Cyber Intrusion Detection System based on Machine Learning Classification Approaches

Roseline Oluwaseun Ogundokun
Department of Multimedia Engineering
Kaunas University of Technology
Kaunas, Lithuania
rosogu@ktu.lt

Sanjay Misra
Department of Computer Sciences and Communication
Ostfold University College
Halden, Norway
ssopam@gmail.com

Akinbowale Nathaniel Babatunde
Department of Computer Science
Kwara State University
Malete, Nigeria
akinbowale.babatunde@kwasu.edu.ng

Sabarathinam Chockalingam
Department of Risk, Safety, and Security
Institute for Energy Technology
Halden, Norway
Sabarathinam.Chockalingam@ife.no

2022 International Conference on Applied Artificial Intelligence (ICAPAI) | 978-1-6654-6781-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICAPAI55158.2022.9801566

***Abstract*— As the internet has advanced over the years, so has the number of cyber-attacks. A sophisticated Intrusion Detection System (IDS) is essential to protect the cyberspace. The goal of IDS is to monitor and evaluate the operations that occur in a network for any signals of probable abnormalities. Although little research has been done in this area, more comprehensive research has yet to be completed. By examining the combinations of most prominent feature extraction (FE) techniques and classifiers, this research offers an IDS for networks based on machine learning (ML) that has a good union of FE techniques and classifiers. This paper introduced a feature extraction (FE) approach for classification issues, using independent component analysis (ICA). We can generate new features independent of each other by utilizing ICA to solve supervised classification issues, and we can also accurately express the output information. A set of significant features is selected from the original collection of features using FE algorithms. The set of significant features is then used to train various types of classifiers to produce the IDS. The proposed methods were evaluated in terms of five different performance measures using the DARPA KDD 99. Finally, it is discovered that the proposed ICA+RF classifier outperforms the others with an accuracy of 99.6%, f-score of 92.6%, and false alarm rate (FAR) value of 0.0029. The result was further compared with state-of-the-art, and it was deduced that our system performed better with higher accuracy and lower FAR.**

Keywords*— *Cyber-attack, Cyber security, Intrusion detection, Intrusion detection system, KDD 99, Machine learning

I. INTRODUCTION

Cyber security is the process of implementing security measures to ensure data confidentiality, integrity, and availability [1, 2]. In the literature, there are several descriptions of cyber security. Canongia and Mandarino referred to cyber security as the art of securing the survival and continuation of a nation's information society, guaranteeing and defending its information, assets, and essential infrastructure in cyberspace [3]. Cyber security is a crucial study subject because all administration, military, commercial, fiscal, and civilian operations obtain, analyze, and accumulate immense volumes of data on computers and other gadgets [1, 3, 4]. Companies must organize their efforts in the entire information system to be on the self-protective side of cyber security. Network security, application security, mobile security, data security, endpoint security, and so on are all components of cyber security [5].

The usage of the Internet and computer applications has increased dramatically in recent years, and they have become an essential element of today's generation. Security is becoming increasingly important as many computer applications and computer networks grow exponentially [5-6]. Attackers may employ various methods to cause havoc in businesses or organizations, including the usage of applications [7]. All of these approaches represent a potential danger that may or may not be significant enough to merit attention [5-6]. Artificial Intelligence (AI) and Machine Learning (ML)-based approaches have become increasingly important in detecting security vulnerabilities as cyber-attacks have increased dramatically. Security-related benchmarks are critical for ensuring that the best security applications are approved, and the required degree of security is achieved. From this perspective, they are crucial for numerous forms of cyber security research, such as Intrusion Detection System (IDS).

By considering the gaps in the field, in this study, we introduced a Feature Extraction (FE) approach for classification issues, which uses Independent Component Analysis (ICA). We can generate new features independent of each other and feasible by utilizing ICA to solve supervised classification issues, and we can also accurately express the output information. The study used four classifiers which are Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbor (KNN), and Decision Tree (DT) with the ICA FE on the IDS benchmark dataset (DARPA KDD 99) for the execution of the experiment.

The remaining section of this study is arranged as follows: section 2 discusses existing IDS systems using ML approaches. The IDS model with the dataset was presented in section 3. The section also discussed the data pre-processing and investigation, proposed methods, and the performance evaluation measures used in the study. Section 4 demonstrates the outcomes, and interpretations of the outcomes presented were also discussed in the section. In section 5, the study was concluded, and future works were suggested.

II. RELATED WORK

This section describes cyber-security as the systems or software that safeguard data, programs, and computer networks against a variety of undesired attacks such as illegal access, alteration, and creation [8]. We emphasized building an IDS to investigate and identify the system's security attacks since traditional security solutions are insufficient for detecting network security attacks [9-11]. Unauthorized action that damages an information system is defined as intrusion [12]. That is, any attack that potentially jeopardizes the privacy, veracity, or accessibility of information is deemed an incursion. Firewalls, access control, and

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

Authorized licensed use limited to: Kaunas University of Technology. Downloaded on June 24,2022 at 08:36:05 UTC from IEEE Xplore. Restrictions apply.

encryption are now the most often employed self-protective techniques for detecting internal threats [13]. On the other hand, IDSs are employed to spot both internal and external threats. Even though signature-based IDS may identify known attacks [12], we will concentrate on anomaly-based IDS in this study. A situation of departure from expected behavior is known as an anomaly. Profiles are generic or desired characteristics derived from tracking user, network connection, and host actions over a set period [14]. The behavior-built IDS is a dynamic method, often known as the anomaly-built approach [14]. Since it is not dependable to recognize undesirable operators' actions in the signature database, the primary benefit of an anomaly-founded IDS is to identify zero-day attacks [15]. There is also an amalgam discovery [16] technology or protocol examination [14] discovery approach that may be used. The amalgam approach benefits a greater discovery rate for attack discovery and a high capableness of discrepancy indicators in detecting the most recent attacks. It improves the detection rate (DR) of earlier identified invasions while lowering the false alarm rate (FAR) of unknown attacks [17]. This research concentrates on a cyber-security data-driven intrusion detection model built using ML approaches.

ML is a technique for making decisions using computers [18, 19]. It is connected to computational statistics and is part of AI. Classification is a type of supervised learning that forecasts the cyber-attack instances based on training security data [18, 20]. As a result, we examine a variety of well-known classifiers, such as the Bayesian approach [21, 22], Tree-based approach [23-25], and Artificial Neural Network-based models [26], which are frequently employed in prognostic analytics [18, 27, 19, 28], to establish a productive data-driven IDS prognostic method for offering intelligent cyber security facilities.

Several pieces of literatures have been discussed and it was deduced that using ML classifiers alone although attained a good percentage of accuracy but it was discovered that introducing dimensionality reduction algorithm (FE) will assist in reducing features of dataset thereby enhancing the classification accuracy, detection rate, recall, f1-score and precision. This therefore led to the motivation of introducing a FE algorithm which is ICA for enhanced classification accuracy. Hence, in this present work, we have used four standard ML classifiers DT, SVM, RF, and KNN, for classifying cyber-attacks and making a comparative evaluation with experiments. We also introduced the use of FE for better performance accuracy and detection rate.

III. IDS Model with Data

This section describes our IDS approach using DARPA KDD 99 datasets on various ML approaches. It integrates numerous phases: data investigation, data processing, and ML-founded security modeling. The phases are being discussed as follows:

A. Data Pre-processing and Investigation

Dataset is an assembly of information archives that include a diversity of behaviors or features, including associated data about the cyber-security model [29]. As a result, understanding cyber-security data, which contains numerous sorts of cyber-attacks and pertinent attributes, is critical. The rationale for this is that unprocessed security data gathered from a pertinent cyber point of supply might be utilized to examine innumerable models of security actions or malevolent conduct to create a data-driven security method that would help us reach our aim. The KDD'99 data from DARPA was utilized in this study [1] to construct prediction models for distinguishing the link between intrusions of multiple attacks. There are 494023 instances in this collection, each having 41 properties. As stated in table 1, attacks are divided into four categories in this dataset. Basic features, content features, time-based traffic features, and host-based traffic features are the four categories of features utilized in this dataset. TCP/IP connections are used to extract feature-based information. Window intervals are used to calculate traffic characteristics. It is broken down into two categories: 'Same Host Features (SHF) and Same Service Features (SSF)'. Both are referred to as time-based features. In the case of probing, the scan may take more than 2 seconds. The connection window recomputes the SHF and SSF to remedy this problem. Then it is referred to as "connection-based features". During a period, DoS and probe might have many contacts to a host(s). Then, based on the requirements, we analyze these characteristics and create the projected ML-built IDS method. This data-driven pattern-built verdict examination aids in delivering data-driven intelligent cyber-security facilities. The training dataset has a size of 345817, whereas the testing dataset has a size of 148206, indicating a 70:30 split.

TABLE 1. FOUR MAIN GROUPS OF CYBER-ATTACKS

Types of Network Attacks	Meaning	Examples
DoS	Denial of Service (DoS) means a genuine individual does not access the system resource.	• apache • smurf • Neptune • ping of death • back • mail bomb • UDP storm
Probe	The adversary aims at the host and makes efforts to acquire records concerning the host	• saint • port sweep • mscan • Nmap
R2L	Remote to Local (R2L) is a type of attack where the adversary makes efforts to gain entry to the victim's system without having an account in it.	• xlock • guest • xnsnoop • phf • sendmail dictionary
U2R	User to Root (U2R) is where an adversary makes an effort to gain rights having local admittance in the victim's system.	• perl • xterm

B. Feature Selection

Feature selection picks out a representative group of qualities from the original set. The learning method takes less time to train and creates a more generic classifier since it removes extraneous and irrelevant qualities from the original collection. Feature selection also aids data visualization and comprehension. The ICA approach, which is utilized in this research, is one of the most prominent feature selection strategies. This is explored briefly below:

Independent Component Analysis (ICA)

Authorized licensed use limited to: Kaunas University of Technology. Downloaded on June 24,2022 at 08:36:05 UTC from IEEE Xplore. Restrictions apply.

ICA is an approach for extracting individual signals from a large number of signals. Its strength stems from the scientifically plausible premise that distinct physical processes produce unrelated signals. Because of the assumption's simplicity and generality, it has been successfully implemented in an extensive variety of study topics.

Despite its broad variety of applications, ICA may be understood in terms of the traditional "cocktail party" problem, which ICA cleverly addresses. Consider a cocktail party where a large number of individuals are conversing simultaneously. A microphone's output is a mix of voices if one is present. ICA detects the mixture's unconnected individual signal components when given such a combination. This is precisely what ICA discovers, assuming that the only unconnected signal components inside the signal mixture are the voices of various persons. (In practice, finding individual signals in a single mixture needs more than one simultaneously recorded mixture.) It is important to note that ICA does not include any information particular to speech signals; for it to operate, the individual voice signals must be unconnected (Tharwat, 2020).

C. ML Classifiers

Classification is a supervised learning approach that is commonly used to simulate cyber intrusions based on a variety of attack categories. Data is always tagged before being used in supervised learning. The classifier learns the labels in the training phase to accurately forecast unknown data in the test phase. We employ prominent machine learning techniques in our study for various goals. The following are some of the most prevalent classification methods used in this research field.

SVM

SVM is grounded on the concept of structural risk minimization, which provides speed and scalability advantages [15]. The main idea behind the SVM is to discover a decision boundary in a multidimensional space that divides unseen patterns into distinct groups.

KNN

This is the simplest among all ML approaches where the output is computed built on k-closest neighbors or k-training pattern [1]. The output computation differs based on the work at hand. For example, when categorizing an unknown pattern, the pattern is allocated to the class that appears most frequently among the k-closest training patterns.

DT

The DT algorithm is a standard classification and prediction tool. It creates a model that forecasts a pattern's output depending on its various input attribute values. The creation of DT does not necessitate any domain expertise or parameter configuration; all required is the learning and modeling of the provided data set [10]. The three essential constituents are the decision node, the edges or branches, and the leaf node [4].

RF

The RF classifier comprises several tree classifiers, each of which is constructed utilizing a random vector sampled separately from the participated vector, and each tree casts an item vote for the utmost well-known class to categorize an input vector (Breiman 2001). The RF classifier utilized in this investigation employs arbitrarily specific characteristics or a mixture of features at each node to create a tree.

TABLE 2. TYPE OF ATTACKS IN THE DARPA KDD'99 DATASET

Attacks	**Attack Names**	**Number of Instances**
DoS	Teardrop, back, pod, smurf, and neptune	3883366
Probe	Satan, ipsweep, portsweep and nmap	41102
R2L	Multihop, ftp write, imap spy, guess password, warez master, phf, warez client	1126
U2R	Perl, buffer overflow, rootkit, and load module	52
Normal		972781

We present our ML-based ID approach, which has four key components:

Attack Class Label: To incorporate these attack classes into the IDS, all of the various threats have been assumed as separate unique class labels. Different forms of attacks, for instance, DoS, U2R, R2L, and PROBE, with their examples are presented in Table 1.

Security Features (SF): These are employed separately to forecast the cyber dangers mentioned above. Protocol kind, facility, length, and fault rate are all examples of features and these are the KDD 99 from DARPA dataset features.

Training and Testing Dataset: There are two types of datasets: training and test. The training dataset is employed to train the IDS method, while the testing dataset is utilized to assess its generalization. We employ a substantial portion of the cybersecurity as mentioned above data to create the IDS technique 70%, with the remainder being used for testing 30%.

D. Proposed System

Figure 1 shows the proposed IDS. FE technique picks out a representative group of qualities from the original set. Data with solely the important features advance the method's suitability with improved accuracy. The substantial group of features was hence employed to train the classifiers. After the training process, the test dataset is then employed to identify whether every instance is normal or attack data.

Authorized licensed use limited to: Kaunas University of Technology. Downloaded on June 24,2022 at 08:36:05 UTC from IEEE Xplore. Restrictions apply.

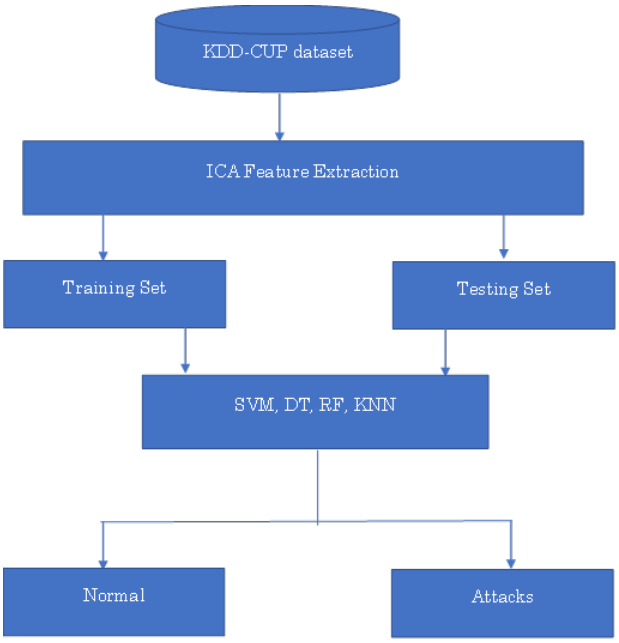


Figure 1. Proposed System Flow Diagram

IV. RESEARCH APPRAISAL

This section specifies intrusion detection performance metrics and examines the results by conducting tests using cybersecurity datasets with various types of attacks. All of the assessment measures are founded on the different features included in the Confusion Matrix [30], which is a 2D matrix that contains information on the Actual and Predicted classes:

True Positive (TP): The cyber-attacks data was properly forecasted as an attack during the experiment

False Negative (FN): The cyber-attacks data erroneously forecasted as normal during the experiment

False Positive (FP): The cyber-attacks data was erroneously categorized as an attack during the experiment

True Negative (TN): The cyber-attacks data properly classified as normal during the experiment

The accurate predictions are represented by the diagonal of the confusion matrix, whereas non-diagonal entries represent the incorrect guesses of a given classifier. Table 3 displays the confusion matrix's characteristics. In addition, the many assessment measures utilized in existing studies are shown in equations (1) to (5) [32].

$$\text{Precision} = TP/(TP+FP) \quad (1)$$

$$\text{Recall} = TP/(TP+FN) \quad (2)$$

$$\text{F-score} = 2*(\text{Precision}*\text{Recall})/(\text{Precision}+\text{Recall}) \quad (3)$$

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN) \quad (4)$$

$$\text{FAR} = FP/(FP+TN) \quad (5)$$

TABLE 3. CONFUSION MATRIX

Predicted Class			
		Attack	**Normal**
Actual Class	**Attack**	TP	FN
	Normal	FP	TN

A. *Investigational Results and Discussion*

In this section, we demonstrate the competency of ML classifiers for identifying intrusions. For this, we examine several well-known classification approaches that include the SVM, DT, KNN, and RF methods. Notably, we have likened numerous well-known classifiers' efficiency, including DT, SVM, RF, and KNN, to examine the IDS developed. To test the IDS, we used 148206 instances while 345817 instances were employed for the training phase of the system. To liken the capableness of the proposed approaches, performance measures, for example, precision, recall, f1-score, accuracy, and FAR are computed as seen in equations (1) to (5).

TABLE 4. CLASSIFIERS CONFUSION MATRIX

Classifiers	TP	TN	FP	FN
ICA+SVM	5842	140148	1680	534
ICA+KNN	3871	143681	503	149
ICA+DT	4846	141212	1590	556
ICA+RF	3546	144091	420	147

Figures 2 and 3 demonstrate a comparison of accuracy, precision, recall, f-score, and FAR to assess the performance of each classifier-based IDS technique. Each classification-based IDS strategy was evaluated using the identical train and testing dataset.

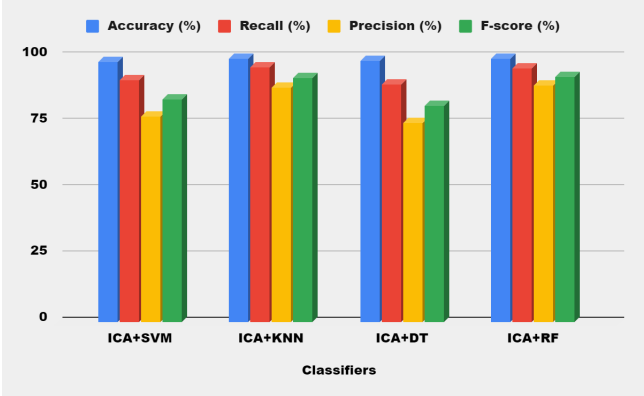


Figure 2. Performance Evaluation Outcome with relation to Accuracy, Precision, F-score, and Recall

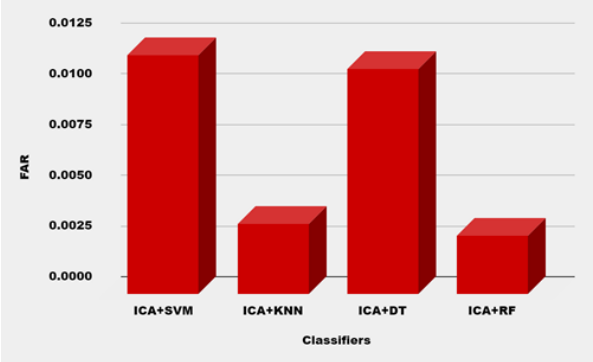


Figure 3. Performance Evaluation Outcome with Relation to FAR

Authorized licensed use limited to: Kaunas University of Technology. Downloaded on June 24,2022 at 08:36:05 UTC from IEEE Xplore. Restrictions apply.

Figures 2 and 3 show that the ICA+RF classifier-based IDS technique steadily performed better than the remaining classifiers concerning intrusion detection. The ICA+RF, in particular, produces the most significant outcomes in relation to the accuracy, precision, f-score, and FAR. The justification for this is that the RF approach generates numerous DT and then presumes a set of rules in the forest. All tree in the RF approach acts as a new ML classification technique, generating more logic rules as a result of considering the bulk vote of these trees when generating the result. As a result, the RF approach has superior precision, f1-score, FAR, and accuracy than other models. Generally, the IDS model based on machine learning classifiers outlined above is data-driven and represents the behavioral patterns of diverse cyber-attacks. Even though we investigate data-driven detection using ML approaches based on the patterns found in a particular dataset, a recency-based approach [32] may be further successful in constructing a data-driven IDS. Furthermore, adding relative data and analyzing it [33, 27] might be useful in establishing an intelligent IDS.

TABLE 5. COMPARATIVE ANALYSIS WITH EXISTING METHODS

Authors (Reference)	Methods	Dataset	Accuracy	FAR
Farnaaz & Jabbar [34]	RF	NSL-KDD	99.67	0.00502
Primaria & Tama [35]	RF	NSL-KDD, UNSW-NB15, GPRS	99.57	0.34
Waskle, Parashar & Singh [36]	PCA+RF	KDD 99	96.78	-----
Iman, & Ahmad [37]	RF	NSL-KDD	99.84	-----
Proposed Method	ICA+RF	DARPA KDD 99	99.68	0.0029

V. CONCLUSIONS AND FUTURE WORK

For security reasons, the possibility and usefulness of ML-based IDS is a major worry for IT professionals, e-commerce developers, and application developers. In general, cyber-security data collection comprises numerous kinds of cyber-attacks and essential characteristics. As a result, depending on a range of attack categories and attributes, specific classifiers may not accomplish sufficient accuracy and real detection rate (DR). We analyzed the efficiency of the data-driven IDS in this research, considering prominent machine learning classification approaches. Precision, f1-score, FAR, and accuracy are some of the performance indicators we looked at. This experiment was accomplished using DARPA KDD 99 intrusion dataset. The full dataset was split into two separate groups: training set 70%, and testing set 30%. The approach was tested on the KDD 99 dataset, which is independent of the full KDD 99 dataset. The simulations demonstrated that our proposed IDS attained an accuracy of 99.68%, a DR of 96.0%, a precision of 89.4%, an f-score of 92.6%, and a lower FAR value of 0.0029. These results represent an enhanced performance in contrast with the existing system that did not employ the use of FE. Furthermore, the application of the ICA in ourresearch has demonstrated that our proposed approach is applicable in cyberspace and network security.

We want to expand the cyber-security datasets in the future and create a data-driven IDS for the cyber-security community to provide automated security services. Deep learning algorithms are similarly projected to be used for the IDS in the future. It is also suggested the ICA FS algorithm is runned on the dataset after it is already splitted into the training and testing parts and then the accuracy is measured. It can also be suggested that the dataset is balalanced using algorithms like SMOTE before been used for classification to achive an enhance accuracy.

ACKNOWLEDGMENT

This paper is partially funded by the Research Council of Norway (RCN) in the INTPART program under the project “Reinforcing Competence in cybersecurity of Critical Infrastructures: A Norway-US Partnership (RECYCIN)” with the project number #30991.

REFERENCES

[1] Dua, S., & Du, X. (2016). Data mining and machine learning in cybersecurity. CRC press.

[2] Azeez, N. A., Bada, T. M., Misra, S., Adewumi, A., Van der Vyver, C., & Ahuja, R. (2020). Intrusion detection and prevention systems: an updated review. Data management, analytics, and innovation, 685-696.

[3] Cruz-Cunha, M. M. (Ed.). (2011). Handbook of research on business social networking: Organizational, managerial, and technological dimensions: Organizational, managerial, and technological dimensions. IGI Global.

[4] Kimani, K., Oduol, V., & Langat, K. (2019). Cyber security challenges for IoT-based smart grid networks. International Journal of Critical Infrastructure Protection, 25, 36-49.

[5] Von Solms, R., & Van Niekerk, J. (2013). From information security to cyber security. computers & security, 38, 97-102.

[6] Fraley, J. B., & Cannady, J. (2017, March). The promise of machine learning in cybersecurity. In SoutheastCon 2017 (pp. 1-6). IEEE.

[7] Ogundokun, R. O., Awotunde, J. B., Sadiku, P., Adeniyi, E. A., Abiodun, M., & Dauda, O. I. (2021). An Enhanced Intrusion Detection System using Particle Swarm Optimization Feature Extraction Technique. Procedia Computer Science, 193, 504-512.

[8] Aftergood, S. (2017). Cybersecurity: The cold war online. Nature, 547(7661), 30-31.

[9] Mohammadi, S., Mirvaziri, H., Ghazizadeh-Ahsaee, M., & Karimipour, H. (2019). Cyber intrusion detection by the combined feature selection algorithm. Journal of information security and applications, 44, 80-88.

[10] Tapiador, J. E., Orfila, A., Ribagorda, A., & Ramos, B. (2013). Key-recovery attacks on KIDS, a keyed anomaly detection system. IEEE Transactions on Dependable and Secure Computing, 12(3), 312-325.

[11] Tavallaee, M., Stakhanova, N., & Ghorbani, A. A. (2010). Toward credible evaluation of anomaly-based intrusion-detection methods. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(5), 516-524.

[12] Khraisat, A., Gondal, I., Vamplew, P., & Kamruzzaman, J. (2019). Survey of intrusion detection systems: techniques, datasets, and challenges. Cybersecurity, 2(1), 1-22.

[13] Anwar, S., Mohamad Zain, J., Zolkipli, M. F., Inayat, Z., Khan, S., Anthony, B., & Chang, V. (2017). From intrusion detection to an intrusion response system: fundamentals, requirements, and future directions. Algorithms, 10(2), 39.

[14] Liao, H. J., Lin, C. H. R., Lin, Y. C., & Tung, K. Y. (2013). Intrusion detection system: A comprehensive review. Journal of Network and Computer Applications, 36(1), 16-24.

[15] Alqahtani, H., Sarker, I. H., Kalim, A., Hossain, S. M. M., Ikhlaq, S., & Hossain, S. (2020, March). Cyber intrusion detection using machine learning classification techniques. In International

Authorized licensed use limited to: Kaunas University of Technology. Downloaded on June 24,2022 at 08:36:05 UTC from IEEE Xplore. Restrictions apply.

Conference on Computing Science, Communication and Security (pp. 121-131). Springer, Singapore.

[16] Viegas, E., Santin, A. O., Franca, A., Jasinski, R., Pedroni, V. A., & Oliveira, L. S. (2016). Towards an energy-efficient anomaly-based intrusion detection engine for embedded systems. IEEE Transactions on Computers, 66(1), 163-177.

[17] Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H., ... & Wang, C. (2018). Machine learning and deep learning methods for cybersecurity. Ieee access, 6, 35365-35381.

[18] Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

[19] Ian, H. W., & Eibe, F. (2005). Data Mining: Practical machine learning tools and techniques.

[20] Sarker, I. H., Kayes, A. S. M., & Watters, P. (2019). Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. Journal of Big Data, 6(1), 1-28.

[21] Sarker, I. H. (2019). A machine learning-based robust prediction model for real-life mobile phone data. Internet of Things, 5, 180-193.

[22] Sarker, I. H., Kabir, M. A., Colman, A., & Han, J. (2017, August). An improved naive Bayes classifier-based noise detection technique for classifying user phone call behavior. In Australasian Conference on Data Mining (pp. 72-85). Springer, Singapore.

[23] Quinlan, J. R. (1986). Induction of decision trees. Mach. learns.

[24] Sarker, I. H., Colman, A., Han, J., Khan, A. I., Abushark, Y. B., & Salah, K. (2020). Behavdt: a behavioral decision tree learning to build a user-centric context-aware predictive model. Mobile Networks and Applications, 25(3), 1151-1161.

[25] Sarker, I. H., Abushark, Y. B., Alsolami, F., & Khan, A. I. (2020). Intrudtree: a machine learning-based cyber security intrusion detection model. Symmetry, 12(5), 754.

[26] Sarker, I. H., Kayes, A. S. M., & Watters, P. (2019). Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. Journal of Big Data, 6(1), 1-28.

[27] Sarker, I. H. (2019). Context-aware rule learning from smartphone data: survey, challenges, and future directions. Journal of Big Data, 6(1), 1-25.

[28] Witten, I. H., Frank, E., Trigg, L. E., Hall, M. A., Holmes, G., & Cunningham, S. J. (1999). Weka: Practical machine learning tools and techniques with Java implementations.

[29] Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: an overview from a machine learning perspective. Journal of Big data, 7(1), 1-29.

[30] Deng, X., Liu, Q., Deng, Y., & Mahadevan, S. (2016). An improved method to construct basic probability assignments based on the confusion matrix for classification problems. Information Sciences, 340, 250-261.

[31] Alabi, K. O., Abdulsalam, S. O., Ogundokun, R. O., & Arowolo, M. O. (2020, November). Credit risk prediction in commercial bank using chi-square with SVM-RBF. In International Conference on Information and Communication Technology and Applications (pp. 158-169). Springer, Cham.

[32] Sarker, I. H., Colman, A., & Han, J. (2019). Recencyminer: mining recency-based personalized behavior from contextual smartphone data. Journal of Big Data, 6(1), 1-21.

[33] Sarker, I. H., Colman, A., Kabir, M. A., & Han, J. (2018). Individualized time-series segmentation for mining mobile phone user behavior. The Computer Journal, 61(3), 349-368.

[34] Farnaaz, N., & Jabbar, M. A. (2016). Random forest modeling for network intrusion detection system. Procedia Computer Science, 89, 213-217.

[35] Primartha, R., & Tama, B. A. (2017, November). Anomaly detection using random forest: A performance revisited. In 2017 International conference on data and software engineering (ICoDSE) (pp. 1-6). IEEE.

[36] Waskle, S., Parashar, L., & Singh, U. (2020, July). Intrusion detection system using PCA with random forest approach. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 803-808). IEEE.

[37] Iman, A. N., & Ahmad, T. (2020, February). Improving intrusion detection system by estimating parameters of random forest in Boruta. In 2020 International Conference on Smart Technology and Applications (ICoSTA) (pp. 1-6).

Authorized licensed use limited to: Kaunas University of Technology. Downloaded on June 24,2022 at 08:36:05 UTC from IEEE Xplore. Restrictions apply.