International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Critical Review - Use of Ensemble Methods in Intrusion Detection System

Indira P. Joshi[1], Dr. Vijaya K. Shandilya[2]

Submitted: 07/02/2024 **Revised**: 15/03/2024 **Accepted**: 21/03/2024

Abstract: IDSs are essential to the security of contemporary ICT systems. IDSs detect and report attacks, which are frequently examined by administrators tasked with thwarting the assault and reducing damage. As a result, it's critical that the IDS's alerts are as thorough as they can be. In this study paper has offered a multi-layered behavior-based IDS that classifies network using ensemble learning approaches. The ensemble has been built using Decision Trees, NB, SVM and Random Forests, these popular and well-liked models. Our solution is made to rapidly filter away traffic that has been identified as benign without further research in order to speed up system response time, while suspicious events are looked into to produce a more precise categorization. According to experimental setup has discussed on the various public datasets, the system can detect nine forms of high performances across all parameters taken into consideration.

Keywords: *Intrusion Detection, Behavior-Based IDS, Ensemble Learning, and Classification.*

1. Introduction

The multiplied accessibility of ICT device offerings over the internet and the ensuing great quantity of statistics that is to be had from anywhere are piqueing the curiosity of unscrupulous customers and imparting unanticipated potentialities for cyberattacks. In truth, systems that offer simple services and are deemed crucial for modern-day civilization, such power [1, 2], transportation, automatic vote casting, and telecommunications, require comprehensive safety on account that, if compromised, they may jeopardize the stableness of a whole nation [3, 4]. Intrusion Detection structures (IDSs) are responsible for the detection in addition to identity of cyberattacks. They alert directors right as soon as when malicious interest is determined on the community, permitting them to take instantaneous movement to contain the damage. In fashionable, an attack is detected and mitigated over the path of 4 steps [5]. network data are gathered in the first phase, accompanied with the aid of the discarding of factors that are not applicable to the identity of malicious traffic in the 2nd and third levels, respectively. The fourth and final phase makes a speciality of the mitigating moves conducted once the attack was observed. The final step isn't always always finished through an automated device; instead, an administrator can also entire it themselves. There are two most important kinds of IDSs, relying on whether or not they monitor the repute of the gadget and through evaluation of the log files or analyze community visitors or even the behavior of particular hosts.

To recover from this restriction and enable the constructing of technology that can robotically recognize whether or not occasions correlate to ordinary or aberrant behaviour, statistical processes inclusive of gadget learning can be applied. A statistical-based totally IDS become counseled in [9]. through inspecting N-grams in HTTP traffic, this IDS can stumble on botnet community traffic. because of the reality that C&Cs appoint similar verbal exchange styles, this approach turned into advanced. [10] offered a detection method primarily based on multidimensional correlation analysis. This system can understand each recognised and unidentified DoS assaults by using reading standard visitors styles. The authors of [11] mimic the immune machine of people using statistical strategies. The counseled IDS consists of layers; the primary recognizes and categorizes adverse communique consistent with its nature, and the second one considers traffic marked by means of the first level as extraordinarily suspicious and finds attributes which can be vital for intrusion detection. The device best does binary category, despite the reality that the dataset used to test it comprise several classifications. other systems use information fashions, which includes ontologies [12], to correctly categorize diverse assault types [13].

Department of Computer Science and Engineering, Sipna College of Engineering and Technology, Amravati, Maharashtra, India.

[1] *ipj.indira@gmail.com*

[2] *vkshandily14@gmail.com*

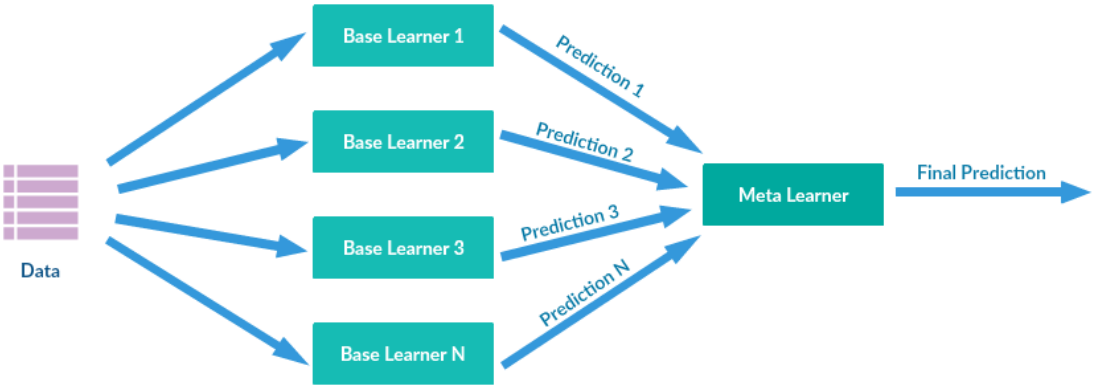

Fig 1. Conventional Ensemble learning technique

Several IDSs use ensemble studying methods for intrusion detection due to the fact they improve classification overall performance and accuracy compared to unmarried vulnerable beginners. for instance, [19] advise the use of an ensemble learning method to discover both and novel DDoS attacks. in this technique, many base models give attention to various factors of incursion. The authors of [20] hire ensemble approaches to especially discover botnet assaults towards diverse IoT community protocols. AdaBoost is particularly applied as an ensemble method the use of choice Tree, Naive Bayes, and artificial Neural community because the weak inexperienced persons. only 8 distinct varieties of botnet attacks can be detected by using the gadget. among the suggested solutions simply classify anomalies the use of a binary scheme. often, administrators are unable to take the important precautions because there's a lack of more data concerning the specific form of attack. it might be beneficial to make use of a popularity device [21, 22] that considers the earlier moves of community nodes, as cautioned by way of [23], a good way to decorate the selection-making method.

In summarized, multiclass IDSs within the literature either best understand a small variety of attack lessons or are overly specialized in that they handiest detect versions of the equal assault. In evaluation, as demonstrated by the checking out findings, our machine efficiently strikes the right balance between the amount of diagnosed lessons and forecast velocity. in this study, we describe a multi-layered intrusion detection gadget that may perceive malicious visitors the usage of gadget studying and ensemble gaining knowledge of techniques. We advise the use of information augmentation methods to balance the dataset being applied and enable the set of rules to as it should be perceive the classes represented by way of few samples. The thorough testing consequences display the splendid performances of our method in terms of accuracy, precision, F1-score, in addition to FNR (fake negative charge).

1.1. Motivation

The majority of earlier research has placed a lot of emphasis on application architectures and ensemble learning techniques. Certain ensemble learning [2, 3], all-encompassing machine learning [1, 19, 21, 22] or specific IDSs application architectures [15–18, 20] have been the focus of certain survey research. Furthermore, because the great majority of studies are not the result of a methodical mapping investigation, their thoroughness and significance are also diminished. To the best of our knowledge, there have been no studies that have thoroughly examined the viability of applying ensemble learning for IDSs. Furthermore, there are no review studies that compare classifier ensemble techniques empirically.

The structure of the essay is as follows. The background synopsis is described in Section 2. The specifics of this study's system overview are explained in Section 3. The experimental setup is shown and explained in section 4. Section 5 offers a summary and conclusion of this study.

2. Background Overview

The background information on IDSs and ensemble learning is presented in this section.

1.2. Intrusion detection systems

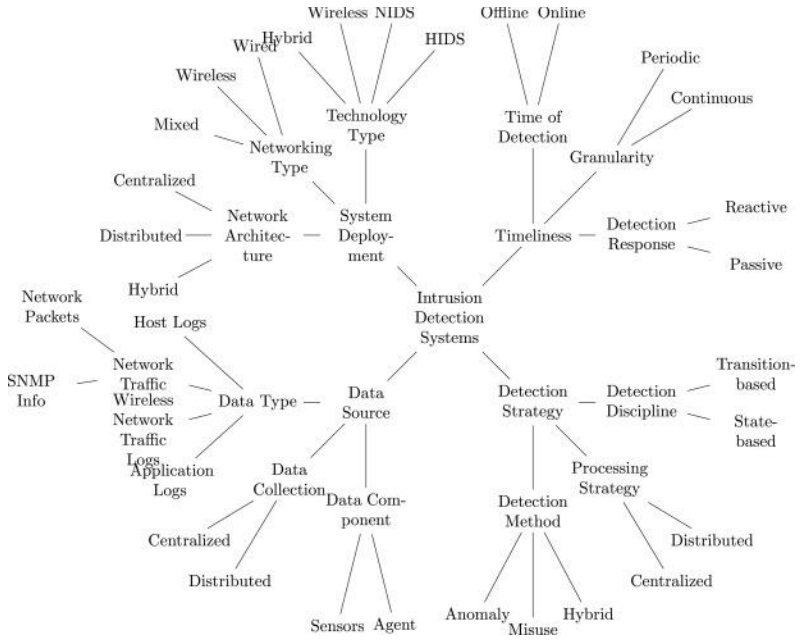

Fig 2. Overview of IDS

As changed into already said, an IDS attempts to keep an eye fixed on the employer's network infrastructure by means of directly identifying adverse hobby. Device deployment, timeliness, detection approach, and facts supply are the 4 key diverse dimensions that Liao et al. [10] Use to categorize IDSs. IDSs can be divided into technological kinds, particularly host-based and community-based totally, primarily based on their deployment technique. Host-based IDSs (HIDS) are designed to keep an eye fixed on activities that take place within a local PC community and sooner or later to inform users of the outcomes. The hash of the file machine is one example determined in HIDS. After evaluating the variations between each the hash cost this is currently being recalculated and that that became formerly recorded within the database, any untrustworthy behavior is diagnosed. Alternatively, community-primarily based IDSs (NIDS) are created to observe community traffic and to find malicious moves inside the community by using searching at inflowing network packets.

Anomaly or abuse are the two lessons of IDSs in terms of detection methodology, and IDSs may be applied in offline or on line modes in phrases of timeliness. In phrases of the facts source used for the analysis, an IDS may also be characterized. This covers the methods used to acquire the records, its categories, and the resources from which it become obtained. network site visitors logs, software logs, or server logs, are some examples of information type classifications. The hypothesized typology of IDSs presented via Liao et al. [10] is depicted in Fig. 1.

1.3. Ensemble learning

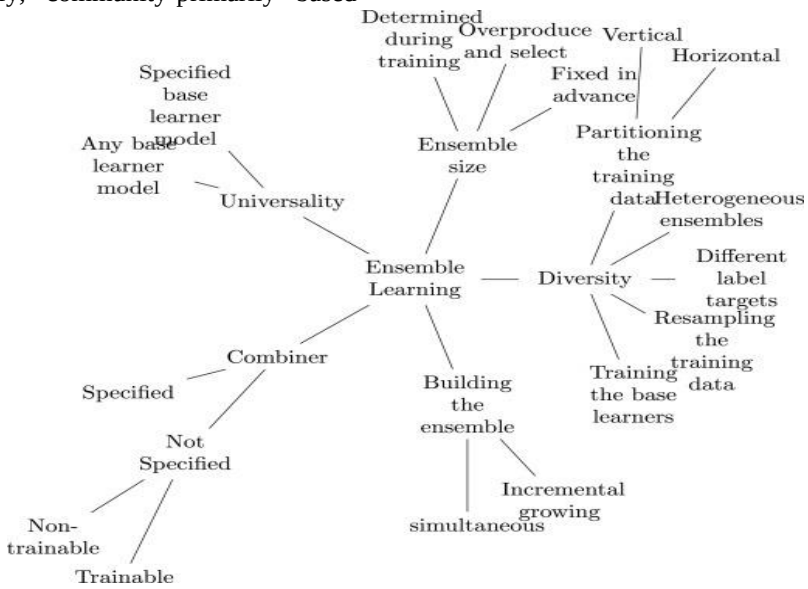

Fig 3. Overview of Ensemble learning

Committee-based totally mastering however alternatively a couple of classifier systems are different names for ensemble studying. A base getting to know technique creates a number of basis newbies that are commonly derived from schooling records [11]. Base learners, which include Bayesian classifiers, selection timber, neural networks, or different styles of gaining knowledge of algorithms, are now and again known as weak beginners because the purpose of an ensemble techniques is to enhance susceptible inexperienced persons. The massive bulk of ensemble methods use a unmarried place learner to create homogeneous ensembles (e.g., the identical styles of inexperienced persons), whilst different ensemble techniques use numerous learner sorts, producing heterogeneous ensembles.

Table 1. Overview Summary of IDS of Ensemble learning.

Ensemble family	Ensemble scheme	Studies/description	Outcomes
Homogeneous	Bagging	Empirical benchmark	Modern ensemble techniques for IDS are addressed.
	Boosting	Empirical benchmark and a scientific mapping take a look at.	Both homogeneous and heterogeneous forms of ensemble methods are taken into account while examining various ensemble and hybrid strategies.
	Random subspace	Ensemble mastering, a scientific mapping examine, and empirical benchmark.	On VANET, attack detection systems are reviewed.
	Rotation forest	A big range of ensemble methods are covered, a scientific mapping observe, and empirical benchmark.	There is a thorough examination of the fundamental principles relating to IDSs.
	Tree ensemble	Ensemble studying, a methodical mapping take a look at, and pragmatic benchmark.	The paper examines a number of current efforts on SDN-based IDS implementations using machine learning techniques.
	Dagging	Ensemble mastering, a methodical mapping examine, and pragmatic benchmark.	By analyzing current protection strategies, the survey categorizes the IoT security risks and difficulties for IoT networks.
Heterogeneous	Stacking	Ensemble gaining knowledge of, pragmatic benchmark, and a methodical mapping observe.	The paper reviews the research on network IDSs based on machine learning that has been published.
	Voting	A methodical mapping have a look at.	In-depth analysis of recent relevant publications that cover a variety of clever approaches and their used IDSs.

3. System Overview

This section focuses on a majority voting ensemble-based classification model for intrusion detection. The ensemble technique was used in this system overview's methodology.

1.4. Ensemble classification

This paper proposes the introduction of an effective excellence award for a gadget for intrusion detection that may skillfully fuse the man or woman classifiers to shape a resilient classifier and that could categorize community attacks. the primary goal of ensemble class is to improve anticipated accuracy over that of any character classifier. This proposed look at applies ensemble-based gadget learning techniques to discover and categorize community attacks. SVM, NB, Logistic Regression, and DT were 4 wonderful gadget gaining knowledge of strategies that we applied to analyze each the original datasets and the UEFFS-decreased datasets. to enhance categorization overall performance, we used a majority balloting technique.

The term "ensemble method" refers to a way for solving a hassle that uses a couple of primary learner or learner version [14]. most of the people of the time, ensemble methods are applied to improve prediction efficiency and accuracy even as overcoming forecast uncertainty. three categories of ensemble methods exist:

1) Bagging: This technique employs the averaging technique for regression and the vote casting technique for classification. The ensemble approach is used to acquire records from the foundational beginners before the labels are voted on. The forecast made with the aid of the device is the one with the maximum votes [15].

2) Boosting: Boosting is indeed an ensemble approach used to enhance the performance of vulnerable classifiers [15]. by way of letting one-of-a-kind iterations of a vulnerable learner function on a given piece of information, it gives a consecutive learning system. The manner is repeated until the preferred outcome is received [16].

The following technique is stacking, which involves combining many classifiers simply at base level with a type on the symbolic stage. One or maybe greater singleton classifiers are utilized on the fundamental classifier level. They paintings with the dataset and benefit understanding from it. The fundamental classifiers' outputs are mixed to offer an enter for the subsequent degree. The meta classifier gets enter and generates an output so as to be the very last type.

Below is the algorithm for the suggested stacking ensemble. Input:

Test and Training Datasets Normal or anomalous output

First, provide the train and test datasets.

Pre-process datasets in step two.

Step 3: Carry out a greedy stepwise search and CFS for feature selection.

Step 4: Train the basic classifiers with the train+ dataset.

Step 5: Create base-level models for the classifiers.

Step 6: Enter the base level's probability as input into the meta level.

Step 7: train the final model using the meta classifier at the meta level.

Step 8: use the test+ dataset to validate the generated model.

Step 9: Sort the dataset's occurrences into categories.

Step 10: Assess the model's effectiveness using the dataset

Step 11: output the predictions

1.5. Classification Algorithms

1) Support vector Machine: (SVM) is a powerful class technique for supervised gaining knowledge of. As seen in Fig. three [17], it makes use of hyper planes to carry out its operations. according to [18], education sets are transformed into better dimensions by using non-linear mapping, and the hyperplane is then utilized to categorize the information. This approach is quick and helps scaling at the same time as having a bad detection price. but, it takes longer to train, the learnt characteristic is tough to comprehend, and it takes a long time for multiclass datasets. [18]; [7].

2) Nave Bayes: one of the essential probabilistic classifiers, the nave bayes (NB) learner is based totally on the robust unbiased presumptions that exist among features. It recognizes that now the provision or absence of one feature is unrelated to the supply or absence of any other. This classifier is brief and effective in recognizing intrusions and has a high charge of accuracy. it's far independent and supported via facts [19]. Naive Bayes, in assessment to other Bayesian category strategies, ignores redundant and needless characteristics within the dataset on account that doing so could gradual down the detection technique and possibly degrade device overall performance [20].

3) Decision trees are strategies made of severa nodes. the edges are frequently used to hyperlink the nodes. A take a look at is accomplished on a node a good way to pick out the next node, and the outcome dictates the brink as a way to be traveled to the subsequent node. A decision node is the node this is being examined [19]. The creation of a woodland for the dataset that has been separated for schooling is required by way of this supervised learning method. A set of rules for decision timber is C4.five. it's

far the selection tree approach used for class in vulnerability scanning this is used the maximum. The dataset is cut up into subsets that contain either a class or some other using the fine characteristics, which can be chosen using C4.5. The cause is to decide the

4) Random forest (RF) is a mixture of several tree prognosticators. The random vector values sampled for the timber are what they rely upon. For generating bushes, forecasters are chosen at random [21]. As visible in Fig. five, random wooded area produces a huge range pf bushes and uses random choice to select the attributes to incorporate into each version. The trees which can be produced, however, are not trimmed [22]. every selection tree's random wooded area is created with the aid of the sampling of randomized subset characteristics. [23] asserts that including randomization to every established tree node improves the precision of random forests. It handles binary statistics, categorical information, and non-stop facts readily, as well as databases with excessive dimensions and missing values.

5) k-nearest Neighbor: This class approach is one of the handiest and oldest. example-based totally reasoning is used. in step with the ok-nearest neighbor (KNN) speculation, two gadgets belonging to the equal elegance have certain characteristics that may be measured by making use of distance metrics to them. as an instance, based totally at the votes of a ok-nearest acquaintances, an item with an unknown magnificence is both categorised into the equal institution as the first nearest neighbor or into the dominant class. The quantity of the friends, k, is typically selected by using empirical cross validation. With k-nearest Neighbor, gadgets are placed in a multi-dimensional feature space as point vectors. Nonparametric ok-Nearest Neighbor is used. This versatility makes it a dependable classifier for figuring out malware in datasets with numerous dimensions [25].

4. Design Of Experimental Setup

1) Dataset: The binary model of network safety Lab's know-how Discovery in Databases (NSL-KDD) dataset changed into used for this look at [3]. This dataset is an up to date model of the 1999 information Discovery with information Mining gear (KDD99) dataset. As illustrated in Fig. four, the NSL-KDD dataset has 41 characteristics with a total of 125,973 occurrences over 67,343 regular times and fifty eight,630 attack cases. The binary intrusion detection dataset's class is either ordinary or anomalous. The traffic for the furnished dataset can be labeled with the aid of type algorithms as both ordinary or normal site visitors. An anomaly is appeared as unlawful site visitors, whereas standard visitors is deemed true. The data that again up the realization

Table 2. CONFUSION MATRIX

	Genuine Positive Class	Genuine Negative Class
Classified Positive Class (True)	(True Positive) TP	(True Negative) TN
Classified Negative Class (False)	(False Positive) FP	(False Negative) FN

$$\frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\frac{TP}{TP+FP} \quad (2)$$

$$\frac{TP}{TP+FN} \quad (3)$$

$$\frac{FP}{FP+TN} \quad (4)$$

These performance indicators were chosen after considering how well the framework performed in addressing the challenges it was intended to address, as well as how it compared to other frameworks, single classifiers, and ensemble approaches. The accuracy metric is widely used in the works under examination, making it simple to compare this framework with theirs.

5. Conclusion

The complex and wide-ranging field of intrusion detection necessitates dynamic solutions. Many researchers have attempted to address the issue of intrusion into a network by using a variety of ways. Data mining is one of the approaches employed. The majority of data mining adopting researchers hardly ever used ensemble approaches. This study demonstrated the effectiveness and dependability of ensemble approaches, notably the IDS etchnique. With a precision of 99.5%, a false positive rate of 0.6%, and an accuracy of 99.5%, the few classification algorithms, such as C4.5 decision trees, demonstrated their dependability and ability to detect intrusions alongside KNN. This framework outperformed single classifiers and the methods suggested in the works reviewed. Moreover, this survey has been discussed the real network set-up to advance vet its concert.

References:

[1] Sadreazami H, Mohammadi A, Asif A, Plataniotis KN (2018) Distributed-graphbased statistical approach for intrusion detection in cyber-physical systems. IEEE Trans Sig Inf Process Netw 4(1):137–147

[2] Bhuyan MH, Bhattacharyya DK, Kalita JK (2014) Network anomaly detection: methods, systems and tools. IEEE Commun Surv Tutor 16(1):303–336

[3] Shafi K, Abbass HA (2013) Evaluation of an adaptive genetic-based signature extraction system for network intrusion detection. Pattern Anal Appl 16(4):549–566

[4] Pasqualetti F, Dörfler F, Bullo F (2013) Attack detection and identification in cyber-physical systems. IEEE Trans Autom Control 58(11):2715–2729

[5] Meshram A, Haas C (2017) Anomaly detection in industrial networks using machine learning: a roadmap. In: Beyerer J, Niggemann O, Kühnert C (eds) Machine learning for cyber physical systems: selected papers from the international conference ML4CPS 2016. Springer, Berlin, pp 65–72

[6] Hoque MAM, Bikas MAN (2012) An implementation of intrusion detection system using genetic algorithm. Int J Netw Secur Appl 4:2

[7] Creech G, Hu J (2014) A semantic approach to host-based intrusion detection systems using contiguous and discontiguous system call patterns. IEEE Trans Comput 63(4):807–819

[8] Alazab A, Hobbs M, Abawajy J, Khraisat A, Alazab M (2014) Using response action with intelligent intrusion detection and prevention system against web application malware. Inf Manag Comput Secur 22(5):431–449

[9] Chebrolu S, Abraham A, Thomas JP (2005) Feature deduction and ensemble design of intrusion detection systems. Comput Secur 24(4):295–307

[10] Koc L, Mazzuchi TA, Sarkani S (2012) A network intrusion detection system based on a hidden Naïve Bayes multiclass classifier. Exp Syst Appl 39(18):13492–13500

[11] Farahnakian F, Heikkonen J (2018) A deep auto-encoder based approach for intrusion detection system. In: 2018 20th international conference on advanced communication technology (ICACT). IEEE, pp 178–183

[12] Hanselmann M, Strauss T, Dormann K, Ulmer H (2020) CANet: an unsupervised intrusion detection system for high dimensional CAN bus data. IEEE Access 8:58194–58205

[13] Boukhalfa A, Abdellaoui A, Hmina N, Chaoui H (2020) LSTM deep learning method for network intrusion detection system. Int J Electr Comput Eng 10(3):2088–8708

[14] Yin C, Zhu Y, Fei J, He X (2017) A deep learning approach for intrusion detection using recurrent neural networks. IEEE Access 5:21954–21961.

[15] Kunang YN, Nurmaini S, Stiawan D, Suprapto BY (2021) Attack classification of an intrusion detection system using deep learning and hyperparameter optimization. J Inf Secur Appl 58:102804

[16] Fatani A, Abd Elaziz M, Dahou A, Al-Qaness MA, Lu S (2021) IoT intrusion detection system using deep learning and enhanced transient search optimization. IEEE Access 9:123448–123464

[17] Kanna PR, Santhi P (2021) Unified deep learning approach for efficient intrusion detection system using integrated spatial-temporal features. Knowl Based Syst 226:107132

[18] Aleesa A, Younis MOHAMMED, Mohammed AA, Sahar N (2021) Deep-intrusion detection system with enhanced unsw-Nb15 dataset based on deep learning techniques. J Eng Sci Technol 16(1):711–727

[19] Lee J, Park K (2021) GAN-based imbalanced data intrusion detection system. Pers Ubiquit Comput 25(1):121–128

[20] Liu C, Gu Z, Wang J (2021) A hybrid intrusion detection system based on scalable K-means+ random forest and deep learning. IEEE Access 9:75729–75740

[21] Ullah I, Mahmoud QH (2021) Design and development of a deep learning-based model for anomaly detection in IoT networks. IEEE Access 9:103906–103926

[22] Aldallal A, Alisa F (2021) Effective intrusion detection system to secure data in cloud using machine learning. Symmetry 13(12):2306

[23] Abusitta A, Bellaiche M, Dagenais M, Halabi T (2019) A deep learning approach for proactive multi-cloud cooperative intrusion detection system. Futur Gener Comput Syst 98:308–318

[24] Zhou X, Liang W, Li W, Yan K, Shimizu S, Kevin I, Wang K (2021) Hierarchical adversarial attacks against graph neural network based IoT network intrusion detection system. IEEE Int Things J

[25] Al Jallad K, Aljnidi M, Desouki MS (2019) Big data analysis and distributed deep learning for next-generation intrusion detection system optimization. J Big Data 6(1):1–18

[26] Mighan SN, Kahani M (2021) A novel scalable intrusion detection system based on deep learning. Int J Inf Secur 20(3):387–403

[27] Vinayakumar R, Alazab M, Soman KP, Poornachandran P, Al-Nemrat A, Venkatraman S (2019) Deep learning approach for intelligent

intrusion detection system. IEEE Access 7:41525–41550

[28] Kasongo SM, Sun Y (2020) A deep learning method with wrapper-based feature extraction for wireless intrusion detection system. Comput Secur 92:101752

[29] Shone N, Ngoc TN, Phai VD, Shi Q (2018) A deep learning approach to network intrusion detection. IEEE Trans Emerg Top Comput Intell 2(1):41–50

[30] Kasongo SM, Sun Y (2019) A deep learning method with filter-based feature engineering for wireless intrusion detection system. IEEE Access 7:38597–38607.