Volodymyr POGORELOV[1], Maryna KOLOMIIETS[2]

Scientific supervisor: Oleksandr KORCHENKO[3]

DOI: https://doi.org/10.53052/9788366249868.17

PRZEGLĄD METOD ANN WYKRYWANIA CYBERATAKÓW NA ZASOBACH SYSTEMU

Streszczenie: Wykazano znaczenie opracowania strategii zapobiegania cyberatakom w oparciu o sieci ANN. Zaproponowano sformalizowanie procedury analizy maszynowej poprzez zdefiniowanie funkcji celu. Opracowano algorytm optymalizacji informacji o zabezpieczeniach SSN i zarządzania zdarzeniami w oparciu o estymację globalnego maksimum funkcji docelowej.
Słowa kluczowe: cyberatak, schemat SIEM, metody ANN, głębokie uczenie, Big Data, funkcje celu, ekstremum globalne.

AN OVERVIEW OF ANN METHODS OF CYBER-ATTACKS DETECTION ON SYSTEM'S RESOURCES

Summary: An importance of development of cyber-attack prevention strategy based on the ANNs was shown in the paper. It was proposed to formalize the machine analysis procedure through the definition of objective functions. There were developed optimization algorithms for ANN security information and event management based on the estimation of the target function global maximum.

Keywords: cyber-attack, SIEM-scheme, ANN methods, deep learning, Big Data, objective functions, global extremum.

1. Introduction

The massive growth of information systems in particular the origin of cloud services [1, 2] and Internet of Things (IoT) concept [3, 4] has led to an increase in

[1] National Aviation University, Faculty of cybersecurity, computer and software engineering, IT-security Academic Department, Assistant, PhD, Volodymyr.Pogorelov@gmail.com.
[2] National Aviation University, Faculty of cybersecurity, computer and software engineering, IT-security Academic Department, Assistant, mv.kolomiiiets@gmail.com
[3] Professor, Dr.habill., National Aviation University, Faculty of cybersecurity, computer and software engineering, IT-security Academic Department, agkorchenko@gmail.com

relevance of cyber-security methods development. Security information and event management scheme (SIEM-scheme) includes both intrusion detection and intrusion prevention. Proper SIEM-scheme is based on analysis of internal/external threats: targets, behavior, pattern and lifecycle as it shown at Figure 1.

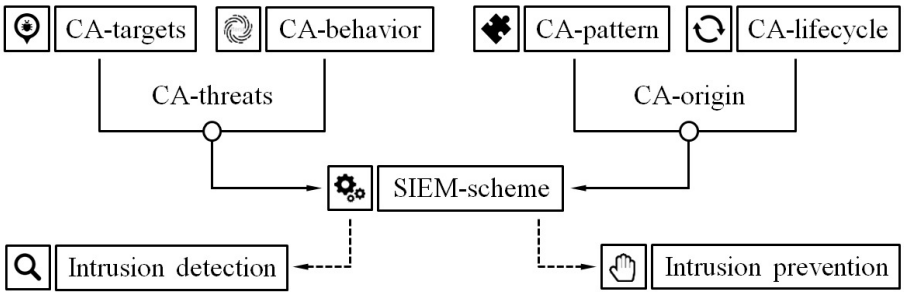

Figure 1. Development and application of SIEM-scheme

Thus the development of the effective cyber-attack prevention strategy implies determination of the cyber-attack (CA) typical targets (hardware nodes, software information, data resources), CA-behavior (system-sabotage, unauthorized access, data leakage), CA-pattern (program code blocks or intrusion consequences) and CA-lifecycle (introduction into the OS/applications or zero-day exploits). The complexity of the task forces to use at SIEM-scheme artificial neural network (ANN) algorithms. This approach makes it possible to extract features of the CA-pattern and CA-behavior by preparing and application of specific training sets. It should be noticed that modern ANN methods show potential extracting high-level features which allows detecting CA with brand new pattern of program code and atypical behavior.

Actual scientific research at the field of ANN methods of cyber-attacks prevention and issues analysis shows priority of deep learning ANN algorithms [5, 6]. There were considered application of:

- Deep Belief Networks' (DBN) group [7, 8]
- Recurrent Neural Networks' (RNN) group [9, 10]
- Convolutional Neural Networks' (CNN) group [11, 12],
- Recursive Neural Networks' (RvNN) group [13, 14]
- Generative Adversarial Networks' (GAN) group [15, 16]
- autoencoders and stacked autoencoders [17, 18].

The analysis carried out indicated the importance of focusing on development of deep learning ANNs for intrusion detection and prevention. At the same time, the development of proper mathematical apparatus will make it possible to reduce optimization problem of the ANN's detection and prevention algorithms to the mathematical problem of finding the global extremums of the objective functions.

2. Development of ANN algorithms fot the CA detection and prevention

ANN methods of CA on system's resources detection include two main groups: shallow machine learning (SL) and deep machine learning (DL). DL-methods are

based on the analysis of large amounts of data (in particular machine analysis in real time); their massive introduction in the field of cyber-defense corresponds to the origin of the Big Data phenomenon (Figure 2).

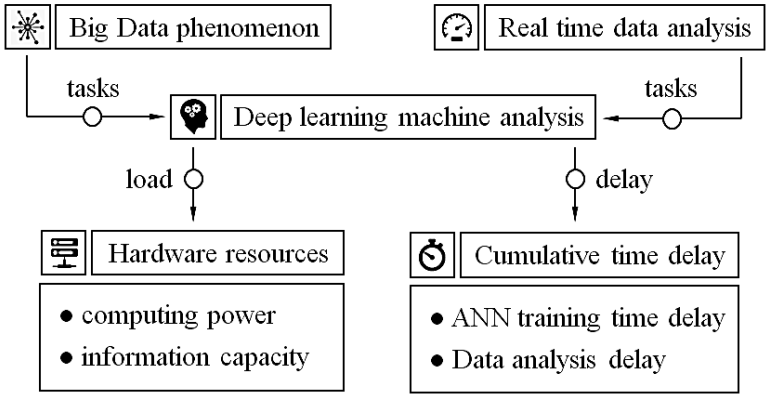

Figure 2. Development of deep learning ANN methods of CA

DL ANN algorithms allow extracting high-level features of the program code which makes possible to detect zero-day exploits of unknown origin, while SL ANN algorithms are specialized in detecting CA of a certain type which reduces their efficiency and increases the role of person-to-machine (P2M) interaction. This comes at a price, which consists in a significant increase in the load on hardware resources, as well as in increase of the training time of the ANN algorithm and analysis of streaming data [8-11]. Nowadays, deep autoencoders, CNN, DBN, GAN, RvNN, RNN are considered as most efficient DL neural network algorithms groups in the field of CA detection and prevention.

The main advantage of using autoencoders is the unsupervised learning application that implies matching of the output and input layers values while dimension of the hidden layers is lower than dimension of the output and input layers. When processing graphic data to provide denoising basic autoencoder architecture is used while for extracting high-level features stacked autoencoder is used. When analyzing the program code for the purpose of CA detecting, as shown in Table 1, both approaches can be used, but the stacked autoencoder shows more stable precision and accuracy values regardless of the amount of input data.

The definition of the DBN group of DL ANN algorithms can be formalized as neural network architecture which includes several hidden layers with no connections between units of each layer or connections between units of any layers other than neighbor layers. Thereby DBN could be modeled as a composition of basic, unsupervised ANNs such as autoencoders coupled with classification layers or restricted Boltzmann machines (RBM). Originally autoencoders were developed to reconstruct the original input through noise reduction, but in the field of CA detection, they are widely used for extraction of code features and preliminary deobfuscation of the input code. When program code machine analysis, a wide range of approaches is effectively used, in particular linear regression DBN, DBN based Probabilistic Neural Networks (PNN) and RBM (Table 1).

CNN architecture is based on the shared-weights of the convolution kernels that during machine analysis procedure slide along input data set and provide translation equivariant responses (feature map). Thereby CNN architecture could be formalized as structure which contains classification layers, pooling layers and convolution layers. To increase the efficiency of the analysis of program code in order to detect CA in the CNN architecture, long short-term memory (LSTM) should be applied (Table 1).

GAN algorithm use implies application of the unsupervised ML ANN contest which could be defined as mathematical zero-sum game. Thereby algorithms includes the generator block (first ML ANN) which produces output data up to the extracted features of the input data and discriminator block (second ML ANN) that analyze results in attempt to distinguish input samples and generated ones by the first ML ANN.

RvNN algorithms are based on applying of the same set of weights recursively to generate a structured prediction up to the structured input. Unlike multilayer perceptrons, RvNN uses internal memory to process sequences of arbitrary length. Therefore, RvNN are highly efficient (Table 1) at the machine analysis of the segments of the input data (potentially dangerous code, streamed data, etc).

RNN-group neural network architecture structure implies connections between ANN's nodes which form a directed graph along a temporal sequence. Thereby RNN algorithms use internal memory to process arbitrary sequences of input. Typical problem of RNN training gradients disappearance (short term memory problem) today is usually solved through the application of LSTM approach (Table 1).

Table 1. Performance of CA detection based on DL machine analysis [7-19]

CA detection method	DL ANN architecture	Method performance	
		Precision	Accuracy
Autoencoder	basic autoencoder architecture	81.9%-87.2%	85.6%-99.2%
	stacked autoencoder	86.1%-86.2%	92.1%-92.2%
DBN	basic DBN architecture	81.3%-92.3%	93.5%-97.5%
	LR-DBN	97.9%-98.1%	99.4%-99.5%
	DBN based PNN	93.2%-93.3%	99.1%-99.2%
	RBM	95.7%-95.8%	93.4%-99.7%
CNN	basic CNN architecture	77.0%-99.0%	78.0%-99.2%
	CNN based on LSTM	85.6%-85.7%	89.4%-89.6%
RvNN	basic RvNN architecture	81.1%-82.6%	83.3%-83.4%
RNN	basic RNN architecture	98.8%-99.3%	95.8%-96.0%

As one can see from the statistical results presented in Table 1, indicators for the effectiveness of using neural network methods in CA detecting vary significantly for most of the methods. It should also be noted that in many cases deep learning methods are not much more effective than SL methods, which is not typical for other areas of ANN algorithms application, such as processing of graphic data, video-fragments and audio-sequences. Factors indicated by the statistical analysis significantly complicate the task of organizing and optimizing of the SIEM-model based on ANNs. Thus, to solve the problem, it is necessary to formalize the procedures for executing ANN algorithms in accordance with target performance indicators.

3. Organizing and optimizing of the SIEM-model

Formalization of the SA detection process at the SIEM-model level includes the analysis of the following key points (Figure 3):

- CA-source: system (system command set or system accounting) and logs data (system logs data or security logs data),
- CA-location: CA based on the host server and CA based on the network infrastructure,
- CA-detection method: anomaly based methods and CA-signature based methods.

CA-detection method analysis is most important stage of SIEM-model development. The analysis shows that the methods based on anomaly detection include:

- data mining methods,
- knowledge base formation,
- methods of statistical analysis
- machine learning methods.

At the same time, the methods based on CA-signature detection can be divided into the following groups:

- expert analysis system,
- high-level features extraction,
- pattern matching,
- state modeling.

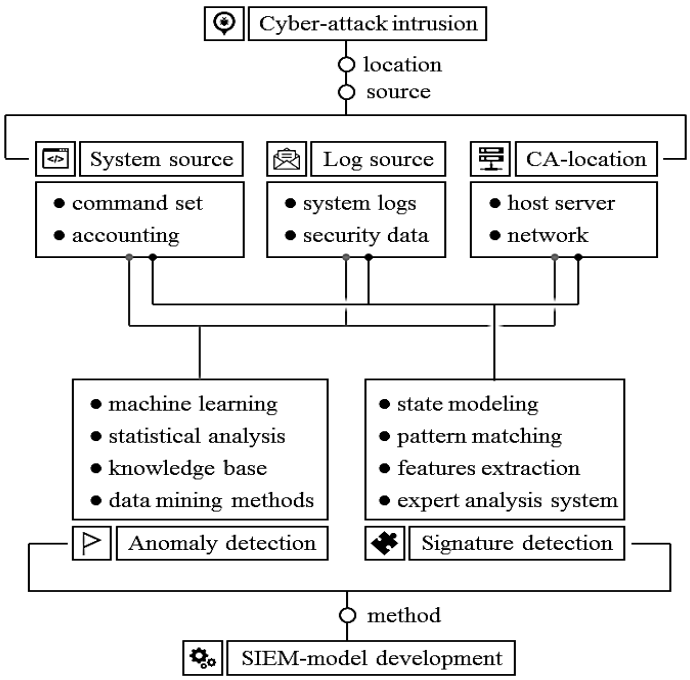

Figure 3. Classification of the CA-source, CA-location and CA detections methods according to the SIEM-model

The procedure for preparing a training dataset can be represented through the introduction of a set $D:\{D_n\}$ of CA-patterns, where $n \in [1;N]$. In turn, D_n can be defined in terms of the pair of labels' and features' sets $\forall D_n:(L_k,F_k)$, where $k \in [1;K_n]$. Finally, L_k and F_k could be formalized as one-dimensional matrices of elements $L_k:\{L_k^i\}$ and $F_k:\{F_k^j\}$, where $i \in [1;I_k]$ and $j \in [1;J_k]$. This approach makes it possible to unambiguously determine the arguments of the objective functions already at the stage of the training set preparation.

Similarly, it is proposed to formalize the following functional elements of the SIEM-model:

- training algorithms: stochastic algorithm, batch algorithm, mini-batch algorithm.
- architecture: the number of hidden layers, the number of neurons in each layer, the number of bias neurons and all connections between neurons, represented through one-dimensional and two-dimensional matrices
- activation function: Ridge activation functions set (linear, ReLU, heaviside, logistic), radial activation functions set (gaussian, multiquadratics), folding activation function, etc.
- training methods: supervised learning, unsupervised learning, reinforcement learning

Following the definition of the complete sets of arguments for the target functions of the CA detection efficiency, it is necessary to determine the target functions, which is also a non-trivial task.

4. CA detection efficiency target functions

Productivity of CA detection efficiency target functions by ANN based SIEM-model can be determined by calculating values N_{TP}, N_{FN}, N_{FP} and N_{TN} as quantity of true positive (TP), false negative (FN), false positive (FP) and true negative (TN) results, respectively. The sum of TP, FN, FP, TN quantities N_Σ is the total number of results as shown in Figure 4.

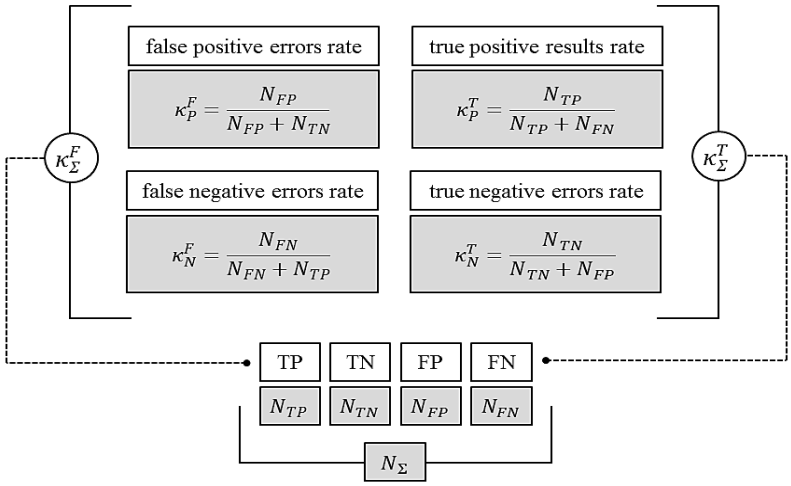

Figure 4. Basic scheme of calculating CA detection error rate based on TP, FN, FP and TN quantities

Based on these values, objective functions can be determined, as well as auxiliary coefficients:

- CA detection error rate $\kappa_{\Sigma}^{F} = \kappa_{P}^{F} + \kappa_{N}^{F}$ as sum of false positive errors rate (FPR) and false negative errors rate (FNR),
- CA detection correct result rate $\kappa_{\Sigma}^{T} = \kappa_{P}^{T} + \kappa_{N}^{T}$ as sum of true positive results rate (TPR) and true negative results rate (TNR),
- κ_{F1} and $\kappa_{\Re}$ coefficients of statistical analysis as F1-score and recall (RE), respectively (Figure 5),
- κ_{ACC} and κ_{PR} functions of CA detection as accuracy (ACC) and precision (PR), respectively (Figure 5).

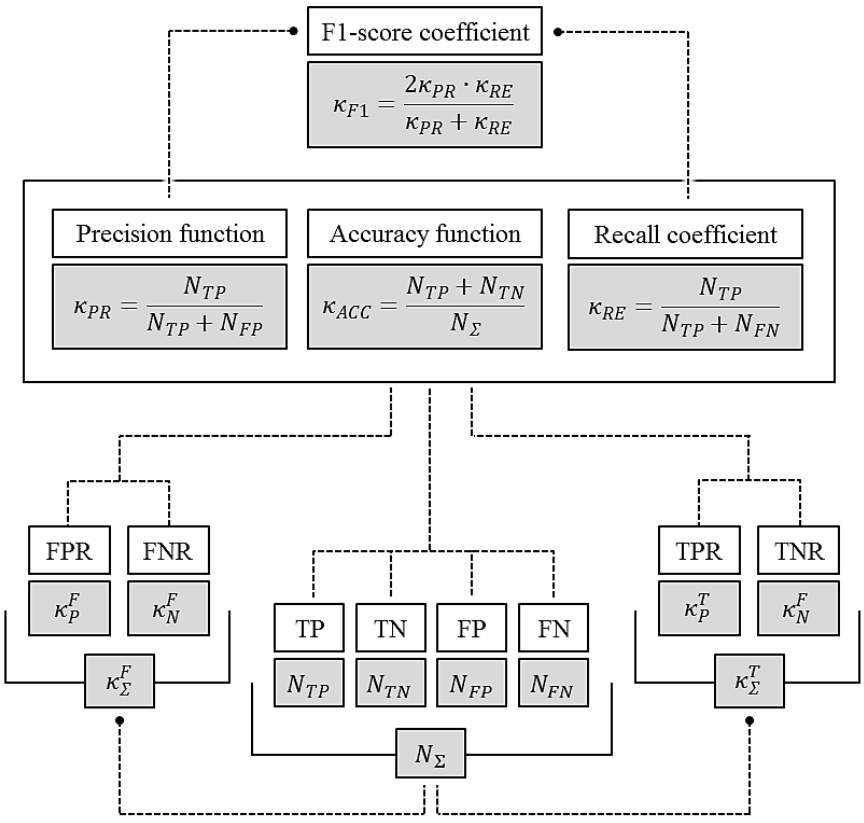

Figure 5. Basic scheme of calculating CA detection efficiency target functions based on TP, FN, FP, TN quantities and FPR, FNR, TPR, TNR values

Thereby, the analysis carried out shows that within the framework of this study estimation of the CA detection efficiency target functions could be based on calculation TP, FN, FP, TN quantities and FPR, FNR, TPR, TNR values as shown in Figure 5.

5. Complex optimization algorithm for SIEM-model

The analysis performed allows us to reduce the optimization problem of CA detection by ANN based SIEM-model to the problem of finding the global maximum or minimum of the objective function . In accordance with the study, the arguments of the objective functions are ANN architecture, training dataset of CA code, activation function, training method and training algorithm as it shown at Figure 6.

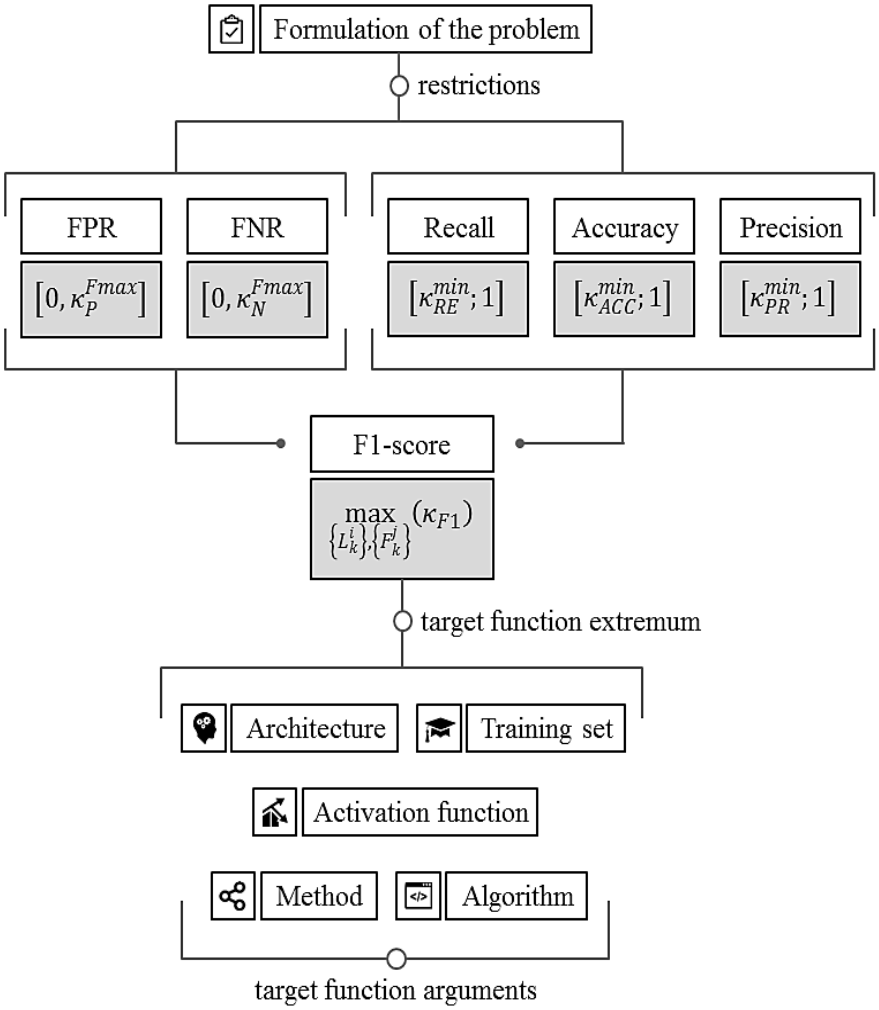

Figure 6. Efficiency optimization algorithm for ANN-based SIEM-model

The question of choosing the objective function is not so obvious; in the previous section, a number of indicators (FPR, FNR, TPR, TNR, F1-score, recall, accuracy and precision) were presented that, depending on the task, can be applied a target. Of course, the extrema of these functions in the general case would be obtained for

different values of the arguments, which does not make it possible to obtain an unambiguous solution for the optimization problem. It is proposed to consider the most universal indicator F1-score as an objective function and for other indicators to introduce restrictions on the maximum or minimum acceptable value as κ_P^{Fmax}, κ_N^{Fmax}, $\kappa_{\Re}^{m\in}$, κ_{ACC}^{min}, κ_{PR}^{min}, respectively (Figure 6). Thus, in accordance with the optimization algorithm for SIEM-model , the ANN architecture, training dataset, activation function, training method and training algorithm optimization is calculated for a maximum of F1-score which is estimated for the corresponding intervals of $\kappa_P^F \in [0, \kappa_P^{Fmax}]$, $\kappa_N^F \in [0, \kappa_N^{Fmax}]$, $\kappa_{\Re} \in \left[\kappa_{\Re}^{min}; 1\right]$, $\kappa_{ACC} \in \left[\kappa_{ACC}^{min}; 1\right]$ and $\kappa_{PR} \in \left[\kappa_{PR}^{min}; 1\right]$.

6. Conclusions

The growth of network resources distributed information systems has led to growth of importance of development of security information and event management system based on neural network analysis of internal and external threats. To solve the problem, within the framework of this study, the following approaches were proposed:

- generalized security information and event management scheme that includes analysis of cyber-attack threats and origin
- development of ANN methods of CA diagram that shows advantages of machine analysis methods, which are based on deep learning
- classification of the cyber-attack types based on the source, location and detections methods according to the security information and event management scheme
- statistical analysis of cyber-attack detection methods based on the neural networks performance indicators
- formalization of the program code machine analysis procedure through the definition objective functions, arguments of objective functions and auxiliary coefficients
- development of optimization algorithm for neural network based security information and event management model.

REFERENCES

1. HUMMER W., SATZGER B., DUSTDAR S.: Elastic stream processing in the cloud. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, **3** (2013)5, 333-345.
2. HEINTZ B., CHANDRA A., SITARAMAN, R.K.: Optimizing Timeliness and Cost in Geo-Distributed Streaming Analytics. IEEE Transactions on Cloud Computing, **8** (2020)1, 232–245.
3. AIYETORO T., OWOLAWI A.: Spectrum Management Schemes for Internet of Remote Things (IoRT) Devices in 5G Networks via GEO Satellite. Future Internet, **11** (2014)12, 17-28.
4. YIN F., LI X., LI X., LI Y.: Task Scheduling for Streaming Applications in a Cloud-Edge System. Security, Privacy, and Anonymity in Computation, Communication, and Storage 2019, 105–114.

5. SARKER, I.H.: Deep cybersecurity: A comprehensive overview from neural network and Deep Learning Perspective. SN Computer Science, **2** (2021)3, 11-14.
6. VIEGAS E., SANTIN A.O., FRANÇA A., JASINSKI R., PEDRONI V.A., OLIVEIRA L.S.: Towards an energy-efficient anomaly-based intrusion detection engine for embedded systems. IEEE Trans. Comput., **66** (2017)1, 163-177.
7. SAISINDHUTHEJA R., SHYAM G.K.: A deep belief network based attack detection using a secure SAAS framework. 2021 International Conference on Innovative Practices in Technology and Management (ICIPTM), 2021 23-29.
8. LI Y., LIU B., ZHAI S., CHEN M.: DDoS attack detection method based on feature extraction of Deep Belief Network. IOP Conference Series: Earth and Environmental Science, 2019, 25-32.
9. SUTSKEVER I.; VINYALS O.; LE Q.V.: Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems; Cambridge, 2014, 3104–3112.
10. BERMAN D.S., BUCZAK A.L., CHAVIS J.S., CORBETT C.L. A Survey of Deep Learning Methods for Cyber Security. Information, **10** (2019), 122-134.
11. SAINATH T.N., MOHAMED A.R., KINGSBURY B, RAMABHADRAN B.: Deep convolutional neural networks for LVCSR. In Proceedings of the 2013 IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP), NY, 2013; 8614–8618.
12. KARACA K.N., CETIN A.: Botnet attack detection using convolutional neural networks in the IOT Environment. 2021 International Conference on INnovations in Intelligent SysTems and Applications, INISTA, 2021, 54-65.
13. PRIYATHARSHINI R., ASWATH RAM. A.S., SHYAM SUNDAR R., NETHAJI NIRMAL G.: Real-time object recognition using region based convolution neural network and recursive neural network. International Journal of Recent Technology and Engineering, **8** (2019)4, 2813–2818.
14. JIANG P., WU H., XIN C.: Huge DeepPOSE: Detecting GPS spoofing attack via deep recurrent neural network. Chongqing University of Posts and Telecommunications, Chongqing, 2021, 16-24.
15. GOODFELLOW I., POUGET-ABADIE J., MIRZA M., XU B.; WARDEFARLEY D., OZAIR S., COURVILLE A., BENGIO Y.: Generative adversarial nets. In Advances in Neural Information Processing Systems, Cambridge, 2014; 2672-2680.
16. MAO X., LI Q.: Generative Adversarial Networks (GANs). Generative Adversarial Networks for Image Generation, 2021, 1–7.
17. YU Y., LONG J., CAI Z.: Network intrusion detection through stacking dilated convolutional autoencoders. Secur. Commun. Netw., 2017, 84-96.
18. MIRSKY Y., DOITSHMAN T., ELOVICI Y., SHABTAI A.: An ensemble of autoencoders for online network intrusion detection, In Proceedings of the IEEE 2017 International Conference on Information Networking (ICOIN), Da Nang, 2017; 712–717.
19. SARKER I.H.: Deep cybersecurity: A comprehensive overview from neural network and Deep Learning Perspective. SN Comp. Science, **2** (2021)3. 51-64.