www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

Vol. 03, Issue 04, April 2023, pp : 932-938

e-ISSN : 2583-1062

Impact Factor : 5.725

PREDICTION OF A BITCOIN HEIST RANSOMWARE ATTACK USING DATA SCIENCE PROCESS

Uma Maheshwari K[1], Suganthi Swarna P[2], Vishnu Aravind R[3], Mrs. Srinidhi S[4]

[1,2,3] Student, Computer science and Engineering, Agni College of Technology,Chennai-600130,TamilNadu,India

[4]Assistant Professor, Computer science and Engineering, Agni College of Technology,Chennai-600130,TamilNadu,India

DOI: https://www.doi.org/10.58257/IJPREMS31061

ABSTRACT

In recent years, ransomware attacks have become a major threat to computer systems. Although these attacks mostly focus on resource-rich general-purpose computing systems, there is a growing need for more efficient predictive algorithms, machine algorithms, for learning. To solve this problem, this paper presents a method for collecting data from artificial intelligence and machine learning to predict ransomware attacks. Data science techniques are used to develop better predictive models. The key steps to building a successful model include analyzing differences and understanding the data.

Keywords: Bitcoin Heist, Ransomware Attack, Voting classifier, Random Forest Classifier, Logistic Regression,

1. INTRODUCTION

The key steps in building a successful model are identifying variables and interpreting the data. The uses several machine learning algorithms for preliminary data to predict the probability of a ransomware attack and compares their accuracy to determine which are most effective. Most modern ransomware uses Bitcoin as payment, but current methods for identifying ransomware transactions rely on limited heuristics or time-consuming data collection techniques. Our research leverages the latest data analysis methods, especially topological data analysis, to obtain transactions associated with malicious Bitcoin addresses and ransomware. Our proposed system can identify new ransomware families or previously unknown ransomware associated with changes. With our approach, we can see that our technology effectively improves the accuracy and recovery of ransomware transaction detection compared to current heuristics.

2. LITERATURE SURVEY

[1] Yitao Li, Cüneyt Gürcan Akcora, Yulia R. Gel and Murat Kantarcıoğlu, 2019 - Bitcoin Heist: Topological Data Analysis for Ransomware Detection on the Bitcoin Blockchain. Ransomware is a virus that encrypts files and other resources for victims and then demands payment to decrypt them. Ransomware can hide or limit access to resources and can be divided into two broad categories. Like PC systems, mobile and IoT devices can be infected with ransomware. Ransomware can be distributed using web-based or email links. The use of exploits to spread ransomware has increased in recent years. For example, Crypto Locker is distributed via spam using the Game over ZeuS botnet. When run, the malware searches the command and control centre. While more ransomware can now use anonymous names, older versions of malware, for example, rely on hard-coded IP addresses and domain names to connect to information. Hide and manage servers.The ransomware then sends a request to send a certain amount of bitcoins to a bitcoin address after locking or encrypting the resource. It can be affected by the number and size of encrypted resources. Victims receive decryption tools after payment. But sometimes ransomware has a flaw, like WannaCry, that makes it impossible to determine who paid the ransom.

[2]Bitcoin Robbery Ransomware Crime Family 2021 - Taxonomy Reviewed by Micheline Al Harrack Due to crime, payments are made using cryptocurrencies and crime is difficult to trace. It is important to check and flag these transactions to identify them as legitimate or illegal transactions of digital currency exchanges and exchanges. Machine learning techniques are used to train computers to recognize certain changes and evaluate whether they are malicious or harmful. I recommend using the Bitcoin Heist dataset to classify various scams. Multiple analyzes of different products to distinguish what is tagged as ransomware or related attacks.I generated a random forest distribution using group learning and split decision trees.

[3]Sabira Karim and Shemitha PA 2021 - examine the detection and distribution of ransomware Bitcoin transactions. Because public information is stored in a highly distributed manner, Bitcoin could be a suburban version of payment. The public record that tracks Bitcoin transactions is called the blockchain and is expanded and managed by anonymous miners. The order of written blocks is called the blockchain. The Bitcoin cryptocurrency market is the most famous.All bitcoin transactions occur digitally and are mostly anonymous. In this case, some scammers have

been prevented from using Bitcoin as a haven for shady deals such as ransomware payments. Payment entry for ransom to be paid is affected by malware known as ransomware. Machine learning can also be used. This study aims to investigate the effectiveness of various machine learning techniques in policing.Ransomware can be a type of malware that encrypts the victim's data and resources and then demands payment to decrypt them. Machine learning algorithms can also be used to evaluate historical transactions based on data to predict who will get ransomware.

[4]J. Hernandez-Castro, A. Cartwright, and E. Cartwright 2021 - Economic analysis of ransomware and its impact on health. This makes the case of Gpcode ransomware even more interesting because it shows how criminals try, make many mistakes and try again and again before they "finally fix the problem"."Cryptolocker was discovered in the wild in 2013 and is one of the first, or even the first, protocols to use a protocol similar to Young and Yung's in a practical way. Cryptolocker has demonstrated the ability to steal a lot of data. It has been monetized using crypto viruses (see below) Ransom since then. the number of families and different types of software (CryptoWall, TorLocker, Fusob, Cerber, and TeslaCrypt, to name a few); strategies valued at $1 billion; financial tools and infectious disease methods they use to steal victims' money.

3. MATERIALS AND METHODS

3.1 EXISTING SYSTEM:

Ransomware is one of the biggest problems in cybercrime, affecting operations, availability, reputation and causing huge financial losses. Ransomware is usually designed to infect a computer. In this study, we propose a method to identify ransomware tactics that exploit this vulnerability. . We do this by using more than 3,000 samples from the latest/confirmed ransomware families to identify the actual suspicious activity that each sample proves. In this article, we present a dynamic analysis technique that can classify ransomware samples based on their targets before they attack.Collect 4,444 signatures/signatures of 3,000+ ransomware samples from 5 major families, run in a sandbox environment, and call 23 pre-attack hijacking APIs to connect to operational data..

3.2 PROPOSED SYSTEM:

The process for developing a model for predicting ransomware attacks is discussed below. Setting the conditions for the goal, such as success and independence, is the first step in the process. Then use the previous method to resolve the missing values. Then use the previous data to build the model, the data layer is split according to a 7:3 ratio, 70% of the data is used for training to let the model learn the model and the remaining 30% is used for testing, so that our performance can be evaluated. Classification models can be used to predict different types of ransomware attacks against Bitcoin.

3.3 WORKING MODEL:

One of the biggest cyber disasters is ransomware, which not only costs a lot of money, but also affects productivity, availability and reputation. Even if the result (encrypted/locked) is required, ransomware often evades detection by tracking the API request (called "paranoid" activity) while searching for the site that handles the appropriate job. In this article, we present a pioneering attempt to use this paranoid behavior to identify different ransomware behaviors. To achieve this goal, we used more than 3000 samples from new/well-known ransomware families to analyze their malicious behavior. In this paper, we propose a dynamic analysis approach to identify ransomware patterns based on pre-emergency paranoid behavior.For the site review, we examined over 3,000 ransomware samples from five major families to record their behavior/signature based on 23 attack API calls. The system aims to create a model that can predict different types of ransomware attacks. The process begins by analyzing variables such as progress and degrees of freedom seen by the system. Troubleshoot the machine first. The previous data is used to build the model, the data is split in a 7:3 ratio, 70% of the data is used for training and the remaining 30% is used for testing so that the model can learn. well project. Classification algorithms can be used to predict various ransomware attacks on Bitcoin.

3.4 HARDWARE USED:

Processor : Intel i3 , Hard disk : minimum 80 GB, RAM : minimum 2 GB.

3.5 SOFTWARE USED:

Operating System : Windows 10 or later , Tool : Anaconda with Jupyter Notebook.

4. CATALOGUE OF MODULES

- Data Pre-processing
- Data visualisation
- Algorithmic implementation
- Deployment Module.

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

Vol. 03, Issue 04, April 2023, pp : 932-938

e-ISSN : 2583-1062

Impact Factor : 5.725

4.1 MODULE 1: DATA PRE-PROCESSING:

Error values for machine learning (ML) models are derived from proven methods and are considered as close as possible to actual error values for the data. Validators will not be needed if the data values are sufficiently representative of the population. However, the use of data models may not reflect the full picture of real-world data. It's important to find additional information about values, missing values, and data types such as floating-point variables or integers. A data array used to evaluate the fit of the model to the training data when updating the model's hyperparameters. Evaluation can be biased when skills derived from evidence are added to the sample set. When this is usually done, the validation process is used to evaluate the model. Machine learning programmers use this information to check the hyperparameters of the model. In the process of using the to-do list, information about the content, quality and organization of the information is collected, analyzed and processed. Understanding your data and its components will help you decide which approach to use for modeling throughout the data analysis process.

Reason for missing data:

• The user is left-handed.

• Data loss during manual conversion from original file.

• A programming error has occurred.

Users may not provide information because they have a preconceived notion of how to use or interpret results. Univariate, bivariate, and multivariate analysis of variance After reading the provided data and loading the library to access and use it, the following data analysis methods are used: copies of data, no values in data frames, no values in boxes Customize data, values in statistical data, data types, and data clusters data analysis; rename and delete dataframes; specify the type of expenditure; add additional lines; rename and delete dataframes.

4.1.1 MODULE FLOW:

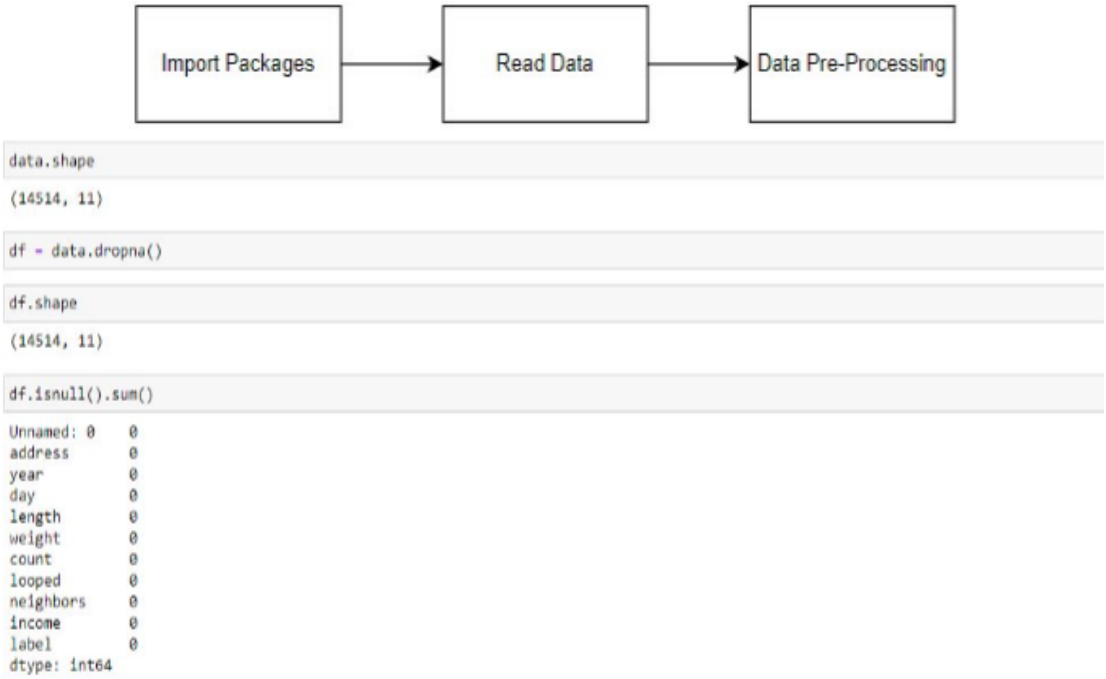
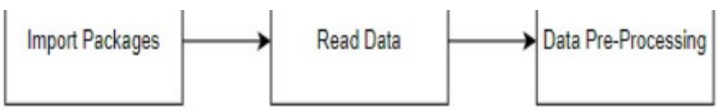


Fig 4.1 Data Pre-processing

4.2 MODULE 2: DATA VISUALIZATION:

In statistics and machine learning, the ability to visualize data is essential. In fact, the main statistical disciplines are statistical forecasting and descriptive statistics. Data visualization provides an important tool for understanding. This includes trends, misinformation, inconsistencies, etc. It is very useful when analyzing and understanding data to find. With a little understanding of the context, data visualization can be used to illustrate and explain relationships in diagrams that are more relevant to people more likely to evaluate organizations or processes.

4.2.1 MODULE FLOW



INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 04, April 2023, pp : 932-938

e-ISSN : 2583-1062

Impact Factor : 5.725

```
df['year'].hist(figsize=(10,4), color='c')
plt.xlabel('YEAR')
plt.ylabel('COUNT')
plt.title('Yearwise Count')
```

```
Text(0.5, 1.0, 'Yearwise Count')
```

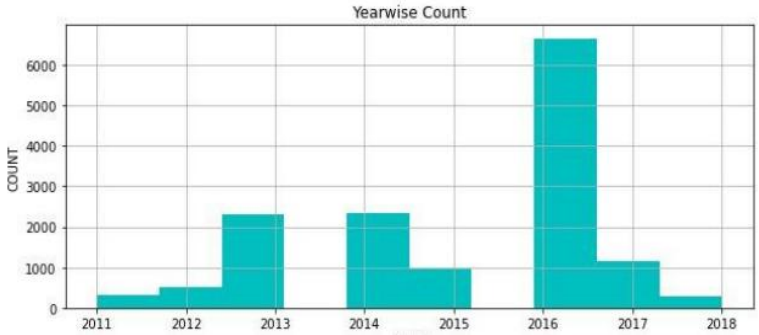


Fig 4.2 Data Visualization

4.3 MODULE 3 : ALGORITHMIC IMPLEMENTATION:

Benchmarks created using Python's scikit-learn can be used to compare the performance of different machine learning algorithms. These benchmarks can be used as a reference for your own machine learning problems or in addition to comparisons with other algorithms. The performance characteristics of each model will be different. Using resampling techniques such as parallel matching, you can calculate a true single sample using unobserved data. It should be possible to select one or two of the best models that you have created with the prediction.

It's a good idea to look at your data differently so you can see new data from different angles. The same logic applies to model selection. You have to check the accuracy of your machine learning algorithm in several ways to pick one or two for the final decision. One of the methods is to check the accuracy, variance, etc. of the distribution model. using a variety of visual aids to show

The four different algorithms listed below are compared:

- Logistic Regression
- Random Forest Classifier
- XGBoost Classifier
- Voting Classifier.

4.3.1 LOGISTIC REGRESSION:

Logistic regression is a powerful machine learning technique used to solve binary classification problems when the target is distributed. Think of logistic regression as the best form of linear regression for solving classification problems. The logistic regression described below (Tolles & Meurer, 2016) basically models binary output variables. This is the main difference between logistic regression and linear regression because the range of logistic regression is limited to values between 0 and 1. Unlike linear regression, logistic regression does not require any correlation of different concepts and products.

```
lr = LogisticRegression()
lr.fit(X_train,y_train)
predicted_lr = lr.predict(X_test)
```

Getting Accuracy

```
accuracy = accuracy_score(y_test,predicted_lr)
print('Accuracy of Logistic Regression is: ',accuracy*100)
```

```
Accuracy of Logistic Regression is:  16.670493685419057
```

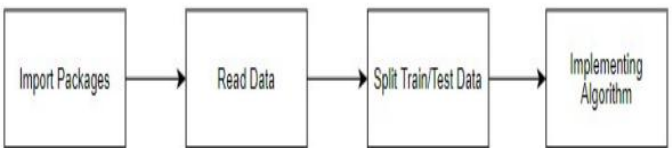


Fig 4.3 Algorithmic implementation (1)

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

Vol. 03, Issue 04, April 2023, pp : 932-938

e-ISSN : 2583-1062

Impact Factor : 5.725

4.3.2 RANDOM FOREST CLASSIFIER:

The algorithm of choice for machine learning An important part of supervised learning is random forests. It can be used in machine learning problems involving retrieval and classification. It is based on the concept of collaborative learning, which is the process of linking various disciplines to solve complex problems and improve performance standards. As the name suggests, Random Forest is a classification system that averages multiple decision trees applied to multiple combinations of given data to improve the accuracy of the data. Random forests take predictions from each tree and predict outcomes based on lots of evidence rather than relying on individual decision trees.

```
rfc = RandomForestClassifier()
rfc.fit(X_train,y_train)
predicted_rfc = rfc.predict(X_test)
```

Getting Accuracy

```
accuracy = accuracy_score(y_test,predicted_rfc)
print('Accuracy of Random Forest Classifier is: ',accuracy*100)
```

```
Accuracy of Random Forest Classifier is:  90.28702640642939
```

Fig 4.4 Algorithmic implementation (2)

4.3.3 XG BOOST CLASSIFIER:

In terms of performance and speed, XG Boost classifiers generally outperform other algorithms designed for supervised learning. Since the library is parallelizable, the main algorithm can be executed on a group of GPUs or even on a computer network. This makes it possible to train high-performance machine learning problems using hundreds of millions of examples.

```
xg = XGBClassifier()
xg.fit(X_train,y_train)
predicted_xg = xg.predict(X_test)
```

Getting Accuracy

```
accuracy = accuracy_score(y_test,predicted_xg)
print('Accuracy of XGBoost Classifier is: ',accuracy*100)
```

```
Accuracy of XGBoost Classifier is:  91.34328358208955
```

Fig 4.5 Algorithmic implementation (3)

4.3.4 VOTING CLASSIFIER:

A machine learning model, called the voter, learned many examples and predicted classes (outputs) based on which classes were most likely to be outputs. The product's rating based on the highest number of votes is estimated by averaging the scores for each provider's shipping options. Instead of building unique models one by one and evaluating their accuracy, create a model that learns from multiple models and predicts results as they are voted on by each team of developers.

```
xg = XGBClassifier()
rf = RandomForestClassifier()
lr = LogisticRegression()
```

```
vc = VotingClassifier(estimators=[('XGBoost', xg), ('RandomForestClassifier', rf), ('LogisticRegression', lr)], voting='hard')
```

```
vc.fit(X_train,y_train)
pred_vc = vc.predict(X_test)
```

Getting Accuracy

```
accuracy = accuracy_score(y_test,pred_vc)
print('Accuracy of Voting Classifier is: ',accuracy*100)
```

```
Accuracy of Voting Classifier is:  91.34328358208955
```

Fig 4.6 Algorithmic implementation (4)

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

Vol. 03, Issue 04, April 2023, pp : 932-938

e-ISSN : 2583-1062

Impact Factor : 5.725

4.4DEPLOYMENTMODULE:

The delivery module allows Bitcoin traders to assess the risk associated with a transaction beforehand. Flask, a highly capable microframework, serves as the web server for this module. Despite being relatively new, Flask boasts a robust community, advanced API, and outstanding extensions. Flask provides fast WSGI features, comprehensive documentation, as well as web and library-level unit testing. For future projects that require additional features and extensions, Flask is definitely worth considering..

5. MODELING AND ANALYSIS

5.1 WORKFLOW DIAGRAM

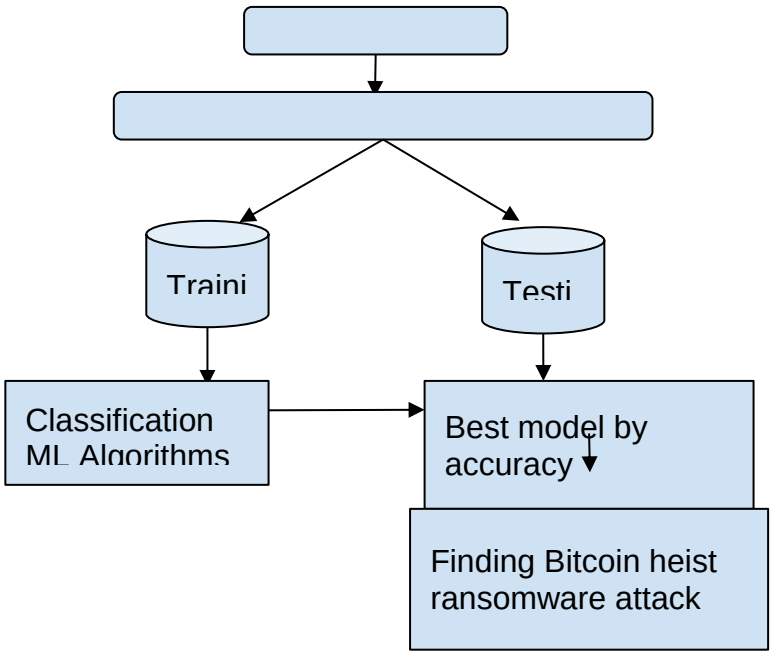


Fig 5.1 Workflow Diagram

5.2 SYSTEM ARCHITECTURE:

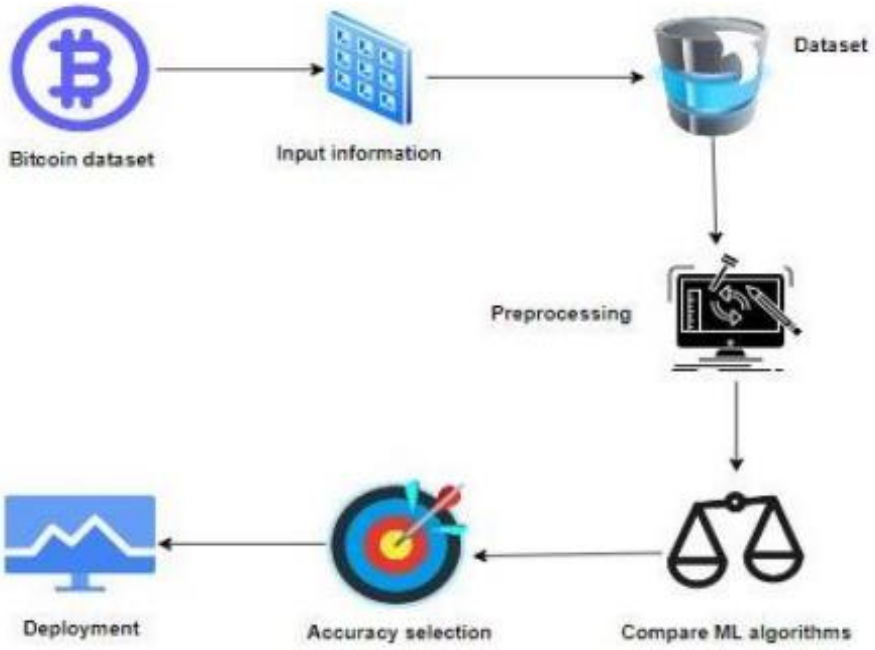


Fig 5.2 System Architecture

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

Vol. 03, Issue 04, April 2023, pp : 932-938

e-ISSN : 2583-1062

Impact Factor : 5.725

6. 6.RESULT AND DISCUSSION

Fig 6.1 Website

7. CONCLUSION

It initiates the cost analysis, research, design and model evaluation review process. The most accurate algorithms will be found in the open process. A program that helps you find a bitcoin heist with the first program created.

8. REFERENCES

[1] Analyzing Ransomware Family Using Qualitative and Dynamic Analysis, K. P. Subedi, D. R. Budhathoki, and D.Dasgupta, Proc. IEEE Security and Privacy Workshop (SPW), 2018, p. 180-185.

[2] Vulnerability and Network Analysis for Ransomware Detection and Classification, by R. Vinayakumar, K. P. Soman,K. K.S. Velan and S. Ganorkar, Proc. IEEE International Meeting.259-265: I.

[3] B. Zhang, W. Xiao, X. Xiao, A. K. Sangaiah, W. Zhang, and J. Zhang, "Ransomware classification using patch-based CNN and self-attention network on embedded N-grams of opcodes," Future Gener. Comput. Syst., vol. 110, pp. 708– 720, Sep. 2020.

[4] A. AlSabeh, H. Safa, E. Bou-Harb, and J. Crichigno, "Exploiting ransomware paranoia for execution prevention," in Proc. IEEE Int. Conf. Commun. (ICC), 2020, pp. 1–6.

[5] J. Yan, G. Yan, and D. Jin, "Classifying malware represented as control flow graphs using deep graph convolutional neural network," in Proc. 49th Annu. IEEE/IFIP Int. Conf. Depend. Syst. Netw. (DSN), 2019, pp. 52–63

[6] H. Daku, P. Zavarsky, and Y. Malik, "Behavioral-based classification and identification of ransomware variants using machine learning," in Proc. 17th IEEE Int. Conf. Trust Security Privacy Comput. Commun. 12th IEEE Int. Conf. Big Data Sci. Eng. (TrustCom/BigDataSE), 2018, pp. 1560– 1564.

[7] H. Zhang, X. Xiao, F. Mercaldo, S. Ni, F. Martinelli, and A. K. Sangaiah, "Classification of ransomware families with machine learning based on N-gram of opcodes," Future Gener. Comput. Syst., vol. 90, pp. 211–221, Jan. 2019.

[8] L. Onwuzurike et al., "MaMaDroid: Detecting Android malware by building Markov Chains of behavioral models (extended version)," ACM Trans. Privacy Security, vol. 22, no. 2, pp. 1–34, 2019.

[9] E. Berrueta, D. Morato, E. Magaña, and M. Izal, "A survey on detection techniques for cryptographic ransomware," IEEE Access, vol. 7, pp. 144925–144944, 2019.

[10] Z. Cohen and G. Sands. Four Key Takeaways on the U.S. Government Response to the Pipeline Ransomware Attack. Accessed: May 2021.