МАТЕМАТИЧЕСКИЕ, СТАТИСТИЧЕСКИЕ И ИНСТРУМЕНТАЛЬНЫЕ МЕТОДЫ В ЭКОНОМИКЕ

DOI: 10.36683/2306-1758/2023-1-43/4-18

УДК (UDC) 330.4

JEL: C45, C60, C80

Ray S.

XA-GANomaly: An Explainable Adaptive Semi-supervised Learning Method for Intrusion Detection using GANomaly in Global Economic dynamic shifts

Samrat Ray
Dr, Dean and head of International Relations
International Institute of Management Studies (IIMS);
Pune, India
e-mail: s.ray@iimspune.edu.in
ORCID: 0000-0002-9845-2974

Рэй Самрат
доцент, декан и глава отдела международных отношений
Международный институт управленческих исследований;
г. Пуна, Индия
e-mail: s.ray@iimspune.edu.in
ORCID: 0000-0002-9845-2974

Economic Intrusion detection is discovering unauthorized network activity and recognizing whether the data is an abnormal network transmission. Recent research has focused on semi-supervised learning mechanisms for identifying abnormal network traffic to deal with both labeled and unlabeled in industry. Economic shift generation is very crucial in creating a balance between hyper sensitive generation and superficial balance. However, real-time training and classifying network traffic remains as challenges, as they can lead to the degradation of the overall dataset and difficulties in preventing attacks. Additionally, existing semi-supervised learning research may not comprehensively analyze the experimental results. To address these issues, we propose XA-GANomaly, a novel technique for explainable adaptive semi-supervised learning using GANomaly, an image anomalous detection model, that sequentially trains small subsets in a dynamic way. First, we introduce a Deep Neural Network (DNN) based GANomaly for semi-supervised learning. Second, we present our proposed adaptive algorithm for DNN-based GANomaly, validated with four subsets of the adaptive dataset. Finally, we demonstrate a monitoring system that incorporating three explainable techniques – SHapley Additive exPlanations (SHAP), reconstruction error visualization, and t-distributed stochastic neighbor embedding (t-SNE) – to respond effectively to attacks toward traffic data at each stage of feature engineering, semi-supervised learning, and adaptive learning. Compared to other single-class classification techniques, the proposed DNN-based GANomaly achieves higher scores than 13% and 8% of F1 scores and accuracy of 4.17% and 11.51% for the NSL-KDD and UNSW-NB15 datasets, respectively. In addition, our adaptive learning experiments show mostly improved results over initial values, and we provide an analysis and monitoring system based on the combination of the three explainable methodologies. As a result, our proposed method has the potential to be applied in the real industry, and future research will explore on handling unbalanced real-time datasets in various scenarios.The essence of such shifts in economic paradigm can be better understood by applying various triggers in functional machine to machine interaction.This can be better understood by creating a supervised paradigm which better understands global economic shifts.

Обнаружение экономических вторжений заключается в нахождении несанкционированной сетевой активности и определении того, являются ли данные ненормальной передачей по Сети. Недавние исследования были сосредоточены на механизмах обучения с частичным контролем для выявления аномального сетевого трафика, чтобы рассматривать как меченый, так и не содержащий метки в промышленности. Генерация экономического сдвига играет очень важную роль в создании баланса между сверхчувствительным поколением и внешним равновесием. Однако обучение в режиме реального времени и классификация сетевого трафика остаются сложными задачами, поскольку они могут привести к ухудшению общего набора данных и трудностям в предотвращении атак. Кроме того, существующие исследования в области обучения с частичным контролем могут не дать всестороннего анализа результатов эксперимента. Чтобы решить эти проблемы, мы предлагаем XA-GANomaly – новую технику объяснимого адаптивного обучения с частичным контролем с использованием GANomaly – модели обнаружения аномалий изображения, которая последовательно обучает небольшие подмножества динамическим способом. Во-первых, мы представляем GANomaly на основе глубокой нейронной сети (DNN) для обучения с частичным контролем. Во-вторых, мы представляем предложенный нами адаптивный алгоритм для GANomaly на основе DNN, проверенный на четырёх подмножествах адаптивного набора данных. Наконец, мы демонстрируем систему мониторинга, которая включает в себя три объяснимых метода: SHapley Additive exPlanations (SHAP), визуализацию ошибок реконструкции и стохастическое вложение соседей с t-распределением (t-SNE) для эффективного предупреждения атак на данные трафика на каждом этапе разработки и проектирования объектов, обучения с частичным контролем и адаптивного обучения. По сравнению с другими методами классификации по одному классу предлагаемая GANomaly на основе DNN позволяет получить более высокие баллы, чем 13 % и 8 % баллов F1, точность 4,17 % и 11,51 % для наборов данных NSL-KDD и UNSW-NB15 соответственно. Кроме того, наши эксперименты по адаптивному обучению показывают в основном улучшенные результаты по сравнению с исходными значениями, и мы предусматриваем систему анализа и мониторинга, основанную на сочетании трёх объяснимых методологий. В результате предло-

женный нами метод имеет потенциал для применения в реальной промышленности, а будущие исследования будут посвящены обработке несбалансированных наборов данных в реальном времени в различных сценариях. Суть таких сдвигов в экономической парадигме можно лучше понять, применив различные триггеры в функциональном взаимодействии от устройства к устройству. Это возможно благодаря созданию контролируемой парадигмы, которая лучше воспринимает глобальные экономические сдвиги.

***Keywords**: intrusion detection system, adaptive learning, semi-supervised learning, explainable artificial intelligence, monitoring system.*

***Ключевые слова**: система обнаружения вторжений, адаптивное обучение, обучение с частичным контролем, объяснимый искусственный интеллект, система мониторинга.*

Ray S. XA-GANomaly: An Explainable Adaptive Semi-supervised Learning Method for Intrusion Detection using GANomaly in Global Economic dynamic shifts. *Economic environment*. 2023; № 1 (43). – С. 4-18. – http://dx.doi.org/10.36683/2306-1758/2023-1-43/4-18.

Рэй С. XA-GANomaly: Объяснимый адаптивный метод обучения с частичным контролем для обнаружения вторжений с использованием GANomaly в условиях глобальных экономических динамических сдвигов // Экономическая среда. – 2023; 1 (43): 4-18. – http://dx.doi.org/10.36683/2306-1758/2023-1-43/4-18.

1. Introduction

As computer networks have developed, the number of users who store critical data in the cloud system has increased, emphasizing the importance of network communication security. However, cyber security threats, such as system infiltration, sneaking network communication, or tampering and stealing essential assets, exist in various ways and forms behind the expansion of network communication, compromising the stability and economic feasibility of institutions and governments [1].

Intrusion detection is a representative way of preventing and responding to malicious behavior in network traffic, which is generally pre-trained network traffic characteristics and analysis results. Intrusion detection systems are categorized as either signature-based or anomaly-based intrusion detection, depending on how they identify attacks [2]. As current attack types diversify, rather than signature methods, anomaly-based intrusion detection approaches that leverage training to identify unknown traffic are deployed as effective technology [3]. One of the most recent abnormal data identification strategies is deep learning-based pattern recognition, such as artificial neural network (ANN)-based traffic classification.

Many researchers have investigated intrusion detection methods using deep learning to prevent and process unlabeled attack data in advance [2]. Since labeling by a security expert is expensive and time-consuming, the significance of semi-supervised learning-based network traffic classification with only a few labeled data has emerged [4]. In terms of practicality, nevertheless, it is problematic to implement conventional semi-supervised learning approaches for intrusion detection, such as predicting through the distribution of reconstruction errors generated from provided data information or class clustering in unlabeled data, is problematic [5]. For example, real-time datasets have low prediction rates because many unlabeled sub-datasets are continuously generated and accumulated, making it challenging to learn changes in the distribution of the dataset in a static way. Furthermore, performance might fluctuate depending on the amount of data or the proportion of labels for each subset [6]. Therefore, a dynamic semi-supervised learning mechanism is required to discover the real-time pattern flow of changing data.

Another issue is validating detection performance on unlabeled real-time data and substantiating the analytic results. Because the post-detection analysis process is crucial for monitoring and determining whether to identify intrusion traffic under the corporate policy, the outcome should be appropriately interpreted [7]. For this reason, tools and libraries for eXplainable Artificial Intelligence (XAI) based on ensemble algorithms such as Local Interpretable Model Agnostic Explanation (LIME) and SHapley Additive exPlanations SHAP have been widely utilized [8, 9]. Unfortunately, existing XAI techniques fall short of explicitly describing the output of generative or Generative Adversarial Network (GAN) based models due to the black box problems of deep learning models and the sluggish expansion of tools.

Therefore, in this paper, we introduce XA-GANomaly, a new intrusion detection model based on adaptive semi-supervised learning and explainable strategies. Specifically, we build the model for

an adaptive learning mechanism using GANomaly, which is invented for image anomaly identification, by simulating real-time data with different subsets and progressively training based on the results of the subsets. Moreover, to ensure the reliability of results, we propose an intrusion detection system (IDS) monitoring method based on XAI methodologies.

The primary contributions of this paper describe as follows:

i. We propose a novel adaptive training algorithm for a semi-supervised learning technique based on the reconstruction error distribution of GANomaly and test it on four small subsets to simulate a real-time dataset.

ii. We present a monitoring technique based on the new explainable adaptive semi-supervised learning method that combines SHAP, reconstruction error distribution, and t-distributed stochastic neighbor embedding (t-SNE).

iii. We evaluate and analyze the proposed model through various experiments, including semi-supervised learning for one-class classification, adaptive learning for the proposed model, and adaptive learning for randomly generated imbalanced datasets, and validate its performance via the proposed explainable methodology.

The remainder of this paper is explained as follows. We explore existing studies on semi-supervised learning and explainable techniques used in intrusion detection tasks in Section 2. We explain our proposed methodology for an explainable adaptive semi-supervised learning method using DNN-based GANomaly in Section 3. We demonstrate our diverse experimental results in Section 4. Finally, we discuss the limitation of our work for future research and summarize the main contributions in Section 5.

2. Related Work

2.1 Artificial Intelligence-based Intrusion Detection Technique

In previous studies, researchers have put efforts to improve the detection accuracy of intrusion detection mechanisms. Various machine learning and deep learning-based technologies have been explored to respond to the increasing varieties of attacks and complicated attack patterns. IDS can be broadly classified into three categories based on their training strategies: supervised, unsupervised, and semi-supervised learning [2].

2.1.1 Supervised and Unsupervised Learning

Supervised learning is a technique that trains data and classifies them into normal and abnormal classes using labeled target data. This approach is commonly used for intrusion detection tasks and involves machine learning algorithms such as support vector machine (SVM), k-Nearest-neighbor (k-NN), and ensemble tree models like Decision Tree (DT) and Random Forest (RF) [10-12, 13]. Deep learning models based on ANNs have also been used by stacking multiple layers in-depth. For example, Rosay et al. [16] presented Multi-Layer Perceptron (MLP)-based deep learning algorithms and achieved the highest result compared to other machine learning techniques. Kim et al. [18] introduced binary and multi-class denial-of-service (DoS) attacks based on Convolutional Neural Networks (CNNs) and showed examples of RGB 2D images transformed features of data. Yin et al. [19] utilized Recurrent Neural Networks (RNNs), one of the time-series models, and found that it performed better than traditional machine learning techniques and MLP. However, research on supervised learning methods has limitations because they have focused on enhancing performance, although it already shows a high accuracy of 99%.

On the other hand, unsupervised learning is an unlabeled training method that involves clustering data based on similar properties or densities. Chapagain et al. [20] researched to enhance the K-means technique by using a feature extraction with Principal Component Analysis (PCA). Generative models have also been suggested for classifying normal and abnormal data based on a threshold of reconstruction error values. Li et al. [22] used Auto-Encoder (AE) based on RF feature engineering, and Yao et al. [21] experimented with Variational Auto-Encoder (VAE) to develop the latent vector representation from AE based on a probabilistic concept. In recent years, Generative Adversarial Networks (GANs) have been suggested to build powerful generators by adversarial training [6]. For instance, Boppana et al. [23] constructed a hybrid model of AE and GAN to deal with IoT security

concerns and showed superior results compared to other one-class classification methods. However, unsupervised learning approaches might be challenging to apply in practice due to the time complexity of training a large amount of unlabeled data.

2.1.2 Semi-supervised Learning

In semi-supervised learning, both labeled and unlabeled data are used for training. Many researchers have explored various ways, including self-training, co-training, and multi-view learning in computer engineering [21]. Meanwhile, modifying supervised and unsupervised learning strategies have been researched for many years to tackle semi-supervised learning in the intrusion detection tasks. For example, Shah et al. [25] incorporate SVM and RF with a semi-supervised algorithm that trains and predicts with a small portion of datasets. However, this mechanism is only suitable for normal and attack-labeled data.

In contrast, generative models can be used to handle normal labeled data, reducing the need for time-consuming manual labeling. Such feature makes them well-suited for use in real-world industries. Hara et al. [26] developed an AE-based model with adversarial training for extracting classification patterns and characteristics from latent vectors. Their results showed better performance compared to traditional machine learning techniques. Similarly, Kale et al. [35] suggested a hybrid architecture that combined K-means, CNN, and GANomaly for each unsupervised, supervised, and semi-supervised learning. This approach resulted in higher accuracy than the one-class classification methods. GANomaly, in particular, is a powerful GAN model for detecting anomalies in images by capturing the underlying distributions of normal images using adversarial training between the generator and discriminator [25]. However, the main drawback of such hybrid models is that they require a significant amount of resources with different label types.

To address these challenges, we focus on GANomaly-based semi-supervised learning, which can handle real-time simulated data, even with a small normal labeled dataset. Furthermore, we aim to allow a single model to adapt and improve over time.

2.2 Adaptive Learning for Intrusion Detection

Recently, some adaptive learning algorithms have been developed to boost the performance of models used in research areas, such as fog computing and IoT [26]. For example, Gao et al. [30] introduced an adaptive MultiTree model that integrates multiple tree models and leverages a weighted voting technique to optimize detection results and enhance result quality. Hu et al. [31] presented a method for generating synthetic data to improve the effectiveness of CNNs in detecting attack traffic. Lin et al. [32] used an online technique to adaptively update the SDN framework to optimize the bagging algorithm and imbalanced network traffic. In contrast, the primary purpose of adaptive learning in the field of computer engineering field has been defined as adapting models to new forms of data [28]. Thus, our goal is to enable the model to adapt to unseen and unlabeled data, which is typically encountered in real-world scenarios.

2.3 Explainable Method Applied to Intrusion Detection

Explainable intrusion detection studies aim to achieve a high detection rate and low false positive rate performance [22]. Explanation skills also allow security professionals to understand the model's construction, guide the development of better models, and make appropriate decisions in identifying attack traffic [30]. For instance, Local Interpretable Model-agnostic Explanation (LIME) based interpretable intrusion detection, which operates regardless of the model type, can observe the result value where the predicted value changes for perturbation. Additionally, there are research cases that explain the built-in feature importance obtained by training a tree-based ensemble algorithm [9].

In contrast, SHAP, a recently introduced explainable AI that utilizes game theory and ranks by contributions, has gained attention. This is because it works well for both tree-based ensemble algorithms and MLP, allowing the measure and analyze contributions related to positive or negative influence [8, 30]. Since SHAP provides explicit interpretation in detail using deep neural networks, this paper aims to develop an explainable model by combining SHAP with two other explainable methods in order to examine the proposed adaptive semi-supervised learning algorithm.

2.4 Monitoring System in IDS

Monitoring systems for network traffic are generally referred to as Network Traffic Monitoring Analysis (NTMA) [31]. Early research has paid attention to the method for supervising the trans-

mission status and speed of network traffic. For example, focusing on cases where the bandwidth is out of the general range, pattern analysis such as File Transfer Protocol (FTP), and flow monitoring [7]. Furthermore, NTMA has been developed to effectively analyze complex traffic transmitted over expanding networks due to cellular and high-capacity video or image [31]. Representative deep learning models for performing specific analysis through traffic classification include CNNs that learn traffic characteristics and RNNs models that learn temporal characteristics [31]. However, the potential limitation of existing monitoring systems in IDS is that it is difficult to ensure network security solely based on the stability and efficiency of the traffics. Hence, this paper focuses on constructing a monitoring methodology using explainable techniques obtained from the proposed model to detect attack traffic.

3 Proposed Methodology

In this section, we introduce our novel intrusion detection framework based on adaptive semi-supervised learning that can dynamically train on small new datasets and a hybrid explainable intrusion detection technique to evolve it into an intrusion detection monitoring system.

3.1 Overall Framework

As illustrated in Fig. 1, the proposed framework is divided into five stages with a combination of three stages generally performed in previous studies [32-33] and the remaining stages developed in this paper. First, in the data preprocessing stage, the categorical data is encoded, the entire dataset is scaled, and essential features are selected. Second, to handle different dataset types, including multiple subsets, we divide datasets into labeled, unlabeled, and adaptive datasets. Third, in the semi-supervised learning phase, we first use DNN-based GANomaly to train a given normal dataset and test labeled abnormal dataset. The reconstruction error distribution is acquired using the gap between the trained normal data and the tested abnormal data. In order to distinguish the attack and non-attack labels, we determine the threshold as the median of reconstruction error from the normal dataset. Fourth, in the adaptive learning phase, we utilize our proposed adaptive algorithm to progressively add normal and abnormal data while assuming that several small subsets are continually accumulated. In the fifth stage, we validate the adaptive semi-supervised learning results using a combination of three explainable strategies: SHAP, reconstruction error visualization, and t-SNE. As a result, we integrate these analysis results into a monitoring dashboard system. The specific methodology is described in Section 3.2.

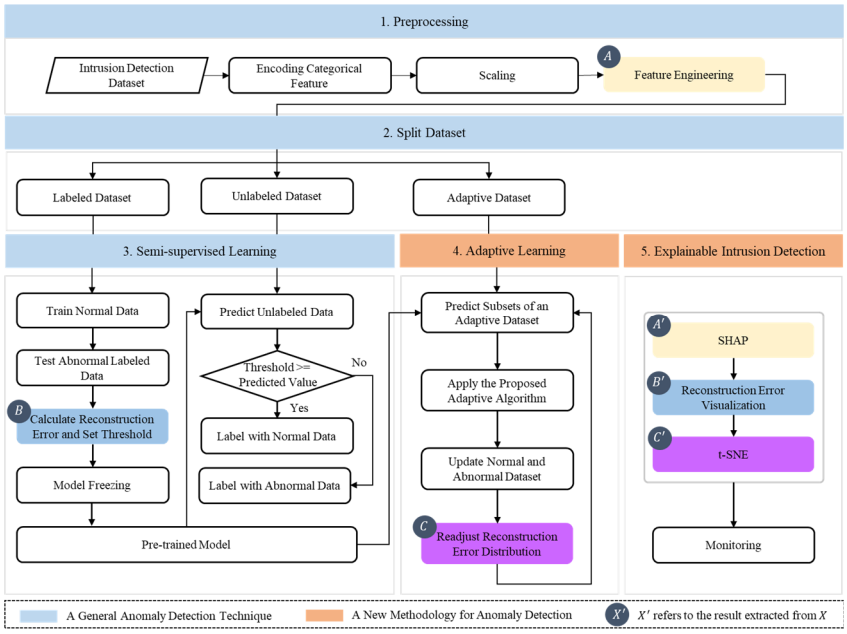


Figure 1 – Overall framework of the explainable adaptive intrusion detection based on the combination of previous semi-supervised learning techniques [32-33] and our proposed techniques

3.2 Explainable Adaptive Semi-supervised Learning Algorithm

To improve the ability of the GANomaly model to detect abnormal classes in images, we modify the DNN-based GANomaly model by training only critical features throughout the feature engineering process. Specifically, we stack fully-connected layers by halving their size using Rectified Linear Units (ReLUs) as the activation function for each layer. Additionally, we employ an L1-regularizer only on the second layer to add weights to the layer's output.

As shown in Fig. 2, the proposed explainable adaptive semi-supervised learning technique uses DNN-based GANomaly and three explainable techniques. The loss function ($\mathcal{L}$) used to optimize the training model is defined as Eq. (1)

$$\mathcal{L} = \|z - \hat{z}\|_2 + \|x - \hat{x}\|_1 + \|D(x) - D(\hat{x})\|_2 , \quad (1)$$

where x denotes as the input,

E is the encoder,

G is the generator,

D is the discriminator, and z is the latent vector extracted from the encoder.

The generator and discriminator are adversarially trained to obtain $\hat{x}$, which simulates the real input data x.

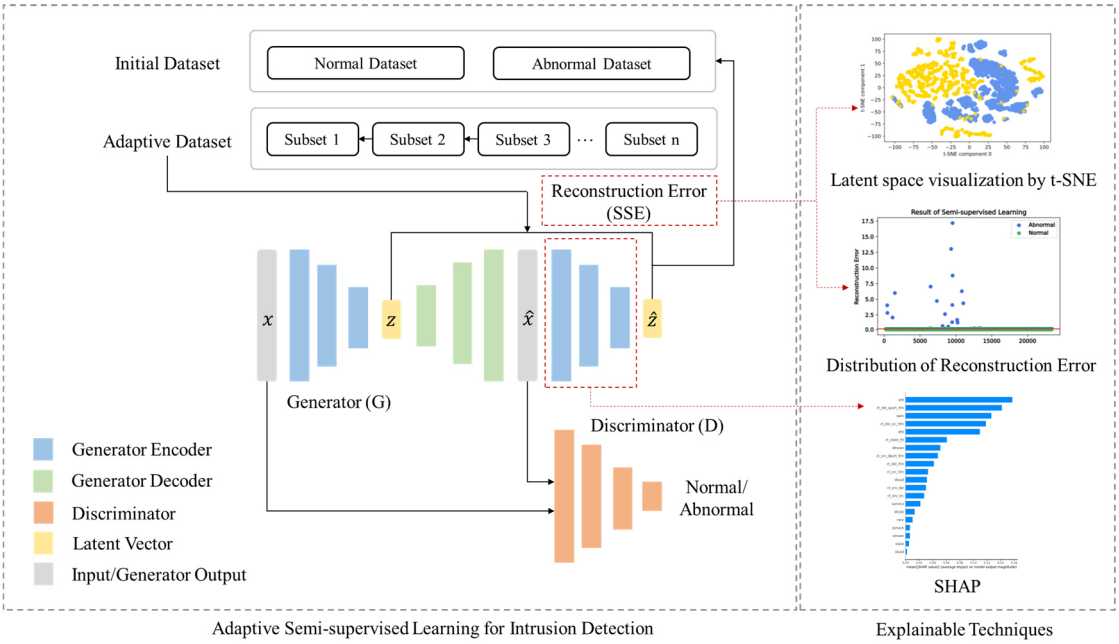


Figure 2 – Proposed explainable adaptive semi-supervised learning technique for intrusion detection using DNN-based GANomaly and three explainable techniques

Algorithm 1: Adaptive semi-supervised learning for intrusion detection

Input: Adaptive Dataset

1. Divide the adaptive dataset into n subsets
2. Get the median threshold from pre-trained DNN-based GANomaly

For each $subset_i$

Predict i^{th} subset from pre-trained G_E(x) and E(G(x))

Score = MinMaxScaler(Reconstruction error computed by SSE)

IF score < $\text{Threshold}_{\text{median}}$

Label to normal

IF score >= ($\text{Threshold}_{\text{max}} \times \alpha$)

Label to abnormal

Merge new normal and abnormal data into the $i-1^{th}$ subset

End For

Output: Adaptively updated dataset

We calculate the reconstruction error using Sum Square Error (SSE), which is defined as Eq. (2), between latent vectors and describe the distribution of normal and abnormal traffic data, focusing on the fact that latent space represents a distribution similar to that of the input [25]. The SSE output is used to update our adaptive dataset with predetermined labeled data based on certain criteria. Algorithm 1, which is based on double thresholds for the distribution that distinguishes normal from abnormal, demonstrates the specific processes and threshold criteria used. The algorithm assumes that data can be trained adaptively when training the difference between normal and abnormal reconstruction rates is significant. Each max and median value of the normal data reconstruction error rates is the criterion for accepting new abnormal and normal data, respectively.

$$SSE = \sum_{i=1}^{n}(z_i - \hat{z}_i)^2 \quad (2)$$

After the adaptive semi-supervised learning process, subsets of the data are collected to create a single dataset log. This log includes provided instances of abnormal behavior and is used to monitor and improve the adaptive learning process. To gain further insights into the data, we propose an adaptive semi-supervised learning method capable of interpreting a combination of SHAP, reconstruction error distribution, and t-SNE.

First, we use SHAP to express the importance of features in rank order based on their contribution, which is useful in the feature engineering stage that follows adaptive learning. Unlike other feature importance techniques, such as ensemble-based feature importance, SHAP takes into account the relationship between features and can provide more accurate results even when multicollinearity between explanatory variables exists. It calculates the contribution of features based on game theory to determine feature importance by aggregating multiple qualities and computing the Shapley Value, which represents the average change in the predictions of the model due to the presence or absence of specific features, as shown in Eq. (3).

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)! \times \left[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)\right]}{|F|!}, \quad (3)$$

Where $\phi_i, F, S, f_{S\cup\{i\}}$, and $f_s(x_s)$ are referred to as Shapley Value for i^{th} data, entire dataset, rest of the data except i^{th} data, entire contributions, and rest of contribution except the i^{th} data, respectively.

Second, we employ reconstruction error-based visualization to show the threshold with normal and abnormal labels by displaying the distribution of reconstruction errors acquired from SSE. This procedure not only helps to alter the position of the threshold set in semi-supervised learning but also helps intuitively confirm the sufficiency of the threshold used within adaptive learning.

Third, we deploy t-SNE, a dimensionality reduction technique, which has been shown to outperform PCA in reducing high- to low-dimensional data. This technique is particularly well-suited for latent vectors extracted from image-based generative models, as it can express high-dimensional latent space. The reason is that vector visualization is conducted by transferring the distance between data into stochastic probability and then applying it to embedding while maintaining the distance between points in the original space.

Since the reconstruction error of our proposed adaptive algorithm means the difference between the latent vectors of the generative model, we use t-SNE. This visualization method can preserve the data distribution of the original dimension. In addition, the better the performance of the

adaptive algorithm, the higher the density between classes of data in t-SNE. Thus, we use t-SNE to demonstrate the SSE results of the proposed adaptive algorithm for each subset, by verifying the density between normal and abnormal classes.

Our ultimate goal is to monitor each stage by integrating these three approaches for each stage, such as feature selection, verifying threshold, and monitoring transforming data distribution. The details are described in Chapter 4 using a monitoring dashboard.

4 Experimental Results

In this section, we presented some experiments, analyzed them to clarify their significance, and compared similar techniques to validate our proposed methodology using two network traffic datasets, NSL-KDD and UNSW-NB15.

4.1 Experimental Environment Settings

All experiments were carried out on Ubuntu 18.04.5 LTS with an Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz and three GeForce RTX 3090s with 256GB of memory. Tensorflow and matplotlib, one of the well-known libraries for deep learning and visualization, were utilized for intrusion detection and monitoring visualization, respectively.

4.2 Datasets and Preprocessing

In this section, we briefly describe NSL-KDD and UNSW-NB15 datasets, which are publicly network intrusion detection datasets employed for performance evaluation in multiple existing studies. In addition, the same preprocessing approach was performed to suggest a universal framework for datasets with two different distributions. Specifically, labels in the dataset were modified to discern binary normal versus abnormal traffic, and categorical features were numerically encoded for training. In addition, among the scalers provided by scikit-learn, scaling was performed using MaxAbsScaler which is sensitive to outliers, to set the two different distributions to be the same. These two datasets were reconstructed based on three types of datasets as described in Section 3 as shown in Tab. 1. In detail, labeled, unlabeled, and adaptive datasets are divided into 50%, 30% and 20% of the total dataset, respectively. In addition, each adaptive dataset is divided into four subsets as represented in Tab. 2.

Table 1 – The information of datasets

Dataset		Labeled dataset	Unlabeled dataset	Adaptive dataset	Total
NSL-KDD	Normal	33,446	23,117	13,448	70,011
	Attack	29,540	20,448	11,747	61,735
	Total	62,986	37,791	25,195	125,972
UNSW-NB15	Normal	23,117	22,215	9,573	54,905
	Attack	18,049	2,484	6,894	27,427
	Total	41,166	24,699	16,467	82,332

Table 2 – The description of subsets of adaptive datasets

Dataset		1st subset	2nd subset	3rd subset	4th subset	Total
NSL-KDD	Normal	3,362	3,429	3,328	3,329	13,448
	Attack	2,936	2,995	2,908	2,908	11,747
	Total	6,304	6,424	6,236	6,237	25,195
UNSW-NB15	Normal	2,393	2,441	2,369	2,370	9,573
	Attack	1,723	1,758	1,707	1,706	6,894
	Total	4,116	4,119	4,076	4,076	16,467

4.2.1 NSL-KDD Dataset

NSL-KDD [34] is a dataset constructed by eliminating duplicate rows and minimizing the size of KDD'99 dataset. The dataset is divided into three subsets, such as KDDTrain+20%, KDDTrain+, and KDDTest+, and we selected KDDTrain+ which had the most records. This dataset contains 41 features: nine basic features derived from packet header information (1st to 9th), twelve payload content features (10th to 21st), and nine-time features representing the number and speed of traffic connections (22nd to 30th), ten statistical features (31st to 40th) related host counts, and a label column [35]. Among them, string types of features, such as protocol type, service, and flag, are converted into numeric values, and normal and abnormal labels are converted to 0 and 1.

4.2.2 UNSW-NB15 Dataset

UNSW-NB15 [36] is a large amount of network traffic data collected by Cyber Range Lab using the IXIA PerfectStorm tool to derive better simulation results than KDD'99 or NSL-KDD. The dataset consists of 49 features, of which 47 are independent variables, excluding the attack type, label, and binary label. In specific, the columns in the dataset are divided into five features related to traffic flow (1st to 5th), twelve intrinsic header features (6th to 18th), seventeen features related to payload content (19th to 26th), eighteen features related to the connection time between the source and target (27th to 35th), and eleven additional features related to various cases of time and connection (36th to 47th). The 48th feature, multi-types of attack, was not used to identify abnormal data, and proto, service and state in string form were encoded in the same way as preprocessing used in the other dataset.

4.2.3 Feature Engineering

Feature selection has been employed as an effective method to learn essential features or reduce the time complexity of training artificial intelligence-based models. In this paper, we selected 21 high-ranked features by evaluating the feature importance of an extra trees classifier, an ensemble algorithm-based tree model known to represent decision boundaries better and provide higher performance than a random forest classifier for classification tasks. The features related to the intrusion detection techniques will be described in Section 4.4.3 in detail.

4.3 Evaluation Metrics

Several metrics are available to evaluate the intrusion detection performance of categorizing normal and abnormal traffic. Accuracy and F1-score are among the most powerful evaluation metrics expressed in Eqs. (4)-(5). Accuracy represents overall performance, whereas the F1-score denotes the probability of a false positive value and a false negative value according to precision and recall mathematics representation according to Eqs. (6)-(7). Since mispredicted abnormal traffic might be harmful to protect network security, the higher the F1-score, the more robust the detector.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + Fasle\ Negative} \tag{4}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

$$Precision = \frac{True\ Positive}{True\ Positive + Fasle\ Positive} \tag{6}$$

$$Recall = \frac{True\ Positive}{True\ Positive + Fasle\ Negative} \tag{7}$$

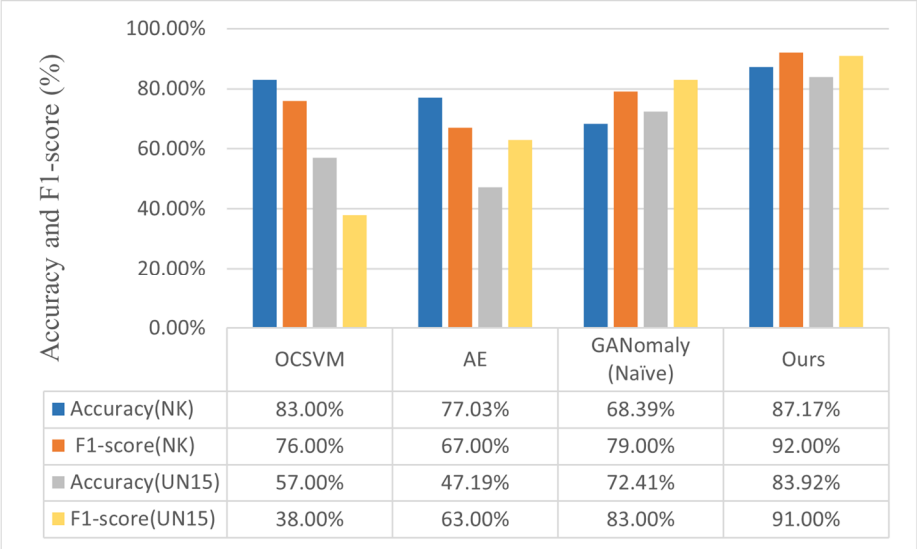
4.4 Performance Evaluation and Analysis

In this section, we provide the experimental comparison of the performance models utilized in our presented adaptive intrusion detection approach, as well as visualizations and analysis of the results obtained from integrating three methodologies for explainable model development. First, we examined the performance of various semi-supervised learning techniques for anomaly detection in order to select a model for our adaptive intrusion detection system. We then presented the outcomes of our proposed adaptive intrusion detection system using the chosen model. Finally, we performed

a detailed analysis of the model from three perspectives: SHAP, t-SNE, and reconstruction error visualization.

4.4.1 Performance of the semi-supervised learning techniques

To establish a baseline model for our proposed adaptive semi-supervised learning technique, we trained the given normal datasets using one-class classification models such as One-class Support Vector Machine (OCSVM), auto-encoder (AE), and CNN-based GANomaly. We then predicted unlabeled datasets, including both normal and abnormal traffic and evaluated their performance using metrics such as accuracy and f1-score. As illustrated in Fig. 3, our proposed DNN-based GANomaly model demonstrates superior performance compared to the other three techniques. Specifically, the results show an improvement of 13% and 8% of f1-score and 4.17% and 11.51% accuracy for the NSL-KDD and UNSW-NB15 datasets, respectively. With decreased false positive and false negative rates, we conclude that our proposed method is a well-trained model for both normal and attack traffic.



	OCSVM	AE	GANomaly (Naïve)	Ours
Accuracy(NK)	83.00%	77.03%	68.39%	87.17%
F1-score(NK)	76.00%	67.00%	79.00%	92.00%
Accuracy(UN15)	57.00%	47.19%	72.41%	83.92%
F1-score(UN15)	38.00%	63.00%	83.00%	91.00%

Figure 3 – Performance comparison of the semi-supervised learning models on the NSL-KDD(NK) and UNSW-NB15(UN15) datasets

4.4.2 Performance of the Adaptive Semi-supervised Learning Algorithm

To evaluate the effectiveness of the proposed adaptive algorithm for the continuous accumulation of subsets, we applied the algorithm to four subsets of data as shown in Tab. 2, and updated the results sequentially. We compared the performance of the proposed DNN-based GANomaly model with that of AE and CNN-based GANomaly, which are not reconstruction error-based classification techniques.

To begin with, we performed initial tests on semi-supervised learning. Then, we applied the proposed adaptive algorithm to the four sequentially incoming traffic subsets. As shown in Tab. 3, the results for the first subset, which is based on the threshold of the one-class classification model, did not show a significant change. However, the results for the second subset improved significantly compared to the initial test results, and further improved for the fourth subset.

Furthermore, we also performed imbalanced dataset classification to evaluate the proposed adaptation algorithm in different scenarios. Each of the four subsets was randomly partitioned, and normal and attack labels were allocated disproportionately. As shown in Fig. 4, the results of the proposed algorithm outperformed those of the initial subset. However, in some of the second subset results, the scores were higher than the result because certain subsets had more data with abnormal data. Therefore, while our proposed algorithm shows high overall performance, it is sensitive to the number of data in unbalanced datasets.

Table 3 – Increased percentage of initial test results when using the proposed adaptive learning algorithm

Dataset	NSL-KDD (%)			UNSW-NB15 (%)		
Method	AE	GANomaly (Naïve)	Ours	AE	GANomaly (Naïve)	Ours
Initial test-set	77.03%	68.39%	87.71%	47.19%	72.41%	83.92%
1st subset	+0.00%	+0.00%	+0.00%	+0.17%	+0.00%	+0.00%
2nd subset	-0.02%	+10.97%	+7.23%	+0.17%	+11.15%	-0.29%
3rd subset	+1.96%	+6.23%	+7.23%	+5.98%	+8.75%	+6.39%
4th subset	+1.70%	+7.95%	+5.46%	+13.68%	+14.45%	+7.95%

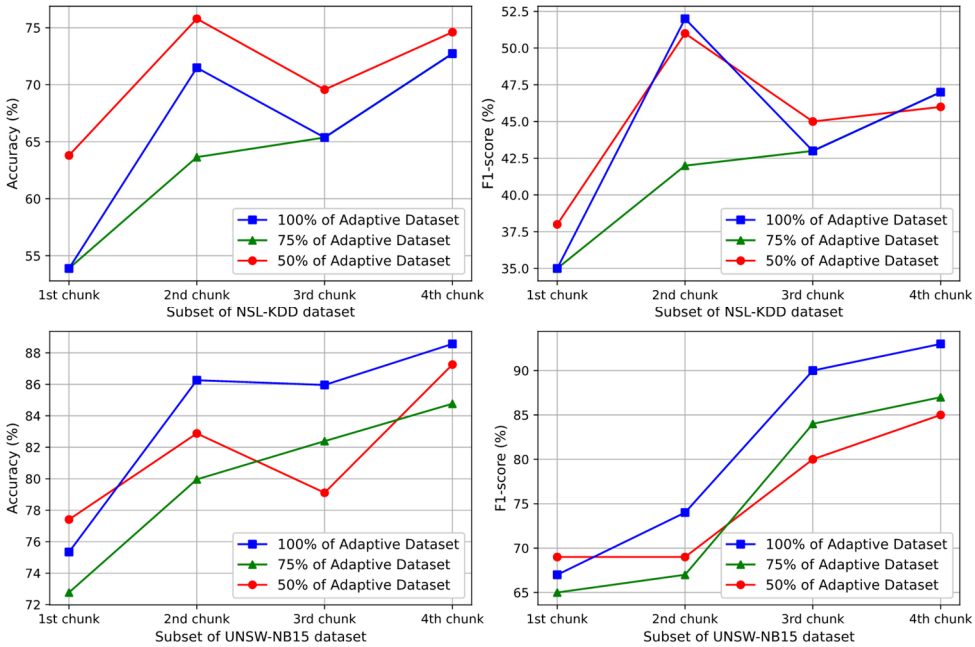


Figure 4 – Intrusion detection performance for 100%, 75%, and 50% of random adaptive datasets

4.4.3 Model Analysis for Hybrid Explainable Intrusion Detection Techniques

In order to perform explainable adaptive semi-supervised intrusion detection, we analyzed the outcomes of using SHAP, reconstruction error distribution, and t-SNE distribution, as shown in Fig. 2. Because SHAP is effective for undertaking feature engineering following adaptive training, we demonstrated feature importance for NSL-KDD and UNSW-NB15 in Fig. 5. The figure shows that traffic-related information being more important than that of basic packet header. Specifically, the NSL-KDD dataset is highly influenced by connection speed, as seen in the top features: "count", "dst host srv serror rate", "level", "dst host same srv rate", "logged in", and "srv serror rate". The UNSW-NB15 dataset is affected by both the TTL value and connection speed, with the top five features being “sttl”, “ct dst sport ltm”, “swin”, “st dst src ltm”, and “dttl”.

The reconstruction error distribution and adaptive t-SNE visualization are displayed in Fig. 6. In the top right corner of Fig. 6, there are a few blue dots above the threshold that represent attack data. Because the distribution of normal and abnormal is so varied, security experts might perceive this model as having been well-trained on the data set. Additionally, security experts could examine the specific range of data that displays a yellow alert, indicating anomalous data that is close to normal data, and track those logs to prepare for future attacks.

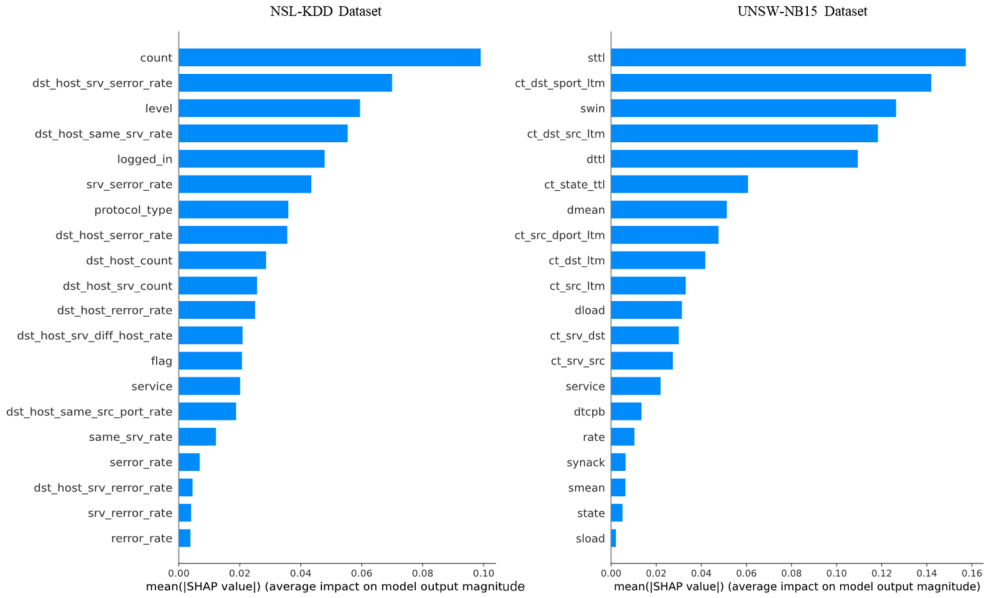


Figure 5 – Feature importance of SHAP for NSL-KDD and UNSW-NB15 dataset

The bottom of the figure shows how the latent space changes dynamically after training with the adaptive method. In the second subset, we observe overlapping portions, whereas in the fourth subset, a clear separation between the normal and abnormal distributions is shown, indicating that the adaptive algorithm's results are satisfactory. As a result, security experts could quantitively examine the performance of the adaptive learning algorithm.

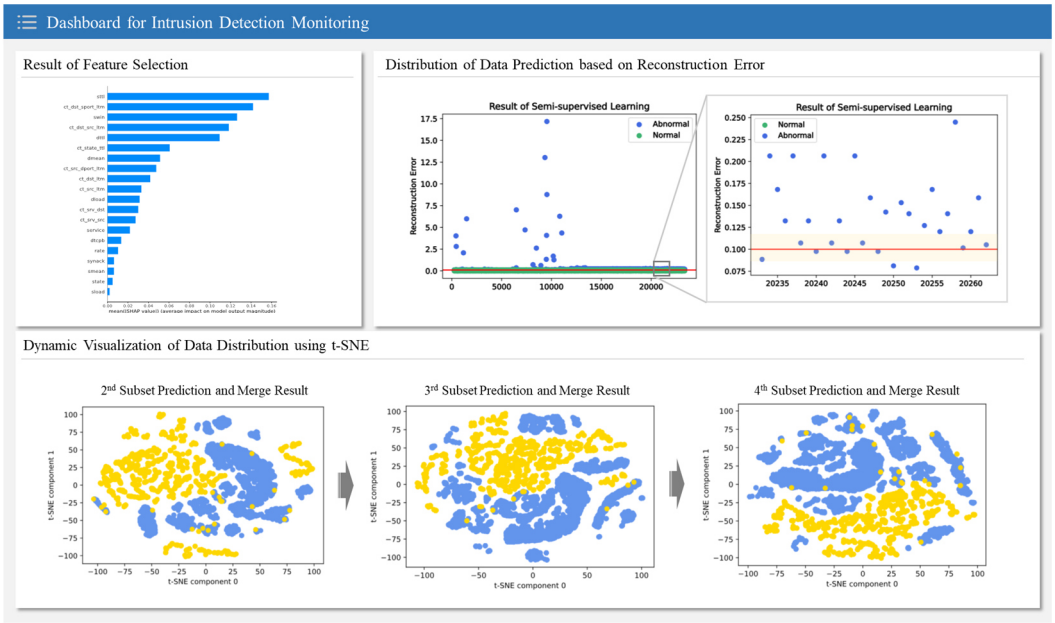


Figure 6 – Example of intrusion detection monitoring screen for NSL-KDD dataset

5 Discussion and Conclusion

The constantly evolving nature of malware and intrusion traffic poses a significant challenge for traditional training models. It is also challenging to understand the results of detection on unlabeled real-time data. This study proposes the XA-GANomaly model, an adaptive semi-supervised learning approach that can learn from small subsets of real-time data and constantly update its detection capabilities to takle these problems. Furthermore, three interpretive strategies for analyzing unlabeled network traffic data were developed: SHAP, reconstruction error distribution, and t-SNE. In

sequence, these can be used to analyze features, semi-supervised learning results, and adaptive learning results, in sequence. As a result, the XA-GANomaly model surpassed other one-class classifications, highlighting the potential for continuously improving detection results through adaptive training. Additionally, a dashboard display was created to provide security professionals with visual insights. However, future work should focus on developing the adaptive algorithm robust to imbalanced datasets. Nonetheless, this study presents a novel adaptive algorithm with the potential to respond quickly to real-time attack traffic and has industrial applicability through its intuitive interpretation and visualization capabilities.

Global economy with supervised learning has seen a shift in advanced integration which can be better understood with values and functionalities.

Acknowledgement: I would like to thank Y. Han and H. Chang for an extended version of their poster [42] "SEEM: A Method for Training Sequentially Enhanced Ensemble Models for Intrusion Detection" submitted to WISA 2022.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References:

1. R. AI Nafea and M.A. Almaiah, "Cyber security threats in cloud: Literature review", In 2021 International Conference on Information Technology (ICIT), Shanghai, China, pp. 779-786, 2021.

2. A. Khraisat, I. Gondal, P. Vamplew and J. Kamruzzaman, "Survey of intrusion detection systems: techniques, datasets and challenges.", Cybersecurity, vol. 2, no. 1, pp. 1-22, 2019.

3. V. V. R. P. V. Jyothsna, R. Prasad and K. M. Prasad, "A review of anomaly based intrusion detection systems.", International Journal of Computer Applications, vol. 28, no. 7, pp. 26-35, 2011.

4. J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning." Machine learning, vol. 109, no. 2, pp. 373-440, 2020.

5. I. Zliobaite, A. Bifet, M. Gaber, B. Gabrys, J. Gama, L. Minku and K. Musial, "Next challenges for adaptive learning system." ACM SIGKDD Explorations Newsletter, vol. 14, no. 1, pp. 48-55, 2012.

6. Z. Lin and D. Hongle "Research on SDN intrusion detection based on online ensemble learning algorithm." In 2020 International Conference on Networking and Network Applications (NaNA), Haikou City, Hainan, China, pp. 114-118, 2020.

7. Thommandru, A., Espinoza-Maguiña, M., Ramirez-Asis, E., Ray, S., Naved, M. and Guzman-Avalos, M., 2021. Role of tourism and hospitality business in economic development. Materials Today: Proceedings.

8. C. So-In, "A survey of network traffic monitoring and analysis tools", Cse 576m computer system analysis project, Washington University in St. Louis., 2009.

9. S. Mane and D. Rao, "Explaining network intrusion detection system using explainable AI framework", arXiv preprint arXiv:2103.07110, 2021.

10. S. Patil, V. Varadarajan, S. M. Mazhar, A. Sahibzada, N. Ahmed, O. Sinha and K. Kotecha, "Explainable Artificial Intelligence for Intrusion Detection System." Electronics, vol. 11, no.19, pp.3079, 2022.

11. J. Jha and L. Ragha, "Intrusion detection system using support vector machine", International Journal of Applied Information Systems (IJAIS), vol. 3, pp. 25-30, 2013.

12. M. A. M. Hasan, M. Nasser, B. Pal and S. Ahmad, "Support vector machine and random forest modeling for intrusion detection system (IDS).", Journal of Intelligent Learning Systems and Applications, vol. 6, no. 1, pp. 1-8, 2014.

13. T. T. H. Le, H. Kim, H. Kang and H. Kim, "Classification and Explanation for Intrusion Detection System Based on Ensemble Trees and SHAP Method.", Sensors, vol. 22, no.3, pp.1154, 2022.

14. I. Abrar, Z. Ayub, F. Masoodi and A. M. Bamhdi, "A machine learning approach for intrusion detection system on NSL-KDD dataset.", In 2020 international conference on smart electronics and communication (ICOSEC), Trichy, Tamil Nadu, India, pp. 919-924, 2020.

15. Sajja, G.S., Jha, S.S., Mhamdi, H., Naved, M., Ray, S. and Phasinam, K., 2021, September. An investigation on crop yield prediction using machine learning. In 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 916-921). IEEE.

16. A. Rosay, F. Carlier and P. Leroux, “MLP4NIDS: An efficient MLP-Based network intrusion detection for CICIDS2017 dataset.” In Machine Learning for Networking: Second IFIP TC 6 International Conference, MLN 2019, Paris, France, pp. 240-254, 2020.

17. Soham, S. and Samrat, R., 2021. Poverty and financial dearth as etiopathogen of psychotic and neurotic diseases. Заметки ученого, (4-1), pp.568-578.

18. J. Kim, J. Kim, H. Kim, M. Shim and E. Choi, “NN-based network intrusion detection against denial-of-service attacks.” Electronics, vol. 9, no. 6, pp. 916, 2020.

19. C. Yin, Y. Zhu, J. Fei and X. He, “A deep learning approach for intrusion detection using recurrent neural networks.” IEEE Access, vol. 5, pp. 21954-21961, 2017.

20. P. Chapagain, A. Timalsina, M. Bhandari and R. Chitrakar, “Intrusion Detection Based on PCA with Improved K-means.”, In Proc. ICEEE, Singapore, pp. 13-27, 2022.

21. R. Yao, C. Liu, L. Zhang and P. Pend, “Unsupervised anomaly detection using variational auto-encoder based feature extraction.” In 2019 IEEE International Conference on Prognostics and Health Management (ICPHM), San Francisco, CA, USA, pp. 1-7, 2019.

22. X. Li, W. Chen, Q. Zhang and L. Wu, “Building auto-encoder intrusion detection system based on random forest feature selection.”, Computers & Security, vol. 95, pp. 101851, 2020.

23. T. K. Boppana and P. Bagade, “GAN-AE: An unsupervised intrusion detection system for MQTT networks”, Engineering Applications of Artificial Intelligence, vol. 119, pp. 105805, 2023.

24. J. E. Van Engelen and H. H. Hoos, “A survey on semi-supervised learning.”, Machine learning, vol. 109. no. 2, pp. 373-440, 2020.

25. S. Shah, P. S. Muhuri, X. Yuan, K. Roy and P. Chatterjee, “Implementing a network intrusion detection system using semi-supervised support vector machine and random forest.”, In Proceedings of the 2021 ACM Southeast Conference, New York, NY, United States, pp. 180-184, 2021.

26. K. Hara and K. Shiomoto, “Intrusion detection system using semi-supervised learning with adversarial auto-encoder.”, In NOMS 2020-2020 IEEE/IFIP Network Operations and Management Symposium, Budapest, Hungary, pp. 1-8, 2020.

27. I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair and Y. Bengio, Generative adversarial networks, Communications of the ACM, vol. 63, no. 11, pp. 139-144, 2020.

28. S. Akcay, A. Atapour-Abarghouei, T. P. Breckon, “Ganomaly: Semi-supervised anomaly detection via adversarial training.”, In Asian conference on computer vision, Perth, Australia, vol. 11364, pp. 622-637, 2018.

29. S. Li, Y. Lu and J. Li, “Cad-ids: a cooperative adaptive distributed intrusion detection system with fog computing.”, In 2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD), Hangzhou, China, pp. 635-640, 2022.

30. X. Gao, C. Shan, C. Hu, Z. Niu and Z. Liu, “An adaptive ensemble machine learning model for intrusion detection.”, IEEE Access, vol. 7, pp. 82512-82521, 2019.

31. Z. Hu, L. Wang, L. Qi, Y. Li and W. Yang, “A novel wireless network intrusion detection method based on adaptive synthetic sampling and an improved convolutional neural network.”, IEEE Access, vol. 8, pp. 195741-195751, 2020.

32. Z. Lin and D. Hongle, “Research on SDN intrusion detection based on online ensemble learning algorithm.”, In 2020 International Conference on Networking and Network Applications (NaNA), Hainan, China, pp. 114-118, 2020.

33. M. Wang, K. Zheng, Y. Yang and X. Wang, “An explainable machine learning framework for intrusion detection systems.”, IEEE Access, vol. 8, pp. 73127-73141, 2020.

34. M. Abbasi, A. Shahraki and A. Taherkordi, “Deep learning for network traffic monitoring and analysis (NTMA): A survey.”, Computer Communications, vol. 170, pp. 19-41, 2021.

35. R. Kale, Z. Lu, W. K. Fok and V. L. Thing, “A hybrid deep learning anomaly detection framework for intrusion detection.”, In 2022 IEEE 8th Intl Conference on Big Data Security on Cloud, Jinan, China, 2022.

36. S. R. Khonde and V. Ulagamuthalvi, “Ensemble-based semi-supervised learning approach for a distributed intrusion detection system.”, Journal of Cyber Security Technology, vol. 3, no. 3, pp. 163-188, 2019.

37. M. Tavallaee, E. Bagheri, and W. Lu, “A detailed analysis of the KDD CUP 99 data set”, In IEEE symposium on computational intelligence for security and defense applications, Verona, Ny, USA, pp. 1–6. 2009.

38. S. Choudhary and N. Kesswani, “Analysis of KDD-Cup’99, NSL-KDD and UNSW-NB15 datasets using deep learning in IoT.”, Procedia Computer Science, vol. 167, pp. 1561-1573, 2020.

39. Rajendran, R., Sharma, P., Saran, N.K., Ray, S., Alanya-Beltran, J. and Tongkachok, K., 2022, February. An exploratory analysis of machine learning adaptability in big data analytics environments: A data aggregation in the age of big data and the internet of things. In 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM) (Vol. 2, pp. 32-36). IEEE.

40. N. Moustafa and J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)", In Military communications and information systems conference, National Convention Centre, Canberra, Australia, pp. 1–6, 2015.

41. Batool, A., Ganguli, S., Almashaqbeh, H.A., Shafiq, M., Vallikannu, A.L., Sankaran, K.S., Ray, S. and Sammy, F., 2022. An IoT and Machine Learning-Based Model to Monitor Perishable Food towards Improving Food Safety and Quality. Journal of Food Quality, 2022.

42. Y. Han and H. Chang, "SEEM: A Method for Training Sequentially Enhanced Ensemble Models for Intrusion Detection.", The 23rd World Conference on Information Security Applications, 2022.

43. Mehbodniya, A., Neware, R., Vyas, S., Kumar, M.R., Ngulube, P. and Ray, S., 2021. Blockchain and IPFS integrated framework in bilevel fog-cloud network for security and privacy of IoMT devices. Computational and Mathematical Methods in Medicine, 2021.

Received 08.02.2023
Статья поступила 08.02.2023 г.