# The Effectiveness of Machine Learning and Deep Learning Algorithms for Cyber Security

January 6, 2023

**MOONIB SARFARAZI, ISMAIL OUBAH, and OUMAIMA HOUMANI**

Department of Computer Engineering
Istanbul Aydin University
Istanbul
Oubah70@gmail.com, houmanii.oumaiima@gmail.com ,Sarfarazimoonib@gmail.com

## *Abstract*

With the growth of the Internet, cyberattacks are evolving quickly, and the state of cyber security is not promising. The main literature reviews on machine learning (ML) and deep learning (DL) techniques for network analysis and intrusion detection are covered in this survey report, along with a brief tutorial description of each ML/DL technique. Based on the temporal or thermal correlations between the papers that represented each method, they were indexed, reviewed, and summarized. We cover some of the frequently used network datasets used in ML/DL, highlight the difficulties of employing ML/DL for cybersecurity, and offer recommendations for future research directions because data are so crucial to ML/DL methodologies.

## I. INTRODUCTION

Security incidents of all kinds have grown exponentially in recent decades as information technology has become more and more prevalent. Businesses and individuals may suffer large financial losses as a result of cybercrime and network assaults.

Cybersecurity is the field of technology and practices that guards against unauthorized access, attacks, and damage to computers, networks, programs, and data.

The shift is being driven by data science, where machine learning, a key component of "Artificial Intelligence," can be crucial in revealing hidden patterns in data.

Several ML techniques, for example feature reduction, regression analysis, unsupervised learning, finding associations or neural network-focused deep learning techniques, can be used to effectively extract the insights or patterns of security incidents. These learning approaches are capable of identifying irregularities, malicious behavior, and data-driven patterns of associated security issues. They can also make intelligent decisions to stop cyberattacks.

In earlier procedures, just cybersecurity algorithms and techniques were used,hence identifying any cubersecurity threats takes human work.This includes even detection of already existing cybersecurity attacks. The effort required to identify an existing attack type is equivalent to that required to identify a new attack type. Given that there are millions of cybersecurity threats occurring worldwide, this task is virtually impossible,hence, it is very important to find cybersecurity threats. And this can be accomplished by utilizing machine learning and deep learning algorithms.

## II. How is Machine Lerning Used in Cybersecurity?

Machine learning is a subset of artificial intelligence, creates algorithms from previous datasets and statistical analysis to understand a computer's behavior. The computer can then modify its activities or even perform functions for which it was not designed. Machine learning is now a key cybersecurity asset as a result of these capabilities.

*1) Supervised ML algorithms*

- **Naïve Bayes (NB)**

These techniques are probabilistic classifiers that operate under the prior assumption that the input dataset's features are unrelated to one another. They are adaptable and don't need big training datasets to yield meaningful results.

- **Logistic Regression (LR)**

They are using a discriminative classification model and are categorical classifiers. The a-priori independence of the input features is an assumption that LR methods share with NB algorithms. The volume of training data has a significant impact on how well they perform. Support Vector Machines (SVM)

- **Support Vector Machines (SVM)**

The objective of these non-probabilistic classifiers is to maximize the distance between each category of samples by mapping data samples in a feature space. They do not make any assumption on the input features, but they perform poorly in multi-class classifications. They should therefore be utilized as binary classifiers. They may have slow processing times due to their restricted scalability.

- **Random Forest (RF)**

A random forest is a collection of decision trees that takes into account the results of each tree before delivering a unified result. Each decision tree functions as a conditional classifier; starting at the top, a given condition is tested against one or more aspects of the analyzed data at each node of the tree. These techniques work well for multiclass issues and large datasets, although deeper trees may result in overfitting.

- **K-Nearest Neighbor (KNN)**

KNNs are used for classification and can be applied to issues with many classes. In order to categorize each test sample, they compare it to all the training examples, which makes both their training and test phases computationally demanding.

2) *Unsupervised SL algorithms*

- **Clustering**

Data points with comparable qualities are grouped together. K-means and hierarchical clustering are well-known strategies. Although clustering approaches have a limited scalability, they offer a flexible approach that is frequently employed as a test run before implementing a supervised algorithm or for anomaly detection.

- **Association**.

They aim to identify unknown patterns between data, making them suitable for prediction purposes. However, they tend to produce an excessive output of not necessary valid rules, hence they must be combined with accurate inspections by a human expert.

## III. Deep learning in Cybersecurity?

Deep learning is a subset of machine learning that involves training artificial neural networks on a large dataset. It has the ability to learn and make decisions on its own, without the need for explicit programming. In the field of cybersecurity, deep learning can be used to identify and classify malicious activity, detect anomalies in network traffic, and recognize patterns that may indicate an attack.

One of many examples on the application of deep learning in cybersecurity is in the detection of network intrusions and anomalies. By training a deep learning model on a large dataset of normal network traffic, the model can learn to recognize patterns that are indicative of normal behavior. If an anomaly is detected, the model can alert security personnel, who can then investigate and take appropriate action.

Overall, deep learning has the potential to significantly improve the effectiveness of cybersecurity measures by enabling the detection of previously unknown threats and anomalies. However, it is important to note that deep learning algorithms are not a silver bullet and should be used in conjunction with other security measures to provide a comprehensive defense against cyber-attacks.

There are a number of different deep learning algorithms that can be used in cybersecurity, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short-term memory (LSTM) networks.

CNNs are particularly useful for image recognition tasks and have been applied to the detection of malware based on the appearance of the code or the behavior of the malware when executed. RNNs, which are well-suited to tasks involving sequential data, have been used for intrusion detection by analyzing network traffic patterns and identifying anomalies. LSTM networks, which are a type of RNN, have also been applied to intrusion detection, as well as to the detection of phishing attacks by analyzing the language used in emails.

Other deep learning algorithms that have been applied to cybersecurity include autoencoders, which can be used for anomaly detection, and generative adversarial networks (GANs), which can be used to generate synthetic data for training and testing purposes.

It is important to note that no single deep learning algorithm is the best choice for all cybersecurity tasks. The appropriate algorithm will depend on the specific problem being addressed and the characteristics of the data being used.

## IV. RELATED WORKS

Here is a table of metrics for a few research studies that have explored the use of machine learning and deep learning algorithms for cybersecurity, along with the ML/DL techniques that were used:

| **Methods** | **Dataset** | **Accuracy** | **Precision** | **Paper** |
|---|---|---|---|---|
| **SMV** | KDD-CUP 99 | 82.31 | 74 | **Pervez &Farid** |
| **Mix-kNN** | KDD-CUP 99 | 98.55 | --- | **E. G. Dada** |
| **DBN** | KDD-CUP 99 | 93.49 | 93.25 | **N. Gao, et al** |
| **RNN** | KDD-CUP 99 | 77.55 | 84.6 | **C.L. Yin, et al** |
| **CNN** | NetFlow | 99.41 | --- | **W. Wang, et a** |
| **DT** | KDD-CUP 99 | 99.89 | --- | **Azad and Jha** |

It is difficult to say which machine learning algorithm is the most commonly used in cybersecurity, as the choice of algorithm depends

on the specific problem being addressed and the characteristics of the data being used. Different machine learning algorithms are suited to different types of problems and data, so the most appropriate algorithm for a given task may vary. this study was based on KDD-CUP 99 dataset and we see that Decision Tree algorithm gives the best performance, it boasted high accuracy of 99.89%. The CNN also gives a good accuracy of 99.41% but on another dataset because CNN is based on image features. Overall, the performance of machine learning algorithms in cybersecurity is promising, but there is still room for improvement and further research is needed to fully understand the capabilities and limitations of these algorithms in this domain.

For our model we tried to implement ML/DL techniques on another aspect of Cybersecurity, in our model we worked on Malware Files (.exe, DLL), we will use Random Forest Algorithm, Logistic Regression and ANN for deep learning, therefore we will see how our techniques will be efficient and we see if it gives a good accuracy.

## V. OUR WORK

As we mentioned earlier, our application based on malware files so we will build a malware detector using the mentioned algorithms (RF, LR, ANN). The first step in searching for a dataset that will met our criteria. After that, we injected the dataset to a Pandas framework within the Spyder IDE running the Scikit-learn and TensorFlow toolkits. But before that we made sure that the dataset is clean and preprocessed well (filling the missing values, drop unnecessary columns, scaling…), after that we split our data to train and test and then we call the fit function, and then predict the output values for the test data, then we score our model and see its precision and if it worked well or not for this specific data. This was for ML algorithms

For ANN we need to define: first layer, hidden layers and output layer then we chose the activation function that could work well in this case we chose ‘relu’ for first and hidden layers and then ‘segmoid’ for the output layer. The final step is the same as ML algorithm with fit and predict functions and then we see the score.

Here is a table of metrics for a few research studies that have explored the use of machine learning and deep learning algorithms for cybersecurity, along with the ML/DL techniques that were used:

| **Methods** | **DataSet** | **Accuracy** |
|---|---|---|
| **RF** | Malware file detection | 0.983 |
| **LR** | Malware file detection | 0.980 |
| **ANN** | Malware file detection | 0.988 |

We can see that ANN algorithm woks perfectly with this dataset with an accuracy of 98,8 % while machine learning algorithms LR and RF got almost the same score of 98%. So, we can say that all the 3 models work perfectly with this dataset, but if we had even larger dataset it’s recommended to use RF algorithm or ANN since they are good at handling large datasets and resistance to overfitting

So we can create a system that works and identify malicious or anomalous files in real-time based on RF, LR, ANN algorithms.

## VI. DISCUSSION

As previously found, the ANN was the best performing algorithm out of the three tested. The scores of our results are seen in the table above, but every method used for implementing an intrusion detection system has its own advantages and disadvantages. There are various methods that can be used in this field. Therefore, it is not possible to choose only one particular method to implement an intrusion detection system.

One advantage of using Random Forest for malware file detection is that it can handle large datasets with high-dimensional features, such as those commonly found in malware detection tasks. Additionally, Random Forest models are relatively easy to train and has a good resist of overfitting since we tried our dataset with scaling and without it, the result were the same for RF , but for ANN and LR the results were bad (around 75% for ANN and 70% for LR)

More than that it can handle unbalanced datasets same as the dataset we used in our model where the number of malicious samples is typically much smaller than the number of benign samples.

There are a few key considerations to keep in mind when using Random Forest Or any other technique for malware detection:

1. Feature selection: It is important to carefully select the features that are used to train the model, as using too many irrelevant or redundant features can degrade the model's performance.
2. Overfitting: Like any machine learning model, Random Forest is prone to overfitting if the model is not properly validated. It is important to use cross-validation and early stopping to prevent overfitting.
3. Model interpretation: While Random Forest models can be highly accurate, they can be difficult to interpret, as the model is essentially a black box that combines the predictions of many decision trees. This can make it difficult to understand how the model is making its predictions and to identify any biases or inconsistencies in the data.

## VII. CONCLUSION

Machine learning and deep learning techniques have the potential to significantly improve cybersecurity by automating the detection and classification of threats. These techniques can analyze large volumes of data, including network traffic and system logs, to identify patterns and anomalies that may indicate the presence of a threat.

There are several advantages to using machine learning and deep learning techniques in cybersecurity, which are

- Scalability (handling large datasets and perform well on high-dimensional data)
- Adaptability (learn from data and adapt to changing circumstances, allowing them to stay up-to-date with the latest threats.)
- Speed (processing data and making predictions much faster than humans)
- Accuracy (Machine learning and deep learning models can achieve high levels of accuracy,and improving the overall effectiveness of cybersecurity defenses.)

However, there are also some challenges to using machine learning and deep learning techniques in cybersecurity, including the need for large amounts of labeled data, the risk of overfitting, and the potential for biased or unfair predictions. It is important to carefully consider these and other factors when implementing machine learning and deep learning techniques in cybersecurity

## References


M. S. Pervez and D. M. Farid, ‘‘Feature selection and intrusion classification in NSL-KDD CUP 99 dataset employing

SVMs,’’ in Proc. 8th Int. Conf. Softw., Knowl., Inf. Manage.Appl. (SKIMA), 2014

] E. G. Dada, ‘‘A hybridized SVM-kNN-pd APSO approach to intrusion detection system,’’

in Proc. Fac. Seminar Ser., 2017.

N.Gao, L.Gao, Q.Gao, H.Wang, ‘‘An intrusion detection model based on deep belief networks,’’ in Proc. 2nd Int. Conf. Adv.

Cloud Big Data,2018

C.L.Yin, Y.F.Zhu, J.L.Fei, X.Z.He, ‘‘A deep learning approach for intrusion detection using recurrent neural networks,’’

IEEE 2017

W. Wang, M. Zhu, X. Zeng, X. Ye, and Y. Sheng, ‘‘Malware traffic classification using convolutional neural network for

representation learning,’’ in Proc. Int. Conf. Inf. Netw. 2017.

C. Azad and V. K. Jha, ‘‘Genetic algorithm to solve the problem of small disjunct in the decision tree-based intrusion

detection system,’’ Int. J. Comput. Netw. Inf. Secur. 2017.