electronics

MDPI

Article

Toward Developing Efficient Conv-AE-Based Intrusion Detection System Using Heterogeneous Dataset

Muhammad Ashfaq Khan and Juntae Kim *

Department of Computer Engineering, Dongguk University, 30 Pildong-ro 1-gil, Jung-gu, Seoul 100-715, Korea; ashfaq_jiskani@dongguk.edu

* Correspondence: jkim@dongguk.edu; Tel.: +82-2-2290-1421

Received: 24 September 2020; Accepted: 22 October 2020; Published: 26 October 2020

Abstract: Recently, due to the rapid development and remarkable result of deep learning (DL) and machine learning (ML) approaches in various domains for several long-standing artificial intelligence (AI) tasks, there has an extreme interest in applying toward network security too. Nowadays, in the information communication technology (ICT) era, the intrusion detection (ID) system has the great potential to be the frontier of security against cyberattacks and plays a vital role in achieving network infrastructure and resources. Conventional ID systems are not strong enough to detect advanced malicious threats. Heterogeneity is one of the important features of big data. Thus, designing an efficient ID system using a heterogeneous dataset is a massive research problem. There are several ID datasets openly existing for more research by the cybersecurity researcher community. However, no existing research has shown a detailed performance evaluation of several ML methods on various publicly available ID datasets. Due to the dynamic nature of malicious attacks with continuously changing attack detection methods, ID datasets are available publicly and are updated systematically. In this research, spark MLlib (machine learning library)-based robust classical ML classifiers for anomaly detection and state of the art DL, such as the convolutional-auto encoder (Conv-AE) for misuse attack, is used to develop an efficient and intelligent ID system to detect and classify unpredictable malicious attacks. To measure the effectiveness of our proposed ID system, we have used several important performance metrics, such as FAR, DR, and accuracy, while experiments are conducted on the publicly existing dataset, specifically the contemporary heterogeneous CSE-CIC-IDS2018 dataset.

Keywords: machine learning; deep learning; intrusion detection; spark MLlib; Conv-AE; big data

1. Introduction

Nowadays, the usage of the internet and its influence on each aspect of society has increased significantly, especially in the business industry. The connectivity to information and communication technology (ICT)-based system offers and allows the corporation to increase its productivity and activity. Recently, the availability of the Internet has increased greatly; therefore, almost every facet of our daily life is integrated with ICT at a comparatively low price. This means anyone can access any network with ease [1]. Along with such kind of improvement, several security problems in the digital world also have been increased due to the democratization of the internet [2], and therefore protecting the computer system from several threats has become a more concerning and vital research topic than before. So, ICT systems want incorporated and concrete security solutions. Regardless of the availability of various primary security solutions, such as firewalls, access control mechanisms, and antivirus, several ICT systems are still exposed to cyber threats that may prevent their functioning,

vulnerable private information, or facing data corruption problems. Although this conventional security mechanism serves as the first line of the security solution, these primary security techniques are inadequate to deal with intrusion skills and techniques. As data is the most important asset of the corporation [3], so there is an excessive need to devise an efficient security mechanism to keep the data secure and make ICT systems more tolerant and resistant to malicious attacks. To this end, we have proposed an efficient ID system, which is a useful security solution appropriate to mitigating malicious network attacks.

In the last two decades, numerous research has been conducted in the ID domain to develop effective NIDS using several approaches, such as statistical learning to conventional ML and current DL techniques. These methods have decent accuracy, aiming to perceive malicious threats, and have improved the speed of network traffic. However, the rapid growth of heterogeneous data has caused a big challenge [4,5]. ID frequently comprises the analysis of big data, which is considered a hot research issue where conventional computing techniques cannot deal with the quantity of data, such as network traffic [6]. Advanced security mechanisms, such as NIDS, must evaluate the gigantic network traffic packets in a real-time environment, as the correspondingly rapid growth of malicious threats can have catastrophic effects on basic security components, such as CIA (confidentiality, integrity, availability). Table 1 summarizes the challenge that is being faced by the ID system in a big data heterogeneous environment.

Table 1. Intrusion detection challenges in a big data heterogeneous environment.

Characteristic of Big Data	Description	Challenges
Volume	Size of the dataset in terabyte or petabyte, etc.	A huge capacity for network traffic is a massive problem for conventional computing approaches, and it is also a big issue for reducing the processing capability.
Velocity	Speed of data generation	Velocity is referred to as the particular speed at which traffic is created; it is also a big issue to handle high-speed network traffic in a real-time environment.
Variety	Dataset complexity, such as structured, unstructured format	Variety denotes the complexity of network traffic, especially when data packets in several formats and from various sources, so attribute selection is not a simple task, and some machine learning approaches are not appropriate for this issue.
Veracity	Data consistency and trustworthiness	Veracity denotes the correctness of data, data quality issues like noise or missing values.
Value	Data statistical, hidden, and unknown values	Value denotes in the sense that if specific data does not afford any meaning (value), it is not considered as big data analysis.

Every day, we experience extraordinary growth of data, which is the main contributor to big data in relation to volume, veracity, variety, velocity, and value [7,8], and brings its specific problems in the field of ID. There are several traditional approaches for ID, such as firewalls, access control, and encryption mechanism. These conventional approaches have a few constraints, particularly when facing a huge amount of suspicious threats like DDOS, and DOS and IDS can get high value of FN and FP attack detection rates. Recently, researchers have used ML and data mining approaches for ID with the desire to improving ID rates as compared to traditional security mechanisms. ID under a heterogeneous data environment has been recognized, and nowadays, the researcher has started to implement efficient big data analytics framework that can evaluate and monitor the network traces professionally [9,10]. The role of the dataset is very crucial; therefore, selecting suitable big data

analytics framework and using an adequate dataset for ID system evolution are two big challenges for developing efficient IDS [11].

Numerous studies on the background problem have highlighted several issues and challenges regarding ID, which need solid and immediate researcher's attention to be addressed. These issues and challenges are

- Intrusion detection algorithms.
- Deficiency or inadequate dataset.
- Integration of several formats of data.
- Poor system design.
- Big data processing framework
- Testing/evaluation of IDS.

However, there are various limitations and problems with ID studies. To start with, finding a huge amount of label data and handmade features is not an easy task, while obtaining unlabeled raw traffic data with a small amount of labeled data is a comparatively easy task. Therefore, the training process of several DL techniques, such as SAEs and DBNs, contains supervised fine-tuning and unsupervised pre-training [12]. In this scenario, a huge amount of unlabeled data and a small amount of labeled data are relatively executed in dual training processes. The outward drawbacks of these fully connected networks are having a huge amount of training parameters due to full connections of units among neighboring layers. As NN layers are limited, it may affect the training process by becoming very slow. Instead, the CNN DL approach decreases the number of training parameters through policies of shared weights and sparse connectivity, while CNN for the supervised learning process requires input labeled data. The real motivation behind this high-performance ID research is to propose an effective and suitable DL method for unsupervised features extraction and benefits of CNN for ID.

With the development of cyber-defense abilities, cyber-attacks have continued to develop to penetrate security defenses like living beings. Assuming the possibility of several enemy attacks, it is essential to choose a proper course of action by proactively evaluating and predicting the effects of a specific security incident. Cyber-attacks, particularly in large-scale military network environments, have a lethal impact on security; therefore, many tests and research must be carried out to create the required preparations. Nowadays, cyber-space is identified as the fifth battlespace after air, sea, space, and land. Therefore, by simply defending the information, cyber warfare can affect the military policies and actions that are specifically related to national security. Although the military is seeking to identify and minimize cyber-attacks to counter them, cyber-attacks are growing frequently, and new forms of threats are continuing to emerge [13,14]. It is important to examine the cyber threat that emerges in different ways to react effectively to it. The consequences of cyber-attacks on the infrastructure should be secured, and the security policies should be developed. Besides, it is important to examine not only current cyber threats but also the possibility of reacting more proactively. Various research has been carried out on cyber-attacks modeling, such as attack tree, attack graph, and cyber kill chain modeling approach, etc. [15,16]. Remember that prior research on cyber-attack modeling has induced some challenges, such as scalability in a large-scale network environment. Recently, cyber-attacks do not simply end with a single attack but have a composite form of several types of attacks. Moreover, new forms of cyber-attacks are taking place continuously. To deal with these challenges, a new approach for modeling method, which is flexible enough to easily add newly developing attack types and model complex attack process systematically, is required [17,18].

In particular, we have developed a better-quality type of ID system, which is built on the MLlib of Spark and deep learning approaches, such as Conv-AE. So, in this research, we have proposed a new ID system that combines the benefits of two systems to increase the performance as associated with the classical system. The important idea of this study is to design an ID system that is based on Spark MLlib and Conv-AE deep learning techniques. It is an innovative approach, which joins shallow and deep learning techniques to achieve their powers and overcome systematic overheads.

- Giving a comprehensive review of the advanced DL approaches in the ID domain.
- We have proposed the ID model, which is based on Spark MLlib and state-of-the-art DL approaches, such as Conv-AE, which concatenate deep and shallow networks to decrease their analytical overheads and exploit their advantages.
- We have analyzed packet capture files (pcap) directly on Spark, while earlier researchers have not assessed the raw packet dataset.
- How to resolve the class imbalance issue that is normally existing in the big data high-speed network?
- We study the performance of our proposed IDS using contemporary heterogeneous real traffic, CSE-CIC-IDS2018.
- We compare the performance of the proposed Spark MLlib and DL approach-based ID system with other classical ML approaches. The experiment outcomes describe that this approach is very efficient for attack detection and detecting misuse intrusions correctly in 98.20% of cases through a 10-fold cross-validation test.

The remainder of the paper is structured as follows: related works to the ID system are described in Section 2. The architecture of the proposed ID system with the framework descriptions are in Section 3. Implementation and experimental outcomes are covered in Section 4 with a comparative analysis with existing methods. Section 5 provides future directions before concluding the paper.

2. Related Work

In this section, we have discussed the current research that is related to our study. As shown in Table 2, various ID approaches have been developed in the last two decades, giving predictive accuracy on various datasets. ID technology is a crucial part of computer security, and the first idea was proposed by James Anderson in 1980 [19], wherein he proposed an ID framework for intrusion classifications to establish a security controlling structure that relies on identifying malicious user behavior. In recent times, comprehensive research has been existing to develop efficient IDs by using various techniques. These ID techniques, ranging from simple statistic algorithms to advanced ML approaches, have been useful in extracting features from network traffic so that abnormal traffic can be distinguished from the normal traffic.

Table 2. Summary of the related works using different approaches.

Reference	Approach	Accuracy	Dataset
Monshizadeh et al. [9]	Linear + Learning Algorithms	87.29	ISCX 2012
-	FS+DT + Variant of RNN	-	-
Monshizadeh et al. [9]	SAE + SVM	80.3	ISCX 2017
-	HAST-IDS	-	-
Naseer et al. [20]	DNN	89	NSL_KDD
Bandyopadhyay et al. [21]	DCNN	84.58	NSL-KDD
Tama et al. [22]	Two-stage ensemble	85.79	NSL-KDD
-	Deep VAE	-	-
Albahar et al. [23]	AMGA2-NB	93.3	UNSW-NB15
Tang et al. [24]	DBN + LR	97	KDD 99
Qatf et al. [25]	SAE + SVM	93.96	KDD 99
Qatf et al. [25]	DL method for IDS	84.96	NSL-KDD
Farahnakian et al. [26]	MCA + EMD	94.71	KDD 99
Thi-Thu et al. [27]	AE, KNN	95.33	ISCX 2012
Pektas et al. [28]	Linear + Learning Algorithms	96.6	ISCX 2012
Mighan et al. [29]	-	90.3	ISCX 2012
Meira et al. [30]	-	95	ISCX 2012
Wang et al. [31]	-	96.6	ISCX 2012

In previous research, Naseer et al. [20], Bandyopadhyay et al. [21], Tama et al. [22], Albahar et al. [23], Tang et al. [24], Qatf et al. [25], Farahnakian et al. [26], Thi-Thu et al. [27], Pektas et al. [28], Mighan et al. [29], Meira et al. [30], Wang et al. [31] used various models, methods, and techniques based on conventional ML supervised and unsupervised approaches have been introduced for ID problems to increase the performance of the ID framework. ML approaches, such as k-NN [32], SVM [33], ANN [34], RF [35,36], and many others, have been extensively used for ID. Laskov et al. [37] provided a study of unsupervised and supervised learning approaches according to their detection accuracy and ability to identify unknown malicious threats. Solanas et al. [38] described the clustering approach for anomaly ID. Ghorbani et al. [39] presented an inclusive review of unsupervised and supervised learning techniques for anomaly IDs. A comprehensive range of anomaly ID systems was described by Kalita et al. [40], and Tavallee et al. [41] analyzed the performance of various classical ML algorithms, including DT, NB, and SVM, etc. In general, ML techniques have brought accuracy and efficiency in the identification of malicious activities in the network traffic. However, few limitations remain in these ML approaches, such as the data preprocessing phase requiring expert knowledge, high FAR value, and low DR value of the attack, etc. ML approaches also require a huge amount of training data for efficient and reliable results, which is not an easy task, particularly in a vigorous and diverse environment.

Due to these deficiencies, DL algorithms have great importance in contemporary research. DL is the cutting-edge field of ML, which can highlight these deficiencies and can solve the problem associated with shallow learning. Earlier, researchers have proved that DL has a better performance as compared to shallow learning due to layer-wise learning features structure [42]. DL algorithms evaluate the network traffic deeply and efficiently recognize the intrusion in the network data. The nomenclature of previous work of DL and shallow learning in the ID domain was summarized by Hodo et al. [43]. Nowadays, the application of DNN for the solution of ID problems is a comparatively hot research area. AE, DBN, RBM, LSTM, CNN have been used for ID. Javaid et al. [44] employed softmax regression with sparse AE on the NSL_KDD ID data, which is an upgraded form of KDD 99 ID data. Fiore et al. [45] used an RBM DL to acquire an accuracy of 85% on the KDD 99 ID dataset. Jihyun et al. [46] utilized the LSTM DL method to identify malicious threats on KDD 99 dataset and claimed that they obtained better accuracy and attack DR as compared to conventional classifiers, such as SVM and KNN. Gao et al. [47] developed an ID architecture using DBN and evaluated the performance of the DL-based ID system using the KDD 99 ID dataset. Aygun et al. [48] proposed denoising the AE-based ID system and claimed to achieve attack classification accuracy up to 88.6% and 88.2% on the NSLKDDTEST+ dataset. Yousefi-Azar et al. [49] proposed AE-based latent features of the generation-based ID architecture using the NSLKDD ID dataset and obtained the ID accuracy up to 83.34%.

Hussain et al. [50,51] developed hybrid NIDS by joining Ada boost with DT and completed an experiment on the NSLKDD ID dataset, which is an upgraded version of the KDD 99 dataset, and the outcomes demonstrate that the hybrid approach is effective in identifying anomaly in the ID system. Nowadays, a substantial amount of research is done in the ID domain. Most of the researchers focus on improving the ID system's ability to identify malicious threats and enhancing the network speed that may be controlled. Ying Chung et al. [52], in his paper, proposed hybrid ID using the SSO approach and achieved attack classification accuracy up to 93% using the KDD99 ID dataset. Ghanem et al. [53] developed another hybrid ID architecture using a metaheuristic technique for an enormous dataset, where ID detection is based on a genetic algorithm and metaheuristic approach. Kim et al. [54] proposed a novel hierarchical ID system that joins anomaly and misuses the ID model via a decomposition structure. The misused ID model is developed using DT, while the anomaly ID model has been created via a one-class SVM method. They evaluated the proposed hybrid ID system in the NSLKDD ID dataset and claimed that it has better performance in terms of ID accuracy and low FPR for both anomaly and misuse attacks. Wang et al. [31] developed a hierarchical spatial-temporal-based ID called HAST-IDS, where low features are detected via CNN, and high features are detected through LSTM deep learning approach. The entire feature learning procedure is accomplished by DNN without

a feature engineering technique. This automatically features a learning process-based ID system evaluated in DARPA98 and ISCX2012 ID datasets and increases the ID accuracy and decreases the FAR as compared to traditional ID techniques. Chencheng et al. [55] proposed a distinct flow of features-based hybrid ID system, where CNN is used to evaluate the sequence of features, and DNN is used to learn various characteristics of high-dimensional features vectors comprising environmental and statistical features. They evaluated the performance of the distinct flow of features-based hybrid ID system by using the ISCX2012 ID dataset.

Monshizadeh et al. [9] combined learning and linear algorithms with a protocol analyzer to identify malicious activities in the network. Their linear and learning architecture is known as a hybrid anomaly detection module (HADM), where linear algorithms extract features, while the learning part of HADM uses these features to identify novel types of attacks. The protocol analyzer is used to filter and categorizes the susceptible protocols to evade a needless computational load. They tested the performance of the HADM ID system by using UNSW-NB15, ISCX2012, ISCX2017 ID datasets.

In the ID domain, most of the researchers use the KDD99 dataset, but being outdated, from this kind of dataset, we are not able to mitigate the threats, which are much new. Therefore, it becomes very substantial that IDS should evaluate and test inefficient and superior datasets [56,57]. So, in our study, we address the problem related to the dataset and evaluate the solution to solve them. Recent research shows that a hybrid approach solves various research problems in different domains, such as sentiment [58], video classification [59], emotion recognition [60], and malicious ID from a video [61]. In a big data environment, heterogeneous data are any data with high variability of data types and format. It may be of poor quality and ambiguous due to noise and missing values. It is a nontrivial task to use heterogeneous data in ID research. Therefore, dealing with a large volume of stream data, ranging from unstructured to structured, text stream to numeric, is also a big issue: real-time data stream, dynamic, and very heterogeneous. So, to solve the aforementioned issues and improve the learning capability and accuracy of IDS, we have developed DL-based IDS. In particular, we have developed an efficient ID system, which is built on Spark MLlib and DL approaches, such as Conv-AE networks. Spark MLlib-based typical ML techniques are useful to detect anomaly network traffic, and Conv-AE assists in detecting misuse network traffic. As we know that ABS and SBS have some restrictions, so we have joined the two systems to reduce their drawbacks. So, in this research, we have proposed a new ID system that combines the advantages of two systems to enhance the performance compared to the conventional system in the well-known modern real-time heterogeneous dataset CSE-CIC-IDS2018.

In a nutshell, the current research attempts to respond to the following research problem:

How to develop a fast, competent ID system to learn the useful features efficiently and automatically from large heterogeneous data by using state-of-the-art Conv-AE DL approaches and identify malicious attacks in the case of both anomaly and misused-based ID system, and how to overcome the FP with better attack detection rate.

3. Proposed Approach

3.1. The Proposed ID System

The proposed framework of the ID system is given in Figure 1. This proposed efficient ID system contains four main stages. The first stage of the proposed approach is preprocessing from the original ID dataset. The second stage is anomaly detection with conventional ML classifiers using Spark MLlib. In the third stage, Conv-AE deep learning approach is used for misuse detection. The final stage is the alarm module of the proposed approach, which detects whether the incoming network traffic is benign or malicious and evaluates the proposed ID system.

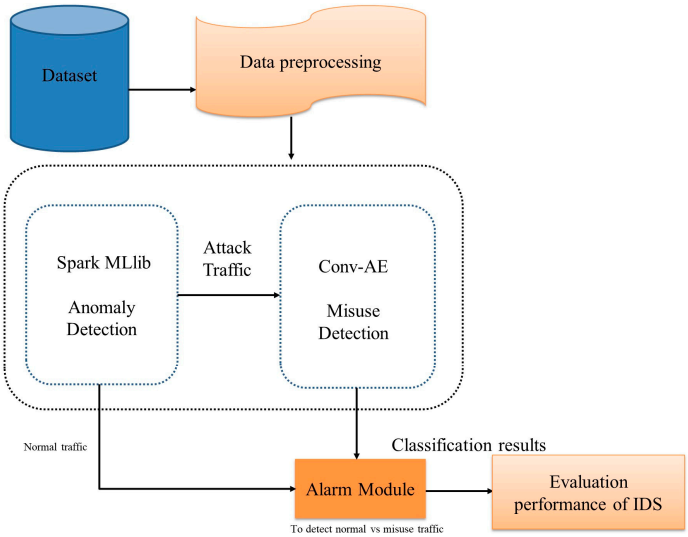

Figure 1. The overview of the proposed ID framework.

The proposed ID system combines benefits from the competent Spark MLlib with DL, utilizing the contemporary real-time heterogeneous CSE-CIC-ID 2018 dataset. The subsequent subsections describe every step-in detail.

3.1.1. Data Preprocessing

This is the first stage of the proposed ID framework. The CSE-CIC-IDS2018 consists of labeled flow for ten days. So, more than 80 attributes can extract from the raw ID dataset by applying CICFlowMeter-V3 and save these features in CSV format, which can be evaluated for the network traffic data. Initially, in CSE-CIC-IDS2018, few attributes have a slight influence on whether network traffic is benign or malicious, such as IP address and time stamp. As the ID system classifies network traffic according to their behavioral attributes, so we have erased this column of the attribute. Besides, the timestamp is not having a high impact on training the network, so we eliminate this attribute. After that, we have divided the dataset into the train test and validation set, which are 70%, 20%, 10%, respectively. The model is trained by using training data; testing is utilized for final assessment, while the validation set is useful for the fast assessment model. We know that CSE-CIC-IDS2018 is a real-world heterogeneous ID data that are usually inadequate: missing features values, missing particular features of interest, or comprising only cumulative data; noisy: covering outliers or errors; inconsistent: covering discrepancies in names or codes. Therefore, to handle the imbalanced issue, we have employed the over-sampling in which we increase the number of instances in the minority class by randomly duplicating them to present a higher representation of the minority class in the sample. Although it has some risk of overfitting the data, no information is lost. Nevertheless, it outperforms the under-sampling technique. The train and test dataset used in this study is given in Table 3.

Table 3. Training and testing data distribution.

Class	Attack Category	Flow Count	Training	Testing
Brute force	SSH-Brute force	230	184	46
-	FTP-Brute Force	611	489	122
-	Brute -Force -XSS	187,589	7504	1876
Web attack	Brute -Force -Web	193,360	15,469	3867
-	SQL-Injections	87	70	17
DOS attacks	DoS-attacks-Hulk	466,664	18,667	4667
-	DoS-attacks-SlowHTTPTest	139,890	55,956	13,989
-	DoS-attacks Slowloris	10,990	4396	1099
DDOS attacks	DDoS attacks-GoldenEye	41508	16,603	4151
-	DDOS-attack-HOIC	686,012	27,441	6860
-	DDOS-attack-LOIC-UDP	1730	1384	346
-	DDOS-attack-LOIC-HTTP	576,191	23,048	5762
Bot	Bot	286,191	11,448	2862
Infilteration	Infilteration	161,934	6478	1620
Benign	-	12,697,719	50,791	12,698
Total	-	15,450,706	231,127	57,782

The significant perception of this research is to test the reliability of the efficient ID system against unknown malicious threats via misuse attack detection technique. Table 4 designates the dataset for the Conv-AE deep learning approach for misuse classification of the testing and training the network.

Table 4. Data distribution for misuse attack classification.

Input	Features	Attack Category
Train set	80	Web attack, DOS attacks, DDOS attacks, Bot, Brute force, Infiltration.
Test set	80	Web attack, DOS attacks, DDOS attacks, Bot, Brute force, Infiltration.

3.1.2. The Anomaly Detection Module

In this stage of the proposed ID system, we have used a machine learning library of SPARK to implement several conventional ML classifiers, such as LR, DT, SVM, and RF, to classify malicious traffic for anomaly detection. In this stage, we have divided the dataset into two subsets—80% for training and 20% for tests. Then, conventional ML classifiers are trained on the training set to detect malicious and normal traffic. This is the overall binary learning stage. The trained conventional ML classifiers are tested on the test dataset. During this stage, the best performing model is selected due to grid search hyperparameter tuning and 10-fold cross-validation.

3.1.3. Misused Detection Using Conv-AE Deep Learning Approach

In this stage, Conv-AE is used for identifying the misused traffic, with an objective to classify the anomalous traffic further into relevant classification policies: DOS attacks, DDOS attacks, bot, brute force. Conv-AE merges the advantages of CNN and unsupervised pretraining AE. The micro overview of Conv-AE is shown in Figure 2. Initially, CNN has two fundamental components: classification and feature extraction. The feature extractor component contains two layers, known as convolutional and pooling layers. In this way, CNN learns features efficiently as output from the extraction component, which is commonly recognized as features map become the input to other components, which is called classification. However, instead of fully connected layers, the encoder consists of the convolutional layers, and the decoder consists of the deconvolutional layers. After the decoding part is fully connected, the softmax classifiers are added to the end for probability distribution over the classes. Here, the trained model is tested to determine whether the behavior of the trained model is malicious or normal, with the test set as one of the inputs.

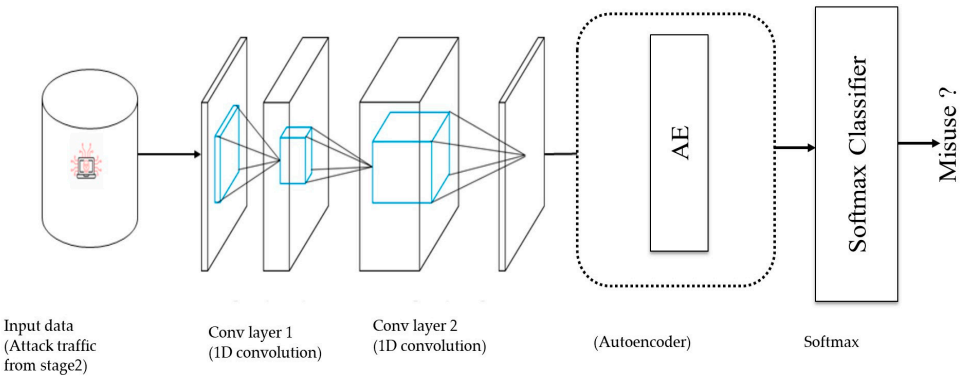

Figure 2. The micro overview Conv-AE.

Then, we randomly divide the data into training and testing—80% and 20%, respectively. The 10% from the training dataset is used for the validation test. During the training, the network is a fine-tune by optimizing AdaMax, Adam, and Ada Gard, with flexible learning rates, and these are optimized with grid search hyperparameters using several combinations and 10-fold cross-validation on 128 batch size. We analyze the network performance by adding a Gaussian noise layer after Conv layers to enhance the overall model generalization ability and overcome the overfitting problem.

3.1.4. Alarm Module

The last stage of the proposed ID system is the alarm module, which interprets the results of the events on both the anomaly and misuse detection stage. It is the final component of the proposed ID system, which helps the administrator or end-user after getting any malicious information that something has happened in the network.

4. Implementation Details

To show the efficacy of the proposed ID system on the contemporary heterogeneous dataset CSE-CIC-IDS2018, we have done various experiments. We have discussed in detail in the below sections.

4.1. Datasets

Since choosing suitable data to test an ID system plays significant roles, we make the ID data before we describe the simulation details of the proposed ID system.

Even though there have been numerous standard ID datasets publicly existing, some of them comprise the undevitrified, old-fashioned, irreproducible, and inflexible intrusion. To reduce the deficiencies of ID datasets, we have used most contemporary heterogeneous ID datasets, such as CSE-CIC-IDS2018, for our proposed high-performance efficient ID system [42]. This dataset is prepared by a collaborative project between the CIC and the CSE. The ID data consist of seven distinct attack states over a huge network for 10 days, such as a botnet, DDOS, brute force, web attack, DOS, infiltration, and heart leech attack.

- Botnet attacks: Many Internet-connected devices are used by a botnet owner to accomplish many tasks. It can be utilized to steal data, send spam, and permit the attacker access to the device and its connection. These kinds of attacks are collected through keylogging and screenshot.
- DDOS attacks: It typically happens when several systems flood the bandwidth or resources of a victim. Such an attack is often the result of many compromised systems (for example, a botnet) flooding the targeted system by making the enormous network traffic. These kinds of attacks use LOIC for TCP, UDP, and HTTP.
- Brute force attacks: This is one of the most widespread attacks that only cannot be used for password breaking but also to discover hidden content and pages in a web application. It is simply

a hit, attempting an attack, and then the victim succeeds. These kinds of attacks are collected through SSH and FTP Patator tools.

- Web attack: These kinds of threats are coming out every day, and now people and organizations take security seriously. It uses the SQL injection, in which an attacker can make a string of SQL commands and then use it to force the database, respond to the information, cross-site scripting (XSS), which is happening when developers don't test their code properly to identify the possibility of script injection, and brute force over HTTP, which can try a list of passwords to find the administrator's password. The web attacks are collected via DVWA and in-house selenium framework (brute force and XSS).
- DOS attacks: The attacker requests to make a computer network resource inaccessible for the time being. It is usually proficient by flooding the intended network resource or machine with superfluous requests in a try to overload systems and avoid few or all authentic requests from being fulfilled. They use the goldeneye, hulk, slow HTTP test, and slow loris to gather these kinds of attacks.
- Infiltration attacks: The infiltration of the network from inside normally takes advantage of a vulnerable application, such as Adobe Acrobat Reader. After effective exploitation, a backdoor is performed on the victim's machine and can conduct diverse attacks on the victim's network. They apply port scan and Nmap to gather these sorts of attacks.
- Heart leech attacks: It comes from a bug in the OpenSSL cryptography library, which is a commonly used transport layer security (TLS) protocol implementation. It is usually exploited by sending a malformed heartbeat request with a small payload and wide length field to the vulnerable party (usually a server) to evoke the victim's response. It is a kind of DOS attack.

There are more than 80 attributes that can extract from the raw ID dataset by applying CICFlowMeter-V3 and save these features as CSV format, which can be evaluated for the network traffic data. To extract novel features of data, the inventive files (logs and pcap) are also accessible, which can be utilized to extract features. Some of the CSE-CIC-IDS2018 features are given in Table 5.

Table 5. List of CSE-CIC-IDS2018 extracted features via CICFlowMeter-V3.

Feature	Explanation	Data Types
fl_dur	It represents the flow duration	String
Dst port	It represents the destination port number	Integer
Time Stamp	It denotes the time stamp particular flow	Integer
protocol	It denotes protocol	Integer
tot_fw_pk	Total number of packets in the onward direction	Integer
tot_bw_pk	Total number of packets in the back direction	Integer
tot_l_fw_pkt	Overall packet size in the onward path	Integer
fw_pkt_ l_max	The network packet maximum size in an onward way	Float
fw_pkt_l_min	The network packet minimum size in a forward way	Float
fw_pkt_l_avg	It represents the network packet average size in an onward way	Float
fw_pkt_l_std	It represents the network packet standard deviation size in onward route	Float
Idle_Max, Min	It represents max and min traffic flow as idle before it becomes active	Float

4.2. Performance Parameters

As models are trained, we have analyzed them by test. Then, performance metrics are computed through the confusion matrix. The predicted and expected classification is represented with the help of the element of the confusion matrix. The outcomes of classification are divided into two classes, such as incorrect class and correct class. There are four critical scenarios to calculate the element of the confusion matrix. We have the confusion matrix in the ID setting as shown in Table 6.

- True-positive (TP)—It is signified by x, and it presents that model is accurate as normal and predicts positive.

- False-negative (FN)—It signifies the wrong prediction and is represented by y. It classifies instances, which are anomalous in certainty, as regular, and the model mistakenly predicts negative.
- False-positive (FP)—It is represented by z and presents that the model incorrectly predicts positive, and in reality, the number of identified attacks is normal.
- True negative (TN)—It is represented by t and states that the instances that are properly detected as an attack predicts negative.

Table 6. Confusion matrix for the proposed ID system.

-	Predicated		
Actual	Normal	TP	FN
	Anomaly	FP	TN

We can calculate the performance of the proposed ID system by using the above conditions of confusion matrix as DR, and TPR and FAR are the two significant and general parameters for the evaluation of IDS. DR means the percentage of anomalous classes recognized by the ID model. FAR means the amount of misclassified normal classes.

$$DR = TP/(TP + FN) = x/(x + y) \quad (1)$$

$$FAR = FP/(TN + FP) = z/(t + z) \quad (2)$$

4.3. Experimental Settings

The initial anomaly detection stage is implemented in Scala with Spark using conventional ML classifiers, while for misuse detection, using Conv-AE is implemented with Keras in python. The experiment is carried on a PC having 64-bit ubuntu14.04 OS with a Core i7 processor and 32 GB RAM. The stack of software consists of Java 1.8 (JDK), Apache Spark v2.3.0, Keras, and Scala 2.11.8. We use 80% of the data for training purposes with 10-fold cross-validation and assess the performance of the trained network, with 20% held over the dataset. The deep learning Conv-AE model is trained on Nvidia TitanX GPU with cuDNN and CUDA in Keras to make the whole training process smooth and fast.

4.4. Evaluation of the ID System

Table 7 illustrates the performance of the proposed ID system for anomaly detection using conventional ML classifiers and for misuse detection using the DL Conv-AE approach. The best results are given in the table, which are obtained through random search only. As results show that LR performs off-color, giving low attack detection accuracy, while RF and SVM perform better, giving attack detection accuracy up to 89%.

Table 7. Performance of the proposed ID system.

Classifier	Precision	Recall	F1-Score	FAR	DR	Stage
LR	0.6630	0.6310	0.64660433	14.47	0.64	2
DT	0.7730	0.82122	0.79638075	10.35	0.82	2
SVM	0.8729	0.8423	0.85732704	9.45	0.85	2
RF	0.9019	0.8845	0.89311526	6.85	0.89	2
Conv-AE	0.9835	0.9820	0.98274943	0.81	0.98	3

The most important improvement in misuse attack detection up to 98.20% is with the Conv-AE approach. The main reason behind the significant enhancement in the performance of the proposed ID system is the superior feature extraction through CNN and AE deep learning approach.

4.5. Overall Analysis

Table 8 is the comparison of the proposed ID system with current solutions on the heterogeneous CSE-CIC-IDS2018 dataset. The CSE-CIC-IDS2018 dataset is produced later as compared to KDD99 and DARPA; therefore, few experimental results are existing for comparison. So, based on existing evaluation outcomes for the comparison, the best one from each study has been chosen relative to the attack detection accuracy.

Table 8. Comparison of the proposed ID system with existing solutions.

Reference	Approach	DR (%)
Ferrag et al. [11]	DNN	96.5
Peng et al. [62]	LSTM	96.2
Ana et al. [63]	BLS	97.0
Lee et al. [64]	SM LSTM	88.69
Chadz et al. [65]	HMM	97.0
Our approach	Spark ML + Conv-AE	98.20

Previous researchers such as Ferrag et al. [11], Peng et al. [62], Ana et al. [63], Lee et al. [64], Chadz et al. [65] used various ML and DL approaches, while we used hybrid approach in in ID domain. It can be noted that our proposed approach for anomaly and misuse attack detection is better as compared to advanced approaches in terms of attack detection accuracy. It is mainly due to the cutting-edge feature selection approach; we implement a machine learning library of Spark and Conv-AE deep learning approach. It is essential to note that this comparison with other approaches is for reference only because various research have used various preprocessing methods and distinct types of traffic proportions, as well as data distribution techniques.

We concentrate on resolving real-life ID problems, using enormous data analysis models (Apache Spark, Apache Hadoop) and AI (ML, DL). Controlling this kind of issue is not a simple task because of time and space restrictions. Big data presently has very huge and increasing volumes but still needs a huge powered computational device to support a learning framework that can handle the data proficiently, using specialized resources. Therefore, we can achieve better security with the proposed ID system using the Spark MLlib data analytics framework with DL Conv-AE against malicious threats.

5. Conclusions and Outlook

In this research, the ID system is developed using Spark and Conv-AE deep learning approach, which is fast, simple, vigorous, and efficient cybersecurity. The proposed ID system based on Conv-AE can automatically and efficiently learn the features representation from the CICIDS2018 heterogeneous dataset. We have implemented our proposed ID system for using various conventional machine classifiers using Spark and for misuse detection using state-of-the-art deep learning approaches, such as Conv-AE. The proposed ID system is better as compared to traditional security approaches in terms of attack detection rate and accuracy and also has less computation complexity. Both deep and machine learning models are assessed with renowned classification parameters, such as attack detection rate, classification accuracy, precision, recall, and F1 score.

We think that our approach can be extended in the future into numerous fields, such as the anomaly and network misuses, which can be recognized on real-time streaming image data, focusing on exploring deep learning as an attribute extraction tool to learn knowledgeable data illustrations in case of other anomaly detection problems in a more modern real-time dataset.

Author Contributions: M.A.K. conceived the research, wrote the paper, designed the framework, and performed the experiments. J.K. assisted with proofreading, revision, and improvements. J.K. supervised the overall research. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT, under Grant NRF-2017M3C4A7083279

Conflicts of Interest: The authors declare no conflicts.

Abbreviations

ML	Machine learning
ICT	Information communication technology
ID	Intrusion detection
DL	Deep learning
Conv-AE	Convolutional-auto encoder
CSE-CIC	Canadian Institute for Cybersecurity (CIC) and the Communications Security Establishment
NIDS	Network intrusion detection system
CIA	Confidentiality, integrity, availability
DDOS	Distributed denial of service
DOS	Denial of service
FN	False-negative
FP	False-positive
SAE	Stack autoencoder
DBN	Deep belief network
CNN	Convolutional neural network
DNN	Deep neural network
MLlib	Machine learning library
ABS	Anomaly-based system
SBS	Signature-based system
pcap	Packet capture files
k-NN	K-nearest neighbor
SVM	Support vector machine
ANN	Artificial neural network
RF	Random forest
DT	Decision tree
NB	Naïve bays
FAR	False alarm rate
DR	Detection rate
RBM	Restricted Boltzmann machine
LSTM	Long short-term memory
KDD	Knowledge discovery in databases
FS	Features selection
HAST	Hierarchical spatial temporal
SSO	Simplified swarm optimization
HADM	Hybrid anomaly detection model
UNSW-NB15	Australian center of cybersecurity at the University of New South Wales in 2015
ISCX2012	Information center of excellence
IP	Internet protocol
HTTP	Hypertext transfer protocol
UDP	User datagram protocol
TCP	Transmission control protocol
LOIC	Low orbit ion Cannon
TP	True-positive
TN	True-negative
JDK	Java development kit

References

1. Terzi, D.S.; Terzi, R.; Sagiroglu, S. A survey on security and privacy issues in big data. In Proceedings of the 10th International Conference for Internet Technology and Secured Transactions (ICITST), London, UK, 14–16 December 2015; pp. 202–207.
2. Chen, C.P.; Zhang, C.-Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Inf. Sci.* **2014**, *275*, 314–347. [CrossRef]
3. Behera, S.; Pradhan, A.; Dash, R. Deep Neural Network Architecture for Anomaly Based Intrusion Detection System. In Proceedings of the 2018 5th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 22–23 February 2018; pp. 270–274.
4. Ibrahim, A.T.H.; Ibrar, Y.; Nor, B.A.; Salimah, M.; Abdullah, G.; Samee, U.K. The rise of "big data" on cloud computing: Review and open research issues. *Infor. Syst.* **2015**, *47*, 98–115.
5. Hassan, M.M.; Gumaei, A.; AlSanad, A.; Alrubaian, M.; Fortino, G. A hybrid deep learning model for efficient intrusion detection in big data environment. *Inf. Sci.* **2020**, *513*, 386–396. [CrossRef]
6. Casas, P.; Soro, F.; Vanerio, J.; Settanni, G.; D'Alconzo, A. Network security and anomaly detection with Big-DAMA, a big data analytics framework. In Proceedings of the 2017 IEEE 6th International Conference on Cloud Networking (CloudNet), Prague, Czech Republic, 25–27 September 2017; pp. 1–7.
7. Al-Garadi, M.A.; Mohamed, A.; Al-Ali, A.K.; Du, X.; Ali, I.; Guizani, M. A Survey of Machine and Deep Learning Methods for Internet of Things (IoT) Security. *IEEE Commun. Surv. Tutor.* **2020**, *22*, 1646–1685. [CrossRef]
8. Casas, P.; D'Alconzo, A.; Settanni, G.; Fiadino, P.; Skopik, F. POSTER: (Semi)-Supervised Machine Learning Approaches for Network Security in High-Dimensional Network Data. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, 24 October 2016; pp. 1805–1807.
9. Monshizadeh, M.; Khatri, V.; Atli, B.G.; Kantola, R.; Yan, Z. Performance Evaluation of a Combined Anomaly Detection Platform. *IEEE Access* **2019**, *7*, 100964–100978. [CrossRef]
10. Trejo, L.A.; Ferman, V.; Medina-Pérez, M.A.; Giacinti, F.M.A.; Monroy, R.; Ramirez-Marquez, J.E. DNS-ADVP: A Machine Learning Anomaly Detection and Visual Platform to Protect Top-Level Domain Name Servers Against DDoS Attacks. *IEEE Access* **2019**, *7*, 116358–116369. [CrossRef]
11. Ferrag, M.A.; Maglaras, L.; Moschoyiannis, S.; Janicke, H. Deep learning for cybersecurity intrusion detection: Approaches, datasets, and comparative study. *J. Infor. Secur. Appl.* **2020**, *50*, 102419.
12. Erhan, D.; Bengio, Y.; Courville, A.; Manzagol, P.-A.; Vincent, P.; Bengio, S. Why does unsupervised pre-training help deep learning? *J. Mach. Learn. Res.* **2020**, *11*, 625–660.
13. Symantec. *Internet Security Threat Report*; Symantec: Mountain View, CA, USA, 2018.
14. AhnLab. *Asec Report*; AhnLab: Gyeonggi-do, Korea, 2018.
15. Jun-Chun, M.; Yong-Jun, W.; Ji-Yin, S.; Chen, S. A Minimum Cost of Network Hardening Model Based on Attack Graphs. *Procedia Eng.* **2011**, *15*, 3227–3233. [CrossRef]
16. Kotenko, I.; Chechulin, A. A cyber attack modeling and impact assessment framework. In Proceedings of the 5th International Conference on Cyber Conflict 2013 (CyCon 2013), IEEE and NATO COE Publications, Tallinn, Estonia, 4–7 June 2013; pp. 119–142.
17. Caltagirone, S.; Pendergast, A.; Betz, C. The diamond model of intrusion analysis. *DTIC Doc. Tech. Rep.* **2013**. Available online: https://www.activeresponse.org/wp-content/uploads/2013/07/diamond.pdf (accessed on 15 September 2020).
18. Hassell, S.; Beraud, P.; Cruz, A.; Ganga, G.; Martin, S.; Toennies, J.; Vazquez, P.; Wright, G.; Gomez, D.; Pietryka, F.; et al. Evaluating network cyber resiliency methods using cyber threat, Vulnerability and Defense Modeling and Simulation. In Proceedings of the MILCOM 2012–2012 IEEE Military Communications Conference, Orlando, FL, USA, 29 October–1 November 2012; pp. 1–6.
19. Anderson, J.P. *Computer Security Threat Monitoring and Surveillance*; Rapport Technique; James P. Anderson Company: Fortwashington, UK, 1980.
20. Naseer, S.; Saleem, Y.; Khalid, S.; Bashir, M.K.; Han, J.; Iqbal, M.M.; Han, K. Enhanced Network Anomaly Detection Based on Deep Neural Networks. *IEEE Access* **2018**, *6*, 48231–48246. [CrossRef]
21. Bandyopadhyay, S.; Ratul, C.; Arindam, R.; Banani, S. A Step Forward to Revolutionize Intrusion Detection System Using Deep Convolution Neural Network. *Preprints* **2020**. [CrossRef]

22. Tama, B.A.; Comuzzi, M.; Rhee, K.-H. TSE-IDS: A Two-Stage Classifier Ensemble for Intelligent Anomaly-Based Intrusion Detection System. *IEEE Access* **2019**, *7*, 94497–94507. [CrossRef]
23. Albahar, M.A.; Binsawad, M.H. Deep Autoencoders and Feedforward Networks Based on a New Regularization for Anomaly Detection. *Secur. Commun. Netw.* **2020**, *2020*, 1–9. [CrossRef]
24. Tang, T.A.; Mhamdi, L.; McLernon, D.; Zaidi, S.A.R.; Ghogho, M. Deep learning approach for Network Intrusion Detection in Software Defined Networking. In Proceedings of the 2016 International Conference on Wireless Networks and Mobile Communications (WINCOM), Fez, Morocco, 26–29 October 2016; pp. 258–263.
25. Al-Qatf, M.; Lasheng, Y.; Al-Habib, M.; Al-Sabahi, K. Deep Learning Approach Combining Sparse Autoencoder With SVM for Network Intrusion Detection. *IEEE Access* **2018**, *6*, 52843–52856. [CrossRef]
26. Farahnakian, F.; Heikkonen, J. A deep auto-encoder based approach for an intrusion detection system. In Proceedings of the 20th International Conference on Advanced Communication Technology (ICACT), Chuncheon-si Gangwon-do, Korea, 11–14 February 2018; p. 1.
27. Le, T.-T.-H.; Kim, Y.; Kim, H. Network Intrusion Detection Based on Novel Feature Selection Model and Various Recurrent Neural Networks. *Appl. Sci.* **2019**, *9*, 1392. [CrossRef]
28. Pektaş, A.; Acarman, T. A deep learning method to detect network intrusion through flow-based features. *Int. J. Netw. Manag.* **2018**, *29*, e2050. [CrossRef]
29. Mighan, S.N.; Kahani, M. Deep Learning Based Latent Feature Extraction for Intrusion Detection. In Proceedings of the Electrical Engineering (ICEE), Mashhad, Iran, 8–10 May 2018; pp. 1511–1516.
30. Meira, J.; Andrade, R.; Praça, I.; Carneiro, J.; Marreiros, G. Comparative Results with Unsupervised Techniques in Cyber Attack Novelty Detection. In *International Symposium on Ambient Intelligence*; Springer: Cham, Switzerland, 2018; pp. 103–112.
31. Wang, W.; Sheng, Y.; Wang, J.; Zeng, X.; Ye, X.; Huang, Y.; Zhu, M. HAST-IDS: Learning Hierarchical Spatial-Temporal Features Using Deep Neural Networks to Improve Intrusion Detection. *IEEE Access* **2017**, *6*, 1792–1806. [CrossRef]
32. Liao, Y.; Vemuri, V. Use of K-Nearest Neighbor classifier for intrusion detection. *Comput. Secur.* **2002**, *21*, 439–448. Available online: http://linkinghub.elsevier.com/retrieve/pii/S016740480200514X (accessed on 15 September 2020). [CrossRef]
33. Mukkamala, S.; Janoski, G.; Sung, A. Intrusion detection using neural networks and support vector machines. In Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No.02CH37290), Honolulu, HI, USA, 12–17 May 2002; p. 17021707. Available online: http://ieeexplore.ieee.org/document/1007774/ (accessed on 15 September 2020).
34. Ingre, B.; Yadav, A. Performance analysis of NSL-KDD dataset using ANN. In Proceedings of the 2015 International Conference on Signal Processing and Communication Engineering Systems, Guntur, India, 2–3 January 2015; pp. 92–96.
35. Farnaaz, N.; Jabbar, M. Random Forest Modeling for Network Intrusion Detection System. *Procedia Comput. Sci.* **2016**, *89*, 213–217. [CrossRef]
36. Zhang, J.; Zulkernine, M.; Haque, A. Random-Forests-Based Network Intrusion Detection Systems. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **2008**, *38*, 649–659. [CrossRef]
37. Laskov, P.; Düssel, P.; Schäfer, C.; Rieck, K. Learning Intrusion Detection: Supervised or Unsupervised? In Proceedings of the 13th International Conference on Image Analysis and Processing (ICIAP), Cagliari, Italy, 6–8 September 2005; pp. 50–57.
38. Solanas, A.; Martinez-Balleste, A. *Advances in Artificial Intelligence for Privacy Protection and Security (Intelligent Information Systems)*; World Scientific: Hackensack, NJ, USA, 2010; Available online: http://site.ebrary.com/id/10421991 (accessed on 18 September 2020).
39. Ghorbani, A.A.; Lu, W.; Tavallaee, M. *Network Intrusion Detection and Prevention (Advances in Information Security)*; Springer: Boston, MA, USA, 2010; Volume 47, Available online: http://link.springer.com/10.1007/978-0-387-88771-5 (accessed on 15 September 2020).
40. Bhattacharyya, D.K.; Kalita, J.K. *Network Anomaly Detection: A Machine Learning Perspective*; CRC Press: Boca Raton, FL, USA, 2013.
41. Tavallaee, M. An Adaptive Hybrid Intrusion Detection System. Ph.D. Thesis, University New Brunswick, Saint John, NB, Canada, 2011.
42. Murugan, P.; Durairaj, S. Regularization and Optimization strategies in Deep Convolutional Neural Network. *arXiv* **2017**, arXiv:1712.04711.

43. Hodo, E.; Bellekens, X.; Hamilton, A.; Tachtatzis, C.; Atkinson, R. Shallow and Deep Networks Intrusion Detection System: A Taxonomy and Survey. *arXiv* **2017**, arXiv:1701.02145.
44. Javaid, A.; Niyaz, Q.; Sun, W.; Alam, M. A Deep Learning Approach for Network Intrusion Detection System. In Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS), New York, NY, USA, 24 May 2016; pp. 21–26.
45. Fiore, U.; Palmieri, F.; Castiglione, A.; De Santis, A. Network anomaly detection with the restricted Boltzmann machine. *Neurocomputing* **2013**, *122*, 13–23. [CrossRef]
46. Kim, J.; Kim, J.; Thu, H.L.T.; Kim, H. Long Short Term Memory Recurrent Neural Network Classifier for Intrusion Detection. In Proceedings of the 2016 International Conference on Platform Technology and Service (PlatCon), Jeju, Korea, 15–17 February 2016; pp. 1–5.
47. Gao, N.; Gao, L.; Gao, Q.; Wang, H. An Intrusion Detection Model Based on Deep Belief Networks. In Proceedings of the 2014 Second International Conference on Advanced Cloud and Big Data, Huangshan, China, 20–22 November 2014; pp. 247–252. Available online: http://ieeexplore.ieee.org/document/7176101/ (accessed on 15 September 2020).
48. Aygun, R.C.; Yavuz, A.G. Network Anomaly Detection with Stochastically Improved Autoencoder Based Models. In Proceedings of the 2017 IEEE 4th International Conference on Cyber Security and Cloud Computing (CSCloud), New York, NY, USA, 26–28 June 2017; pp. 193–198. Available online: http://ieeexplore.ieee.org/document/7987197/ (accessed on 20 September 2020).
49. Yousefi-Azar, M.; Varadharajan, V.; Hamey, L.; Tupakula, U. Autoencoder-based feature learning for cybersecurity applications. In Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 14–19 May 2017; p. 38543861. Available online: http://ieeexplore.ieee.org/abstract/document/7966342/ (accessed on 15 September 2020).
50. Hussain, J.; Lalmuanawma, S. A hybrid approach for determining the efficient network intrusion 583 detection system. *IUP J. Comput. Sci.* **2014**, *8*, 34.
51. Hussain, J.; Lalmuanawma, S. An intelligent hybrid decision approach with feature selection for 585 anomaly network intrusion detection systems. In Proceedings of the 5th International Conference on Internet Technologies and Society 586, Taipei, Taiwan, 1 December 2014; pp. 3–10.
52. Chung, Y.Y.; Wahid, N. A hybrid network intrusion detection system using simplified swarm optimization (SSO). *Appl. Soft Comput.* **2012**, *12*, 3014–3022. [CrossRef]
53. Ghanem, T.F.; Elkilani, W.S.; Abdul-Kader, H.M. A hybrid approach for efficient anomaly detection using metaheuristic methods. *J. Adv. Res.* **2014**, *6*, 609–619. [CrossRef]
54. Kim, G.; Lee, S.; Kim, S. A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. *Expert Syst. Appl.* **2014**, *41*, 1690–1700. [CrossRef]
55. Ma, C.; Du, X.; Cao, L. Analysis of Multi-Types of Flow Features Based on Hybrid Neural Network for Improving Network Anomaly Detection. *IEEE Access* **2019**, *7*, 148363–148380. [CrossRef]
56. Soheily-Khah, S.; Marteau, P.-F.; Béchet, N. Intrusion detection in network systems through hybrid supervised and unsupervised mining process- a detailed case study on the ISCX benchmark dataset. In Proceedings of the 1st International Conference on Data Intelligence and Security (ICDIS), South Padre Island, TX, USA, 8–10 April 2018.
57. CSE-CIC-IDS2018 on AWS: A collaborative project between the Communications Security Establishment (CSE) & the Canadian Institute for Cybersecurity (CIC). Available online: https://www.unb.ca/cic/datasets/ids-2018.html (accessed on 15 September 2020).
58. Tang, D.; Qin, B.; Liu, T. Document Modeling with Gated Recurrent Neural Network for Sentiment Classification. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1422–1432.
59. Wu, Z.; Wang, X.; Jiang, Y.-G.; Ye, H.; Xue, X. Modeling Spatial-Temporal Clues in a Hybrid Deep Learning Framework for Video Classification. In Proceedings of the 23rd ACM International Conference on Multimedia, Brisbane, Australia, 26–30 October 2015; pp. 461–470.
60. Fan, Y.; Lu, X.; Li, D.; Liu, Y. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016; pp. 445–450.

61. Vignesh, K.; Yadav, G.; Sethi, A. Abnormal Event Detection on BMTT-PETS 2017 Surveillance Challenge. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 2161–2168.
62. Lin, P.; Ye, K.; Xu, C.-Z. Dynamic Network Anomaly Detection System by Using Deep Learning Techniques. In *International Conference on Cloud Computing*; Springer: Cham, Switzerland, 2019; pp. 161–176.
63. Rios, A.L.G.; Li, Z.; Bekshentayeva, K.; Trajkovic, L. Detection of Denial of Service Attacks in Communication Networks. In Proceedings of the 2020 IEEE International Symposium on Circuits and Systems (ISCAS), Sevilla, Spain, 10–21 October 2020; pp. 1–5.
64. Lee, M.-W. LSTM Model based on Session Management for Network Intrusion Detection. *J. Inst. Internet Broadcast. Commun.* **2020**, *20*, 1–7.
65. Chadza, T.; Kyriakopoulos, K.G.; Lambotharan, S. Contemporary Sequential Network Attacks Prediction using Hidden Markov Model. In Proceedings of the 2019 17th International Conference on Privacy, Security and Trust (PST), Fredericton, NB, Canada, 26–28 August 2019; pp. 1–3.