Neural Network Based Anomaly Detection Method for Network Datasets

Bilal Zahid Hussain[1], Yusuf Hasan[2], and Irfan Khan[1]

[1]Texas A&M University
[2]Aligarh Muslim University

February 27, 2024

Neural Network Based Anomaly Detection Method for Network Datasets

B Zahid Hussain, *Student Member, IEEE,* Yusuf Hasan, *Student Member, IEEE* and Irfan Ahmad Khan, *Member, IEEE*

***Abstract*—This research paper presents a comprehensive investigation into the development of an innovative and novel custom neural network model for intrusion detection systems (IDS). In the current era of rapid data transfer facilitated by the internet and advancements in communication technologies, the security of sensitive information is of paramount concern. As attackers continuously devise new methodologies to steal or tamper with data, IDSs face significant challenges in effectively detecting and mitigating intrusions. While extensive research has been conducted to enhance IDS capabilities, the need for improved detection accuracy and reduced false alarm rates remains a pressing issue. Moreover, the identification of zero-day attacks continues to pose a formidable obstacle. In contrast to conventional IDS approaches that heavily rely on statistical methodologies and rule-based expert systems, this study embraces data mining techniques, specifically Neural Networks (NNs), to overcome the limitations associated with large datasets. This research paper proposes a meticulously designed custom neural network model that leverages machine learning (ML) algorithms to analyze contemporary host activity and cloud service data. The paper extensively discusses the utilized dataset, meticulously evaluates the performance of various classifiers, and introduces our innovative neural network model. Emphasizing the significance of our model in anomaly detection, the findings underscore the importance of robust ML models to ensure the efficacy and longevity of deployed defensive systems. By capitalizing on its innovative design and leveraging the power of ML algorithms, our model not only addresses the limitations of traditional IDS approaches but also paves the way for enhanced accuracy, reduced false alarms, and improved resilience against zero-day attacks. This research contributes to the advancement of the field, shedding light on the novel possibilities and remarkable innovation offered by our custom neural network model in safeguarding critical information in an increasingly hostile digital landscape.**

***Index Terms*—Intrusion Detection System (IDS), Deep Learning, Lightweight Models.**

I. Introduction

With the increasing reliance on the Internet in our daily lives, network security has emerged as a fundamental concern for various web applications, including online auctions and retail sales. Detecting and preventing intrusions is a crucial aspect of network security, as it helps safeguard data integrity and confidentiality. However, the proliferation of networks, escalating data transfer rates, and evolving internet usage patterns have introduced new challenges related to anomaly detection. To address these challenges, researchers are continuously striving to develop more reliable, effective, and self-monitoring systems that can identify and mitigate security threats without human intervention. Intrusion Detection Systems (IDS) play a vital role in network security by identifying and alerting against unauthorized access attempts and malicious activities [1]. Traditionally, IDS utilized statistical approaches and rule-based expert systems, but they often fell short when confronted with large datasets. To overcome these limitations, researchers turned to data mining techniques, including various machine learning paradigms such as Linear Genetic Programming (LGP) [2], neural networks, Bayesian networks, Support Vector Machines (SVM) [3], Fuzzy Inference Systems (FISs), and Multivariate Adaptive Regression Splines (MARS). Among these paradigms, Neural Networks (NN) [4] have proven to be particularly effective in resolving complex practical problems and have found successful applications in IDS [5].

However, Neural Network-based IDS still face certain drawbacks that need to be addressed. One of the key challenges is the ability to detect anomalies and adapt to shifts in data distributions. As ML models are deployed in real-world scenarios [6], anomalous data points and changes in the data landscape are inevitable. This is especially crucial in the realm of cybersecurity, where anomalies and dataset shifts can result from both defensive advancements and adversarial attacks. The development of robust ML models becomes imperative to ensure the performance, protection, and longevity of deployed defensive systems. Existing research on the robustness of ML models has primarily focused on widely used datasets within the ML community, often centered around image and text datasets. However, there is a lack of understanding regarding how well these methods, particularly those based on deep learning, generalize beyond these specific input modalities.

Anomaly data points and changes in the data distribution are unavoidable when ML models are applied in the real world. These anomalies and dataset changes are caused by both defensive and adversarial development from the perspective of cyber security. Therefore, the creation of strong models is essential to the effectiveness, protection, and durability of deployed defensive systems in order to resist the expense of crucial system failure.

The novelty of our custom neural network model lies in its ability to address the challenges faced by traditional anomaly detection systems and existing machine learning based IDS models. With the increasing volume of transferred data and the

B Zahid Hussain is with the Dept. of Electrical & Computer Engineering, Texas A&M University, USA, e-mail: zahidhussain909@tamu.edu.

Yusuf Hasan is with the Department of Computer Engineering, Aligarh Muslim University, India, e-mail: yusufhasan1209@gmail.com.

Dr. Irfan Ahmad Khan is with the Dept. of Marine Engineering Technology in a joint courtesy appointment with Dept. of Electrical & Computer Engineering and Dept. of Computer Science and Engineering, Texas A&M University, USA, e-mail: irfankhan@tamu.edu.

constant emergence of new attack techniques, the safety of our systems is at risk. While many researchers have explored novel IDS systems, they still struggle with detection accuracy and false alarm rates, as well as the identification of attacks. Our research paper proposes a methodology that utilizes machine learning models, particularly our innovative custom neural network, to analyze modern host activity and cloud service data.

Contributions:

The main contributions of this work are as follows:

• The ML models are utilized to analyze the modern host activity and activity collected from cloud services, relevant for current real-world deployments as it captures real activity rather than artificial user activity. It includes both kernel-process and network logs, enabling a comprehensive perspective on malicious behavior.

• Propose a novel, custom neural network model that represents a significant advancement in the field of anomaly detection systems. Unlike traditional approaches and existing neural network-based IDS models, our innovative solution effectively tackles the challenges posed by the growing volume of transferred data and ever-evolving attack techniques. Our custom neural network demonstrates exceptional performance in detecting anomalies and adapting to shifts in data distributions. Its novel approach and robustness make it a groundbreaking tool for ensuring the safety and durability of defensive systems in real-world cybersecurity scenarios.

The remainder of the paper is structured as follows. Related works are described in Section II. Section III gives an overview about proposed methodology which includes a brief about the dataset and a detailed view on the methodology approach.

II. Related Work

The security and intrusion risks for systems will considerably increase as the technology advances. A variety of machine learning techniques are included in the study on intrusion detection systems. Increasing detection rates, decreasing false positives, and identifying unknown intrusions remain challenging tasks for current IDS. Researchers have looked into the integration of machine learning into IDSs to address current problems. It is possible to automatically distinguish between normal and abnormal data by employing hybrid-based machine learning techniques.

The authors in [7] proposes a novel hybrid method that combines swarm intelligence and evolutionary algorithms for feature selection, named PSO-GA (PSO-based GA). This method is applied to the CICIDS-2017 dataset prior to model training. To assess the performance of the proposed framework, the author evaluates the model using ELM-BA (Extreme Learning Machine with Bootstrap Aggregating), which employs bootstrap resampling to increase the reliability of ELM. Notably, the results demonstrate exceptional accuracy, achieving a perfect 100% detection rate for PortScan, SQL injection, and brute force attacks. These findings highlight the effectiveness of the proposed model and its potential for deployment in real-world cybersecurity applications.

The research paper [8] provides a comprehensive review of the current research trends in network-based intrusion detection systems (NIDS), including the different approaches employed and the datasets commonly used for evaluating IDS models. The analysis presented in this paper is based on several factors, such as the number of citations received by published articles, the total count of articles published on intrusion detection in a given year, and the most highly cited research articles in journals and conferences specific to the intrusion detection system. By examining articles published in the intrusion detection field over the past 15 years, this paper discusses the state-of-the-art techniques in NIDS, commonly utilized NIDS, benchmark datasets based on citation analysis, and the various NIDS techniques employed for intrusion detection. Furthermore, the paper includes a comparative analysis based on citations and publications to quantitatively assess the popularity of different approaches. This study aims to provide guidance and insights to both newcomers and researchers interested in evaluating the research trends in NIDS and its associated applications.

The authors in [9] suggests that the rapid growth of data transmission on the Internet has led to increased attacks by hackers seeking to exploit or manipulate this data. Intrusion detection systems (IDS) play a crucial role in network security by detecting and preventing unauthorized access. However, current IDS solutions need improvement in terms of detection accuracy and handling zero-day attacks. The paper also presents the primary metrics used to assess the performance of IDS and conducts a comprehensive review of recent IDS solutions that employ machine learning techniques. Each solution's strengths and weaknesses are outlined to provide a comprehensive evaluation. Additionally, the paper discusses the datasets used in these studies, highlighting their relevance and significance. The accuracy of the results obtained from the reviewed works is thoroughly examined.

An ensemble strategy was suggested by Rohit et al. [10] to identify infiltration. They run three tests to demonstrate how their strategy suggested improved outcomes. They used a correlation approach to do feature selection after first normalising the KDD Cup99 dataset. Finally, they adopt an ensemble technique that combines three algorithms: Naive Bayes, PART, and Adaptive Boost for the feature selection process. Information gain was used as a deciding criterion in this process. The outcome is then determined by averaging the outputs of the several algorithms or by the majority of votes. They also employ the bagging technique to lessen variance error. Using their method, they were able to achieve an accuracy of 99.9732% on the KDD Cup99 dataset.

In order to detect cloud invasions, Kanimozhi et al. [11] suggested employing oppositional tunicate fuzzy C-means. To create two datasets—one for training and one for testing—they first pre-processed the data and performed a normalisation on it. They employed the OPTSA and FCM clustering model and did a feature selection using logistic regression to maintain the more pertinent features. The fuzzy C-means algorithm is used to divide the dataset into C clusters. They did a cluster expansion and integration after the data was clustered in order to eliminate redundant clusters. On various datasets, including

CICIDS2017, they tested their method and got an accuracy of 80%.

III. Proposed Methodology

Intrusion detection systems (IDS) play a critical role in mitigating the ever-evolving nature of attacks. To effectively combat these evolving threats, regular updates are essential. Anomaly-based network intrusion detection systems have emerged as a prevalent technology in the detection stage. Considering the need for high accuracy and precision, various machine learning techniques have been chosen for this purpose. Subsequent sections of this paper delve into the dataset employed, the recommended operational framework, the proposed network design, and the evaluation metrics utilized to assess the performance of the IDS. These components collectively contribute to the comprehensive understanding and implementation of intrusion detection strategies in the face of constantly evolving attacks.

A. Dataset

BPF-extended tracking honeypot(BETH)[1] [12] is a cybersecurity dataset for out-of-distribution analysis and anomaly detection. The dataset contains over eight million data points tracking 23 hosts with true "anomalies" gathered using a new tracking technique. Each host has recorded only benign behaviour and, at most, one attack, making behavioural analysis easier. BETH not only offers one of the most up-to-date and comprehensive cybersecurity datasets on the market, but it also makes it possible to create anomaly detection algorithms on heterogeneously structured real-world data with obvious downstream uses. The dataset consists of two sensor logs: network traffic and kernel-level process calls. Initially, only process logs were included in the benchmark subset. 14 unprocessed features and 2 manually created labels make up each process call.

The dataset consists of two sensor logs: network traffic and kernel-level process calls. Two labels and 14 raw features make up each process call, labelled suspicious (sus) or evil to aid analysis. Unusual activity or data distribution outliers, such as an external userId with a systemd process, infrequent daemon process calls, etc., are indicated by suspicious markings in the logs. Evil denotes a malevolent outside influence that is not a natural part of the system, such as a bash execution call to list the contents of the computer's memory.

The dataset was then divided into training, validation, and testing sets with about a 60/20/20 split according to the host, the volume of logs created, and the activity logged; only the test set contains an attack.

Fig. 1 and Fig. 2 show the dataset distribution of sus and evil across train and test set respectively.

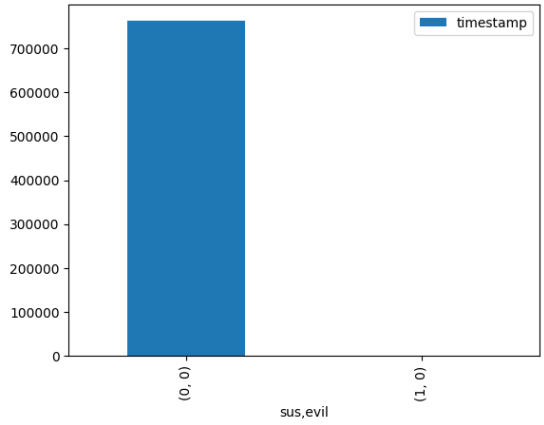

Fig. 1. Dataset distribution of sus and evil across train set

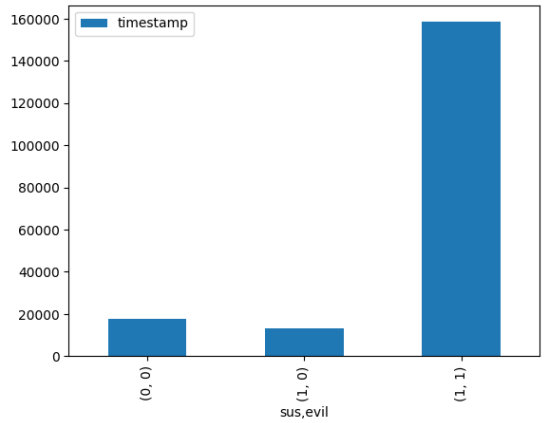

Fig. 2. Dataset distribution of sus and evil across test set

B. Evaluation Metrics

When evaluating different machine learning models, it is essential to compare their performance using various metrics. One common approach is to analyze the confusion matrix, which provides information about true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN) values.

TP represents instances correctly predicted as the positive class, while TN indicates instances correctly classified as the negative class. FP occurs when the model wrongly predicts the positive class, and FN represents instances mistakenly classified as the negative class.

Accuracy measures overall prediction correctness and is calculated by dividing the sum of TP and TN by the total number of instances. Precision assesses the proportion of correctly predicted positives out of all instances predicted as positive, while recall measures the proportion of correctly predicted positives out of all actual positive instances. Precision and recall can be calculated using specific formulas.

Precision is a metric that measures the precision of positive projections. It calculates the fraction of accurately anticipated positive instances out of all positive instances predicted. Precision is determined by dividing TP by the product of TP and FP.

To calculate accuracy and precision one can use the following formulas:

[1][Online] Available at: https://www.kaggle.com/datasets/katehighnam/beth-dataset

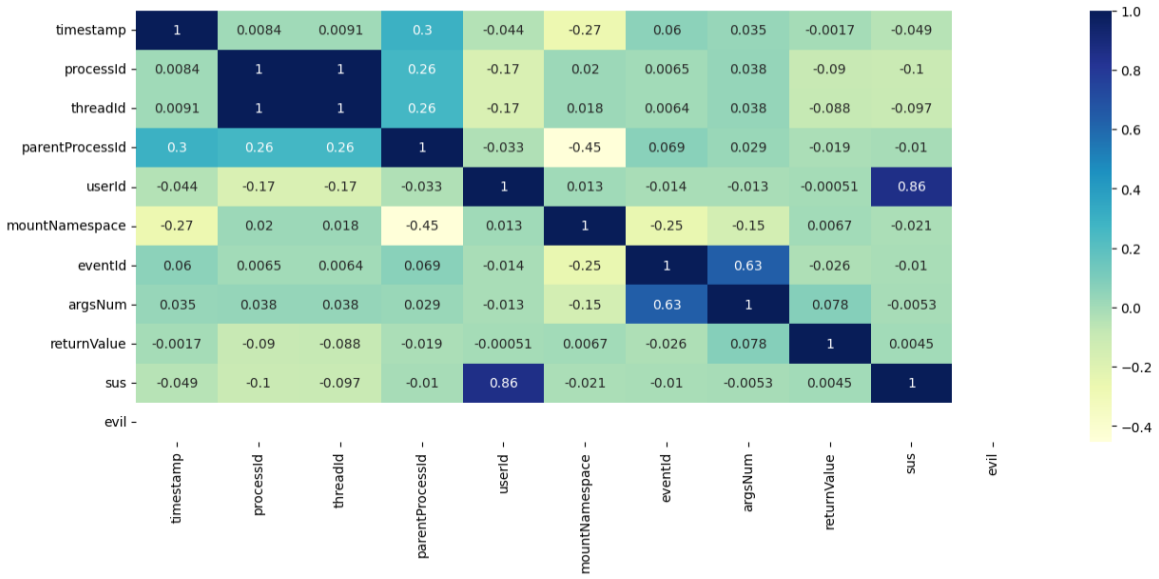

Fig. 3. Correlation plot for train set showing heavy correlation between userid and the associated labels which feature in the dataset

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

$$Precision = \frac{TP}{(TP + FP)} \quad (2)$$

C. Methodology and Approach

1) Random Forest [13]: Random Forest is a versatile and widely used machine learning technique that may be utilised for classification and regression tasks. It combines numerous decision tree to improve accuracy and robustness. Each tree is trained on a portion of the training data that is chosen at random. A random subset of features is considered at each node during tree construction to make splits. This random feature selection minimizes tree correlation while increasing ensemble diversity.

Random Forest uses majority voting to combine predictions of individual trees for classification tasks. Each tree classifies the input data independently, and the class with the highest votes becomes the final prediction. The outputs of the trees are averaged to give the final prediction in regression tasks.

2) Light GBM [14]: Light GBM (Light Gradient Boosting Machine) is a gradient boosting framework designed for supervised machine learning tasks. It sequentially integrates weak prediction models to generate a powerful predictive model. Gradient-based One-Side Sampling (GOSS), which selects instances with greater gradients to enhance training time and handle imbalanced datasets, and Exclusive Feature Bundling (EFB), which minimises the number of split points for categorical features, are two unique characteristics of Light GBM. For faster training and increased accuracy, Light GBM employs a leaf-wise tree construction strategy, prioritising nodes that result in the greatest loss reduction.Light GBM is fast, scalable, and well-suited for large-scale datasets. Its uses include online advertising, recommendation systems, and financial modelling, all of which require speed and precision.

3) Isolation Forest [15]: Isolation Forest (IF) is a famous example of a semi-supervised ensemble learning model with different features. In contrast to other classic algorithms such as LOF, IF takes a more relaxed approach to data features. It divides the data space into orthogonal lines and assigns higher scores to data points that require fewer splits to isolate. Outliers can be effectively identified using this scoring mechanism. The use of orthogonal lines by IF, as well as its scoring system based on isolation divides, contribute to its effectiveness in spotting outliers. The growing number of trees improves its stability, and its distribution-agnostic nature enables it to handle a wide range of datasets. However, when using IF to complex high-dimensional data with a large presence of local outliers, care should be taken because performance may suffer.

4) Custom Deep Neural Network Model: The model architecture is made up of several layers, including dense layers with ReLU activation, batch normalisation layers for stability, and dropout regularisation for overfitting prevention. The model's goal is to process the input data and predict outcomes using a binary classification job. The model architecture is made up of numerous levels that each serve a different purpose. The first layer, 'Flatten,' is in charge of flattening the input data, which has the shape (5,) in this case. To preparation for further processing, it converts the input into a one-dimensional array. A 'BatchNormalization' layer is added after the initial 'Dense' layer. Batch normalisation normalises the previous layer's outputs, making the optimisation process more stable and speeding up training. Another 'Dense' layer with 180 units and ReLU activation follows. This layer goes over the data again, extracting pertinent features. After the second 'Dense' layer, another 'BatchNormalization' layer is added to ensure the model's training process is stable. After the second 'BatchNormalization' layer, a 'Dropout' layer with a dropout rate of 0.2 is introduced to prevent overfitting and increase generalisation. During training, dropout randomly sets

a fraction of the input units to 0, which helps to reduce over-reliance on individual characteristics and pushes the model to learn more resilient representations. Finally, there is a 'Dense' layer with a single unit and sigmoid activation as the final layer. This layer generates the model's final output, which is a probability between 0 and 1, signifying the likelihood of the input belonging to a specific class. For binary classification tasks, the sigmoid activation function is often utilised.

Overall, the model has a total of 69,481 parameters, with 68,401 of them being trainable. The remaining 1,080 parameters are non-trainable and likely related to the internal operations of the layers.

IV. Results

The table I presents the results of different models in terms of accuracy and precision. The custom Neural Network model achieved the highest accuracy and precision among all the models listed in the table. With an accuracy of 99.35%, it correctly predicted the outcome for 99.35% of the instances. The precision of 99.57% indicates that when the model predicted a positive outcome, it was correct 99.57% of the time. The custom Neural Network model demonstrates superior performance in terms of accuracy and precision compared to the other models in the table.

Our research introduces a highly innovative and novel custom neural network model for anomaly detection. A key distinguishing feature of our model is its exceptional lightweight design, characterized by a remarkably reduced number of parameters. This lightweight architecture allows for efficient and resource-friendly deployment, enabling real-time inference and significantly enhancing reaction time. Despite its compact size, the model maintains superior performance in terms of accuracy and precision, surpassing conventional IDS approaches. This breakthrough in lightweight neural network design offers a unique and efficient solution for intrusion detection, presenting new possibilities for the field and demonstrating our commitment to pushing the boundaries of innovation in cybersecurity.

The deployment of our custom neural network model on devices heralds a multitude of unparalleled advantages, magnifying its groundbreaking essence. With an exceptional degree of accuracy, our model enables real-time inference, endowing it with the remarkable ability to generate faster and astoundingly precise predictions. Furthermore, meticulous optimization empowers our model to achieve efficient inference, reducing dependence on network connectivity and markedly enhancing reaction time. In addition, the model's autonomous offline operation proves indispensable in scenarios characterized by limited or unreliable connectivity, ensuring uninterrupted functionality and seamless performance. By residing directly on the device itself, our model guarantees the utmost levels of privacy and security, rendering the transmission of sensitive data to external servers entirely obsolete.

In summary, the custom Neural Network model has the highest accuracy and precision, followed by the Light GBM, Random Forest, and Isolation Forest models. This implies that the custom Neural Network model is the most accurate and precise in predicting the outcome compared to the other models in the evaluation.

TABLE I
COMPARISON OF ACCURACY AND PRECISION OF VARIOUS MACHINE LEARNING METHODS

Model	Accuracy(%)	Precision(%)
Isolation Forest	93.15	89.5
Random Forest	95.05	96.24
Light GBM	97.76	98.00
Custom Neural Network	**99.35**	**99.57**

V. Conclusion

In our research, we address the challenges of intrusion detection by proposing a novel custom neural network model specifically designed to detect intrusions in a lightweight manner. Unlike the prevailing trend of employing complex deep learning methods that demand substantial computing resources, our model offers an efficient alternative. We recognize the growing importance of deep learning techniques in anomaly detection, but also acknowledge the need for lightweight solutions to ensure practicality and scalability. By leveraging our custom neural network model, we strike a balance between accuracy and resource efficiency, enabling effective intrusion detection while minimizing computational requirements. This innovative approach allows for the detection of intrusions in real-world scenarios, leveraging both kernel-level process events and DNS network traffic from the BETH dataset. Our research contributes to advancing the field of intrusion detection systems by presenting a novel and lightweight neural network model that combines accuracy with efficiency in the detection of intrusions. Numerous classification techniques from machine learning have been used to compare and accentuate the effectiveness of our novel custom model, as detailed in this work for anomaly detection using the BETH dataset. The dataset contains both kernel-level process events and DNS network traffic. It contains real-world attacks in the presence of benign modern OS and cloud provider traffic.

References

[1] M. Tahreeem, I. Andleeb, B. Z. Hussain, and A. Hameed, "Machine learning-based android intrusion detection systems," 12 2022.

[2] M. Willis, H. Hiden, P. Marenbach, B. McKay, and G. Montague, "Genetic programming: An introduction and survey of applications," 10 1997, pp. 314 – 319.

[3] Y. Zhang, "Support vector machine classification algorithm and its application," in *Information Computing and Applications*, C. Liu, L. Wang, and A. Yang, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 179–186.

[4] B. Z. Hussain, I. Andleeb, M. S. Ansari, and N. Kanwal, "Lightweight deep learning model for automated covid-19 diagnosis from cxr images," in *2021 IEEE International Conference on Computing (ICOCO)*. IEEE, 2021, pp. 218–223.

[5] U. S. Musa, M. Chhabra, A. Ali, and M. Kaur, "Intrusion detection system using machine learning techniques: A review," in *2020 International Conference on Smart Electronics and Communication (ICOSEC)*, 2020, pp. 149–155.

[6] B. Z. Hussain, I. Andleeb, M. S. Ansari, A. M. Joshi, and N. Kanwal, "Wasserstein gan based chest x-ray dataset augmentation for deep learning models: Covid-19 detection use-case," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2022, pp. 2058–2061.

[7] L. Ashiku and C. Dagli, "Network intrusion detection system using deep learning," *Procedia Computer Science*, vol. 185, pp. 239–247, 2021, big Data, IoT, and AI for a Smarter Future. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050921011078

[8] S. Kumar, S. Gupta, and S. Arora, "Research trends in network-based intrusion detection systems: A review," *IEEE Access*, vol. 9, pp. 157 761–157 779, 2021.

[9] P. Vanin, T. Newe, L. L. Dhirani, E. O'Connell, D. O'Shea, B. Lee, and M. Rao, "A study of network intrusion detection systems using artificial intelligence/machine learning," *Applied Sciences*, vol. 12, no. 22, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/22/11752

[10] R. K. S. Gautam and E. A. Doegar, "An ensemble approach for intrusion detection system using machine learning algorithms," in *2018 8th International conference on cloud computing, data science & engineering (confluence)*. IEEE, 2018, pp. 14–15.

[11] P. Kanimozhi and T. Aruldoss Albert Victoire, "Oppositional tunicate fuzzy c-means algorithm and logistic regression for intrusion detection on cloud," *Concurrency and Computation: Practice and Experience*, vol. 34, no. 4, p. e6624, 2022.

[12] K. Highnam, K. Arulkumaran, Z. D. Hanif, and N. R. Jennings, "Beth dataset: Real cybersecurity data for anomaly detection research," 2021.

[13] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random forests and decision trees," *International Journal of Computer Science Issues(IJCSI)*, vol. 9, 09 2012.

[14] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *NIPS*, 2017.

[15] F. T. Liu, K. Ting, and Z.-H. Zhou, "Isolation forest," 01 2009, pp. 413 – 422.