**How to Cite**:
Madhavi, S., Divyadharshini, D., & Divya, S. (2022). A machine learning model to predict cyber attack. *International Journal of Health Sciences*, *6*(S6), 1341–1349.
https://doi.org/10.53730/ijhs.v6nS6.9745

# A machine learning model to predict cyber attack

**Dr. S. Madhavi**
Professor, HoD, Dept. of Computer Science and Engineering, K.S. Rangasamy College of Technology
Email: Tiruchengodemadhavis@ksrct.ac.in

**D. Divyadharshini**
UG Students, Dept. of Computer Science and Engineering, K.S. Rangasamy College of Technology, Tiruchengode
Email: divyadharshini1045@gmail.com

**S. Divya**
UG Students, Dept. of Computer Science and Engineering, K.S. Rangasamy College of Technology, Tiruchengode.
Email: 2001divya2001@gmail.com

***Abstract*---**All users place a high value on information security. Over the last few decades, cyber security has been one of the most fascinating and essential topics in cyber re-search. The goal of this work is to propose a machine learning method for the problem of cyber security threats. As a result, a multi-model machine learning approach was developed to predict the top ten traits that can be used to detect cyber threats. This work finds the importance of each feature in from the dataset using machine learning. this model trained with random forest model and exclude the features which are least important to the model.

***Keywords*---**cyber-attacks, data, machine learning, dataset training, model building, prediction of accuracy, literature review.

## Introduction

Recognizing the offenders of cybercrime and understanding the techniques of attack are critical components in the battle against crime and criminals. It is tough to detect and avert cyber-attacks. However, academics have lately solved these issues by creating security models and generating predictions using machine learning. There are several approaches for predicting offender-induced attacks in the literature. On the other hand, they are unable to forecast cyber-

crime and cyber-attack strategies. Using actual data, this problem may be addressed by identifying an assault and the culprit of such an attack. The information includes the sort of crime, the gender of the criminal, the amount of damage, and the techniques of assault. The data can be obtained through the applications of people who have been subjected to cyber-attacks to the forensic units. In this research, we use machine-learning approaches to assess cybercrime in two separate models and estimate the influence of the stated attributes on detecting the cyber-attack method and the culprit.

Huge volumes of data are exchanged between networked devices, and the security of that data deserves our attention. Intruders steal confidential communications while moving from one device to another. It is essential to protect data and information from intruders. Machine learning can quickly give insights hidden in cyber data for fast, automatic answers and smarter conclusions as the study of computer algorithms improves via experience. The main goal of this project is to safeguard the network by modeling the application and detecting and analyzing cyber-attacks in the network using the ML technique. We also aim to protect data transfer in the network and increase security to the greatest extent possible. The suggested approach attempts to prevent unauthorized/malicious users from gaining access to personal information and, as a result, to fight threats. Network security measures and protection schemes have also evolved in response to user wants and challenges.

**Aim of the project**

- The main goal of this project is to safeguard the network by using machine learning to model the application and detect assaults. To investigate a network cyber attack.
- To ensure the data transmission in the network is secure.
- To increase the number of security personnel as much as possible.

The Contribution of this paper is as follows: Section 2 briefly discussed related works. Section 3 recounts the plan for the algorithm as well as the implementation of modules. Section 4 Contains the testing and Section 5 contains results and Section 6 Conclusion and Future enhancements and Section 7 contains References.

**Related Works**

In this paper [1], a survey on all 4 machine learning algorithms for prediction of cyber attacks were studied and the 4 algorithm that were used are random forest, Decision tree, SVM and K- nearest neighbour algorithm. Out of these algorithm, Random forest produces the highest accuracy. In this paper [2], a framework called Information Foraging for algorithm discovery is addressed and the main use of this framework is, that this framework has a set of standards that has to be followed to prevent the cyber attacks. In this paper [3], Cyber attack detection model is explained based on the dataset and the accuracy it produces. In this paper [4], the authors explained about the 3 main types of network attacks. The first one is single stage network attack, the second one is double stage network

attack and the final one is multiple stage network attack. Also in this paper, the features that causes these attacks are also identified.
The cybersecurity issue has become more serious day by day not only in IT field, also in all the domains. In this paper [5], the harmful effects of cyber-attacks are studied and an oversampling approach is used to predict the cyber attack. In this paper [6], a dataset that contains features to predict the cyber attacks is taken and this dataset is preprocessed and the features are studied to predict the top features. In this paper [7] , a real time case study is shown. SABU platform were affected with SQL injection. To sort those attacks, 1 Million cyber security alerts were sent to the platform. In this paper [8], the internal security issues that can happen in any computer was discussed. The issues including firewall issue, and antivirus issue and the only solution to prevent this issue is by setting up a strong internal security software and proper updation of OS. The various cyber attacks like SQL injection, DoS attacks and DDoS are studied and their method and nature of attack is also studied. These attacks mostly deals with sensitive data [9]. The various cyber attacks were studied based on false positive and true positive. In this paper, the algorithm that is used is tf-idf and Linear SVC model. The tf-idf produced an accuracy of 60% and the Linear SVC model produced an accuracy of 62 % [10].

**Implementation of modules**

To identify cyberattacks, we propose developing a multi-model supervised learning-based system using Random Forest,. The prediction model will assist us in identifying the top ten characteristics that may be used to anticipate cyber assaults. The fundamental advantage of integrating numerous approaches is that each method may benefit from the complimentary predictive properties of the others.Multi-component algorithms use multi-layer and deep architectures to progressively extract data's fundamental properties from the lowest to highest levels, and they can also uncover diverse patterns in enormous amounts of data. The implementation of our approach is given below –

- Exploratory Data Analysis, which is required for spotting hidden patterns, detecting outliers, identifying relevant variables, and identifying any irregularities in data.
- Model Building Using Multiple Algorithms, we build the model using several algorithms
- Detection of top 10 features from the dataset using our model.

**EDA (exploratory data analysis)**

The following are EDA's objectives in a nutshell:

- Gain as much insight into the database as possible/understand its structure;
- Create a visual representation of the possible links (direction and magnitude) between exposure and outcome variables.
- Look for outliers and anomalies (numbers that deviate significantly from the rest of the data);
- Build parsimonious models (a predictive or explanatory model that performs

with the fewest potential exposure variables) or make a preliminary model selection;
- Generate clinically significant variables by extracting and creating them.

Exploratory data analysis is a crucial step in studying and analyzing diverse data sets, as well as summarizing their key properties. The application of EDA may aid in the discovery of hidden patterns in datasets, and its importance in data science cannot be overstated. Exploratory data analysis (EDA) is often required for spotting hidden patterns, detecting outliers, identifying relevant variables, and identifying any irregularities in data. At this point, the data is cleaned and pre-processed, with missing and null value records being removed. To achieve an accurate result, preprocessing identifies and eliminates or substitutes missing values in the dataset that only make up a tiny fraction of the total data. After that, the dataset is pre-processed, and the cleaned data is utilized for training. The cleaned data will then be examined, and all essential steps will be taken, such as eliminating noise from the image and setting all resolutions to the same. There are four stages to data preparation.

- The first phase is data cleaning, which involves fixing or removing duplicate, poorly formatted, or damaged data.
- The second phase is data integration, which involves combining data from several sources into a single perspective.
- The data reduction stage is the third step, in which the data is encoded, scaled, and sorted if necessary.
- The data transformation is the last phase, in which the data is turned into the desired format.

**Building the model**

The second module leverages the use of supervised machine learning algorithms to predict the top 10 features from the dataset that can predict cyberattacks**–**

**Model building – Random Forest**

Steps involved in random forest algorithm:

- Step 1: n number of random records are taken from the data set having k number of records.
  In our dataset, there are more than 150 features that are present. From that 150 features, a certain number of records will be taken for the next step.
- Step 2: Individual decision trees are constructed for each sample.
  Then a decision tree will be constructed for all the records that were taken in the first step.The algorithm operates by constructing a multitude of decision trees at training time and outputting the mean/mode of prediction of the individual trees
- Step 3: Each decision tree will generate an output.
  The job of each decision tree is to generate a probability of all the combinations. These probabilities are the outputs of each decision tree.

- Step 4 : Selecting the top 10 features from the datasets
  Our Algorithm builds a forest in the form of an ensemble of decision trees which adds more randomness while growing the trees. While splitting a node, it searches for the best features from the random subset of features which adds more diversity, thereby resulting in a better model.
- Step 5 : Improving accuracy
  Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, our algorithm takes the prediction from each tree.

**Hyper parameter tuning – Random Forest**

Implementing hyper parameter tuning will be helpful in determining few things that are related to our project.

- It helps to identify the depth of the random forest algorithm that we are implementing?
- (It helps to find the minimum number of samples required at a leaf node in our decision tree
- It will also find how many trees should be implemented in a random forest tree.

The acquired accuracy for the hyperparameter tuning is 86% with 20 splits.

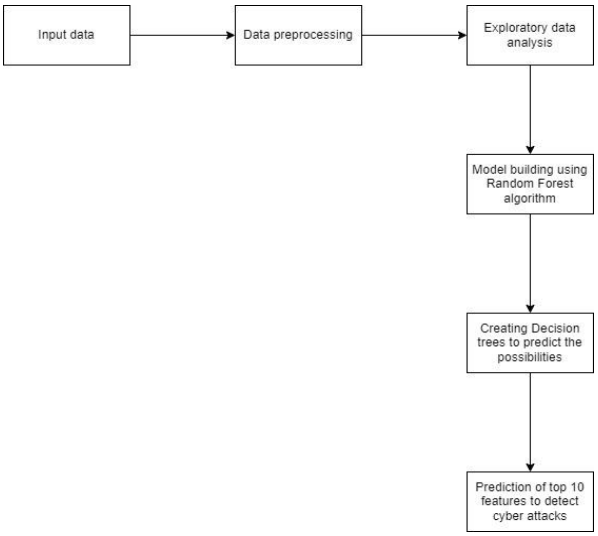
**Detection of top 10 features from the dataset**

Following the classification by the algorithms, the final module detects the top ten characteristics from the dataset. The results of the testing were outstanding in terms of the system's usability and learning impact. Our method achieved 99 percent training accuracy and 90.04 percent testing accuracy. This study was completed with the assistance and consultation of cybersecurity data sources.

**Implementation**
**System requirements**

- **Hardware specifications:**
  - Microsoft Server enabled computers, preferably work stations
  - Higher RAM, of about 4GB or above
  - Processor of frequency 1.5GHz or above
- **Software specifications:**
  - Python 3.6 and higher
  - Anaconda software
  - Jupyter Notebook

**System Architecture**




Initially, a dataset is taken and the preprocessing steps are done to the dataset. The preprocessing steps include data cleaning, data transformation. After the dataset preprocessing, the dataset is visualized and the exploratory data analysis process will be carried out. Then we will start to build our model using random forest algorithm. After the model building, the top 10 features that detect cyber attacks will be detected.

**Results**

To identify cyber assaults and risks to online information, we evaluated a multi-model supervised machine learning classification-based system that predicts the top 10 characteristics from a given dataset. To evaluate and classify the state-of-the-art in the area, a review of cutting-edge multi-model machine learning-based techniques was provided, spanning from the state estimation-based sensor to the new moving target defense methods. As technologies grow more ubiquitous and more data becomes accessible online, substantial vulnerabilities in network levels are developed, as are a multitude of possibilities and difficulties, mostly relating to the danger of cyberattacks by unscrupulous users. We assessed the effectiveness of our technique in predicting top attributes that may identify cyber threats. The above-mentioned finding is simply offered for theoretical reasons. A multi-model strategy employing supervised classifiers outperforms the other known techniques. The results of the testing were outstanding in terms of the system's usability and learning effect.

**Top 10 Features by using Random Forest Model**

Table 5.1
Top 10 Features

| **Features** |
|---|
| Entry Point |
| 'NEUTRAL' |
| 'MSVCRT.DLL' |
| 'WININET.DLL' |
| 'WS2_32.DLL' |
| 'USER32.DLL' |
| 'ENGLISH US' |
| 'SHELL32.DLL' |
| 'KERNEL32.DLL' |
| 'CRYPT32.DLL' |

**Comparison of the existing and proposed system**

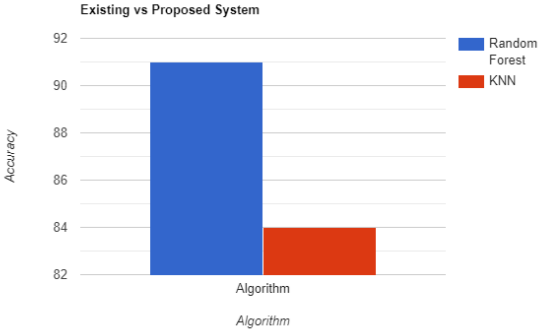



Figure 5.1 Comparison of existing and proposed system

As the existing system used just KNN algorithm to predict the top features that can help predict cyber attacks, the accuracy is low and the system is not fast processing. Such existing systems are effective, but show a low accuracy rate and are not very precise. The accuracy of the existing systems for cyber attacks prediction tends to decrease when there is a high gap between the training and the testing data. The existing systems are no robust in identifying and predicting different types of attacks, which proves to be inefficient and pose a high risk of data leakage and malicious threats to the data. The existing has low efficiency in terms of loading time and implementation time. Also, the testing and training is not done with the proper test-train split ratio.

The proposed system is done in real time and also the accuracy is high in the proposed system. The proposed system loading speed and execution speed is really fast when compares with the existing system. The proposed system is highly efficient and scalable and also further improved for complex use cases. To detect the cyber attacks, we consider building aefficient system using Random

Forest Algorithm. The predictive model will help us detect the top 10 features that can be used to predict cyber attacks. The main advantage of using a random forest algorithm is, that it will use many decision trees to predict the probability and deep architectures to gradually extract theinherent characteristics of data fromthe lowest to highest levels, and they may also discover various structures in a large amount of data. The proposed framework has better accuracy and efficiency than state-of-the-art methods.

**Conclusion and Future Enhancements**

We proposed a random forest framework for predicting cyber attack features. The framework can accommodate complex phenomena exhibited by datasets, including long-range dependence and highly nonlinearity. Using datasets from kaggle, we showed that the framework significantly outperforms the other prediction approaches in terms of prediction accuracy. We hope the present work will inspire more research in deploying deep learning to prediction tasks in the cybersecurity domain. This project can be developed in a web application and also this project can be deployed in a cloud platforms like AWS and Digital ocean so that this project willbe accessible by anyone

**References**

1. Abel Yeboah-Ofori; Charles Boachie,"Malware Attack Predictive Analytics in a Cyber Supply Chain Context Using Machine Learning", 2019 International Conference on Cyber Security and Internet of Things (ICSIoT)
2. Adam Dalton; Bonnie Dorr; Leon Liang; Kristy Hollingshead,"Improving cyber-attack predictions through information foraging",2019 IEEE International Conference on Big Data (Big Data)
3. Fahima Hossain; MarzanaAkter; Mohammed Nasir Uddin,"Cyber Attack Detection Model (CADM) Based on Machine Learning Approach",2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)
4. Haitian Liu; Rong Jiang; Bin Zhou; Xing Rong; Juan Li; Aiping Li,"A Survey of Cyber Security Approaches for Prediction", 2021 IEEE Sixth International Conference on Data Science in Cyberspace (DSC).
5. Hamzah Al Najada; Imad Mahgoub; Imran Mohammed, "Cyber Intrusion Prediction and Taxonomy System Using Deep Learning And Distributed Big Data Processing", 2018 IEEE Symposium Series on Computational Intelligence (SSCI).
6. Kulvinder Singh; Sudan Jha, "Cyber Threat Analysis And Prediction Using Machine Learning",2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)
7. Lo'ai Tawalbeh; Fadi Muheidat; Mais Tawalbeh; Muhannad Quwaider; Gokay Saldamli, "Predicting and Preventing Cyber Attacks During COVID-19 Time Using Data Analysis and Proposed Secure IoT layered Model", 2020 Fourth International Conference on Multimedia Computing, Networking and Applications (MCNA).
8. Martin Husák; Jaroslav Kašpar,"Towards Predicting Cyber Attacks Using Information Exchange and Data Mining",2018 14th International Wireless Communications & Mobile Computing Conference (IWCMC).

9. Md Anisur Rahman; Yeslam Al-Saggaf; Tanveer Zia,"A Data Mining Framework to Predict Cyber Attack for Cyber Security", 2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA).
10. Prerit Datta; Natalie Lodinger; Akbar SiamiNamin; Keith S. Jones,"Predicting Consequences of Cyber-Attacks", 2020 IEEE International Conference on Big Data (Big Data).
11. Kanya, N., Rani, P. S., Geetha, S., Rajkumar, M., & Sandhiya, G. (2021). An efficient damage relief system based on image processing and deep learning techniques. REVISTA GEINTEC-GESTAO INOVACAO E TECNOLOGIAS, 11(2), 2124-2131.
12. Arnawa, I.K., Sapanca, P.L.Y., Martini, L.K.B., Udayana, I.G.B., Suryasa, W. (2019). Food security program towards community food consumption. *Journal of Advanced Research in Dynamical and Control Systems, 11*(2), 1198-1210.
13. Gede Budasi, I. & Wayan Suryasa, I. (2021). The cultural view of North Bali community towards Ngidih marriage reflected from its lexicons. *Journal of Language and Linguistic Studies*, *17*(3), 1484–1497
14. Lopez, M. M. L., Herrera, J. C. E., Figueroa, Y. G. M., & Sanchez, P. K. M. (2019). Neuroscience role in education. International Journal of Health & Medical Sciences, 3(1), 21-28. https://doi.org/10.31295/ijhms.v3n1.109