	0.88	0.95	0.50	0.05
	SSH	0.11	0.86	0.86	0.20	0.14
DoS	GoldenEye	0.23	0.97	0.89	0.38	0.11
	Hulk	0.88	0.94	0.91	0.91	0.09
	SlowHTTPTest	0.11	0.92	0.85	0.19	0.15
	Slowloris	0.12	0.93	0.87	0.22	0.13
DDoS		0.83	0.90	0.91	0.86	0.09
Heartbleed		0.0004	0.91	0.90	0.001	0.10
Infiltration		0.001	0.94	0.91	0.003	0.09
Port Scan		0.91	0.88	0.92	0.89	0.08
Web	Brute Force	0.05	0.94	0.90	0.09	0.10
	SQLi	0.001	0.86	0.96	0.003	0.04
	XSS	0.02	0.97	0.88	0.04	0.12

Finally, considering the model performance in detecting unknown attacks using the CIC-IDS-2018 dataset shown in Table 11, the precision and F1-score were noticeably poor when predicting attacks with a low number of instances. The results showed that the precision ranged between 0.001 and 0.94, with a 0.4 average and a very high standard deviation of 0.39 without outliers. On the other hand, the recall was considerably high, ranging between 0.95 and 0.98 with a 0.96 average and 0.01 standard deviation with a single outlier at 0.95. Also, the accuracy results were interestingly high, with an average of 0.97, a standard deviation of 0.01, with the lowest value being 0.95 and the highest 0.98 and no outliers. The F1-score had a medium average of 0.45, where the lowest value was near zero (i.e., 0.003), and the maximum value 0.96 with a very high standard deviation of 0.42 and no outliers. The classification error rate was interestingly promising, with an average of 0.03 and a standard deviation of 0.01, where the lowest value was 0.02 and the highest 0.05.

Table 11. Results of detecting unknown attacks using the CIC-IDS-2018 dataset.

Class	Subclass	Precision	Recall	Accuracy	F1	Error
Botnet		0.85	0.96	0.98	0.90	0.02
Brute Force	FTP	0.001	0.96	0.97	0.003	0.03
	SSH	0.64	0.96	0.96	0.77	0.04
DDoS	HOIC	0.79	0.98	0.96	0.87	0.04
	LOIC HTTP	0.94	0.98	0.97	0.96	0.03
	LOIC UDP	0.03	0.96	0.95	0.06	0.05
DoS	GoldenEye	0.47	0.97	0.96	0.64	0.04
	Hulk	0.77	0.96	0.96	0.86	0.04
	SlowHTTPTest	0.001	0.95	0.97	0.003	0.03
	Slowloris	0.22	0.95	0.97	0.35	0.03
Infiltration		0.80	0.96	0.97	0.87	0.03
Web	Brute Force	0.01	0.97	0.97	0.03	0.03
	SQLi	0.002	0.96	0.97	0.005	0.03
	XSS	0.005	0.96	0.96	0.01	0.04

4.3 Known Type

In this subsection, the results of testing the proposed architecture on known data are presented and discussed per dataset.

When examining the performance of the proposed architecture in detecting known attacks, the results show high precision for all the attacks included in the four studied datasets, where the average precision was 0.85 for the IoT-23 dataset, 0.84 for the Bot-IoT dataset, 0.71 for the CIC-IDS-2017 dataset, and 0.76 for the CIC-IDS-2018 dataset. Although those numbers might be considered lower than other proposed models in the literature; however, the proposed architecture is designed to maximise the recall of detecting unknown attacks, which is evident by the high recall values across all datasets, ranging between 0.88 and 0.96. More importantly, the average classification error rate for known attacks was less than 13% across all four datasets, where the highest value was 20% for the Mirai DDoS attack in the IoT-23 dataset, and the lowest value was around 2% for detecting several types of attacks in the CIC-IDS-2018 dataset.

Tables 12, 13, 14, and 15 show the testing results for detecting known attacks using the IoT-23, Bot-IoT, CIC-IDS-2017, and the CIC-IDS-2018 datasets, respectively. A detailed discussion about the performance on the different datasets is provided next.

Table 12. Results of detecting known attacks using the IoT-23 dataset.

Class	Subclass	Precision	Recall	Accuracy	F1	Error
Gagfyt	CC-HeartBeat	0.84	0.96	0.88	0.90	0.12
	DDoS	0.80	0.88	0.84	0.84	0.16
Hakai	CC	0.85	0.94	0.88	0.89	0.12
HideAndSeek	CC	0.88	0.87	0.87	0.87	0.13
	PartOfAHorizontalPortScan	0.94	0.83	0.92	0.88	0.08
IRCBot	Attack	0.92	0.84	0.88	0.88	0.12
	CC	0.90	0.86	0.88	0.88	0.12
	PartOfAHorizontalPortScan	0.90	0.84	0.88	0.87	0.12
Kenjiro	Attack	0.90	0.93	0.91	0.92	0.09
	CC-HeartBeat	0.88	0.85	0.86	0.86	0.14
	DDoS	0.88	0.79	0.84	0.83	0.16
	Okiru	0.83	0.89	0.85	0.86	0.15
	PartOfAHorizontalPortScan	0.94	0.94	0.94	0.94	0.06
	PartOfAHorizontalPortScan	0.93	0.84	0.89	0.89	0.11
Linux.Hajime	PartOfAHorizontalPortScan	0.90	0.89	0.91	0.90	0.09
Linux.Mirai	CC-HeartBeat	0.86	0.88	0.86	0.87	0.14
	DDoS	0.85	0.80	0.83	0.82	0.17
	Okiru	0.89	0.82	0.86	0.85	0.14
Mirai	Attack	0.81	0.84	0.82	0.83	0.18
	CC	0.88	0.80	0.84	0.84	0.16
	CC-FileDownload	0.87	0.85	0.86	0.86	0.14
	CC-HeartBeat-Attack	0.87	0.95	0.90	0.91	0.10
	CC-HeartBeat-FileDownload	0.94	0.86	0.89	0.89	0.11
	CC-Mirai	0.89	0.98	0.92	0.93	0.08
	CC-PartOfAHorizontalPortScan	0.93	0.90	0.91	0.91	0.09
	DDoS	0.29	0.82	0.80	0.43	0.20
	FileDownload	0.90	0.91	0.90	0.90	0.10
	Okiru	0.84	0.98	0.89	0.90	0.11
	PartOfAHorizontalPortScan	0.41	0.89	0.84	0.56	0.16
Muhstik	Attack	0.84	0.94	0.87	0.88	0.13
	CC	0.86	0.95	0.90	0.91	0.10
	PartOfAHorizontalPortScan	0.78	0.85	0.83	0.82	0.17
Okiru	CC-HeartBeat	0.96	0.83	0.89	0.89	0.11
	Okiru	0.98	0.94	0.96	0.96	0.04
	Okiru-Attack	0.94	0.98	0.96	0.96	0.04
Torii	CC-Torii	0.84	0.84	0.83	0.84	0.17
Trojan	CC-FileDownload	0.87	0.96	0.91	0.91	0.09
	FileDownload	0.89	0.91	0.89	0.90	0.11

4.3.1 Table 12: Performance on IoT-23 dataset. This table presents the evaluation of our proposed architecture on the IoT-23 dataset, focusing on unknown attack detection. The precision values varied significantly, indicating a diverse effectiveness in identifying specific attack types accurately. A high recall across most attack types suggests

Table 13. Results of detecting known attacks using the Bot-IoT dataset.

Class	Subclass	Precision	Recall	Accuracy	F1	Error
DoS	HTTP	0.73	0.86	0.88	0.79	0.12
	TCP	0.86	0.86	0.93	0.86	0.07
	UDP	0.86	0.96	0.95	0.91	0.05
DDoS	HTTP	0.74	0.91	0.89	0.82	0.11
	TCP	0.74	0.88	0.89	0.80	0.11
	UDP	0.93	0.89	0.95	0.91	0.05
Reconnaissance	Data Exfiltration	0.91	0.86	0.90	0.88	0.10
	Keylogging	0.90	0.97	0.94	0.93	0.06
Theft	OS Fingerprint	0.93	0.90	0.96	0.92	0.04
	Service Scan	0.83	0.95	0.93	0.88	0.07

Table 14. Results of detecting known attacks using the CIC-IDS-2017 dataset.

Class	Subclass	Precision	Recall	Accuracy	F1	Error
Botnet		0.66	0.90	0.89	0.76	0.11
Brute Force	FTP	0.89	0.90	0.96	0.90	0.04
	SSH	0.64	0.90	0.89	0.75	0.11
DoS	GoldenEye	0.61	0.91	0.87	0.73	0.13
	Hulk	0.44	0.94	0.92	0.60	0.08
	SlowHTTPTest	0.76	0.90	0.93	0.82	0.07
	Slowloris	0.85	0.85	0.94	0.85	0.06
DDoS		0.66	0.92	0.94	0.77	0.06
Heartbleed		0.85	0.86	0.94	0.86	0.06
Infiltration		0.73	0.85	0.91	0.79	0.09
Port Scan		0.60	0.92	0.94	0.73	0.06
Web	Brute Force	0.75	0.86	0.92	0.80	0.08
	SQLi	0.76	0.85	0.92	0.80	0.08
	XSS	0.69	0.91	0.90	0.78	0.10

Table 15. Results of detecting known attacks using the CIC-IDS-2018 dataset.

Class	Subclass	Precision	Recall	Accuracy	F1	Error
Botnet		0.68	0.97	0.96	0.80	0.04
Brute Force	FTP	0.84	0.97	0.98	0.90	0.02
	SSH	0.72	0.96	0.96	0.83	0.04
DDoS	HOIC	0.70	0.96	0.96	0.81	0.04
	LOIC HTTP	0.65	0.96	0.98	0.78	0.02
	LOIC UDP	0.73	0.98	0.96	0.83	0.04
DoS	GoldenEye	0.85	0.95	0.98	0.90	0.02
	Hulk	0.74	0.95	0.97	0.83	0.03
	SlowHTTPTest	0.84	0.97	0.98	0.90	0.02
	Slowloris	0.78	0.95	0.96	0.86	0.04
Infiltration		0.79	0.97	0.97	0.87	0.03
Web	Brute Force	0.79	0.98	0.97	0.88	0.03
	SQLi	0.75	0.97	0.96	0.85	0.04
	XSS	0.73	0.96	0.96	0.83	0.04

the architecture's strength in ensuring minimal false negatives, a critical aspect in security applications. The overall accuracy remained robust, affirming the model's general applicability to IoT environments. The variation in F1-scores across attack types highlights areas for future refinement, particularly in balancing precision and recall.

4.3.2 Table 13: Performance on Bot-IoT dataset. The results in Table 13 showcase the architecture's performance on the Bot-IoT dataset, underscoring a consistently high precision and recall. This dataset, rich in botnet attack data, demonstrates the architecture's adeptness at handling voluminous and varied attack simulations. Notably, the high accuracy across different attack simulations reflects the model's robustness and its potential for practical security solutions in IoT settings. The lower classification error rates further validate the model's efficacy, suggesting it as a viable tool for detecting known and unknown botnet-related activities.

4.3.3 Table 14: Performance on CIC-IDS-2017 dataset. Table 14 details the model's evaluation against the CIC-IDS-2017 dataset, highlighting its performance in a more traditional network setting. The data reveals a challenge in achieving high precision for certain rare attack types, indicating potential overfitting or a lack of generalizability in those cases. However, the consistently high recall rates underscore the model's sensitivity to attack detection, a paramount feature for IDS. This performance points to the necessity for further model tuning to improve precision without compromising on recall, ensuring balanced performance across all attack categories.

4.3.4 Table 15: Performance on CIC-IDS-2018 dataset. Finally, Table 15 examines the architecture's application to the CIC-IDS-2018 dataset, emphasizing its performance in detecting a broad spectrum of attack types within contemporary network environments. Similar to the CIC-IDS-2017 results, we observe challenges in precision for low-frequency attack types, suggesting areas for model enhancement. Nevertheless, the high recall rates across the board reaffirm the model's utility in minimizing false negatives. The accuracy metrics across various attack types attest to the architecture's overall effectiveness and its suitability for modern network intrusion detection scenarios.

4.4 Benign Type

In this subsection, the results of testing the proposed architecture on benign data are presented and discussed per dataset.

Given that the proposed architecture considers all traffic as malicious unless proven otherwise, its performance in distinguishing benign traffic from the malicious one is as important as its ability to detect unknown attacks. The proposed architecture showed significant ability to detect benign traffic with an average precision between 0.86 and 0.98 across all four datasets with a single outlier at 0.48 using the IoT-23 dataset. Moreover, the recall value was also considerably high, ranging between 0.90 and 0.93 with no outliers across all datasets. Therefore, the F1-score was notably significant at an average between 0.88 and 0.95, with a single outlier at 0.60 using the IoT-23 dataset. The classification error rate across all the datasets was at an acceptable low rate, where the maximum value was at 20% when trying to detect C&C Mirai attacks using the IoT-23 dataset. However, the average classification error rate across ranged between 0.08 and 0.11.

Tables 16, 17, 18, and 19 show the testing results for detecting benign traffic using the IoT-23, Bot-IoT, CIC-IDS-2017, and the CIC-IDS-2018 datasets, respectively.

4.5 Overall Model Performance

In this subsection, we evaluate the key performance indicators to show the possible cases of wrong and correct classification. The overall results of testing the proposed architecture are presented and discussed per dataset. Moreover, the proposed architecture is compared to the other models in the literature in terms of error generalisation metrics.

When considering the overall average performance of the proposed architecture, it was clear that it has a high accuracy and recall value than precision and F1-score, whereas the classification error rate was consistently meagre. The overall average accuracy and the recall for all datasets ranged between 0.88 and 0.95, while the precision ranged between 0.62 and 0.84, and the F1-score between 0.64 and 0.87. The performance of the proposed architecture in terms of classification error rate is considered extremely good, where it is averaged between 0.05 and 0.12. Table 20 shows the average overall performance of the proposed architecture across all datasets and traffic type (i.e., benign, known and unknown).

Most importantly, when considering only the unknown attacks, the overall average performance of the proposed architecture shows good recall and accuracy measures (i.e., 0.92), with an excellent classification error rate (i.e., 0.08) across all datasets. Conversely, the precision and F1-score were moderate, 0.4 and 0.44, respectively.

Table 16. Results of detecting benign traffic using the IoT-23 dataset.

Class	Subclass	Precision	Recall	Accuracy	F1	Error
Gagfyt	CC-HeartBeat	0.84	0.96	0.89	0.90	0.11
	DDoS	0.77	0.81	0.80	0.79	0.20
Hakai	CC	0.97	0.84	0.91	0.90	0.09
HideAndSeek	CC	0.88	0.93	0.90	0.90	0.10
	PartOfAHorizontalPortScan	0.92	0.89	0.93	0.91	0.07
IRCBot	Attack	0.93	0.96	0.94	0.94	0.06
	CC	0.93	0.89	0.92	0.91	0.08
	PartOfAHorizontalPortScan	0.78	0.81	0.81	0.80	0.19
Kenjiro	Attack	0.94	0.89	0.92	0.91	0.08
	CC-HeartBeat	0.82	0.87	0.85	0.85	0.15
	DDoS	0.93	0.95	0.94	0.94	0.06
	Okiru	0.97	0.95	0.96	0.96	0.04
	PartOfAHorizontalPortScan	0.89	0.93	0.91	0.91	0.09
	PartOfAHorizontalPortScan-Attack	0.94	0.81	0.88	0.87	0.12
Linux.Hajime	PartOfAHorizontalPortScan	0.79	0.97	0.88	0.87	0.12
Linux.Mirai	CC-HeartBeat	0.78	0.82	0.81	0.80	0.19
	DDoS	0.85	0.93	0.90	0.89	0.10
	Okiru	0.83	0.87	0.85	0.85	0.15
Mirai	Attack	0.84	0.98	0.90	0.91	0.10
	CC	0.93	0.97	0.95	0.95	0.05
	CC-FileDownload	0.84	0.95	0.89	0.89	0.11
	CC-HeartBeat-Attack	0.96	0.88	0.92	0.92	0.08
	CC-HeartBeat-FileDownload	0.85	0.84	0.85	0.85	0.15
	CC-Mirai	0.78	0.80	0.80	0.79	0.20
	CC-PartOfAHorizontalPortScan	0.96	0.80	0.89	0.87	0.11
	DDoS	0.48	0.80	0.84	0.60	0.16
	FileDownload	0.80	0.93	0.86	0.86	0.14
	Okiru	0.85	0.97	0.90	0.91	0.10
	PartOfAHorizontalPortScan	0.84	0.90	0.95	0.87	0.05
Muhstik	Attack	0.91	0.95	0.93	0.93	0.07
	CC	0.86	0.97	0.91	0.91	0.09
	PartOfAHorizontalPortScan	0.92	0.97	0.95	0.95	0.05
Okiru	CC-HeartBeat	0.81	0.86	0.84	0.84	0.16
	Okiru	0.94	0.98	0.96	0.96	0.04
	Okiru-Attack	0.90	0.80	0.86	0.85	0.14
Torii	CC-Torii	0.83	0.98	0.89	0.90	0.11
Trojan	CC-FileDownload	0.80	0.85	0.82	0.82	0.18
	FileDownload	0.92	0.81	0.88	0.86	0.12

Table 21 shows the average overall performance of the proposed architecture in detecting unknown attacks across all datasets.

When comparing the proposed architecture in detecting unknown attacks to those proposed in the literature, it shows significant improvement in terms of accuracy, recall and classification error rate, as shown in Table 22. It is important to note that since our proposed architecture operates by completely removing entire classes and subclasses from the training set and reintroducing them in the testing set to simulate unknown attacks, comparing our results with other approaches that do not adopt the same methodology for defining unknown attacks would be inappropriate. Such a comparison would be analogous to comparing apples and oranges.

Table 17. Results of detecting benign traffic using the Bot-IoT dataset.

Class	Subclass	Precision	Recall	Accuracy	F1	Error
DoS	HTTP	0.88	0.97	0.94	0.92	0.06
	TCP	0.90	0.88	0.92	0.89	0.08
	UDP	0.87	0.95	0.93	0.91	0.07
DDoS	HTTP	0.77	0.88	0.86	0.82	0.14
	TCP	0.82	0.95	0.90	0.88	0.10
	UDP	0.94	0.94	0.96	0.94	0.04
Reconnaissance	Data Exfiltration	0.93	0.92	0.92	0.92	0.08
	Keylogging	0.87	0.96	0.91	0.92	0.09
Theft	OS Fingerprint	0.77	0.86	0.86	0.82	0.14
	Service Scan	0.95	0.95	0.96	0.95	0.04

Table 18. Results of detecting benign traffic using the CIC-IDS-2017 dataset.

Class	Subclass	Precision	Recall	Accuracy	F1	Error
Botnet		0.99	0.86	0.88	0.92	0.12
Brute Force	FTP	0.97	0.86	0.87	0.91	0.13
	SSH	0.98	0.92	0.92	0.95	0.08
DoS	GoldenEye	0.97	0.89	0.89	0.93	0.11
	Hulk	0.89	0.91	0.91	0.90	0.09
	SlowHTTPTest	0.97	0.91	0.90	0.94	0.10
	Slowloris	0.98	0.87	0.88	0.92	0.12
DDoS		0.97	0.89	0.92	0.93	0.08
Heartbleed		0.97	0.94	0.93	0.96	0.07
Infiltration		0.97	0.96	0.95	0.97	0.05
Port Scan		0.91	0.87	0.89	0.89	0.11
Web	Brute Force	0.99	0.87	0.88	0.92	0.12
	SQLi	0.98	0.96	0.95	0.97	0.05
	XSS	0.98	0.93	0.93	0.95	0.07

Table 19. Results of detecting benign traffic using the CIC-IDS-2018 dataset.

Class	Subclass	Precision	Recall	Accuracy	F1	Error
Botnet		0.96	0.95	0.94	0.96	0.06
Brute Force	FTP	0.99	0.93	0.92	0.96	0.08
	SSH	0.99	0.89	0.90	0.94	0.10
DDoS	HOIC	0.97	0.95	0.94	0.96	0.06
	LOIC HTTP	0.96	0.93	0.93	0.94	0.07
	LOIC UDP	0.99	0.87	0.88	0.93	0.12
DoS	GoldenEye	0.99	0.89	0.90	0.94	0.10
	Hulk	0.99	0.86	0.88	0.92	0.12
	SlowHTTPTest	1.00	0.93	0.94	0.96	0.06
	Slowloris	0.99	0.92	0.92	0.95	0.08
Infiltration		0.96	0.89	0.89	0.93	0.11
Web	Brute Force	0.99	0.97	0.96	0.98	0.04
	SQLi	0.99	0.95	0.95	0.97	0.05
	XSS	0.98	0.86	0.86	0.91	0.14

Table 20. Average overall performance of proposed architecture across all datasets.

Dataset	Precision	Recall	Accuracy	F1	Error
IoT-23	0.62	0.88	0.88	0.64	0.12
Bot-IoT	0.84	0.92	0.92	0.87	0.08
CIC-IDS-2017	0.64	0.90	0.91	0.68	0.09
CIC-IDS-2018	0.71	0.95	0.95	0.75	0.05
Total	**0.70**	**0.91**	**0.92**	**0.73**	**0.08**

5 Discussion

In this section, the performance and real-world applications of the proposed architecture are presented and discussed. It is well-established that the time required to detect cyberattacks directly influences the average cost of an attack. According to IBM's 2024 Cost of a Data Breach Report, the mean time to identify and contain (MTTI

Table 21. Average overall performance in detecting unknown attacks across all datasets.

Dataset	Precision	Recall	Accuracy	F1	Error
IoT-23	0.15	0.86	0.88	0.18	0.12
Bot-IoT	0.79	0.93	0.93	0.83	0.07
CIC-IDS-2017	0.26	0.91	0.90	0.31	0.10
CIC-IDS-2018	0.40	0.96	0.97	0.45	0.03
Total	**0.40**	**0.92**	**0.92**	**0.44**	**0.08**

Table 22. Performance comparison in detecting unknown attacks.

Model	Precision	Recall	Accuracy	F1	Error
Proposed	0.40	**0.92**	**0.92**	0.44	**0.08**
[5]	**0.92**	0.56	0.56	**0.65**	0.44
[6]	0.88	0.44	0.44	0.53	0.56

and MTTC) a cyber data breach reached a seven-year low, attributed to the increased adoption of AI systems. For instance, the use of AI in detection reduced the average cost of a data breach by 32.98%, from $5.7 million to $3.82 million. Furthermore, the MTTI and MTTC decreased by 33.77%, from 308 days to 204 days, when AI systems were employed. These findings underscore the critical role of early detection techniques leveraging AI in mitigating the financial impact of cyberattacks, particularly when addressing unknown attacks. This further emphasizes the significance of the proposed architecture.

5.1 Complexity Analysis and Strategic Dataset Selection

To analyze the complexity of our proposed zero-trust IDS architecture, we consider both space and time complexities, which are crucial for understanding its scalability and performance in real-world scenarios. The space complexity of our architecture is primarily influenced by the size of the datasets used for training and the complexity of the machine learning models deployed. Specifically, the deep learning models and the unsupervised clustering algorithm (DBSCAN) require memory proportional to the size of the input data and the number of parameters in the models.

The time complexity is determined by the computational cost of training and running the machine learning models across multiple stages and layers. The training phase of deep learning models involves iterative backpropagation, which has a time complexity of $O(n \cdot m \cdot i)$, where n is the number of samples, m is the number of features, and i is the number of iterations. The DBSCAN algorithm, used in the unsupervised stage, has a time complexity of $O(n^2)$ in the worst case, but optimizations can reduce it to $O(n \log n)$ for many datasets.

In practical terms, the architecture's performance and scalability will depend on the specific hardware and software environment, the efficiency of the implementations, and the complexity of the network traffic being analyzed. Future work will include detailed empirical evaluations to quantify the architecture's space and time requirements under various conditions, aiming to optimize both for efficient deployment in diverse network settings.

Further, our research specifically targets the detection of unknown attacks in IoT and traditional networks, guiding our choice of datasets to align with this focus. While some other well-known datasets such as the Ton dataset[1] is valuable, it was not included because we aimed to address unique challenges posed by unknown attacks. This strategic selection ensures that our study remains concentrated on emerging threats, demonstrating the effectiveness of our architecture in varied network environments.

[1]https://dx.doi.org/10.21227/fesz-dm97

5.2 Practical Implementation and Real-World Applicability

In a real-world scenario, the proposed architecture could be deployed in a smart manufacturing facility with interconnected IoT devices and traditional networks. The system would be integrated at critical points, such as IoT gateways and network routers, to monitor incoming and outgoing traffic. Traffic data would undergo preprocessing, and the supervised models would be trained on known attack patterns, while unsupervised clustering techniques would identify deviations indicative of unknown attacks. The system would enforce a zero trust policy, flagging all traffic as malicious unless proven otherwise, ensuring a comprehensive security posture.

The deployment benefits include enhanced detection of unknown type-A and type-B attacks, scalability across diverse network environments, and reduced false positives due to the multi-layer filtering approach. By continuously updating the supervised models with new attack patterns, the system ensures adaptability to evolving threats, providing robust network security and improving the detection capabilities of IoT and traditional networks alike.

Last, in large-scale enterprise deployments, ensuring continuous performance and accuracy of the proposed architecture requires adaptive learning mechanisms, periodic model retraining with updated datasets, and dynamic tuning of unsupervised components like DBSCAN. A feedback loop to incorporate newly detected anomalies into the training data enhances adaptability to evolving threats. To address network topology changes and dynamic environments, the architecture can leverage software-defined networking (SDN) for flexible traffic management, while automated monitoring and real-time anomaly detection ensure resilience against emerging attack techniques. Regular audits and simulations further validate its robustness in operational settings.

6 Conclusion

Unknown attacks have been ranking in the top three attack types since 2014, remaining an open research issue to be solved. This research work addressed the issue of detecting unknown attacks by proposing and evaluating a multi-stage multi-layer zero trust architecture for detecting unknown attacks. It assumes all traffic is malicious unless proven otherwise; thus, zero-trust. It simulates a multi-layer filtering funnel, where traffic passes through multiple filters designed to take advantage of the concept of type-A and type-B unknown attacks. Malicious traffic is identified using a combination of supervised and unsupervised machine learning algorithms. The overall performance results showed significant improvement in the accuracy, recall and error classification rate in detecting unknown attacks.

As our work considers all traffic malicious unless proven otherwise, our future work will focus on empirically testing our ZTA against real-world IoT traffic and botnet attack simulations. This approach will help us validate the efficiency of our model in practical scenarios, ensuring it is capable of handling high traffic volumes and diverse cyber threats.

References

[1] 2024. CrowdStrike 2024 Global Threat Report. In CrowdStrike. https://www.crowdstrike.com/blog/crowdstrike-2024-global-threat-report/

[2] I. Ahmad, M. Basheri, M. J. Iqbal, and A. Rahim. 2018. Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection. *IEEE Access* 6 (2018), 33789–33795. doi:10.1109/ACCESS.2018.2841987

[3] Mohssine El Ajjouri et al. 2016. LnaCBR:Case Based Reasoning Architecture for Intrusion Detection to Learning New Attacks. *Revue Méditerranéenne des Télécommunications* 6, 1 (2016), 54–59.

[4] Malek Al-Zewairi et al. 2017. Experimental Evaluation of a Multi-layer Feed-Forward Artificial Neural Network Classifier for Network Intrusion Detection System. In *2017 1st International Conference on new Trends in Computing Sciences (ICTCS)*. IEEE, 167–172. doi:10.1109/ictcs.2017.29

[5] Malek Al-Zewairi et al. 2020. Unknown Security Attack Detection Using Shallow and Deep ANN Classifiers. *Electronics* 9, 12 (Nov 2020), 2006. doi:10.3390/electronics9122006

[6] Malek Al-Zewairi et al. 2021. Discovering Unknown Botnet Attacks on IoT Devices Using Supervised Shallow and Deep Learning Classifiers. *Journal of Theoretical and Applied Information Technology* 99, 14 (July 2021).
[7] Shadi Aljawarneh et al. 2017. Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model. *Journal of Computational Science* 25 (Mar 2017), 152–160. doi:10.1016/j.jocs.2017.03.006
[8] Fatima Alwahedi et al. 2024. Machine learning techniques for IoT security: Current research and future vision with generative AI and large language models. *Internet of Things and Cyber-Physical Systems* (2024).
[9] Flora Amato et al. 2018. Smart Intrusion Detection with Expert Systems. In *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*. Springer International Publishing, 148–159. doi:10.1007/978-3-030-02607-3_14
[10] M Arunkumar and K Ashok Kumar. 2022. Malicious attack detection approach in cloud computing using machine learning techniques. *Soft Computing* 26, 23 (2022), 13097–13107.
[11] Da Bao et al. 2015. Predicting New Attacks for Information Security. In *Computer Science and its Applications*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1353–1358. doi:10.1007/978-3-662-45402-2_188
[12] Mansi Bhavsar et al. 2023. Anomaly-based intrusion detection system for IoT application. *Discover Internet of Things* 3, 1 (2023), 5.
[13] Anna Drewek-Ossowicka et al. 2021. A survey of neural networks usage for intrusion detection systems. *Journal of Ambient Intelligence and Humanized Computing* 12, 1 (2021), 497–514.
[14] Martin Ester et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*. 226–231.
[15] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (Portland, Oregon) *(KDD'96)*. AAAI Press, 226–231.
[16] Xian Guo et al. 2023. An intelligent zero trust secure framework for software defined networking. *PeerJ Computer Science* 9 (2023), e1674.
[17] Arash Heidari and Mohammad Ali Jabraeil Jamali. 2023. Internet of Things intrusion detection systems: a comprehensive review and future directions. *Cluster Computing* 26, 6 (2023), 3753–3780.
[18] Wooyeon Jo et al. 2020. Packet Preprocessing in CNN-Based Network Intrusion Detection System. *Electronics* 9, 7 (Jul 2020), 1151. doi:10.3390/electronics9071151
[19] Neelu Khare et al. 2020. SMO-DNN: Spider Monkey Optimization and Deep Neural Network Hybrid Classifier Model for Intrusion Detection. *Electronics* 9, 4 (Apr 2020), 692. doi:10.3390/electronics9040692
[20] Ansam Khraisat et al. 2020. Hybrid Intrusion Detection System Based on the Stacking Ensemble of C5 Decision Tree Classifier and One Class Support Vector Machine. *Electronics* 9, 1 (Jan 2020), 173. doi:10.3390/electronics9010173
[21] Jiyeon Kim et al. 2020. CNN-Based Network Intrusion Detection against Denial-of-Service Attacks. *Electronics* 9, 6 (Jun 2020), 916. doi:10.3390/electronics9060916
[22] K. Kizzee. 2023. Cyber Attack Statistics to Know in 2023. In Parachute. https://parachute.cloud/cyber-attack-statistics-data-and-trends/
[23] Nickolaos Koroniotis et al. 2019. Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset. *Future Generation Computer Systems* 100 (Nov 2019), 779–796. doi:10.1016/j.future.2019.05.041
[24] Przemyslaw Kukielka and Zbigniew Kotulski. 2014. New unknown attack detection with the neural network-based ids. *The State of the Art in Intrusion Prevention and Detection* (2014), 259–284.
[25] Sang-Woong Lee et al. 2021. Towards secure intrusion detection systems using deep learning techniques: Comprehensive analysis and review. *Journal of Network and Computer Applications* 187 (2021), 103111.
[26] Mohammed Maithem and Ghadaa A Al-Sultany. 2021. Network intrusion detection system using deep neural networks. In *Journal of Physics: Conference Series*, Vol. 1804. IOP Publishing, 012138.
[27] Jorge Meira et al. 2018. Comparative Results with Unsupervised Techniques in Cyber Attack Novelty Detection. In *Advances in Intelligent Systems and Computing*. Springer International Publishing, 103–112. doi:10.1007/978-3-030-01746-0_12
[28] Nour Moustafa, Jiankun Hu, and Jill Slay. 2019. A holistic review of Network Anomaly Detection Systems: A comprehensive survey. *Journal of Network and Computer Applications* 128 (Feb 2019), 33–55. doi:10.1016/j.jnca.2018.12.006
[29] Agustin Parmisano et al. 2020. A labeled dataset with malicious and benign IoT network traffic.
[30] Paolo Passeri. 2022. 2022 Cyber Attacks Statistics. In HACKMAGEDDON. https://www.hackmageddon.com/2023/01/24/2022-cyber-attacks-statistics/
[31] Paolo Passeri. 2023. Q2 2023 Cyber Attacks Statistics. In HACKMAGEDDON. https://www.hackmageddon.com/2023/08/08/q2-2023-cyber-attacks-statistics/
[32] Kitsuchart Pasupa and Wisuwat Sunhem. 2016. A comparison between shallow and deep architecture classifiers on small dataset. In *ICITEE*. IEEE, 390–395. doi:10.1109/iciteed.2016.7863293
[33] Ayyaz-Ul-Haq Qureshi et al. 2019. RNN-ABC: A New Swarm Optimization Based Technique for Anomaly Detection. *Computers* 8, 3 (Aug 2019), 59. doi:10.3390/computers8030059
[34] Pranesh Santikellur, Tahreem Haque, Malek Al-Zewairi, and Rajat Subhra Chakraborty. 2019. Optimized Multi-Layer Hierarchical Network Intrusion Detection System with Genetic Algorithms. In *2019 2nd International Conference on new Trends in Computing Sciences*

(ICTCS). IEEE, 1–7. doi:10.1109/ictcs.2019.8923067
[35] Jakob Michael Schoenborn and Klaus-Dieter Althoff. 2023. A Multi-agent Case-Based Reasoning Intrusion Detection System Prototype. In *ICCBR*. Springer, 359–374.
[36] Hichem Sedjelmaci and Nirwan Ansari. 2023. Zero trust architecture empowered attack detection framework to secure 6G edge computing. *IEEE Network* (2023).
[37] Lynda Sellami et al. 2016. Detection of New Attacks on Ubiquitous Services in Cloud Computing and Against Measure. *Advanced Science Letters* 22, 10 (Oct 2016), 3168–3172. doi:10.1166/asl.2016.7991
[38] Iman Sharafaldin et al. 2018. A detailed analysis of the cicids2017 data set. In *ICISSP*. Springer, 172–188.
[39] Iman Sharafaldin et al. 2018. Toward generating a new intrusion detection dataset and intrusion traffic characterization. *ICISSP* 1 (2018), 108–116.
[40] Bhawana Sharma et al. 2023. Anomaly based network intrusion detection for IoT attacks using deep learning technique. *Computers and Electrical Engineering* 107 (2023), 108626.
[41] Rahul Sharma et al. 2023. Probabilistic distributed intrusion detection for zero-trust multi-access edge computing. In *NOMS/IFIP*. IEEE, 1–9.
[42] Tuan Anh Tang et al. 2020. DeepIDS: Deep learning approach for intrusion detection in software defined networking. *Electronics* 9, 9 (2020), 1533.
[43] Ayyaz ul Haq Qureshi et al. 2019. A Heuristic Intrusion Detection System for Internet-of-Things (IoT). In *Advances in Intelligent Systems and Computing*. Springer International Publishing, 86–98. doi:10.1007/978-3-030-22871-2_7
[44] Monika Vishwakarma and Nishtha Kesswani. 2022. DIDS: A Deep Neural Network based real-time Intrusion detection system for IoT. *Decision Analytics Journal* 5 (2022), 100–142.
[45] Na Xing et al. 2023. A Dynamic Intrusion Detection System Capable of Detecting Unknown Attacks. *International Journal of Advanced Computer Science and Applications* 14, 7 (2023).
[46] Huan Yang, Liang Cheng, and Mooi Choo Chuah. 2019. Deep-learning-based network intrusion detection for SCADA systems. In *CNS*. IEEE, 1–7.
[47] Li Yang et al. 2024. Enabling AutoML for Zero-Touch Network Security: Use-Case Driven Analysis. *IEEE Transactions on Network and Service Management* (2024).

Received 18 May 2024; revised 27 December 2024; accepted 15 March 2025