# Cyberattack Prediction Using Machine Learning Techniques for Network Security

**Aman Goyal**
B.Tech Student, Department of CSE, Global Institute of Technology, Jaipur, Rajasthan, India
23egjcs017@gitjaipur.com

**Aaditya Gupta**
B.Tech Student, Department of CSE, Global Institute of Technology, Jaipur, Rajasthan, India
23egjad002@gitjaipur.com

**Dr. Gaurav Kumar Jain**
Associate Professor, Department of CSE, Global Institute of Technology, Jaipur, Rajasthan, India
gaurav.jain@gitjaipur.com

**ABSTRACT:** Cyberattacks pose a significant threat to the confidentiality, integrity, and availability of information in modern digital environments. With the rapid expansion of cyberspace and the increasing volume of data generated through internet-connected devices, identifying and preventing malicious activities has become more challenging. Although numerous cybersecurity models and algorithms exist, there remains a strong need for advanced machine learning-based approaches capable of effectively analyzing large-scale and unstructured network data. This research models cyberattack prediction as a classification problem, utilizing supervised machine learning techniques to detect and categorize network intrusions. The dataset is analyzed through various statistical and machine learning methods, including variance analysis, bivariate analysis, and performance evaluation metrics. A comparative study is conducted to assess multiple machine learning algorithms in classifying four major categories of cyberattacks: Denial of Service (DoS), Remote to Local (R2L), User to Root (U2R), and Probe attacks. Experimental results demonstrate the effectiveness of the proposed system, with performance evaluated through accuracy, precision, recall, F1-score, sensitivity, specificity, and entropy measures. The study highlights the significance of machine learning models in enhancing cyber defense strategies and improving early attack detection for network security.

**KEYWORDS:** Cyber Attack, Machine Learning, Deep Neural Network, Auto Encoder, SWAT, Network Security, Data Security.

## 1. INTRODUCTION

The Cyberspace Attacks: When businesses use cyberspace to disrupt or damage it, machine learning relies on past data to predict the future. Machine learning (ML) is a type of Artificial intelligence (AI) in the form of machine learning (ML) enables computers to learn without being specifically programmed [1-3]. The fundamental objective of machine learning is to build a computer that can adapt while researching new ideas and machine learning concepts and employing basic machine learning techniques in Python. Algorithms with specialized functions are used during the training and estimate phase. The algorithm receives the training data and utilizes it to create predictions about the fresh test data [4-5]. Three categories can be used to categorize machine learning.

Education under supervision, education without supervision and further education. A system for learning contains an input as well as a linked collection of data from which to draw that was previously developed by humans [6-7].

Without supervision, unguided instruction it gives the learning algorithm data. The set of input data must be analyzed by the algorithm [8]. Last but not least, dynamic learning promotes engagement with its surroundings and uses both positive and negative feedback to enhance it [9-10]. Data scientists in Python utilize a variety of machine learning methods to find patterns that produce insightful conclusions. Based on "learning" data to create predictions, these many high-level techniques can be classified into two groups: supervised learning and unsupervised learning. It is possible to describe classification as the process of estimating the classes of a set of data [11]. Targets/tags or groups are other names for classes.

The job of approximating a function from an input variable (X) to an output variable (y) is the definition of probability distribution. Classification is a supervised learning process in machine learning and statistics where a computer learns from data that is provided to it and then applies that learning to categorize fresh observations. This data may be multivariate or strictly binary (for example, determining whether a person is male or female or whether an email is spam). Language recognition, typing skills, biometric data, and data categorization are a few examples of classification issues. The majority of organizations employs a preventive and attack strategy to combat cyberattacks. Threats are only eliminated and examined after they have been identified.

At this point, damage has been done-the network has been breached and important data has been compromised.

Detection and blocking of access, such as antivirus software and firewalls, and access control, such as passwords, are tools and measures used by many organizations. However, given the complexity and diversity of cybercrime in recent years, and the number of attacks that haven't reached the tabloids, it's safe to say that the answer has been incredibly damaging to the best and most ineffective administration. However, the future of cybersecurity is not as bleak as we think [12-13]. With the quick development of artificial intelligence, machine learning and quantum encryption technology, many new ways to deal with cyber-attacks have appeared endlessly. Cybersecurity plays an important role in the future of business and government. Armed with this understanding, many of our best thinkers have tackled this problem and succeeded.

## 2. LITERATURE SURVEY

The forecasting process is based on past and present circumstances, and it employs a technique or technology to find or support an unidentified, unseen, or challenging process environment before evaluating the outcomes.

As stated in the prior warning, the primary objective of network attack and defence operations is the precise prediction of DoS attacks. An efficient technique for identifying DoS assaults is anomaly-based detection. Many studies focus on DoS attacks from various perspectives. However, this process requires prior knowledge and it is difficult to distinguish between collisions and DoS traffic. They also need lots of historical data that cannot predict such an attack. Based on the traffic search data and data entry, first a common method based on genetic algorithm and Bayesian method and a probability estimation of the DOS attack distribution is planned, then the joint process is optimized using a genetic algorithm. By the agreement of the sample data, we obtained several batches of the relationship between traffic

and attacks, and then created several prediction sub-models about DoS attacks. In addition, the probability distribution of the DoS attack was obtained by subtracting the decision probability of each model according to the Bayesian method. This article explains a standard technique based on genetic optimization to categorize data against DoS, starting with the correlation between network traffic data and the amount of DoS attacks. This method first appropriately classifies the relationship between network connectivity and DoS attack volume according to group agreement and creates a DoS attack prediction sub-model. At the same time, the Bayesian method is used to calculate the results of connected devices for each model and then to obtain the distribution of DoS attack volume over a certain range. Thanks to the emergence of the Internet on mobile phones, phone calls, blogs, opinions, reviews, reviews, documentaries, social media, social media, Wikipedia etc. many documents [14]. It is possible to determine whether someone is engaging with a user maliciously or adversely by closely examining these patterns, which can also cause social difficulties.

From the above CDR simulations, it can be concluded that there are many predictions that can be made if this experiment is used in an Internet-based network and it can retrieve information that can be changed in the process conversion and emission matrices made. Doing so will help us decide. [15] Intrusion Detection Systems (IDS) are used to detect malicious attacks on IT systems. Attacks can be detected through monitoring and analysis of IT system operations. Ideally, IDS will generate alerts for all malicious detections and store them in the IDS database. Some alerts stored in the IDS database are relevant. A social stimulus is the opposite of a social response equivalent to the same anti-social response. A similar relationship means that two warnings were generated for the same offense. Being closely related to the same event means that two related alerts are produced due to related acts of violence [16]. An attack or multi-step attack is a group of malicious activities performed by the same attacker to achieve a specific goal. The relationship of violence to the same situation is violence. Causality means that the results of the current malicious campaign are a prerequisite for running the next malicious campaign [17].

System architecture is a model_based framework that defines the structure, behavior, and multiple perspectives of a system. Modeling is the description and representation of work in a way that supports thinking about patterns and behavior of systems.Users will provide input data that will be processed and then analyzed using different machine learning algorithms before using the previous data. The high accuracy of the algorithm will be given as GUI output. . On other networks, performance might be challenging and subpar. There are no comparisons of machine learning algorithms or performance indicators such recall F1 scores [18]. Process Analysis / Data Validation Process Development The error of the learning model (ML), which may be thought of as being near to the real error of the data, is obtained via machine learning. If the data volume is large enough to adequately reflect the population, validation might not be required. However, utilizing a data model that cannot accurately depict the population of the supplied data in reality. Identify causes for missing values, equal values, and data kinds, including integers and floating-point variables. When modifying the model's hyperparameters, the data model is utilized to give an objective evaluation of the model's fit to the training data. As evidentiary skills are ingrained in the organizational structure, evaluations may be biased. The validation procedure is employed for repeated examination of a certain model. As a specialist in machine learning, he utilizes this information to adjust model hyperparameters.

The process of data collection, data analysis and reporting on data content, performance and standards can be expanded over time using the worklist. It helps you understand your data and its properties as you search for data. Choosing algorithms and designs can benefit from

this information. Regression algorithms, for instance, may be used to analyze time series data, while classification algorithms can be used to analyze other types of data. (such as seeing the file format for a particular dataset) [19] Process for validating, cleansing, and preparing data Load the provided dataset after importing the library package. Determine the variations in data, such as data, data type, and measurement missing values, and then assign more values.

The acceptance certificate is an example of the information held by the training model to measure the model's skills when editing the model, and the process you can use to validate and evaluate data when measuring the sample. By renaming a specific dataset and posting rows etc. clear / prepare data. Identify different, distinct, and diverse processes. The steps and procedures to clear data vary by dataset. [20] To improve the value of data in analysis and decision-making, data cleaning's primary goal is to find and remove mistakes and inaccuracies.

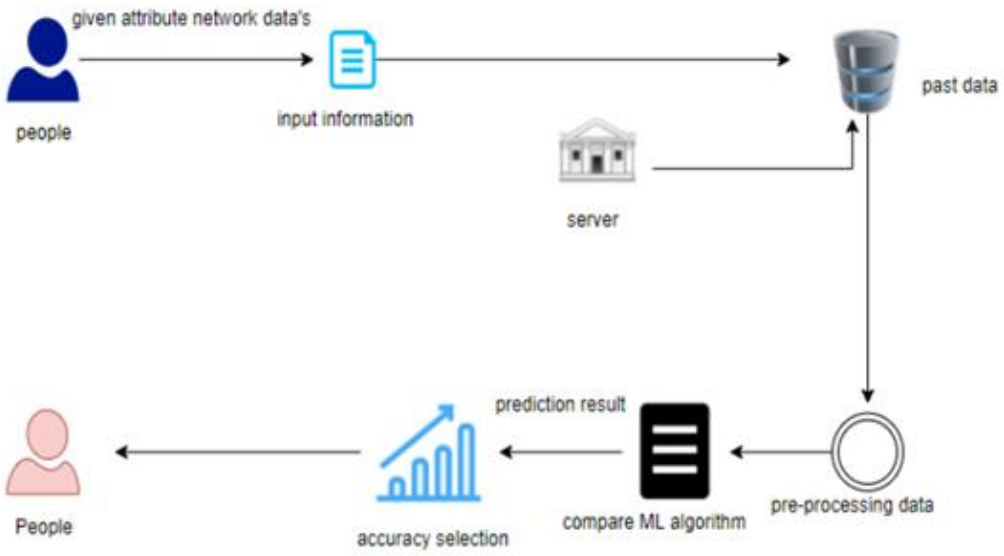




**Figure 1: System Architecture Diagram**

Data Visualization and Analysis The ability to visualize data is crucial for both statistics and machine learning. The quantitative description and estimate of data are truly the emphasis of statistics [21].

This is useful for searching and understanding data, helping identify patterns, data errors, inconsistencies, and more. With a little understanding, visual information can be presented and presented in a diagram that is more intuitive and useful than correlation or measurement or values in visual images such as tables and charts. Quickly visualizing patterns in data and other data is important in statistics and machine learning. It will show you many of the graphs you need to know about how to visualize data in Python and use them to better understand your data. The breadth and distribution of important behavior in the input data are factors that the majority of machine learning algorithms are sensitive to.

Outliers in input data can affect and interrupt the learning process of machine learning algorithms, causing additional training, incorrect models, and incomplete results. An attacker could skew the, training data before the prediction model was prepared and misinterpret the collected data accordingly. Outliers can compress the data volume by affecting the distribution of important features in statistics such as mean and standard deviation, and in

graphs such as histograms and scatter plots [22]. Finally, outliers may represent sample data related to computer security concerns such as fraud and inconsistency. It is not possible to fit the model to the training data, and it is not possible to say that the model is valid on real data.

To do this we need to make sure that our model captures the data well and does not make too much noise. Cross-validation is a process where we train a model using one dataset and then evaluate it using another dataset. The three steps involved in authentication are as follows:

- Keep part of the specified data structure.
- Demonstrate the model using additional information.
- Evaluate the model using a subset of the retained data. Advantages of the practice / split test
- This runs K times faster than onetime cross validation because K-fold cross validation recurrent the matching/partition K times.
- For the testing process it is easier to check the main information.

## 3. VALIDATION ADVANTAGES

- A more accurate estimate of the accuracy of the error.
- Use of more "more" data as all observations are used for training and testing.

Data preprocessing means transforming data before it is fed into an algorithm. The process of conversion of raw data into clean data is known as data processing .When data is collected from various sources, it is written in a raw form that cannot be analyzed.To get better results from using patterns in machine learning from data, it needs to be done right. Some specific learning models require information in certain formats; for example, the values are not supported by Random Forest algorithm. Hence, it is necessary to check for nulls the raw data from the original to perform the random forest algorithm. Another issue is that the data should be in a format where multiple machine learning and deep learning algorithms can be made on the given data. Anaconda Navigator is a desktop graphical user interface (GUI) included with the Anaconda® distribution that allows you to launch applications and easily manage conda packages, environments and channels without using the command line.

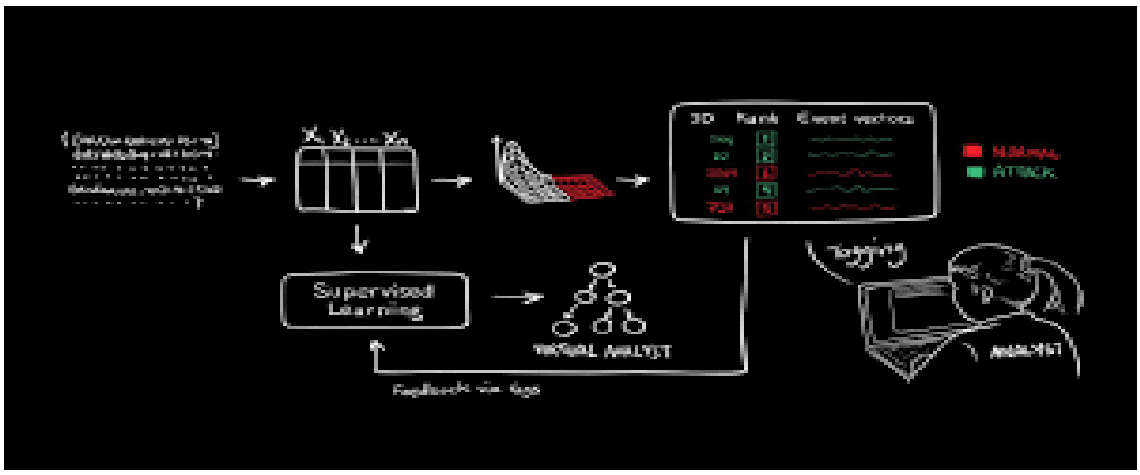


**Figure 2: Operational Diagram**

Using either the local Anaconda repository or Anaconda.org, Navigator may search for packages.

**Test Measures:** The logistic regression technique also makes estimates of values using independent linear regression. Any number between zero and infinity can be used as the estimated value. The algorithm's output has to be categorized using several pieces of information. The model with the highest predictive value is the logistic regression model.

**False Positive (FP):** It is assumed that the payer is an unauthorized individual. when the predicted class is true but the actual class is false. An illustration. If the prediction class claims that the passenger will live but the real class claims that the passenger is dead.

**False Negatives (FN):** Individuals who assume. When the actual class is offered but not the projected class an illustration. If the estimated level predicts that the driver will die but the real number indicates that the driver is still alive.

**True Positive (TP):** Anyone who refuses to pay is automatically assumed to be at fault. These positive numbers indicate that the prediction is accurate, which indicates that the answer is "yes" for both the real class and the anticipated class. The projected level will inform you of the same thing if the real number suggests that the driver is still alive.

**True Negatives (TN):** It is anticipated for defaulters to default. These are negative if the prediction was accurate, indicating that both the real class and the projected class include nulls an illustration. The approximate class will inform you of the same information if the genuine class indicates that the passenger is not living.

The True Positive Rate (TPR) is defined as the ratio of true positives to false negatives+ true positives.

False Positive Rate (FPR) is calculated as follows:

False Positive / (False Positive + True Negative).

**Accuracy:** The proportion of predictions for which the model correctly predicts defaulters and non-defaulters.

Accuracy is calculated as follows:

Accuracy = (True Positive + True Negative) / (True Positive + True Negative + False Positive +False Negative).

Accuracy is the most practical measurement. The model can be taken into consideration as a superior model if truth value is high. Yes, accuracy is a useful metric, but only if your data is reliable and both negative and negative numbers are about equal.

Precision is the percentage of accurate estimates.

Precision = True Positive / (False Positive + True Positive)

The proportion of the right prediction to all reliable observations is known as precision. All passengers made it out of this test alive. How many made it? False positives are linked to those with high levels. Our accuracy was 0.788, which is rather respectable. Remember: The portion of positive observations that was accurately anticipated.

(Proportion of actual defaulters at which the model will be validated) Recall = True Positive / (True Positive + False Negative)

**Recall (Precision):** Recall provides a precise estimation of the observed quality discovered for each observation made in the space. The weighted average of Precision and Recall is the F1 score. As a result, this score considers both true and false negatives. Although F1 is frequently more useful than the truth, especially when distribution classes are not uniform, the truth is difficult to explain intuitively. The most effective true, fake, and negative stories all share a common feature.It may be wise to consider both Precision and Recall if the outcomes of negative and negative are considerably different.

F-Measure is defined as 2TP divided by (2TP plus FP plus FN).

Formula for calculating the F1 score: F1 Score = 2*(Recall * Precision) / (Recall + Precision).

The four different algorithms: Logistic Regression [LR], Decision Tree [DT], Random Forest [RF], Support Vector Machine [SWM]. After continuous analysis of all the attacks of the algorithm, it has been proven that the random forest algorithm and the support vector machine have the highest percentage among all comparisons.

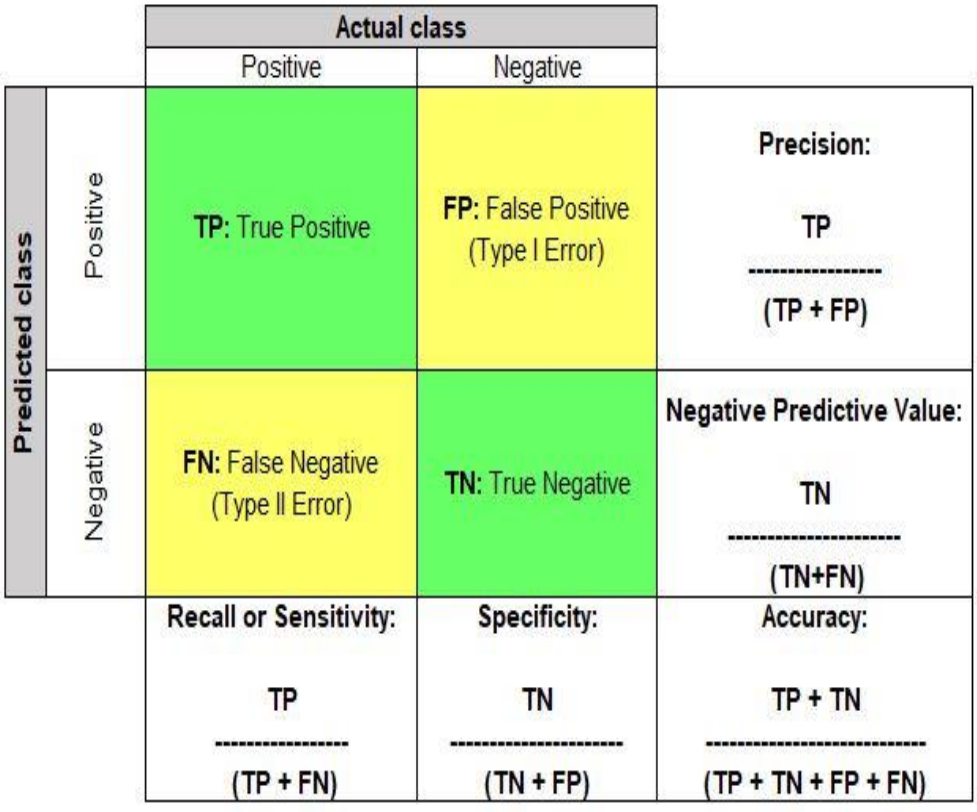


**Figure 3: Confusion Matrix**

## 4. CONCLUSION

Investigation is the first step in the analytical process, which finishes with design and interpretation rather than necessarily with data preparation and processing. Comparing each technique with various network attacks yields the public test's highest accuracy rating, which may then be used to forecast the effects of discovering the best connection in the future. The

information on diagnosing cyberattacks for each new connection is derived from this. Publish artificial intelligence-assisted divination models to increase human accuracy and offer prompt aid. It is clear from this model that the use of regional analysis and machine learning has contributed to the development of predictive models that will aid the cyber sector in shortening the time required for identifying and excluding all human mistakes.

Future network development departments will like to see real-time packet forwarding via port. By seeing the outcomes in a web or desktop application, you may automate this procedure. Improve work that was done in a simulated setting.

## REFERENCES


**[1]** K. Ahuja, H. Sekhawat, S. Mishra, and P. Jha, "Machine learning in artificial intelligence: Toward a common understanding," Turkish Online Journal of Qualitative Inquiry, vol. 12, no. 8, pp. 1143–1152, 2021.

**[2]** R. Ajmera and N. Saxena, "Face detection in digital images using color spaces and edge detection techniques", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, no. 6, pp. 718–725, 2013.

**[3]** M. Kumar, R. Ajmera, and D. Kumar, "Statistical analysis and accuracy assessment of improved machine learning based opinion mining framework," Advances in Nonlinear Variational Inequalities, vol. 27, no. 1, 2024.

**[4]** H. Sharma and R. Ajmera, "Comprehensive review and analysis on machine learning based Twitter opinion mining framework," Journal of Propulsion Technology, vol. 44, No. 5, 2023.

**[5]** R. Misra, "Cloud Computing: Fundamentals, Services and Security", International Conference on Engineering & Design (ICED), 2021.

**[6]** S. Pachauri, D. Sharma, and R. Misra, "Role of computer education in Indian schools," International Journal of Recent Research and Review, vol. XV, no. 3, pp. 15–18, 2022.

**[7]** K. K. Gautam, S. Prakash, R. K Dwivedi, "Patients medical record monitoring using IoT based biometrics blockchain security system", 2023 International Conference on IoT, Communication and Automation Technology (ICICAT), pp. 1-6, 2023.

**[8]** R. Misra and R. Sahay, "Evaluation of student performance prediction models with two-class using data mining approach," International Journal of Recent Research and Review, vol. XI, no. 1, pp. 71–79, 2018.

**[9]** P. Jha, K. K. Sharma, B. Jain, V. Sharma, "Digital Image Encryption Using AES Algorithm", EIJO Journal of Engineering, Technology And Innovative Research (EIJO–JETIR), Vol. 4, Issue. 2, 2019.

**[10]** R. Misra and R. Sahay, "Evaluation of five-class student model based on hybrid feature subsets," International Journal of Recent Research and Review, vol. XI, no. 1, pp. 80–86, 2018.

**[11]** S. Sharma, D. Arora, G. Shankar, P. Sharma, and V. Motwani, "House price prediction using machine learning algorithm," in Proceedings of the IEEE 7th International Conference on Computing Methodologies and Communication (ICCMC), pp. 982–986, 2023.

**[12]** H. Arora, T. Manglani, G. Bakshi, and S. Choudhary, "Cyber security challenges and trends on recent technologies," in Proceedings of the IEEE 6th International Conference on Computing Methodologies and Communication (ICCMC), pp. 115–118, 2022.

**[13]** K. Ahuja, Khushi, D. Sharma, and N. Sharma, "Cyber security threats and their connection with Twitter," in Proceedings of the IEEE 2nd International Conference on Artificial Intelligence and Smart Energy (ICAIS), pp. 1458–1463, 2022.

**[14]** W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," Bulletin of Mathematical Biophysics, vol. 5, pp. 115–133, 1943.

**[15]** Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," IEEE Transactions on Neural Networks, vol. 5, no. 2, pp. 157–166, 1994.

**[16]** N. Moustafa, J. Hu, and J. Slay, "A holistic review of network anomaly detection systems: A comprehensive survey," Journal of Network and Computer Applications, vol. 128, pp. 33–55, 2019.

**[17]** A. Agarwal, R. Joshi, H. Arora, and R. Kaushik, "Privacy and security of healthcare data in cloud based on the blockchain technology," in Proceedings of the IEEE 7th International Conference on Computing Methodologies and Communication (ICCMC), pp. 87–92, 2023.

**[18]** A. M. Samar, V. Bartos, and B. Lee, "Shallow and deep learning approaches for network intrusion alert prediction," Procedia Computer Science, vol. 171, pp. 644–653, 2020.

**[19]** D. S. Berman, "Survey of deep learning methods for cyber security," Information, vol. 10, no. 4, p. 122, 2019.

**[20]** V. Bartos, M. Zadnik, S. M. Habib, and E. Vasilomanolakis, "Network entity characterization and attack prediction," Future Generation Computer Systems, vol. 97, pp. 674–686, 2019.

**[21]** A. Aldweesh, A. Derhab, and A. Z. Emam, "Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues," Knowledge-Based Systems, vol. 189, 2020.

**[22]** S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.