





Research Article

An Improved Big Data Analytics Architecture for Intruder Classification Using Machine Learning

Muhammad Babar ¹, Sarah Kaleem ^{2,3}, Adnan Sohail,² Muhammad Asim ^{3,4}
and Muhammad Usman Tariq ⁵

¹Robotics and Internet of Things Lab, Prince Sultan University, Riyadh 11586, Saudi Arabia

²Computing and Technology Department, Iqra University, Islamabad 44000, Pakistan

³ELAS Data Science Lab, Prince Sultan University, Riyadh 11586, Saudi Arabia

⁴School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China

⁵Abu Dhabi University, Abu Dhabi 59911, UAE

Correspondence should be addressed to Sarah Kaleem; sarahkaleem33887@iqraisb.edu.pk

Received 18 October 2022; Revised 21 January 2023; Accepted 11 April 2023; Published 4 December 2023

Academic Editor: Chien-Ming Chen

Copyright © 2023 Muhammad Babar et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The approval of retrieving information on the Internet originates several network securities matters. Intrusion recognition is a critical study in network security to spot unauthorized admission or occurrences on protected networks. Intrusion detection has a fully-fledged reputation in the current era. Research emphasizes several datasets to upsurge system precision and lessen the false-positive proportion. This article proposes a new intrusion detection system using big data analytics and deep learning to address some of the misuse and irregularity detection limitations. The proposed method could identify any odd activities in a network to recognize malicious or unauthorized action and permit a response during a confidentiality break. The proposed system utilizes the big data analytics platform based on parallel and distributed mechanisms. The parallel and distributed platforms improve the training time along with the accuracy. The experimentation appropriately classifies the information as either normal or abnormal. The proposed system has a recognition proportion of 96.11% that pointedly expands overall recognition accuracy related to existing strategies.

1. Introduction

The eminence and number of cyberattacks have amplified pointedly as the services upsurge due to the Internet [1, 2]. The demand and the execution of safety actions are needed to avoid and comprehend the cyberattacks, i.e., many dangerous software packages such as trojans, malware, viruses, and other unknown hacking tactics [3, 4]. Hence, there is a need for a method that must drastically improve the security of the system's networks to avoid cyberattacks. Data integrity, confidentiality, and availability will not be limited as an upshot [5]. Every network may be considered a target, and all the systems are vulnerable. As a result, it is susceptible to illegal access and disclosing private and erogenous data [6, 7]. A firewall is a versatile and essential

component of any security system. It configures the security policy but cannot protect us from malicious activities. Only the text of a packet's caption is examined in this firewall situation, whereas both materials and captions of a package are read in an intrusion detection system (IDS) [8]. The IDS is a more dynamic method for defending sensitive and private data.

According to the definition, an intrusion is any act that compromises data integrity, confidentiality, or availability [9]. IDS still needs to be fully formed and considered a comprehensive safeguard despite playing a vibrant role in scheming and safeguarding a secure structure. An IDS distinguishes intrusions and issues a warning in the form of an alert to guarantee that resources are not compromised [10]. An IDS acts as a deterrent to illegal access to

information. It offers a user-friendly interface for nonexpert workers to activate the systems professionally. In current years, online fraud has become more predominant in China. However, a community clamor exploded the previous year after a phone scam targeted a college student. An IDS is a network security system formerly intended to distinguish vulnerability against a sole application. The IDS is out-of-band on the network structure and is not in the correct communication path between the transmitter and receiver of data [11]. Instead, IDS on edge would engage a TAP or SPAN port to check a copy of the inline traffic flow stream [12]. The IDS was built in this manner as the complexity of the analysis essential for intrusion detection could not be steered at a rate that could persevere the equipment on the network [13]. As previously noted, the IDS is also a listen-only device, as the attackers can exploit vulnerabilities quickly after gaining network access.

Deep learning (DL) is a class of ML methods considered by some computation layers that permit an algorithm to learn suitable predictive features [14, 15]. The advent of deep learning has squeezed many machine learning-based applications. The victory is based on two foremost benefits: (1) it affords the capability to learn nonlinear relationships between parameters and (2) it permits one to leverage information from unlabelled data that does not belong to the problem. Furthermore, DL practices could be used in cases, where massive or complex data processing challenges ML or conventional data analysis methods [16]. Currently, the researchers are relying upon the DL technique. Deep learning may be employed along with other automation techniques, e.g., rule- and heuristics-based and machine learning techniques [17].

Unauthorized movements on a computer network are referred to as intrusion. It was either passively or aggressively successful. Data collection and eavesdropping are used in a passive intrusion. In the event of full of life, however, intrusion occurs and is accomplished through destructive packet forwarding, packet dropping, and whole attacks [18]. An IDS aims to classify an intruder before it imposes actual harm to the system. Currently, the current security methods that are unrealistic are utilized to avoid security faults. Hence, the misuse recognition method cannot classify unidentified attacks. Anomaly-based detection is realized by increasing the accuracy rate of intrusion detection using deep learning methods. In this research, an improved system is proposed for addressing the precise recognition of the intrusion. The classical LSTM is modified to classify, investigate, and produce estimates of the intruders based on time series data.

The remaining of the paper is organized as follows. Section 2 elaborates the detailed literature review along with related work. Section 3 represents the proposed methodology. Afterward, Section 4 discusses the results and discusses on results. Finally, Section 5 concludes the manuscript.

2. Related Work

An IDS is observed as a hazardous element in shielding systems that store critical information, intellectual property,

and other digital resources. The complete system could rapidly crumble and cast doubt on the long-term viability if a private party has access to the information [19]. An IDS has conventionally maintained administrators in detecting intrusions and managing risks [20]. On the other hand, the act of IDS is steadily endangered. The technology hackers use to hijack a network and the counter-technology administrators are deployed to combat these attacks. This has outstripped the opportunity and measure of IDS [21]. Soft computing is one of the policies that aid in reducing detection costs [22]. Because of their learning and flexibility, capabilities are used to construct utilities in the intrusion detection industry. The ability to detect threats and attacks is critical to their prevention, and accurate threat detection is vibrant. Numerous intrusion detection models have been proposed in the context of security. In the current research, the neural network method has become one of the most often utilized soft computing strategies [23].

IDSs have been presented based on anomaly detection using an unsupervised ANN [24]. A hybrid data mining method combined k-means clustering with the SMO classification algorithm for classifying network intrusions [25]. The NSL-KDD dataset was utilized with k-means clustering to reduce the dimensionality of the training dataset. SMO has completed the classification process to classify the intruders. The suggested approach (k-mean + SMO) achieved a 94.48 percent positive detection rate while lowering the false alarm rate to 0. A novel edge-up methodology called cluster center and nearest neighbor was created and implemented that computed two distances [26]. The KDD Cup 99 dataset was utilized in the tests. It trains the dataset before it is divided into multiple subsets. Normal and abnormal CANN profiles are created using SVM that do not outperform K-NN in classifying R2L and U2L attacks [27]. A new hybrid approach, DT-SVM, was presented, and a foundation classifier that uses two classifiers, including SVM and a decision tree was used. This mixed technique aimed to increase detection accuracy while minimizing computational complexity.

A hierarchical hybrid detection method was proposed that integrates misuse and anomaly detection algorithms [28]. This hybrid method outperforms the traditional models. Multicategory classifiers were employed to upsurge intrusion detection accuracy [29]. The study's main goal was to use all the classifiers' excellent features. The results of the tests show that the accuracy of detecting denial of service attacks has increased. A hybrid method with several metrics is proposed and applied for better accuracy [30]. Finally, the hybrid model's results were compared to other primary algorithms that better detect intrusion. The SVM algorithm accuracy is 94%, and the result also shows a significant percentage of (FP), (TP), (FN), and (TN) alarms when using a hybrid model. A novel IDS based on a data gain criterion and unique SVM was proposed for extracting noteworthy features from the network traffic archives domain to categorize and detect future intrusions [31]. Machine learning techniques have acknowledged considerable attention in the middle of the intrusion detection scholars to report facts-based weaknesses in detecting methods [32]. Unsupervised

techniques like k-means, SOM, and one-class SVM outperformed the others, although their ability to correctly identify all attack types needed to be more consistent.

Deep learning has glimmered much attention in academia and industry as a newfangled hotspot in neural networks. Deep knowledge has formed excellent results in the field of intrusion detection. It is claimed and proved that an LSTM recurrent neural network could detect intrusions realistically. The suggested classifier successfully distinguished between DOS attacks and network probes, each with a time series of events. With a 93.82 percent accuracy rate, LSTM surpassed the winning entries in the KDD Cup 99 competition, according to experiments. Machine learning employs a compound architecture or a series of nonlinear operations to obtain high-level abstractions in data. LSTM is applied to an RNN in this article. The NSL-KDD dataset is used to train the model and measure its performance.

3. Proposed Methodology

The proposed system performs processing in a parallel fashion. It utilizes the big datasets using the big data analytics platform. The proposed approach utilizes the LSTM model of learning. It preprocesses information with an unsupervised filter and classifies it with the LSTM algorithm. The proposed system is fabricated from many processes that process data to classify irregularities in network traffic. The multiple functions are processed in a parallel manner. The proposed system adds various layers to address the limitation of intrusion detection accuracy. The block model of the proposed scheme is depicted in Figure 1, where the detailed structure is provided in Figure 2.

The proposed hybrid approach combines two leading deep learning convolution-based models. The proposed technique adds three convolutional layers to design the hybrid method. The endangered gradient delinquent, where the neural network's efficiency diminishes due to inadequate training, is one limitation of widespread RNNs. Usual RNNs with a gradient-based learning method reduce as their significance and intricacy rise. Amending the sceneries professionally at the initial stages is time consuming and computationally exhaustive. RNNs might practice the obligation to discover to remember the critical data and then loop back to the network if the data are not recognized. Although network traffic collection comprises several impractical, imprecise unrelated data, and noisy data that affect the result across normal and abnormal network traffic categorization. As a result, preprocessing is essential to expand the recognition capacities of classifiers.

Preprocessing permits standardizing and cleaning of the data; the preprocessing is carried out so that the inappropriate information does not encumber classifier accuracy, and redundant data are removed. The big datasets are preprocessed before being translated into the format required for the proposed improved modified LSTM model. The proposed modified LSTM model will be a multilayered sequential model, including two LSTM layers followed by a thick layer that predicts the detection rate. The sequential class came from the Keras Models library, while the

Embedding, Masking, LSTM, Dropout, and Dense types came from the Keras Layers library. Initially, an instance of the Sequential class is created to utilize as a proposed modified model. In addition, the Embedding, Masking, LSTM, Dropout, and Dense layers are added to it.

The LSTM layer is added to the sequential model to start the modeling building. The embedding layer for data form follows it, and the masking layer for pretrained embeddings. The first parameter is the number of neurons or nodes required for the LSTM layer. A dropout layer is added to the proposed layer using the second parameter. Finally, a thick layer is added at the end to make the model more resilient. The convolutional layers accomplish the operation over the input K. The output of the coating can be calculated as equation (1). To compute the more diverse and rich representation of the input, multiple filters have been used.

$$CN(Y_{i,j}) = \sum_{x=-a/2}^{a/2} \sum_{l=-n/2}^{n/2} F_z(l,m)Z_{i-l,j-m}. \quad (1)$$

Later the rectified linear unit is processed that is applied after convolution operation. ReLU can be calculated as follows:

$$RECT(Z) = \max(O, Z). \quad (2)$$

ReLU provides faster convergence during training and performs better than other activation functions like Sigmoid because it overcomes the vanishing gradient problem since the gradient is linear function. Finally, the polling layer is processed. There are different types of polling such as average, maximum, and minimum polling. Maximum polling is the most popular pooling technique, which takes the largest value over the input x . Let p be the size of the polling filter, and the output of the polling operation is computed as follows:

$$M(Z_i) = \max\{Z_{i+n,i+m}\}. \quad (3)$$

4. Experimental Results and Discussion

The experimental results and discussion are discussed in this section. Big data processing is based on the parallel and distributed mechanism, where the big data are divided into various chunks (blocks), and each piece is loaded and processed in parallel. It is imperative to decide the number of parallelisms that how many blocks (nodes) are required to be loaded at a time for computations and processing. This concept is known as the level of equality. We utilized the Apache Spark big data open-source platform.

RNN can be used to solve sequence problems. LSTM is being exploited to manage sequence problems. The recommended techniques based on LSTM with selecting 41 features in an NSL-KDD dataset are adopted to increase the accuracy. The dataset is divided into 70-30 training and testing rats. The text labels in the target data are first encoded into an integer. The same preprocessing technique turns all text inputs in the training and test data into integers. Instead of memorizing the data displayed during training, the

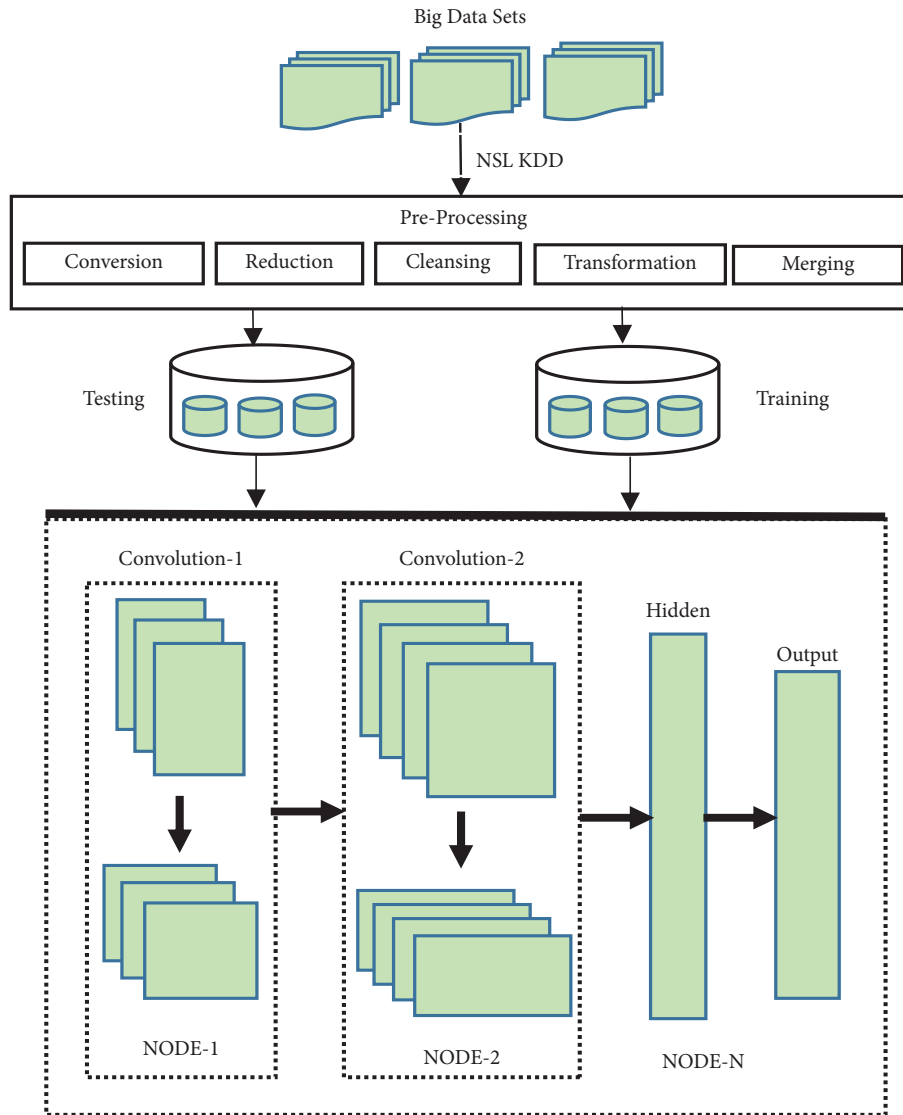


FIGURE 1: Proposed block model.

machine learning model aims to discover patterns that generalize well to new data. The proposed model performs well on unseen instances that were not used to train the model. Figure 3 shows the model prediction on the evaluation dataset (held-out data). Figure 4 depicts the loss, which is reasonably encouraging when compared to earlier traditional procedures.

It is evident from the graph that the proposed classifier accuracy is improved. The classifier accurately recognises incursions with 20 epochs.

The model loss is also reported for 20 epochs with a batch size of 64. Compared to previous strategies, the model converges after 20 cycles which is a level-headedly short time. Figure 5 demonstrates the model building time over several epochs. The accuracy analysis of different models is also shown in Figure 6.

The confusion matrix is used to generate most performance metrics. When making predictions, the classification model becomes perplexed. To obtain a logic of precise and error in prediction, the normalization process is carried out.

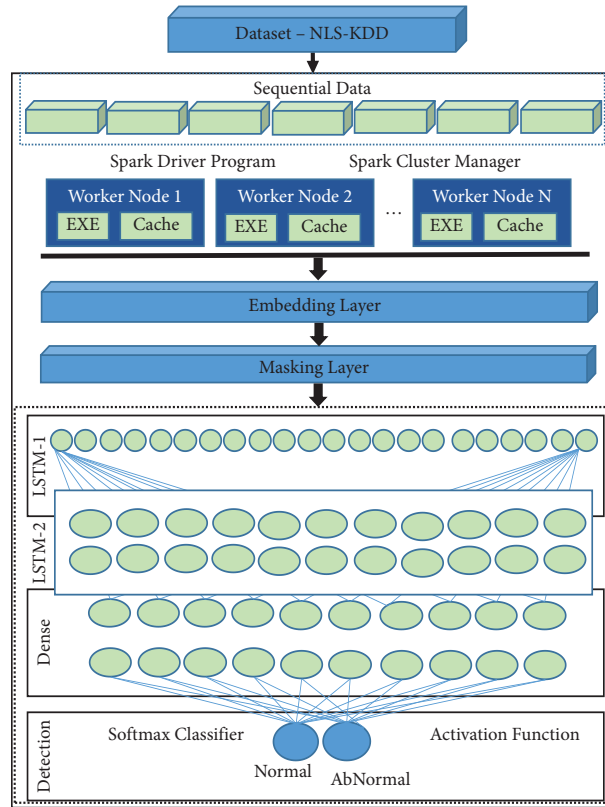


FIGURE 2: Proposed model structure.

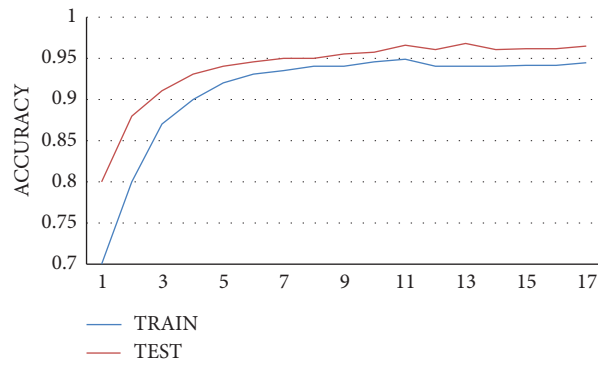


FIGURE 3: Proposed modified LSTM model accuracy rate.

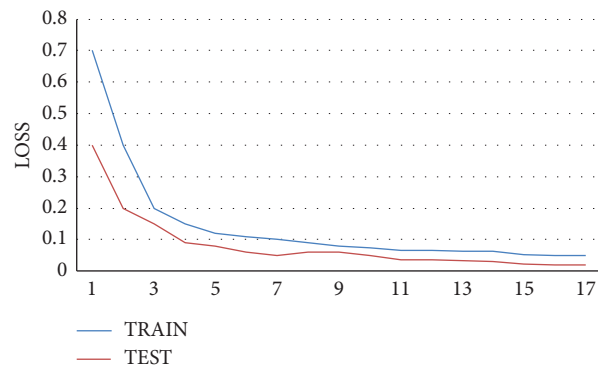


FIGURE 4: Proposed modified LSTM model loss rate.

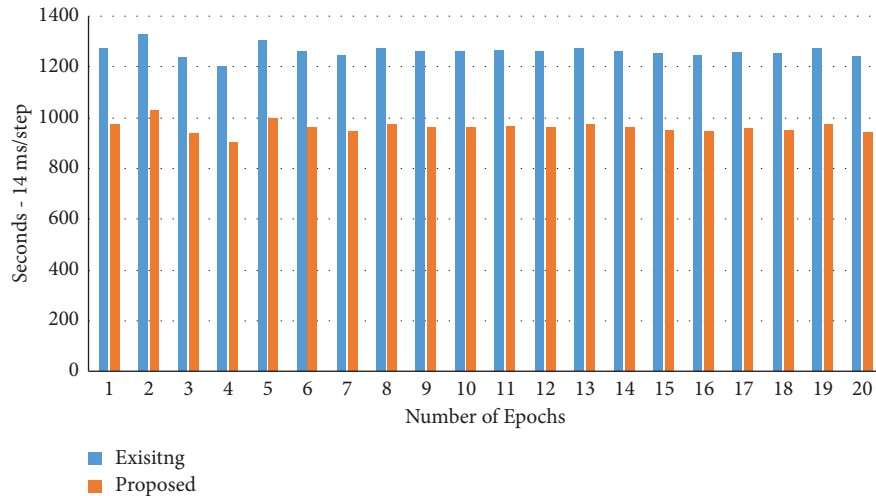


FIGURE 5: Model building time.

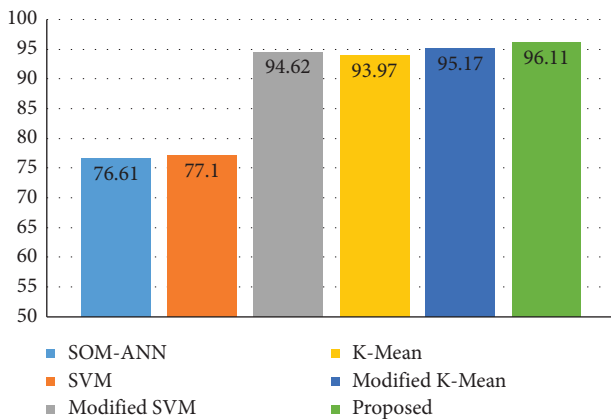


FIGURE 6: Accuracy analysis of different models.

5. Conclusion

The current security methods could be more efficient in avoiding security faults. An intrusion detection process has become an important part of network security. Hence, the misuse recognition method cannot classify unidentified attacks. The irregularity recognition practice is utilized to detect anomalies. Anomaly-based detection is realized by increasing the accuracy rate of intrusion detection using deep learning methods. In this article, an improved methodology is proposed for accurately detecting the intrusion using a modified LSTM algorithm. The proposed system utilizes the big data analytics platform based on parallel and distributed mechanisms. The parallel and distributed platform improves the training time along with the accuracy. The classical LSTM is modified to classify, investigate, and produce estimates of the intruders based on time series data. The proposed modified LSTM is compared with traditional algorithms and their modified versions. It is evident from the results that the proposed system outperforms the existing approaches. The proposed method works well when it comes to detecting assaults. In terms of detection rate, the proposed method knocks existing techniques. The proposed system

accurately classifies data as normal or abnormal. The proposed method has a detection rate of 96.11 percent, which is implausible. Filter layers like dropout ten eliminate non-sensical, noisy, and irrelevant data from the source data.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Muhammad Babar proposed the methodology, wrote the original draft, and validated the study. Sarah Kaleem proposed the methodology, wrote the original draft, and validated the study. Adnan Sohail reviewed and edited the manuscript and formally analysed the study. Muhammad Asim reviewed and edited the manuscript and acquired the funding. Muhammad Usman Tariq reviewed and edited the manuscript and visualized the study.

Acknowledgments

The authors would like to acknowledge the support of Prince Sultan University for paying the Article Processing Charges.

References

- [1] M. K. Kagita, N. Thilakarathne, T. R. Gadekallu, P. K. R. Maddikunta, and S. Singh, "A review on cybercrimes on the Internet of Things," in *Deep Learning for Security and Privacy Preservation in IoT, Signals and Communication Technology*, pp. 83–98, Springer, Singapore, 2022.
- [2] I. Almomani, M. Ahmed, and L. Maglaras, "Cybersecurity maturity assessment framework for higher education institutions in Saudi Arabia," *PeerJ Computer Science*, vol. 7, p. e703, 2021.

- [3] A. Sedik, O. S. Faragallah, H. S. El-sayed et al., "An efficient cybersecurity framework for facial video forensics detection based on multimodal deep learning," *Neural Computing and Applications*, vol. 34, no. 2, pp. 1251–1268, 2022.
- [4] R. Deepalakshmi, R. Vijayalakshmi, C. Sam Ruben, R. Pandiya Rajan, and J. Pradeep, "Application of artificial intelligence in cybersecurity: a detailed survey on intrusion detection systems," in *An Interdisciplinary Approach to Modern Network Security*, pp. 1–22, CRC Press, Boca Raton, FL, USA, 2022.
- [5] E. N. Witanto, Y. E. Oktian, and S.-G. Lee, "Toward data integrity architecture for cloud-based AI systems," *Symmetry*, vol. 14, no. 2, p. 273, 2022.
- [6] M. I. Talukdar, R. Hassan, M. S. Hossen, K. Ahmad, F. Qamar, and A. S. Ahmed, "Performance improvements of AODV by black hole attack detection using IDS and digital signature," *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 6693316, 13 pages, 2021.
- [7] S. Sanober, I. Alam, S. Pande et al., "An enhanced secure deep learning algorithm for fraud detection in wireless communication," *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 6079582, 14 pages, 2021.
- [8] H. A. Hassan, E. E. Hemdan, W. El-Shafai, M. Shokair, and F. E. A. El-Samie, "Intrusion detection systems for the internet of thing: a survey study," *Wireless Personal Communications*, vol. 128, no. 4, pp. 2753–2778, 2022.
- [9] A. Thakkar and R. Lohiya, "A survey on intrusion detection system: feature selection, model, performance measures, application perspective, challenges, and future research directions," *Artificial Intelligence Review*, vol. 55, no. 1, pp. 453–563, 2022.
- [10] A. Kim, M. Park, and D. H. Lee, "AI-IDS: application of deep learning to real-time Web intrusion detection," *IEEE Access*, vol. 8, pp. 70245–70261, 2020.
- [11] A. Palshikar, "What distinguishes binary from multi-class intrusion detection systems: observations from experiments," *International Journal of Information Management Data Insights*, vol. 2, no. 2, Article ID 100125, 2022.
- [12] D. A. Kumar and S. Venugopalan, "Intrusion detection systems: a review," *International Journal of Advanced Research in Computer Science*, vol. 8, no. 8, pp. 356–370, 2017.
- [13] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman, "Deep learning approach for intelligent intrusion detection system," *IEEE Access*, vol. 7, pp. 41525–41550, 2019.
- [14] I. Idrissi, M. Azizi, and O. Moussaoui, "Accelerating the update of a DL-based IDS for IoT using deep transfer learning," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 23, no. 2, pp. 1059–1067, 2021.
- [15] Z. Ahmad, A. Shahid Khan, C. Wai Shiang, J. Abdullah, and F. Ahmad, "Network intrusion detection system: a systematic study of machine learning and deep learning approaches," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 1, Article ID e4150, 2021.
- [16] S. E. Whang, Y. Roh, H. Song, and J.-G. Lee, "Data collection and quality challenges in deep learning: a data-centric ai perspective," *The VLDB Journal*, vol. 32, no. 4, pp. 791–813, 2023.
- [17] A. Khan, S. H. Khan, M. Saif, A. Batool, A. Sohail, and M. Waleed Khan, "A survey of deep learning techniques for the analysis of COVID-19 and their usability for detecting omicron," *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 2023, pp. 1–43, 2023.
- [18] S. Sharma and A. Kaul, "A survey on Intrusion Detection Systems and Honeypot based proactive security mechanisms in VANETs and VANET Cloud," *Vehicular Communications*, vol. 12, pp. 138–164, 2018.
- [19] R. Yang, R. Wakefield, S. Lyu et al., "Public and private blockchain in construction business process and information integration," *Automation in Construction*, vol. 118, Article ID 103276, 2020.
- [20] A. Stewart, *The Community Defense Approach: A Human Approach to Cybersecurity for Industrial and Manufacturing Systems*, University of Cincinnati, Cincinnati, OH, USA, 2019.
- [21] L. Balke, "China's new cybersecurity law and US-China cybersecurity issues," *Santa Clara Law Review*, vol. 58, p. 137, 2018.
- [22] A. H. Hamamoto, L. F. Carvalho, L. D. H. Sampaio, T. Abrão, and M. L. Proença, "Network anomaly detection system using genetic algorithm and fuzzy logic," *Expert Systems with Applications*, vol. 92, pp. 390–402, 2018.
- [23] V. Gowdhaman and R. Dhanapal, "An intrusion detection system for wireless sensor networks using deep neural network," *Soft Computing*, vol. 26, no. 23, pp. 13059–13067, 2021.
- [24] M. Ozkan-Okay, R. Samet, Ö. Aslan, and D. Gupta, "A comprehensive systematic literature review on intrusion detection systems," *IEEE Access*, vol. 9, pp. 157727–157760, 2021.
- [25] Gadai, S. M. Ali Mohamed, and R. A. Mokhtar, "Anomaly detection approach using hybrid algorithm of data mining technique," in *Proceedings of the 2017 International Conference on Communication, Control, Computing and Electronics Engineering (ICCCCEE)*, pp. 1–6, IEEE, Khartoum, Sudan, January, 2017.
- [26] W.-C. Lin, S.-W. Ke, and C.-F. Tsai, "CANN: an intrusion detection system based on combining cluster centers and nearest neighbors," *Knowledge-Based Systems*, vol. 78, pp. 13–21, 2015.
- [27] S. Peddabachigari, A. Abraham, C. Grosan, and J. Thomas, "Modeling intrusion detection system using hybrid intelligent systems," *Journal of Network and Computer Applications*, vol. 30, no. 1, pp. 114–132, 2007.
- [28] V. Hajisalem and S. Babaie, "A hybrid intrusion detection system based on ABC-AFS algorithm for misuse and anomaly detection," *Computer Networks*, vol. 136, pp. 37–50, 2018.
- [29] L. S. Liu and H. Y. Fu, "Intrusion detection model based on multi-category feature fusion," in *Proceedings of the International Conference on Frontiers of Electronics, Information and Computation Technologies*, pp. 1–5, Yangzhou, China, May, 2021.
- [30] Y. Mirsky, D. Tomer, Y. Elovici, and A. Shabtai, "Kitsune: an ensemble of autoencoders for online network intrusion detection," 2018, <https://arxiv.org/abs/1802.09089>.
- [31] G. Bovenzi, G. Aceto, D. Ciunzo, V. Persico, and A. Pescapé, "A hierarchical hybrid intrusion detection approach in IoT scenarios," in *Proceedings of the GLOBECOM 2020-2020 IEEE Global Communications Conference*, pp. 1–7, IEEE, Taipei, Taiwan, December, 2020.
- [32] I. Guarino, G. Bovenzi, D. Di Monda, G. Aceto, D. Ciunzo, and A. Pescapé, "On the use of machine learning approaches for the early classification in network intrusion detection," in *Proceeding of the 2022 IEEE International Symposium on Measurements & Networking (M&N)*, Padua, Italy, July, 2022.