Software Impacts 21 (2024) 100681

Contents lists available at ScienceDirect

Software Impacts

journal homepage: www.journals.elsevier.com/software-impacts

Original software publication

Adversarial attack defense analysis: An empirical approach in cybersecurity perspective

Kousik Barik [a], Sanjay Misra [b,c,*]

[a] Department of Computer Science, University of Alcala, Madrid, Spain
[b] Department of Computer Science and Communication, Østfold University College, Halden, Norway
[c] Department of Applied Data Science, Institute for Energy Technology, Halden, Norway

ARTICLE INFO

Keywords:
Adversarial attack
Cybersecurity
Adversarial machine learning
Adversarial defense
Deep learning

ABSTRACT

Advancements in artificial intelligence in the cybersecurity domain introduce significant security challenges. A critical concern is the exposure of deep learning techniques to adversarial attacks. Adversary users intentionally attempt to mislead the techniques by infiltrating adversarial samples to mislead the prediction of security devices. The study presents extensive experimentation of defense methods using Python-based open-source code with two benchmark datasets, and the outcomes are demonstrated using evaluation metrics. This code library can be easily utilized and reproduced for cybersecurity research on countering adversarial attacks. Exploring strategies for protecting against adversarial attacks is significant in enhancing the resilience of deep learning techniques.

Code metadata

Current code version	*1.0*
Permanent link to code/repository used for this code version	https://github.com/SoftwareImpacts/SIMPAC-2024-147
Permanent link to reproducible capsule	
Legal code license	*MIT License*
Code versioning system used	*git*
Software code languages, tools and services used	*Python, Jupyter Notebook*
Compilation requirements, operating environments and dependencies	*Python, Scikit-learn, ART, DeepFool, Limited-memory BFGS*
If available, link to developer documentation/manual	https://github.com/kousikbarik/Adversarial-attack-defense-analysis/blob/main/README.md
Support email for questions	sanjay.misra@ife.no

1. Introduction to DL and adversarial attacks

The remarkable advances in Artificial Intelligence (AI) find extensive broad use in various fields such as cybersecurity [1,2], sentiment analysis [3,4], medical [5], digital forensics [6,7], etc., underscore the necessity to endure the protection and trustworthiness of deep learning (DL)-based solutions. DL models have lately been found vulnerable to adversarial attacks, where crackers attempt to trick the system by introducing incorrect adversarial concern inputs. For instance, in the field of cyber protection, adversarial attacks specifically target the weaknesses present in security systems, which are designed using DL techniques. These attacks affect the manipulation of input packet information by introducing small malicious noise, which results in false model predictions [8]. These illustrations emphasize the need to analyze and improve AI applications' stability and safety. Given the significant influence of AI on different aspects of our lives, it is crucial to prioritize resolving these security challenges to guarantee the dependability and credibility of AI-driven techniques [9]. It is crucial to ensure the security of DL techniques. Also, extensive studies have been performed to enhance the implementation of deep learning (DL) techniques and improve their outcome measures [10,11]. Nevertheless, it is imperative to consider the generality and resilience capacity in the present day, given the constant emergence of unfamiliar cyber security risks [12].

As DL-based IDS techniques become more common, adversarial attacks become more significant. Therefore, it is crucial to develop effective defense strategies to reduce negative outcomes and ensure

* Corresponding author at: Department of Computer Science and Communication, Østfold University College, Halden, Norway.
E-mail addresses: kousik.kousik@edu.uah.es (K. Barik), sanjay.misra@ife.no (S. Misra).

https://doi.org/10.1016/j.simpa.2024.100681
Received 13 June 2024; Accepted 28 June 2024

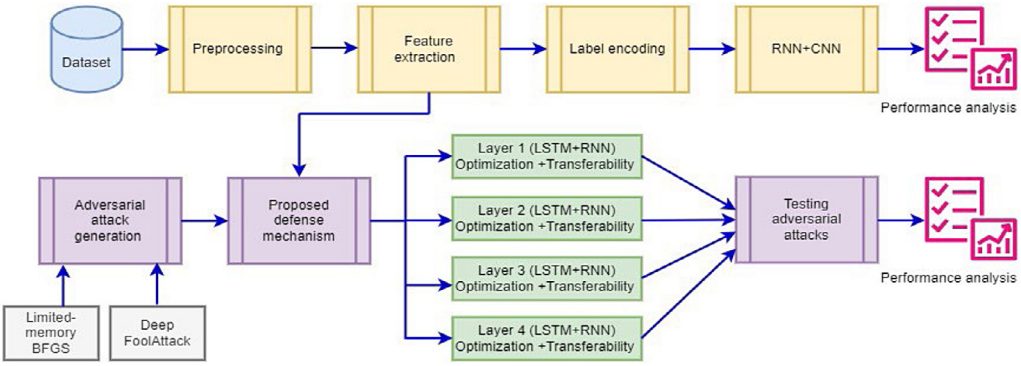

Fig. 1. An outline of the proposed model.

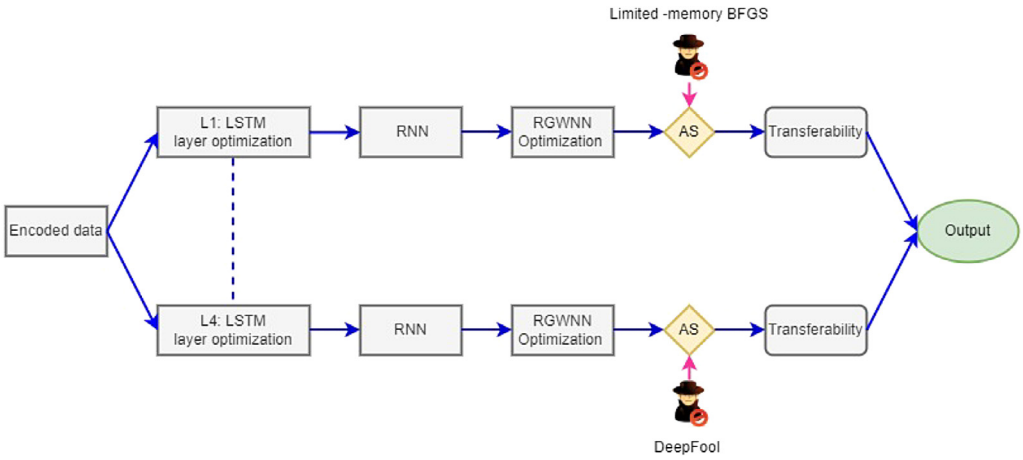

Fig. 2. Architecture of the proposed model.

reliability and stability against adversarial attacks [13]. This presented study explores the susceptibility of the DL-based technique against two adversarial attack techniques, i.e., Limited-memory BFGS [14] and DeepFool [15]. The study illustrates a DL-based defense technique, i.e., a combined RNN and LSTM optimization defense mechanism to counter adversarial attacks. In preprocessing, RobustScaler is utilized, and in feature extraction, Linear Discriminant Analysis (LDA) is employed [16]. It uses four LSTM architecture layers, i.e., Layer 1, Layer 2, Layer 3, and Layer 4. The Grey Wolf Optimization (GWO) [17] is used to fine-tune the LSTM model's hyperparameters to achieve better optimization performance, and RNN is used as a source of LSTM layers [18]. Transferability learning is used to the model's robustness when subjected to adversarial attacks, such as those generated by the DeepFool and Limited-memory BFGS methods. This process involves creating adversarial examples, inputs designed to deceive the model into making incorrect predictions. The Adam optimizer is known for its significance in addressing sparse gradients on noisy data. The loss function utilized is absolute cross-entropy, suitable for forecast studies where the model predicts probabilities across multiple classes. The pre-trained model is tested against these adversarial examples to evaluate its resilience and robustness. This step is vital for understanding the model's security and reliability, as adversarial attacks could be a threat. The model's outcome is evaluated using a testing dataset.

2. Code functionality and adversarial attacks

The software enables researchers to design defense techniques to protect against adversarial attacks using DL-based models. The presented study addresses the research questions mentioned below.

- How can LSTM architecture layer optimization be performed, and can RNN be used as a source of LSTM layers, a useful classifier, and improve the attack detection rate?
- How can DeepFool and Limited-memory BFGS be employed to develop adversarial attacks?
- What is the importance of transfer learning in adversarial attacks?
- What techniques can be used to enhance the ability of DL models to identify adversarial attacks by combining hybrid defense techniques?

Fig. 1 displays the outline of the presented defense model. It incorporates the implementations for identifying adversarial attacks in DL-based techniques, GWO, and adversarial sample creation using DeepFool and Limited-memory BFGS. The software is designed in Python, and various libraries are employed: ART (Adversarial Robustness Toolbox) library [19], Scikit-learn, NumPy, Tensorflow, DeepFool, and Limited-memory BFGS attack. Two publicly available datasets are used, i.e., CIC-IDS-2017 [20] and CSE-CIC-IDS2018 [21]. Fig. 2. Presents the design components of the proposed model. The source code includes the four files mentioned below.

1. Limited-memory BFGS_2017.ipynb: The CIC-IDS-2017 dataset is processed using RobustScaler. Linear discriminant analysis (LDA) is used to extract features. The proposed model uses a combination of LSTM and RNN classifiers for prediction. The adversarial attacks are created employing the Limited-memory BFGS technique. It uses four LSTM architecture layers. The GWO is employed to optimize the LSTM model's hyperparameters for better performance, and RNN is used as an input for LSTM layers. The ReLU (Rectified Linear Unit) and Adam activation functions are employed. Transferability learning enhances the

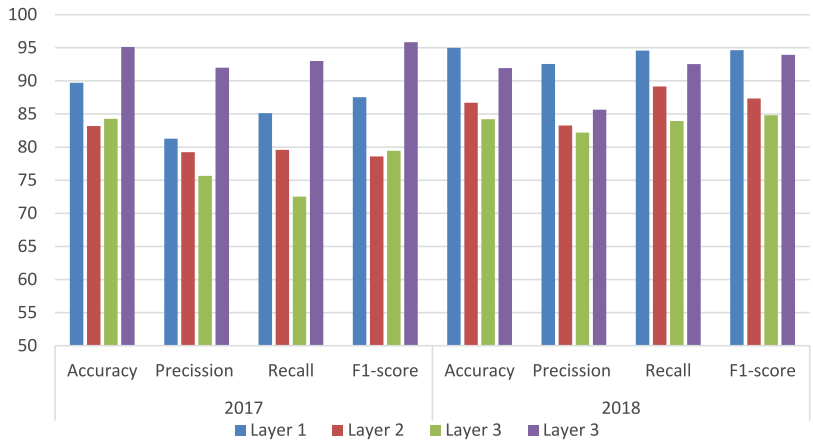

Fig. 3. Comparative outcome using Limited-memory BFGS attack.

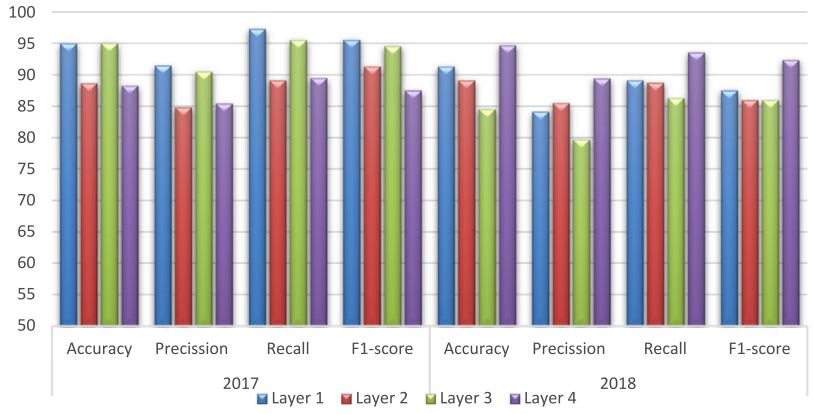

Fig. 4. Comparative outcome using DeepFool attack.

model's performance and creates adversarial examples, inputs designed to deceive the model into making inaccurate predictions. The presented defense model is trained with 80% of the training dataset and tested with 20% of the testing data with four different architectures, i.e., Layer 1, Layer 2, Layer 3, and Layer 4, to compile and predict the outcome.

2. Deepfool_2017: This scenario uses the CIC-IDS-2017, and other methods are followed as described earlier. The adversarial samples are developed using the DeepFool technique. The outcome of the model is demonstrated.
3. Limited-memory BFGS_2018.ipynb: This method employs the CSE-CIC-IDS2018; other procedures were followed earlier. The proposed model's outcome is evaluated using a limited-memory BFGS attack to validate its robustness.
4. Deepfool_2018.ipynb: This scenario utilizes the CSE-CIC-IDS2018 dataset and follows the previously specified methodologies. The adversarial samples are generated using the DeepFool method. The outcome is demonstrated to validate the proposed model's effectiveness.

Figs. 3 and 4 summarize the relative results of the proposed defense model using two different datasets, including different evaluation metrics against Limited-memory BFGS and DeepFool adversarial attacks.

Figs. 5 and 6 outline the outcomes of the presented defense model using three additional important evaluation parameters: FNR (False Negative Rate), FPR (False Positive Rate), and ASR (Attack Success Rate).

The presented defense model only partially nullifies them. This signifies the potential to enhance the model's capability to oppose and protect against adversarial attacks by increasing its resilience. More related research is needed. It is imperative to recognize that the presented model has specific constraints. While it is difficult to achieve an impenetrable defense, the presented model significantly increased the amount of time and computational resources a potential attacker would need to carry out an attack successfully. In real-world scenarios, the longer time and higher commuting costs associated with an attack can make it ineffectual or economically unfeasible, acting as an obstruction and improving safety benchmarks.

3. Software impacts

The presented defense model, which focuses on deploying defense mechanisms to counter adversarial attacks, is designed with user-friendliness. While a few research projects currently exist on deep learning (DL) based adversarial attack detection, there is a lack of publicly accessible code that includes security approaches designed to counter against adversarial attacks. The code is accessible to the public. It can be used to expand upon the existing study on DL-based adversarial attacks. This user-friendly software, developed in Python, comes with detailed explanations.

The software presented is not merely a tool but a pragmatic solution that aids scientists in comprehending the ramifications of adversarial attacks on DL techniques. The software showcases two methods, Limited-memory BFGS and DeepFool, for developing adversarial attacks. Our software not only offers a comprehensive understanding of adversarial attacks but also provides three robust defense mechanisms. These include a combination of LSTM and RNN with four LSTM architecture layer optimizations, the GWO global optimum in a search space, and transfer learning, which creates adversarial examples. The

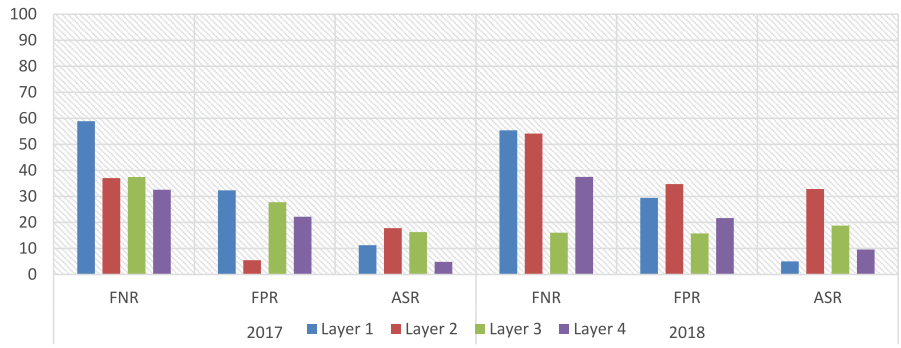

Fig. 5. Comparative outcome in terms of FNR, FPR, and ASR using Limited-memory BFGS attack.

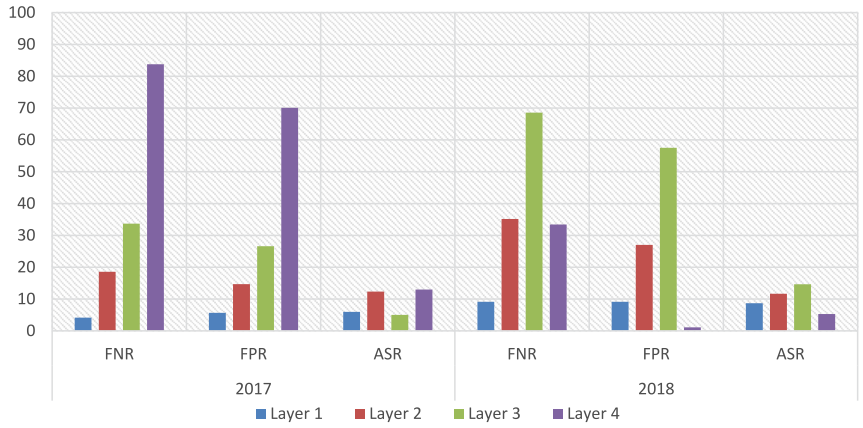

Fig. 6. Comparative outcome in terms of FNR, FPR, and ASR using DeepFool attack.

software's experimental study of two public standard datasets indicates its capability to control and effectively determine adversarial attacks. Using the presented software, researchers can design proficiency design and significantly improve DL models against adversarial attacks.

The presented defense model has supported our studies on mitigating cyber threats. This model can be applied in practice, expanding the analysis of adversarial attacks. Its primary focus is developing and implementing defensive measures against adversarial attacks on the DL model. Three research papers that were published and were supported by this practical process include:

1. The study to design a cybersecurity dataset and a presented framework employing deep learning-based approaches to identify cyberattacks was published in [22].
2. The paper, published in [23], aimed to design adversarial samples using various approaches and present an enhanced model with a protection strategy to evaluate the effectiveness of IDS.
3. IDS-Anta is a publicly available code that utilizes hybrid protection approaches to identify adversarial attacks in IDS [24].

This software delivered a similar and significant contribution to the current studies [25–27]. These papers proposed a defense strategy for DL methods to protect against adversarial attacks. The presented model's defense mechanism, characterized by its versatility, enables researchers to explore several different fields, including data mining and computer vision.

4. Conclusions and future work

Cyberattacks are growing exponentially, and their magnitude and intricacy are rising. It is imperative to discern various forms of attacks and comprehend methodologies. The software provides a user-friendly implementation to counter adversarial attacks. Due to its concise presentation, scientists can employ this software to counter adversarial attacks. This study's insights indicate various potential research and development opportunities, which can be expanded and enriched in four primary future directions.

1. The presented study used LSTM and RNN as classifiers, and additional standard datasets can explore other DL-based hybrid modes with optimization strategies.
2. The study was carried out under controlled development testing conditions. However, new challenges can be investigated when the proposed approach is applied in a real-world setting.
3. By combining the heuristic and signature-based methods, it is possible to comprehensively analyze potential risks and reduce the occurrence of false alarms.
4. Subsequent research can investigate the ramifications of evolving adversarial strategies for IDS. Staying updated on evolving attack methodologies is crucial to enhancing the robustness and efficiency of DL. Therefore, there is a growing demand for more research on adversarial attacks, incorporating novel techniques and defense measures.

CRediT authorship contribution statement

Kousik Barik: Conceptualization, Investigation, Methodology, Writing – original draft. **Sanjay Misra:** Conceptualization, Methodology, Project administration, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Miranda-García, A.Z. Rego, I. Pastor-López, B. Sanz, A. Tellaeche, J. Gaviria, P.G. Bringas, Deep learning applications on cybersecurity: A practical approach, Neurocomputing 563 (2024) http://dx.doi.org/10.1016/j.neucom.2023.126904.

[2] S.M. JS, M. Thirunavukkarasu, N. Kumaran, D. Thamaraiselvi, Deep learning with blockchain based cyber security threat intelligence and situational awareness system for intrusion alert prediction, Sustain. Comput. Inform. Syst. 42 (2024) http://dx.doi.org/10.1016/j.suscom.2023.100955.

[3] K. Barik, S. Misra, Analysis of customer reviews with an improved VADER lexicon classifier, J. Big Data 11 (1) (2024) 10, http://dx.doi.org/10.1186/s40537-023-00861-x.

[4] K. Barik, S. Misra, A.K. Ray, A. Bokolo, LSTM-DGWO-based sentiment analysis framework for analyzing online customer reviews, Comput. Intell. Neurosci. 2023 (1) (2023) 6348831, http://dx.doi.org/10.1155/2023/6348831.

[5] K. Barik, S. Misra, S. Chockalingam, M. Hoffmann, Data analytics, digital transformation, and cybersecurity perspectives in healthcare, in: International Workshop on Secure and Resilient Digital Transformation of Healthcare, Springer Nature Switzerland, Cham, 2023, pp. 71–89, http://dx.doi.org/10.1007/978-3-031-55829-0_5.

[6] K. Barik, A. Abirami, K. Konar, S. Das, Research perspective on digital forensic tools and investigation process, Illum. Artif. Intell. Cybersecur. Forensics (2022) 71–95, http://dx.doi.org/10.1007/978-3-030-93453-8_4.

[7] K. Barik, S. Das, K. Konar, B.C. Banik, A. Banerjee, Exploring user requirements of network forensic tools, Glob. Transitions Proc. 2 (2) (2021) 350–354, http://dx.doi.org/10.1016/j.gltp.2021.08.043.

[8] T. Huang, Q. Zhang, J. Liu, R. Hou, X. Wang, Y. Li, Adversarial attacks on deep-learning-based SAR image target recognition, J. Netw. Comput. Appl. 162 (2020) 102632, http://dx.doi.org/10.1016/j.jnca.2020.102632.

[9] A. Habbal, M.K. Ali, M.A. Abuzaraida, Artificial intelligence trust, risk and security management (AI TRiSM): Frameworks, applications, challenges and future research directions, Expert Syst. Appl. 240 (2024) 122442, http://dx.doi.org/10.1016/j.eswa.2023.122442.

[10] Z. Ahmad, A. Shahid Khan, C. Wai Shiang, J. Abdullah, F. Ahmad, Network intrusion detection system: A systematic study of machine learning and deep learning approaches, Trans. Emerg. Telecommun. Technol. 32 (1) (2021) e4150, http://dx.doi.org/10.1002/ett.4150.

[11] S.M. Kasongo, A deep learning technique for intrusion detection system using a Recurrent Neural Networks based framework, Comput. Commun. 199 (2023) 113–125, http://dx.doi.org/10.1016/j.comcom.2022.12.010.

[12] M. Dacorogna, N. Debbabi, M. Kratz, Building up cyber resilience by better grasping cyber risk via a new algorithm for modelling heavy-tailed data, European J. Oper. Res. 311 (2) (2023) 708–729, http://dx.doi.org/10.1016/j.ejor.2023.05.003.

[13] B. Sharma, L. Sharma, C. Lal, S. Roy, Anomaly based network intrusion detection for IoT attacks using deep learning technique, Comput. Electr. Eng. 107 (2023) 108626, http://dx.doi.org/10.1016/j.compeleceng.2023.108626.

[14] J. Zhang, W. Qian, J. Cao, D. Xu, LP-BFGS attack: An adversarial attack based on the Hessian with limited pixels, Comput. Secur. (2024) 103746, http://dx.doi.org/10.1016/j.cose.2024.103746.

[15] X. Yuan, S. Han, W. Huang, H. Ye, X. Kong, F. Zhang, A simple framework to enhance the adversarial robustness of deep learning-based intrusion detection system, Comput. Secur. 137 (2024) 103644, http://dx.doi.org/10.1016/j.cose.2023.103644.

[16] O.H. Abdulganiyu, T. Ait Tchakoucht, Y.K. Saheed, A systematic literature review for network intrusion detection system (IDS), Int. J. Inf. Secur. 22 (5) (2023) 1125–1162, http://dx.doi.org/10.1007/s10207-023-00682-2.

[17] Y.K. Saheed, S. Misra, A voting gray wolf optimizer-based ensemble learning models for intrusion detection in the Internet of Things, Int. J. Inf. Secur. (2024) 1–25, http://dx.doi.org/10.1007/s10207-023-00803-x.

[18] A.Q. Md, S. Kapoor, C.J. AV, A.K. Sivaraman, K.F. Tee, H. Sabireen, N. Janakiraman, Novel optimization approach for stock price forecasting using multi-layered sequential LSTM, Appl. Soft Comput. 134 (2023) 109830, http://dx.doi.org/10.1016/j.asoc.2022.109830.

[19] C. Eleftheriadis, A. Symeonidis, P. Katsaros, Adversarial robustness improvement for deep neural networks, Mach. Vis. Appl. 35 (3) (2024) 35, http://dx.doi.org/10.1007/s00138-024-01519-1.

[20] A. Thakkar, R. Lohiya, A review of the advancement in intrusion detection datasets, Procedia Comput. Sci. 167 (2020) 636–645, http://dx.doi.org/10.1016/j.procs.2020.03.330.

[21] J.L. Leevy, T.M. Khoshgoftaar, A survey and analysis of intrusion detection models based on cse-cic-ids2018 big data, J. Big Data 7 (2020) 1–19, http://dx.doi.org/10.1186/s40537-020-00382-x.

[22] K. Barik, S. Misra, K. Konar, L. Fernandez-Sanz, M. Koyuncu, Cybersecurity deep: Approaches, attacks dataset, and comparative study, Appl. Artif. Intell. 36 (1) (2022) 2055399, http://dx.doi.org/10.1080/08839514.2022.2055399.

[23] K. Barik, S. Misra, L. Fernandez-Sanz, Adversarial attack detection framework based on optimized weighted conditional stepwise adversarial network, Int. J. Inf. Secur. (2024) 1–24, http://dx.doi.org/10.1007/s10207-024-00844-w.

[24] K. Barik, S. Misra, IDS-anta: An open-source code with a defence mechanism to detect adversarial attacks for intrusion detection system, Software Impacts (2024) 100664, http://dx.doi.org/10.1016/j.simpa.2024.100664.

[25] H. Mohammadian, A.A. Ghorbani, A.H. Lashkari, A gradient-based approach for adversarial attack on deep learning-based network intrusion detection systems, Appl. Soft Comput. 137 (2023) 110173, http://dx.doi.org/10.1016/j.asoc.2023.110173.

[26] M. Aurangzeb, Y. Wang, S. Iqbal, A. Naveed, Z. Ahmed, M. Alenezi, M. Shouran, Enhancing cybersecurity in smart grids: Deep black box adversarial attacks and quantum voting ensemble models for blockchain privacy-preserving storage, Energy Rep. 11 (2024) 2493–2515, http://dx.doi.org/10.1016/j.egyr.2024.02.010.

[27] A. McCarthy, E. Ghadafi, P. Andriotis, P. Legg, Defending against adversarial machine learning attacks using hierarchical learning: A case study on network traffic attack classification, J. Inf. Secur. Appl. 72 (2023) 103398, http://dx.doi.org/10.1016/j.jisa.2022.103398.