

Using Machine Learning Techniques to Increase the Effectiveness of Cybersecurity

Vasyl Buhas¹, Ihor Ponomarenko¹, Valeriy Bugas¹, Andrii Ramskyi²,
and Volodymyr Sokolov²

¹ Kyiv National University of Technologies and Design, 2 Nemyrovycha-Danchenka str., Kyiv, 01011, Ukraine

² Borys Grinchenko Kyiv University, 18/2 Bulvarno-Kudriavska str., Kyiv, 04053, Ukraine

Abstract

In today's world, a great number of organizations generate and accumulate large amounts of information, which is of great value to owners, and is also considered by attackers as a valuable resource for enrichment. Any data storage system has vulnerabilities that will be exploited during cyberattacks. The inability to build a system secure enough against unauthorized access to data, forces companies to respond on an ongoing basis to evolving technologies of misappropriation of information by developing more effective methods of identifying and combating cyberattacks. This article examines the features of the use of machine learning methods to identify illegal access by third parties to the information of individuals and legal entities with economic and reputational damage. The study considers methods of processing various types of data (numerical values, textual information, video and audio content, images) that can be used to build an effective cybersecurity system. Obtaining a high level of identification of unauthorized access to data and combating their theft is possible through the implementation of modern machine learning approaches, which are constantly improving by creating innovative data processing algorithms and the use of powerful cloud computing services, acting as an element to counter rapidly evolving technologies.

Keywords

Cybersecurity, machine learning, neural networks, image recognition, optimization, information, dataset.

1. Introduction

In the context of digitalization, the number of Internet users is growing rapidly both in the private sector and in the business environment. The reorientation to the digital environment is associated with the intensive development of advanced information technologies that simplify the implementation of economic, technological and social processes. Respectively, the demand for innovative products is growing. In this aspect, it is important to pay special attention to the development of cloud technologies, which allow to accumulate large amounts of structured, semi-structured and unstructured information in the mode 24/7 [1]. At the same time, the methods of processing the generated information are actively evolving, which, owing to powerful and capacious servers, make it possible to speed up the data processing by using cloud computing. The existence of significant competition in the market of data collection and processing leads to an increase in the level of availability of cloud services with appropriate software solutions for most users. If at the beginning of the introduction of this technology the users were only TNCs and organizations with the support of national governments, in modern conditions a large number of small and medium-sized companies can use cloud services to optimize their business processes. Due to the variety of such services, users also apply cloud technologies to ensure the performance of certain works and to access certain services (e-mail, mobile banking, personal accounts, etc.) [2].

CPITS-II-2021: Cybersecurity Providing in Information and Telecommunication Systems, October 26, 2021, Kyiv, Ukraine
EMAIL: buhas.vv@knu.edu.ua (V. Buhas); ponomarenko.iv@knu.edu.ua (I. Ponomarenko); bugas.vv@knu.edu.ua (V. Bugas);
a.ramskyi@kubg.edu.ua (A. Ramskyi); v.sokolov@kubg.edu.ua (V. Sokolov)
ORCID: 0000-0001-8317-3350 (V. Buhas); 0000-0003-3532-8332 (I. Ponomarenko); 0000-0003-1046-9737 (V. Bugas); 0000-0001-7368-697X (A. Ramskyi); 0000-0002-9349-7946 (V. Sokolov)



© 2022 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

In many cases, the information generated in cloud services is of interest not only to data owners and related institutions, but also to third parties who are interested in obtaining certain illegal benefits as a result of the acquisition of personal information. The purpose of data theft may be to obtain trade secrets of companies, access to private user information, misappropriation of funds in bank accounts, interference with software to disrupt the management and production processes of various private and public organizations, access to state secrets, illegal access to servers supporting web resources of public administration bodies, etc. In any case, it is an illegal access to information in order to obtain certain benefits and cause significant harm to data owners. Due to the intensification of digitalization processes, the number of fraudulent actions with information in the global environment is gradually increasing, causing significant damage to international and national economic systems.

The spread of cybercrime compels various national and international institutions to develop and implement strategies to ensure a sufficient level of safety of valuable information resources, to involve relevant specialists and specialized software solutions. Building an effective cybersecurity system minimizes the risks of information loss and the use of illegally obtained data to cause economic and reputational losses to owners [3]. The process of ensuring data retention requires constant improvement of the technologies and techniques used, because due to the evolution of the information environment and the emergence of new approaches, fraudsters are getting more effective tools for illegal access to data.

Periodic cases of successful theft of information force companies to act quickly, leveling the vulnerabilities identified in the protection procedures. Improving the cybersecurity system involves the use of various approaches that are implemented in software, hardware and organizational solutions of specialized companies.

One of the effective ways to ensure data storage is machine learning methods, which are implemented as part of Data science. Thanks to the use of specialized algorithms it is possible with a high level of probability to identify fraudulent actions and limit unauthorized access to private information. The prospects of using machine learning methods as an important element of cybersecurity are based on the possibility of their improvement through the accumulation of relevant information, which allows to increase the accuracy of identification on the principle of “friend or foe” and to detect illegal actions. Various data (numerical values, textual information, video and audio content, images) can be used as sources of information for constructing models [4].

2. The Aim

The importance of detecting cybercrime in the digital environment and addressing existing challenges is vital to ensuring the stability of individual institutions and the system as a whole. Identifying serious threats requires significant financial resources to support scientific and practical developments in the field of cybersecurity. Modern cybercriminals use such approaches as phishing, host intrusion, malware integration, etc. to commit illegal acts [5]. Based on the existing needs in modern conditions, a large number of methods are developed and various scientific papers are published to ensure the protection of information through the introduction of innovations. The analysis of publications shows a significant interest of scientists in the use of machine learning methods to build modern and effective information security systems.

The presented research is devoted to the study of advanced methods of machine learning, the introduction of which allows to increase the efficiency and resilience of cybersecurity systems. Significant importance is attached to the use of various types of data in the process of building machine learning models. First of all, it is advisable to pay attention to the use of images to build neural networks as a way to raise the effectiveness of cybersecurity.

It should be noted that it is advisable to conduct research in the field of using machine learning methods to improve the efficiency of information security systems on a permanent basis. Confirmation of the reliability and prospects of this approach was made by a group of scientists in the implementation of algorithms that detect fake images of faces created by fraudsters through the use of ceramic masks with foil in certain areas to model the uneven heat in different parts of the head (Spoofing) [6]. In addition, scientists are addressing the issue of countering *Deepfake* technology, which allows original photos of the owners of certain information resources to create "live" and close to the original videos. Because of the improvement of Deepfake algorithms, it is possible to mislead the facial identification system and gain access to valuable information resources. Respectively, scientists and practitioners will constantly test various models for detecting fake photos and videos [7].

3. Models and Methods

Machine learning methods have become widespread in the vast majority of areas of human activity due to the ability to process large amounts of information and use the results of modeling to optimize the relevant processes. Constructing an effective cybersecurity system involves the implementation of a comprehensive strategy, which should include the use of machine learning methods to identify illegal interferences in the information system. Based on the peculiarities of building a system to combat fraud and the specifics of the available information, specialists in the field of Data Science use a variety of methods of machine learning. It should be noted that to obtain results with a high level of reliability several methods of machine learning can be used, among which the best in a particular situation will be selected not only from the standpoint of accuracy, but also based on the time spent on the process of modeling. Here are some of the most common machine learning techniques used to improve cybersecurity.

3.1. Using Neural Networks in Image Recognition

Images in modern conditions are very often used for cybersecurity due to the active use of webcams, smartphones and other specialized devices for tracking and identification. The development of effective neural networks through the use of images involves the implementation of several successive stages, but only their accuracy makes it possible to get a quality result. The specifics of the implementation of this model of machine learning involves the transformation of the graphic image into a digital form as a basis for modeling. In practice, there are two main approaches to image transformation: the 2D function and the 3D function. The 2D function is a function with x and y coordinates in space. The image in digital format is presented for calculations as the amplitude of the function F with finite values of x and y . When using the 3D function to transform images, the spatial coordinates x , y and z are entered. This approach to digital image conversion is called RGB (Red, Green, Blue).

The transformation of data into digital form has its own specifics and in some way may have a negative impact on the simulation results. A key disadvantage of using RGB color space is the inability to separate color data from other information. The use of three channels in the implementation of the RGB approach significantly slows down the process of computing within the corresponding neural network. The HSV (Hue, Saturation, Value) color space approach is devoid of the above drawback because it transforms the image into a single H (Hue) channel [9]. The practice of implementing RGB and HSV approaches shows the feasibility of choosing the best method of digitalization and the implementation of the corresponding neural network based on a set of factors. When modeling with neural models in some cases it is possible to achieve a better result due to RGB, and in others - due to the HSV approach.

Considering a set of algorithms for image processing while building an effective cybersecurity system, alongside with neural networks, such tools as Edge Detection in Image Processing, Fourier Transform in Image Processing, Gaussian Image Processing, Morphological Image Processing, Wavelet Image Processing are also used [10]. Based on the specifics of the problem, the features of the image dataset, the level of competence of the analyst, different algorithms can be used, but at the present stage of science development neural networks have the greatest prospects for implementation in the direction of graphic image recognition.

Neural networks are constructed as the basic elements of information processing (neurons), which are combined into a complex system with a certain number of layers. The principle of the neural network function is based on approaches to the functioning of the human brain: data are obtained from the environment, the process of modeling and learning to identify implicit patterns in information is realized by connecting neurons. The last stage involves obtaining predictive values or assigning an object to a specific group.

A typical neural network model has the following layers:

- An input layer.
- A hidden layer.
- An output layer.

The basic structure of the neural network contains input layers, hidden layers and output layers (Fig. 1). The input layers are used to enter the primary transformed information into the neural

network. At the next stage, the calculation process is implemented, which involves the activation of a certain number of neurons in accordance with the selected probabilities and the number of hidden layers used. The choice of architecture with a certain number of hidden layers is based on the analyst's experience and the specifics of the primary data. In the process of neural network implementation, a number of iterations are performed, which allow to track the level of accuracy of the implemented algorithm and to adjust the number of hidden layers to achieve an acceptable level of model quality.

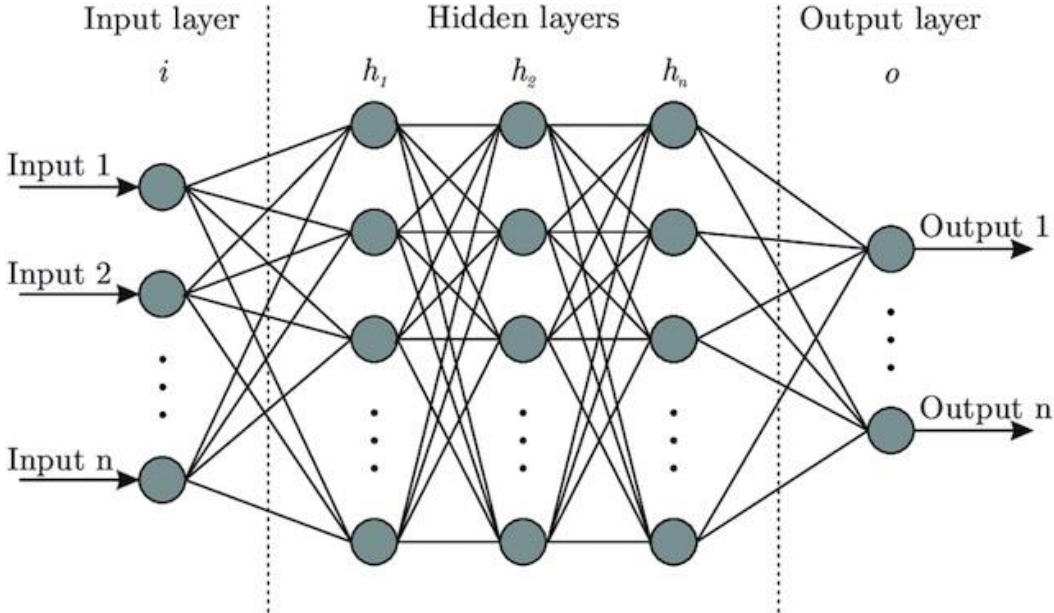


Figure 1: Basic structure of the neural network [11]

The algorithm for implementing the neural network in image processing is as follows:

1. A specific image according to the chosen approach is divided into pixels, which act as neurons of the first layer.
2. Each channel in accordance with scientifically sound approaches is assigned a certain probability (from 0 to 1).
3. Weighted sums are calculated by multiplying the weights by the corresponding input data, the calculated value is used as input to the hidden layers of the neural network.
4. A certain activation function is set for the source data, based on the specifics of the data, the neuron is activated or this channel is blocked.
5. Activated neurons act as data propagators to the next layers of the neural network.
6. The output neuron on the layer is selected automatically according to the maximum probability.
7. To assess the optimality, the error is calculated by subtracting from the expected value of the actual output. To approximate the optimal result, the calculated values are inversely propagated through the network to the previous layers.
8. The learning process involves the implementation of a certain number of iterations of direct and reverse propagation of data, at each stage there is a change in weights. The neural network stops the learning process at the stage of achieving optimal value. Fig. 3 illustrates a typical operation for a single neuron that is part of a neural network, where a_i – is the i -th input, w_i – is the i -th weight, z is the output, and g is a specific activation function.

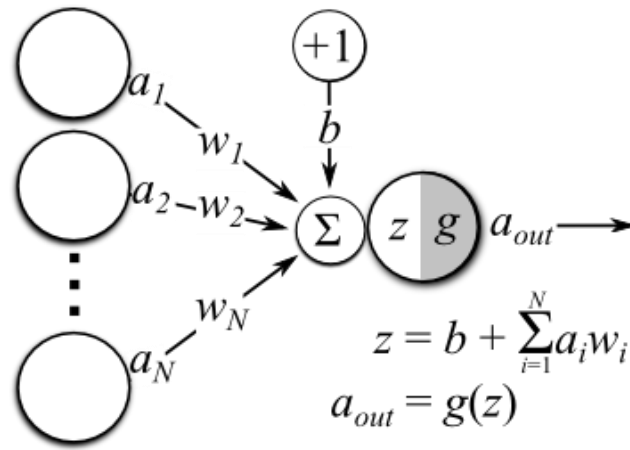


Figure 2: Operations on the neuron of the neural network [12]

The need to select the activation function for the correct implementation of the neural network was mentioned above. There are a large number of activation functions characterized by different specifics of neuronal activation (Fig. 3) [13]. It should be noted that in modern conditions the ReLU activation function (linear equalizer with "leakage") has become widespread [14].

3.2. Application of cluster analysis for identification of suspicious transactions

Constructing an effective cybersecurity system in the presence of digital and attributive indicators can be done through the integration of a classification system that provides for the identification of suspicious transactions. One of the approaches, which involves the division of operations into several groups without prior selection of them, is cluster analysis. In conditions of uncertainty, the presented method of machine learning, for example, allows you to assign individual records in network traffic to certain groups, which have characteristic features. At the next stage, the analysis of each of the groups is carried out to identify atypical records, which in the studied aggregates look like emissions [15]. The process of implementing cluster analysis based on available information involves the implementation of the following stages:

1. Recoding of textual information into digital form for the use of attributive indicators in the division of the aggregate into groups using this method of machine learning.
2. Standardization of data or use of the method of expert assessments. The data generated for the purposes of cluster analysis are contained in indicators characterized by different dimensions. The use of actual data leads to distortion of the results of cluster formation due to indicators of large dimension, while the impact of indicators of small dimension will not be significant. Due to the standardization of primary data, the indicators lead to a moderate form of influence on the process of cluster construction. In addition, the method of expert assessments is expected to be used when it is necessary to provide different weights to a certain indicator, assigning greater weights to more significant from the point of view of the researcher indicators. Owing to the process of standardization, we move to a certain same-type description of all the indicators used, i.e. a new conditional unit of measurement is calculated, which allows a formal comparison of objects [16]. Standardization of indicators is carried out according to the following formula:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j} \quad (1)$$

When allocating indicators of direct and reverse action, it is advisable to divide the indicators into stimulants and destimulators, respectively. The following formulas are used to standardize each of the types of indicators:

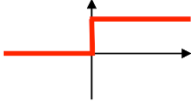
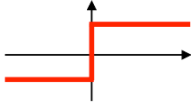
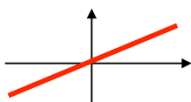


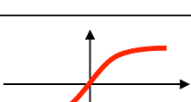
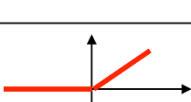
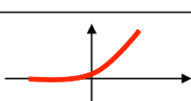
$$z_{ij} = \frac{x_{ij} - x_{\min}}{x_{\max} - x_{\min}} \quad (\text{direct direction of action}), \quad (2)$$

$$z_{ij} = \frac{x_{\max} - x_{ij}}{x_{\max} - x_{\min}} \quad (\text{reverse direction of action}), \quad (3)$$

where x_{ij} is the value of the i^{th} indicator in the j^{th} element;

x_{\min} is the minimum value of the i^{th} indicator;

x_{\max} is the maximum value of the i^{th} indicator.

Activation function	Equation	Example	1D Graph
Unit step (Heaviside)	$\phi(z) = \begin{cases} 0, & z < 0, \\ 0.5, & z = 0, \\ 1, & z > 0, \end{cases}$	Perceptron variant	
Sign (Signum)	$\phi(z) = \begin{cases} -1, & z < 0, \\ 0, & z = 0, \\ 1, & z > 0, \end{cases}$	Perceptron variant	
Linear	$\phi(z) = z$	Adaline, linear regression	
Piece-wise linear	$\phi(z) = \begin{cases} 1, & z \geq \frac{1}{2}, \\ z + \frac{1}{2}, & -\frac{1}{2} < z < \frac{1}{2}, \\ 0, & z \leq -\frac{1}{2}, \end{cases}$	Support vector machine	
Logistic (sigmoid)	$\phi(z) = \frac{1}{1 + e^{-z}}$	Logistic regression, Multi-layer NN	
Hyperbolic tangent	$\phi(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$	Multi-layer Neural Networks	
Rectifier, ReLU (Rectified Linear Unit)	$\phi(z) = \max(0, z)$	Multi-layer Neural Networks	
Rectifier, softplus	$\phi(z) = \ln(1 + e^z)$	Multi-layer Neural Networks	

Copyright © Sebastian Raschka 2016
(<http://sebastianraschka.com>)

Figure 3: Main functions of the activation [13]

Among the available indicators it is advisable to use only statistically significant indicators by calculating p-value. In order to improve the results, it is also logical in the process of clustering to

carry out checking of any group of sets by overriding various combinations of indicators in the process of modeling implementation.

3. Determining the number of clusters is carried out using one of the methods: graph of silhouette width, graph of GAP-statistics as well as the “elbow” method. To determine the optimal number of clusters, it is also necessary to analyze the obtained groups, because in some situations the formation of a group with only one unit of the population takes place. The outlined situation indicates the need to use another method to identify the number of clusters [17].

4. Choice of clustering method. Among clustering methods, hierarchical and non-hierarchical cluster analysis are the most popular. In the process of choosing the optimal approach to the formation of groups, it is possible to conduct experiments alternately using the above clustering methods. Among the available indicators, it is advisable to use only statistically significant indicators for clustering, calculating the p-value. To improve the results, it is also advisable in the process of clustering to check any groups of the set by searching for various combinations of indicators in the modeling process.

The main parameters for the formation of clusters are the measure of distance and the rule of aggregation, which allows direct grouping of aggregate units into appropriate groups according to the formed system of indicators. The measure of distance (tree clustering method) allows us to select individual clusters based on available indicators. It should be noted that the distance between the formed clusters is measured in any dimension. An example of a tree clustering method is the Euclidean distance, which is calculated by the following formula:

$$d_{ij} = \sqrt{\sum_{k=1}^p (z_{ik} - z_{jk})^2}, \quad (4)$$

where d_{ij} is the distance between the objects i and j , and z_{ik} is standardized value of k variable for i^{th} object.

Non-hierarchical clustering is characterized by a certain flexibility in the redistribution of aggregate units between clusters in the optimization process. At the first stage, the centers of clusters are identified in accordance with their number. At the next stage of modeling, the distances of each of the elements to the existing centers are estimated and assigned to the nearest cluster according to the established threshold distances.

5. Validation of clusters. Evaluation of the adequacy of the obtained clusters is as follows:

- External validation is implemented by conducting a comparative analysis of the results of cluster analysis with the reference result.
- Relative validation involves the study of the structure of clusters, provided that the values of the parameters are used in the implementation of a separate method of cluster analysis.
- Internal validation is carried out on the basis of internal information on the implementation of the cluster formation procedure.
- Assessment of the stability of clustering involves the implementation of cluster analysis algorithms based on various samples [10, 18, 19].

6. Selection of the optimal clustering method. Examining the results obtained after the implementation of various approaches to cluster analysis, the optimal option is selected in accordance with the needs. The selection process can be based both on the system of qualitative assessments of the formed clusters and on the basis of the obtained visualizations of the formed groups.

4. Further Research

The results obtained in the study show the effectiveness of the use of machine learning methods to identify illegal actions of fraudsters to obtain access to information through the use of various data (numerical values, textual information, video and audio content, images). Further research should be focused on improving various approaches in the sphere of Data science in order to perfect the cybersecurity system and constantly bring data protection in line with existing realities. Optimization of machine learning algorithms in the field of cyberattack involves the use of such specialized programming languages as Python with the connection of appropriate libraries (Keras, PyTorch,

Scikit-learn, TensorFlow, etc.) [20]. The development of deep learning technologies leads to the emergence of more complex neural networks, which allow to optimize the cybersecurity system. The implementation of comprehensive research in the field of deep learning will test a variety of models and offer the best solutions to the market.

5 Conclusions

The introduction of innovative technologies expands the company's ability to collect, process and use large amounts of information to optimize key processes. Comprehensive databases of companies arouse the interest of outsiders, which leads to the creation of various tools for illegal acquisition of information. To counter cyberattacks, effective security systems for storing valuable information and a multi-level algorithm for access to relevant resources are created. Advanced approaches to the implementation of effective and robust cybersecurity systems involve the use of machine learning methods, which are implemented by building appropriate models based on structured, semi-structured and unstructured data. Practice shows the effectiveness of the use of neural networks in the process of combating spoofing, as fraudsters are trying to seize someone else's data in order to create high-quality forged images. Cluster analysis allows you to segment objects based on a system of various metrics, identifying specific groups and distinguishing emissions that are likely to be considered suspicious transactions.

6 References

- [1] J. Xie, et al., Efficient Indexing Mechanism for Unstructured Data Sharing Systems in Edge Computing, in: IEEE Conference on Computer Communications, 820–828, 2019. <https://doi.org/10.1109/infocom.2019.8737617>
- [2] Y. Kravchenko, et al., Evaluating the Effectiveness of Cloud Services, 2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT), 120–124, 2019. <https://doi.org/10.1109/atit49449.2019.9030430>
- [3] A. Corallo, M. Lazoi, M. Lezzi, Cybersecurity in the context of industry 4.0: A structured classification of critical assets and business impacts, Computers in Industry, vol. 114, 103–165, 2020. <https://doi.org/10.1016/j.compind.2019.103165>
- [4] I. H. Sarker, et al. Cybersecurity data science: an overview from machine learning perspective, J Big Data, 7, 41, 2020. <https://doi.org/10.1186/s40537-020-00318-5>
- [5] J. E. Thomas, Individual cyber security: Empowering employees to resist spear phishing to prevent identity theft and ransomware attacks. International Journal of Business Management, 12(3), 1–23, 2018. <https://doi.org/10.5539/ijbm.v13n6p1>
- [6] Z. Yu, et al., Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 5295–5305, 2020.
- [7] M. H. Maras, A. Alexandrou, Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos. International Journal of Evidence & Proof, 23(3), 255–262, 2019. <https://doi.org/10.1177/1365712718807226>
- [8] M. Kubanek, J. Bobulski, J. Kulawik, A Method of Speech Coding for Speech Recognition Using a Convolutional Neural Network. Symmetry, 11, 1185, 2019. <https://doi.org/10.3390/sym11091185>
- [9] A. Radovan, Z. Ban, Prediction of HSV color model parameter values of cloud movement picture based on artificial neural networks, 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 1110–1114, 2018. <https://doi.org/10.23919/mipro.2018.8400202>
- [10] M. Gandhi, J. Kamdar, M. Shah, Preprocessing of Non-symmetrical Images for Edge Detection. Augment Hum Res 5, 10, 2020. <https://doi.org/10.1007/s41133-019-0030-5>
- [11] Introduction to Different Activation Functions for Deep Learning. <https://medium.com/@shrutijadon10104776/survey-on-activation-functions-for-deep-learning-9689331ba092>
- [12] Everything you need to know about Neural Networks. <https://hackernoon.com/everything-you-need-to-know-about-neural-networks-8988c3ee4491>

- [13] Activation Functions for Artificial Neural Networks. http://rasbt.github.io/mlxtend/user_guide/general_concepts/activation-functions/
- [14] D. Zou, et al. Gradient descent optimizes over-parameterized deep ReLU networks. *Mach Learn* 109, 467–492 (2020). <https://doi.org/10.1007/s10994-019-05839-6>
- [15] K. Demertzis, L. Iliadis, S. Spartalis, A spiking one-class anomaly detection framework for cyber-security on industrial control systems. In: *International Conference on Engineering Applications of Neural Networks*, pp. 122–134. Springer, Cham (2017).
- [16] W. Liang, et al., An Industrial Network Intrusion Detection Algorithm Based on Multifeature Data Clustering Optimization Model, in: *IEEE Transactions on Industrial Informatics*, vol. 16, no. 3, 2063-2071, 2020, <https://doi.org/10.1109/TII.2019.2946791>.
- [17] Y. Chunhui, H. Yang, Research on K-Value Selection Method of K-Means Clustering Algorithm, 2, no. 2: 226–235, 2019. <https://doi.org/10.3390/j2020016>
- [18] O. Romanovskyi, et al., Automated Pipeline for Training Dataset Creation from Unlabeled Audios for Automatic Speech Recognition, in: *Advances in Computer Science for Engineering and Education IV*, 25–36, 2021. https://doi.org/10.1007/978-3-030-80472-5_3
- [19] Z. B. Hu, et al., Authentication System by Human Brainwaves Using Machine Learning and Artificial Intelligence, in: *Advances in Computer Science for Engineering and Education IV*, 374–388, 2021. https://doi.org/10.1007/978-3-030-80472-5_31
- [20] C. D. Costa, *Python libraries for modern machine learning models & projects*, 2020. <https://towardsdatascience.com/best-python-libraries-for-machine-learning-and-deep-learning-b0bd40c7e8c>