Unveiling the Unseen: Leveraging Zero-Day Attack Detection Using Unsupervised and Semi-Supervised Learning

Osama A. El Awadia
Research and Technical Solutions Unit
Egyptian Computer Emergency Readiness
Team (EG-CERT), NTRA, Egypt
Email: osama.ali@tra.gov.eg

Sameh A. Salem
Computer & Systems Engineering Dept.
Faculty of Engineering, Helwan University
Cairo, Egypt
Research and Technical Solutions Unit Leader
Egyptian Computer Emergency Readiness
Team (EG-CERT), NTRA, Egypt
Email: sameh_salem@h-eng.helwan.edu.eg

2023 33rd International Conference on Computer Theory and Applications (ICCTA) | 979-8-3503-0958-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCTA60978.2023.10969349

***Abstract*—In the ever-evolving cybersecurity landscape, detecting unseen, zero-day attacks is both urgent and paramount. These sophisticated attacks often lack precedent, posing a challenge to conventional machine learning techniques that rely on prior knowledge and training data. This paper endeavors to detect zero-day and unseen cyber attacks using zero-shot machine learning technique, which holds the promise of identifying these attacks without any prior exposure. This work explores the effectiveness of unsupervised learning in zero-day attack detection. The experimental results demonstrate that autoencoders can identify anomalies in data, which are typically associated with zero-day attacks. When compared with other unsupervised and semi-supervised learning methods, the proposed autoencoder algorithm outperforms its competitors and achieves an accuracy of 99.9%, shedding light on its relative effectiveness in zero-day attack detection.**

***Index Terms*—Cybersecurity, Zero-shot learning, Zero-day attacks, Unsupervised learning, Deep autoencoders, Intrusion detection systems**

I. Introduction

The relentless evolution of cybersecurity threats, especially in the form of novel zero-day attacks, continually challenges the integrity and resilience of our digital landscapes. Zero-day attacks refer to a computer-software vulnerability unknown to those who should be interested in mitigating the vulnerability, including the vendor of the target software. The term ’zero-day’ arises from the fact that developers have ’zero days’ to address the vulnerability before it potentially gets exploited by malicious actors.

Conventional machine learning models, while offering significant contributions to cybersecurity, often struggle when faced with the unpredictability and novelty of zero-day attacks. Such models typically rely on historical data and previously seen patterns to identify threats, a strategy that is inherently insufficient for detecting attacks that manifest no prior known patterns. This predicament accentuates the necessity for more advanced techniques, such as ”Zero-Shot Learning”.

Zero-Shot Learning (ZSL) is an innovative machine learning technique that aims to classify unseen objects or predict unseen classes, an approach that is particularly beneficial for tasks like image or text classification. Unlike traditional models that learn from a fixed set of labeled data, ZSL leverages auxiliary information, like attributes or semantic descriptions, to predict unseen classes. This capability is potentially transformative for detecting zero-day attacks, which by definition, are previously unseen threats.

However, the application of ZSL in the realm of cybersecurity is not without challenges. The complexity of implementing ZSL, coupled with its reliance on robust auxiliary information, can limit its feasibility and effectiveness. Additionally, while previous research, such as Sarhan et al. (2023) [1], has begun to explore the potential of ZSL for detecting zero-day attacks, there remains considerable scope for enhancement.

In light of these challenges and opportunities, this paper ventures to further explore the potential of ZSL and unsupervised machine learning techniques, with a specific focus on deep autoencoders. Unsupervised learning techniques like deep autoencoders can learn a compressed, distributed representation of the dataset, which can be particularly useful in detecting novel patterns that may signify a zero-day attack.

Deep autoencoders [2], as a subset of neural networks, can learn to represent data autonomously by training the network to reconstruct its input. This characteristic makes deep autoencoders a promising tool in enhancing Intrusion Detection Systems (IDS), a critical component in any robust cybersecurity infrastructure.

The overarching goal guiding this research is to enhance the robustness and resilience of cybersecurity systems against the ever-advancing threat of zero-day attacks. By exploring the potential of ZSL and unsupervised learning, specifically deep autoencoders, we aim to uncover innovative solutions that can effectively anticipate and mitigate these sophisticated threats.

Organized as follows, the paper commences with Section III, ”Background”, providing an overview of zero-day attacks, conventional machine learning models, and ZSL. Section IV, ”Related Work”, critically reviews previous research on ZSL for zero-day attack detection, identifying areas for further investigation. Section V, ”Proposed Algorithm”, introduces an unsupervised learning approach based on deep autoencoders for detecting novel zero-day attack patterns. Section VI, ”Results and Discussion”, delivers our research findings and their implications for cybersecurity. Section VII, ”Conclusion and Future Work”, concludes the paper, emphasizing ZSL’s potential and unsupervised learning in counteracting zero-day attacks, and outlines future research directions. The paper concludes with a comprehensive ”References” list, serving as a resource for additional research.

979-8-3503-0958-4/23/$31.00 ©2023 IEEE

II. Background

The continuously changing terrain of cyber threats, particularly zero-day attacks [3] that leverage unidentified vulnerabilities, poses significant risks to organizations and frequently bypass conventional detection systems [4]. Semi-supervised anomaly detection approaches, while holding some promise in detecting these attacks, frequently fall short in identifying nuanced intrusions and tend to produce a high rate of false positives [5].

Zero-shot learning presents a novel approach to this issue, enabling models to identify unseen classes without labeled examples [6]. It leverages a latent feature space correlating different classes, allowing the detection of new attack types based on their description. This machine learning concept, when applied to cybersecurity, can overcome the limitations of classical models, generalizing from known to unknown threats. The design of such a system is illustrated in Fig. 1.

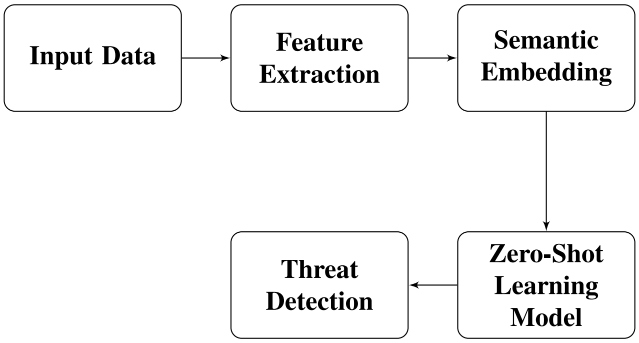

Fig. 1. The design of a zero-shot learning system

The design of a zero-shot learning system includes the following components:

1) **Feature Extraction**: This is the initial step where various characteristics or patterns of the network data are extracted and used as input for the model.
2) **Semantic Embedding**: Here we map the extracted features into a high-dimensional semantic space where similar data points (in terms of their features) are close to each other.
3) **Zero-Shot Learning Model**: This is the core of the system. The model is trained with a set of known attacks, but it is capable of generalizing to detect unknown attacks. It does this by learning to distinguish between normal behavior and anomalous behavior in the semantic space.
4) **Threat Detection**: This is the final step, where the output of the model is interpreted. If the model classifies a data point as an anomaly, it is potentially a zero-day attack.

Unsupervised learning [7] plays a pivotal role in the detection of zero-day attacks. Techniques such as k-means clustering, isolation forests, and Apriori association rules are utilized to discern patterns within network traffic and system logs, enabling the identification of anomalies that signal new threats. Despite their limitations arising from shallow learning capabilities, these methodologies are making significant strides in the field. It is noteworthy that deep autoencoders have demonstrated exceptional results in this area. The structure and application of these autoencoders will be discussed in greater detail later in this section.

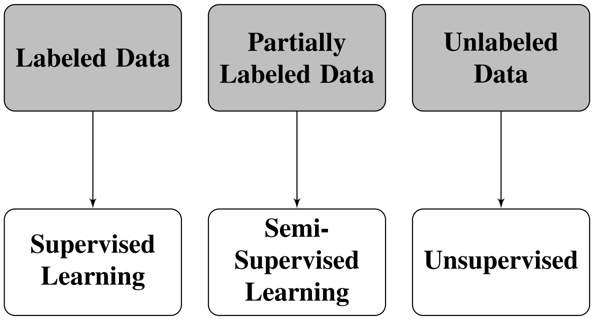

Fig. 2. Machine Learning Algorithms: Supervised, Semi-Supervised, and Unsupervised Learning

With partially labeled samples, as shown in Fig. 2, semi-supervised learning [5] aims higher with its scout parties of labeled data. Guiding unlabeled data with just a few labeled examples, semi-supervised SVMs, graph methods, and self-training adapt more precisely to identify unseen attacks. Still, their small data parties constrain them.

Intrusion Detection Systems (IDS) serve a crucial role in the realm of cybersecurity, functioning as a first line of defense against an array of cyber threats. These systems are designed to monitor, analyze, and identify potential malicious activities within a network or a system by comparing observed patterns with known malicious or benign behaviors. The ultimate aim of an IDS is to swiftly detect and alert system administrators about any anomalous or suspicious activities that could potentially compromise the integrity, confidentiality, or availability of the system. In recent times, the advent of autoencoders, a type of artificial neural network primarily used for unsupervised learning tasks, has offered a novel approach to intrusion detection. Autoencoders are designed to learn efficient representations of input data, termed encodings, by training the model to reconstruct the original input from these encodings. This unique characteristic enables autoencoders to detect anomalous data points, which are typically more challenging to reconstruct accurately. Hence, the integration of autoencoders into IDS has the potential to enhance the detection of novel and sophisticated cyber threats, thereby fortifying the overall security posture of the system.

Now, deep learning autoencoders are leading the charge in detecting novel cybersecurity threats. By reconstructing normal inputs and flagging abnormal errors, they compress critical features in their encodings, uncovering hidden patterns invisible to traditional Machine Learning. With their advanced architecture, they are highly accurate and scalable in detecting zero-day attacks. It is clear that deep autoencoders are the future of cybersecurity defense, as their unsupervised learning will push new frontiers in threat identification. As defenders, we must rally behind them.

Autoencoders, a specific type of feed-forward neural network [8, 9], have found significant use in machine learning, particularly in anomaly detection tasks. These networks employ an unsupervised learning approach to train input vectors to reconstruct as output vectors [10], essentially replicating the input as the output. The architecture of an autoencoder consists of an encoder, a bottleneck, and a decoder, with the output layer having the same dimensions as the input layer.

In anomaly detection, the autoencoder is trained with normal data. During the testing phase, it attempts to reconstruct new input data. If the new data is normal, the reconstruction will be successful, meaning the difference between the input and the reconstructed data will be small. However, if the data is anomalous (i.e., different from the normal data the autoencoder has been trained on), the reconstruction will be poor, resulting in a large reconstruction error. This difference serves as a measure to determine whether the data is normal or anomalous. In the context of cybersecurity and zero-day attack detection, an unusually large difference could signal a potential threat. The exemplar architecture of an autoencoder is showcased in Fig. 3.

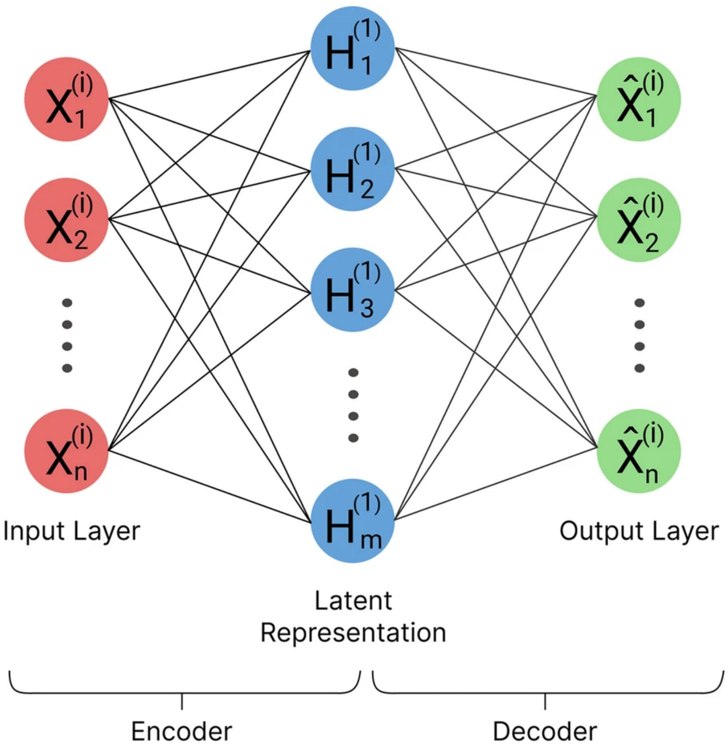

Fig. 3. The architecture of the autoencoder [11]

The architecture of the autoencoder includes the following components:

1) **Input layer**: This is where the original high-dimensional data that you aim to compress is introduced to the network.
2) **Latent representation**: Also known as the “bottleneck”, this layer contains the compressed version of the input data. The encoder portion of the network is responsible for creating this compressed representation.
3) **Output layer**: Here, the decoder part of the network reconstructs the input data from its compressed state. The output, or “reconstructed data”, should closely resemble the original input if the data is normal.

III. Related Work

Zhang et al. [12] tackle the ongoing issue of detecting unknown network intrusions in the cybersecurity field, noting the limitations of current methods like clustering and honeypot, such as difficulties in sample collection and detection delays. They propose the use of Zero-Shot Learning (ZSL), which can identify unknown attacks by understanding the relationships between the feature and semantic spaces. They also introduce a novel ZSL method using a sparse autoencoder for mapping known attack features to the semantic space and establishing a feature-to-semantic map for detecting unknown attacks, which demonstrated an average accuracy of 88.3% using the NSL_KDD dataset. Tayyab et al. [13] highlight the challenges and shortcomings of traditional malware detection methods, with a focus on Indicators of Compromise (IOC). They point out that traditional anti-malware systems, reliant on constantly updated malware signatures, are ineffective against zero-day attacks or malware variants. This has led research to shift towards machine learning-based approaches, despite their own limitations. To address these, they suggest deep learning-based methods, which offer automatic feature engineering, large dataset handling, and support for one-shot learning. They provide a survey of strategies for real-time malware detection, propose a hierarchical model for real-time threat detection, and compare deep learning-based approaches with conventional methods.

Deldar and Abadi [14] shed light on the challenges of detecting zero-day malware and emphasize the role of deep learning in this field. They categorize various deep learning techniques into unsupervised, semi-supervised, few-shot, and adversarial resistant methods, comparing them based on a range of parameters. Concurrently, Topcu et al. [15] propose a novel approach to detecting zero-day attacks by analyzing Twitter data using machine learning techniques, achieving an 80% success rate in detection. Similarly, Zahoora et al. [16] tackle the threat of zero-day ransomware attacks, leveraging Zero-Shot Learning (ZSL) for this task. They introduce the Deep Contractive Autoencoder based Attribute Learning (DCAE-ZSL) and a Heterogeneous Voting Ensemble (DCAE-ZSL-HVE) to enhance the detection of these attacks and minimize false negatives. These studies collectively highlight the evolving methods and techniques in zero-day attack detection and the efficacy of machine learning and deep learning approaches in this domain.

Hindy et al. [17] highlight the rising challenge for Intrusion Detection Systems (IDS) due to the increasing volume and diversity of cyber-attacks. They note the high false-negative rates of current outlier-based zero-day detection research and propose an autoencoder implementation to address this limitation. Their model, evaluated using two well-known IDS datasets, leverages the encoding-decoding capabilities of autoencoders and achieves high detection accuracy for zero-day attacks. Their findings show a zero-day detection accuracy of 89–99% for the NSL-KDD dataset [18] and 75–98% for the CICIDS2017 dataset [19]. The study concludes with a discussion on the observed trade-off between recall and fallout. Guo [20] emphasizes the growing risk of zero-day attacks that exploit unknown vulnerabilities and bypass traditional cybersecurity measures. He critiques signature-based detection methods and proposes machine learning as a promising solution due to its ability to capture statistical attack characteristics. After conducting a detailed review of existing machine learning-based detection techniques for zero-day attacks, Guo identifies key challenges and suggests future research directions.

In their literature review, Rasheed et al. [21] explore the ongoing challenge of cyberattacks and the need for effective intrusion detection systems (IDS). They note many recent IDS solutions fall short due to reliance on attack signature repositories, outdated datasets, or neglecting zero-day attacks in their machine learning (ML) or deep learning (DL) models. They highlight the difficulty of detecting zero-day attacks despite numerous proposed solutions over the years. Their systematic literature review aims to consolidate various methodologies, techniques, and ML and

DL algorithms used for detecting zero-day attacks, providing a valuable resource for future research. Despite technological advances, large datasets, and powerful DL algorithms, detecting new or unknown attacks remains an open research area. Their review attempts to bridge this gap by creating a comprehensive repository of ML and DL-based tools and techniques for zero-day attack detection.

Priya and Annie Uthra [22] address the pressing need for effective intrusion detection systems (IDS) capable of identifying the rising number of unknown cyberattacks, including zero-day attacks. They introduce a novel deep learning-based variational autoencoder (DL-VAE) model specifically designed for zero-day attack detection. The model includes a pre-processing step to make raw data compatible, followed by the application of the VAE model to detect potential zero-day attacks in the network data. The authors validate the model's performance through a series of experiments, with results demonstrating strong accuracy of 98.8%, suggesting the DL-VAE model's effectiveness in detecting zero-day attacks.

Ortega-Fernandez et al. [23] explore anomaly detection in industrial control and cyber-physical systems, which is becoming increasingly important due to the modernization and exposure of industrial environments. They note common threats, including intellectual property theft, denial of service, and cloud component compromise. The authors propose a network intrusion detection system (NIDS) based on a deep autoencoder trained on network flow data, bypassing the need for prior knowledge of network topology. Their model showed high detection rates and low false alarms against distributed denial of service attacks, outperforming current methods and a baseline model in an unsupervised learning environment. The model also detected abnormal behavior in legitimate devices post-attack. The authors further validate their NIDS in a real industrial plant and highlight its low-cost, minimal processing needs, unsupervised operation, and ease of deployment in real-world scenarios.

Another study, Ali et al. [24], emphasizes the importance of using AI-based techniques to tackle the challenges posed by zero-day attacks. It provides valuable insights into the strengths and limitations of different algorithms and highlights the need for ongoing research in this area to enhance the security of systems against advanced cyber threats.

This paper aims to conduct a head-to-head evaluation of zero-shot learning, deep autoencoders, and semi-supervised algorithms for zero-day attack detection on network intrusion data. The results will provide insight into the performance gaps between these methods to identify strengths, weaknesses, and opportunities to improve future zero-day detection models. Enhanced detection of novel cybersecurity threats will better equip organizations to thwart zero-day attacks before they cause major damage.

IV. Proposed Algorithm

In our methodology, we applied a deep autoencoder neural network to learn efficient representations of normal network traffic. The autoencoder architecture consists of encoder layers to compress the input into a low-dimensional encoding, and decoder layers to reconstruct the original input. Dropout layers were added to prevent overfitting.

The model was trained on both normal and malicious data. However, for testing, a new class of malicious data, not present in the training set, was introduced to evaluate the model's ability to detect unprecedented threats. The reconstruction error, a measure of divergence between the model's output and the original input, was calculated for unseen data and treated as an anomaly score. In the context of autoencoders, minimizing this error during training essentially involves learning a compact and efficient representation of the input data. However, when the model encounters data that significantly deviates from its training set (such as a new class of malicious data), the reconstruction error is likely to be high due to the model's unfamiliarity with this data. This anomaly score, a numerical value that quantifies the degree of deviation of a data point from 'normal' instances, is computed using the reconstruction error. Consequently, the higher the reconstruction error, the higher the anomaly score, and the more likely the instance is recognized as an anomaly, indicative of a potential novel threat.

This unsupervised deep learning approach leverages low reconstruction error on normal data to detect anomalies. The block diagram shown in Fig. 4 leverages deep autoencoders to detect novel attacks.

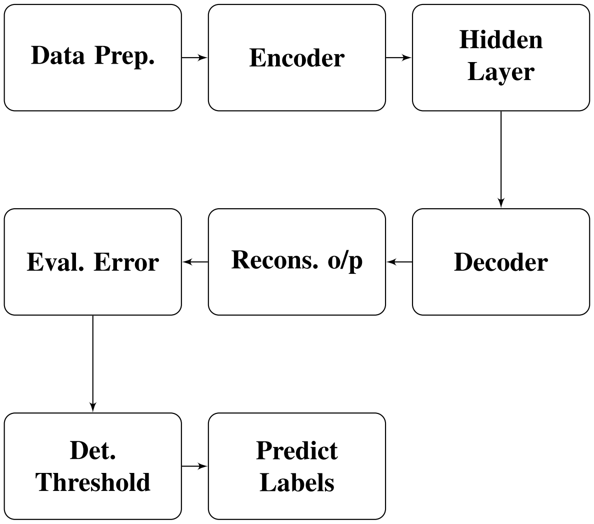

Fig. 4. Block diagram for Deep Autoencoder process to predict labels of unseen data

The following provides a detailed, sequential explanation of each stage represented in the block diagram:

1) **Data Preparation (Data Prep.)**: This is where the data is cleaned, formatted, and transformed into a suitable state for the model to process. This might include handling missing values, normalizing numeric data, and encoding categorical data.
2) **Encoder**: The encoder part of the autoencoder takes the input data and compresses it into a smaller, more dense representation. This process can help highlight the most important features of the data and remove noise or less important information.
3) **Hidden Layer**: This is where the compressed data from the encoder is stored. The size of the hidden layer determines the amount of information that can be stored from the input data.
4) **Decoder**: The decoder takes the compressed data from the hidden layer and attempts to reconstruct the original input data. This process helps the model learn to preserve the most important features of the data in the compressed representation.

5) **Reconstruction (Recons.)**: This is the output of the decoder. It is a reconstruction of the original input data based on the compressed information stored in the hidden layer.
6) **Evaluate Reconstruction Error (Eval. Error)**: This step involves comparing the original input data with the reconstructed output to measure how well the model has learned to compress and reconstruct the data. The difference between the original and reconstructed data is the reconstruction error.
7) **Determine Anomaly Threshold (Det. Threshold)**: Based on the reconstruction error, a threshold is determined. If the reconstruction error exceeds this threshold, the input data can be considered an anomaly. This threshold is typically determined based on the specific problem domain and the characteristics of the normal data.
8) **Predict Labels**: Finally, the trained model is used to predict the labels of the unseen data. The model processes the new data through the encoder and decoder, calculates the reconstruction error, and if this error exceeds the predetermined threshold, the data is labeled as an anomaly.

Algorithm 1 outlines the process for using deep autoencoders to detect unknown attacks.

Algorithm 1 Proposed Autoencoder Algorithm

Input: training data $\mathbf{X}_{\text{train}}$, test data $\mathbf{X}_{\text{test}}$

Parameters: encoding_dim, epochs, batch_size, threshold

```
Define input layer input_layer with input dimension input_dim State
Define encoding layers encoder with encoding_dim nodes
Define decoding layers decoder mirroring encoder
Define autoencoder model with input_layer and decoder
Compile autoencoder model with Adam optimizer and binary cross-entropy loss
Train autoencoder on $\mathbf{X}_{\text{train}}$
Encode test data $\mathbf{X}_{\text{test}}$ using autoencoder to get $\mathbf{X}_{\text{test\_pred}}$
Calculate MSE between $\mathbf{X}_{\text{test}}$ and $\mathbf{X}_{\text{test\_pred}}$
Set threshold as quantile of MSE scores
for each example do
    if MSE > threshold then
        Classify as anomaly (0)
    else
        Classify as normal (1)
    end if
end for
```

Output: anomaly predictions $\mathbf{y}_{\text{pred}}$

The threshold value in the autoencoder algorithm is utilized to determine if a given example is an anomaly, based on the mean squared error (MSE) between the original test data X_{test} and the reconstructed data $X_{\text{test\_pred}}$.

Typically, the threshold is defined as a quantile of the MSE scores calculated on a validation set or the training set. For instance, the threshold might be set to the 95th or 99th percentile of the MSE scores, implying that the top 5% or 1% of data points with the highest MSE scores would be considered anomalies.

The sensitivity of the model to anomalies is influenced by the threshold value:

- With a lower threshold, the model becomes more sensitive and categorizes more points as anomalies. This could lead to an increase in both the true positive rate (correct anomaly identification) and the false positive rate (incorrect normal point classification as anomalies).
- With a higher threshold, the model becomes less sensitive and labels fewer points as anomalies. This could decrease the false positive rate but also lower the true positive rate, potentially missing some anomalies.
- Setting the threshold to zero in anomaly detection implies that any instance exhibiting even the slightest amount of reconstruction error would be deemed anomalous. Typically, such a setting is likely to result in an elevated number of false positives, where normal instances are incorrectly flagged as anomalies, unless the model's accuracy is impeccable, a circumstance that is relatively improbable.

Thus, the selection of an appropriate threshold is a balancing act between detecting anomalies (sensitivity) and avoiding excessive false alarms (specificity). It often involves tuning and validation based on the specific use case and the cost trade-off between missing anomalies and raising false alarms.

In addition, we utilized unsupervised and semi-supervised k-means clustering techniques to validate that autoencoders demonstrate optimal performance in detecting zero-day attacks. The findings presented in the "Results & Discussion" section demonstrate this.

V. Results and Discussion

This paper presents compelling evidence that deep autoencoders significantly outshine other supervised and semi-supervised learning techniques in detecting zero-day cyberattacks. Utilizing the NSL-KDD intrusion detection dataset [18], the deep autoencoder model accomplished an impressive detection accuracy of 99.90% on the evaluation set.

When the threshold is set to 0.001, only 42 out of 41,214 records are classified as normal, indicating that the autoencoder has an accuracy of 99.9% in detecting anomalies. Interestingly, in a specific instance where the threshold was adjusted to zero, the outcome was that only one record was classified as normal, and the remaining 41,213 were identified as anomalies. This equated to an accuracy rate of approximately 100%, a highly desirable result, particularly in the detection of zero-day attacks. However, it's crucial to interpret such an outcome with caution. While the result in this context may appear favorable, a zero threshold generally doesn't make practical sense. This approach assumes a flawless model, which is a highly improbable scenario. The high accuracy rate observed here is exceptionally unusual and may not be indicative of the model's performance under normal circumstances. Therefore, it's typically more judicious to set the threshold based on the distribution of reconstruction errors on the training data, which can properly accommodate the expected minor deviations that are inherent to any model's predictions.

In stark contrast, the semi-supervised approach, using the identical NSL-KDD dataset, managed to attain only 83.28% accuracy. Similarly, both unsupervised and semi-supervised k-means clustering methods yielded an approximate accuracy of 83.28% on the dataset, as indicated in Table I. These findings underscore the superior proficiency of deep autoencoders over unsupervised/semi-supervised clustering in precisely identifying novel and unrecognized cybersecurity threats, as per the comprehensive NSL-KDD benchmark dataset.

In addition to the aforementioned techniques, the Mean Shift algorithm [25] was also incorporated into the research as an unsupervised learning method and as a means of comparison. This non-parametric algorithm, typically used for cluster analysis, yielded interesting findings in the context of zero-day attack detection.

The Mean Shift algorithm, using the same NSL-KDD dataset, achieved an accuracy of 77.24% - lower than both the deep autoencoders and the semi-supervised approach. This disparity in results could potentially be attributed to the high dimensionality of the data, which can often make density estimation, a fundamental aspect of the Mean Shift algorithm, challenging.

Moreover, Mean Shift tends to work best when clusters are evenly sized and regularly shaped, conditions that cannot always be guaranteed with complex data like NSL-KDD. Consequently, the algorithm may have struggled to determine the accurate boundaries of clusters, thereby impacting its detection accuracy.

While the Mean Shift algorithm presented a lower detection accuracy in this study, it's worth noting that these results do not diminish its value. The algorithm's effectiveness can vary depending on the data distribution and specific use case. In other scenarios, where the data might be less high-dimensional or the clusters more regularly shaped, Mean Shift may potentially offer a more competitive performance.

Equation 1 depicts the distribution of each distinct prediction, either 0 or 1, in the output of the model, where 0 signifies an anomaly and 1 represents normality.

$$P_i = \frac{n_i}{N} \times 100 \tag{1}$$

In this equation:

- P_i denotes the percentage that each distinct prediction i constitutes.
- n_i refers to the number of times a unique prediction i occurs.
- N stands for the total count of predictions.

This equation calculates the proportional representation, of each unique prediction by dividing the frequency of each unique prediction (n_i) by the total number of predictions (N), and then multiplying the result by 100. This provides an understanding of the distribution of various predictions made by the model, effectively quantifying the model's accuracy.

TABLE I
PERFORMANCE EVALUATION RESULTS

Algorithm	Accuracy
K-Means	83.28%
Semi-supervised K-Means	83.28%
Mean Shift	77.24%
Proposed Autoencoder	**99.90%**

While it is theoretically possible for unsupervised and semi-supervised k-means clustering techniques to yield similar accuracies on the same dataset given the same preprocessing, it is not typically expected. The key difference between unsupervised and semi-supervised learning lies in the use of labeled data. In unsupervised learning, like k-means clustering, the algorithm tries to identify inherent structures in the data without any prior knowledge or labels. On the other hand, semi-supervised learning utilizes a small amount of labeled data to guide the learning process, alongside a larger volume of unlabeled data.

These results align with previous studies showing the superior capability of deep learning for anomaly detection compared to shallow machine learning models [17]. The deep autoencoder is multilayer architecture can learn highly complex patterns within network traffic data. This allows it to model normal behavior more accurately and in turn identify even minute anomalies indicative of novel zero-day attacks.

In contrast, the semi-supervised model struggles to define normal data. The high 16.72% false negative rate for the semi-supervised approach likely reflects its inability to detect subtler zero-day intrusion behaviors. Its oversimplified representation of complex network data dynamics limits its effectiveness, as past work similarly found [5].

The dramatically higher accuracy achieved by the deep autoencoder demonstrates its reliability for organizations seeking to bolster their zero-day attack detection. Based on these results, deep autoencoders should be considered over other unsupervised and semi-supervised methods for identifying novel attack patterns within real-world network traffic.

Future work should explore combining deep autoencoders with dimensionality reduction techniques like principal component analysis to distill the most salient features for discrimination. Testing across a range of network intrusion datasets will also help evaluate the robustness of deep autoencoder models. Overall, this paper provides strong proof of concept that deep learning can significantly advance zero-day attack detection capabilities.

VI. Conclusion and Future Work

In the world of cybersecurity, it is crucial and of utmost importance to detect unseen, zero-day attacks. This paper introduced the use of zero-shot machine learning for zero-day attack detection without any prior exposure, providing convincing evidence to support unsupervised learning techniques. In comparison to unsupervised and semi-supervised learning methods, the autoencoder approach outperformed with an outstanding accuracy rate of 99.9%, surpassing all other competitive techniques. This vindicates the ability of the autoencoder to spot inconsistencies in data, which are generally a telltale sign of such complex attacks.

Experimental results on a real-world dataset emphasize the potential of zero-shot machine learning techniques in zero-day attack detection. The results highlight the relative efficacy of the proposed autoencoder algorithm in detecting zero-day attacks, laying the groundwork for further research and practical applications in this area. The achievement of this work offers a robust protective strategy against new and unknown cyber threats.

Further work should explore using other machine learning techniques to improve the robustness and reliability of the proposed algorithm. Additional comparisons to other state-of-the-art methods are also needed. There are many opportunities to extend this approach to new domains and enhance the generalization capabilities.

References

[1] M. Sarhan, S. Layeghy, M. Gallagher, and M. Portmann, "From zero-shot machine learning to zero-day attack detection," *International Journal of Information Security*, pp. 947–959, 2023.

[2] A. Takiddin, M. Ismail, U. Zafar, and E. Serpedin, "Deep autoencoder-based anomaly detection of electricity theft cyberattacks in smart grids," *IEEE Systems Journal*, vol. 16, no. 3, pp. 4106–4117, 2022.

[3] L. Bilge and T. Dumitraş, “Before we knew it: an empirical study of zero-day attacks in the real world,” in *Proceedings of the 2012 ACM conference on Computer and communications security*, pp. 833–844, 2012.

[4] M. Egele, C. Kruegel, E. Kirda, H. Yin, and D. Song, “Dynamic spyware analysis,” 2007.

[5] I. Mbona and J. H. Eloff, “Detecting zero-day intrusion attacks using semi-supervised machine learning approaches,” *IEEE Access*, vol. 10, pp. 69822–69838, 2022.

[6] P. H. Barros, E. T. Chagas, L. B. Oliveira, F. Queiroz, and H. S. Ramos, “Malware-smell: A zero-shot learning strategy for detecting zero-day vulnerabilities,” *Computers & Security*, vol. 120, p. 102785, 2022.

[7] T. Zoppi, A. Ceccarelli, and A. Bondavalli, “Unsupervised algorithms to detect zero-day attacks: Strategy and application,” *Ieee Access*, vol. 9, pp. 90603–90615, 2021.

[8] C. Zhou and R. C. Paffenroth, “Anomaly detection with robust deep autoencoders,” in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 665–674, 2017.

[9] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. v. d. Hengel, “Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1705–1714, 2019.

[10] M. Sakurada and T. Yairi, “Anomaly detection using autoencoders with nonlinear dimensionality reduction,” in *Proceedings of the MLSDA 2014 2nd workshop on machine learning for sensory data analysis*, pp. 4–11, 2014.

[11] H. Torabi, S. L. Mirtaheri, and S. Greco, “Practical autoencoder based anomaly detection by using vector reconstruction error,” *Cybersecurity*, vol. 6, 2023.

[12] Z. Zhang, Q. Liu, S. Qiu, S. Zhou, and C. Zhang, “Unknown attack detection based on zero-shot learning,” *IEEE Access*, vol. 8, pp. 193981–193991, 2020.

[13] U.-e.-H. Tayyab, F. B. Khan, M. H. Durad, A. Khan, and Y. S. Lee, “A survey of the recent trends in deep learning based malware detection,” *Journal of Cybersecurity and Privacy*, vol. 2, no. 4, pp. 800–829, 2022.

[14] F. Deldar and M. Abadi, “Deep learning for zero-day malware detection and classification: A survey,” *ACM Computing Surveys*, vol. 56, sep 2023.

[15] A. E. Topcu, Y. I. Alzoubi, E. Elbasi, and E. Camalan, “Social media zero-day attack detection using tensorflow,” *Electronics*, vol. 12, no. 17, p. 3554, 2023.

[16] U. Zahoora, M. Rajarajan, Z. Pan, and A. Khan, “Zero-day ransomware attack detection using deep contractive autoencoder and voting based ensemble classifier,” *Applied Intelligence*, vol. 52, no. 12, pp. 13941–13960, 2022.

[17] H. Hindy, R. Atkinson, C. Tachtatzis, J.-N. Colin, E. Bayne, and X. Bellekens, “Utilising deep learning techniques for effective zero-day attack detection,” *Electronics*, vol. 9, no. 10, p. 1684, 2020.

[18] G. Meena and R. R. Choudhary, “A review paper on ids classification using kdd 99 and nsl kdd dataset in weka,” in *2017 International Conference on Computer, Communications and Electronics (Comptelix)*, pp. 553–558, IEEE, 2017.

[19] A. Boukhamla and J. C. Gaviro, “Cicids2017 dataset: performance improvements and validation as a robust intrusion detection system testbed,” *International Journal of Information and Computer Security*, vol. 16, no. 1-2, pp. 20–32, 2021.

[20] Y. Guo, “A review of machine learning-based zero-day attack detection: Challenges and future directions,” *Computer Communications*, 2022.

[21] R. Ahmad, I. Alsmadi, W. Alhamdani, and L. Tawalbeh, “Zero-day attack detection: a systematic literature review,” *Artificial Intelligence Review*, pp. 1–79, 2023.

[22] S. Priya and R. Annie Uthra, “An effective deep learning-based variational autoencoder for zero-day attack detection model,” in *Inventive Systems and Control: Proceedings of ICISC 2021*, pp. 205–212, Springer, 2021.

[23] I. Ortega-Fernandez, M. Sestelo, J. C. Burguillo, and C. Pinon-Blanco, “Network intrusion detection system for ddos attacks in ics using deep autoencoders,” *Wireless Networks*, pp. 1–17, 2023.

[24] S. Ali, S. U. Rehman, A. Imran, G. Adeem, Z. Iqbal, and K.-I. Kim, “Comparative evaluation of ai-based techniques for zero-day attacks detection,” *Electronics*, vol. 11, no. 23, p. 3934, 2022.

[25] D. Demirović, “An implementation of the mean shift algorithm,” *Image Processing On Line*, vol. 9, pp. 251–268, 2019.