BITCOIN HEIST RANSOMWARE ATTACK PREDICTION USING DATA SCIENCE PROCESS

Harshavarthini R Dr T Jacklin

*Department of Computer Science and Engineering Associate Professor*

*Panimalar Engineering College Department of Computer Science and Engineering*

Chennai *Panimalar Engineering College*

[harshavarthini0412@gmail.com](mailto:harshavarthini0412@gmail.com) Chennai

[karthijackulin@gmail.com](mailto:karthijackulin@gmail.com)

Isabella Mary J Muthupriya M

*Department of Computer Science and Engineering Department of Computer Science and Engineering Panimalar Engineering College Panimalar Engineering College*

Chennai Chennai

[isainfanta27@gmail.com](mailto:isainfanta27@gmail.com) [muthupriya138@gmail.com](mailto:muthupriya138@gmail.com)

**ABSTRACT--Ransomware assaults rank among the most disruptive cyber dangers, with a negative impact on reputation, productivity, and accessibility in addition to incurring large monetary losses. Despite their intended use (locking or encrypting), ransomware are frequently created to execute a sequence of "paranoia" activities, or pre-attack API calls, to find a suitable execution environment while avoiding detection. In this paper, we describe a ground breaking attempt to use such paranoia actions for identifying distinguishing ransomware tendencies. To accomplish this, In our project more than 14K samples from recent/notable ransomware families to identify the specific paranoia-inspiring activities that each sample represents. It classifies ransomware using several machine and deep learning algorithms while modelling ransomware-generated evasion API queries using techniques rooted in Natural Language Processing (NLP), such as Occurrence of Words (OoW).**

***Keywords:Ransomware analysis, Machine learning, Phishing, Cyber-attacks, Paranoia***

**I. INTRODUCTION**

Ransomware is extortion-style software that is intended to lock or encrypt user data on targeted devices. Researchers have suggested many methods to identify utilising several static/dynamic analysis techniques to comprehend malware's behaviour and avoid ransomware attacks code architecture, the things it does throughout its life cycle, and the things it does after an infection. To combat this threat, researchers used a variety of ML and DL techniques. Despite prior efforts, protecting against the growing number of ransomware assaults is still seen to be a difficult undertaking due to ignorance of newly discovered malware and the continually changing families and variants .In fact, the prevalence and severity of these attacks show that there aren't any reliable ways to identify ransomware. In order to overcome these difficulties, it is necessary to research efficient methods for identifying and

thwarting ransomware attacks by examining their behavioural traits and attributing them to known bad whenever possible, families/variants.The examination of recently discovered ransomware families revealed that they are programmed to elude detection by carrying out a series of tasks and actions intended to sense the execution environment before executing the malicious payload.

Among the malicious usage of cryptocurrencies, ransomware payments continue to attract an ever increasing attention. Although encrypting files and resources for ransom has a long history, receiving ransom payments securely had never been simple until the emergence of Bitcoin. Proliferation of cryptocurrencies (e.g., Bitcoin) that allow pseudo-anonymous transactions, has made it easier for ran somware developers to demand ransom by encrypting sensitive user data. The recently revealed strikes of ransomware attacks have already resulted in significant economic losses and societal harm across different sectors, ranging from local governments to health care. Most modern ransomware use Bitcoin for payments. However, although Bitcoin transactions are permanently recorded and publicly available, current approaches for detecting ransomware depend only on a couple of heuristics and/or tedious information gathering steps (e.g., running ransomware to collect ransomware related Bitcoin addresses). Bitcoin transactions can be created anonymously, and participation in the network does not require identity verification. A payment can be requested by delivering a public Bitcoin address (i.e., a short string) to a sender by using anonymity networks.

The remaining work is arranged as follows: Section II, introduce related works in the context of ransomware family classification. The proposed methodology is detailed in Section III, including data collection, visualization, feature extraction/selection, classifier implementation / evaluation, and limitations of the proposed

methodology. Section IV describes the conducted experiments and the results obtained from each classifier, discussing them simultaneously. This Section presents a comparison between the classifiers’ attained outcomes. Finally, the future work and all the work is concluded in Section V.

**II. LITERATURE REVIEW**

For the purpose of ransomware detection and classification, several studies have used static malware analysis techniques to extract features that can be combined with ML/DL techniques. For example, Ricardo Misael [1] carried out one of the most troublesome forms of cybercrime is ransomware, which affects productivity, accessibility, and reputation in addition to costing a lot of money. Despite having encryption or locking as one of its final goals, ransomware is frequently built to avoid detection by making a sequence of pre-attack API requests, or "paranoia" activities, to find a suitable execution environment. In this paper, a novel approach to using such paranoid behaviours to characterise different ransomware behaviours is provided. By calculating the Term Frequency-Inverse document of the N-grams that are changed from ransomware opcodes, extensive experiments on real-world datasets are performed. They conducted analysis utilising several N-gram feature dimensions and assessed their strategy using five various machine learning methods. Similar to this, the authors used retrieved N-grams features from opcodes in their classification models. While implementing numerous ML classification models, they also tested their framework using DL models such the Self-Attention Convolutional Neural Network (SA-CNN), which performed well for lengthy opcode sequences. Suleiman Ali [2] carries out one of the most hazardous linked crimes in the cryptocurrency business is ransomware attacks. Early ransomware detection appears to be important in order to raise the difficulty of defending against the attack. The suggested method uses three supervised machine learning techniques—logistic regression (LR), random forest (RF), and extreme gradient boosting—to identify distinctive patterns in Bitcoin payment transactions (XGBoost). In terms of classification accuracy and other evaluation metrics including confusion matrix, recall, and F1-score, these M L-based predictive models on the Bitcoin Heist ransomware dataset. This article of Hernandes Castro[3] presents an economic study of ransomware, a recent type of cyber-enabled extortion. The methods used by the criminals, such as exact pricing and price discrimination, will determine the illicit gains they make. Furthermore provided are the findings of a pilot poll that serve as proof of concept for calculating the costs of ransomware attacks. The many tactics that were examined have already been seen in malware that is

now in use, as well as its propensity to be used in the future. This research will give some helpful insights into how ransomware might change in the future. Hardi Shababh concludes that [4] an instance of cybercrime and a category of virus known as ransomware encrypts files, stops users from accessing their data or systems, and requests money in exchange for the decryption and restoration of access to the users' files. It is challenging to classify

ransomware data using current data mining and ML techniques because predictions aren't always accurate. This seeks to develop two models that successfully handle these issues and can precisely detect and categorise Ransomware assaults, then

compare the models' performance. The Rule based approaches in detecting data is advantageous because the algorithms efficiently classify non-linear datasets.Article of Abdullah [5] shows that attacks using ransomware have been one of the largest hazards to a variety of users worldwide, especially vital cyber-physical systems. The evasive techniques these assaults use to evade detection frequently render the existing fixes useless. This survey is focused on examining and assessing the most recent advancements in ransomware attack detection in order to support the research community's efforts to stop this extremely serious and growing ransomware problem. The focus is on crypto ransomware because it is the most common, harmful, and difficult type.

Sabira [6] came to the conclusion that crypto-currency transactions using ransomware are common. The majority of current ransomware accepts Bitcoin as payment. The researchers have used sophisticated data analytics techniques to automatically identify ransomware-related transactions and malicious Bitcoin addresses based on the collected ransomware-related Bitcoin addresses. In order to recognise and report assaults, machine learning methods are assessed based on the patterns that distinguish such cybercrime activities from regular bitcoin transactions. By implementing certain early preventions and tactics to stop these kinds of harmful transactions, this endeavour would be the ideal opportunity to coordinate and expand research efforts in the future. Several ransomware prediction algorithms have been put forth, but more effective ones are needed for restricted heterogeneous IoT systems, according to [7] PV Lakshmi. Resource-constrained IoT systems utilise fewer resources when attack information are used. The context ontology presented in this research is used to extract information features (such as connection requests and software upgrades) and Artificial Intelligence and Machine Learning techniques are used to forecast ransomware. The focus and foundation of the suggested solutions is on the early prediction and penetration attempts into IoT devices with limited resources. Harrack came to the conclusion that ransomware assaults [8] are on the rise attackers are stealing important data from

various crucial infrastructures, and organisations are being forced to pay ransoms to have their files decrypted. To identify a classifier label among those that have been classified as ransomware or connected to harmful activity, the various transactional aspects are studied. To teach the computer to identify particular transactions and determine whether they were malicious or benign, machine learning techniques are used. In order to create a random forest classifier, this work uses ensemble learning and decision tree classifiers. Kirat Jadhav came to the conclusion [9] that cryptocurrency trading in the online world had undergone a revolution. Thousands of cryptocurrencies have been introduced in the roughly ten years since the first block of bitcoin was introduced. Cybercriminals were drawn to cryptocurrencies because of the anonymity they offered. In order to effectively identify ransomware payments made to the operators via bitcoin transactions, this research looks at various machine learning techniques. To recognise and report attacks, ML models may be created based on features that distinguish such cybercrime activities from typical bitcoin transactions. The dataset for bitcoin ransomware is used to evaluate the machine learning techniques. According to experimental findings, Gradient Boosting and XGBoost algorithms outperformed k-Nearest Neighbor, Random Forest, Naïve Bayes and Multilayer Perceptron approaches in terms of detection rate, precision, recall, and F-measure rates. According to Jason E, there are several possible target permutations, messages, and value propositions, making phishing a challenging issue to solve. Spear phishing [10] is linked to social engineering, which can be challenging for even knowledgeable or educated personnel to recognise. Because of this, users are the crucial point of entry for criminals looking to commit cybercrimes like identity theft and ransomware propagation, which result in annual losses of billions of dollars. Researchers are looking into a variety of solutions to this issue, including educating users and making them aware of the consequences of falling for phishing. The goal of this study[10] was to speak with security experts to learn more about how to prevent users and staff from falling victim to phishing. Attackers are aware of all other digital resources are crucial to the ongoing operation and expansion of any firm, according to Navneeth Kaur Popli. And because the business values these digital assets so highly, holding them all to ransom is the greatest and fastest method to make a lot of money. The largest threat to businesses is ransomware, which has the power to halt operations and result in significant economic losses. In this essay, the behaviours of the most recent ransomwares that have targeted both organisations and individuals are studied analytically. This is accomplished by analysing their assault process, file system analysis, persistence analysis, and network-level analysis while executing them in a simulated environment. This behaviour analysis step [11] is completed using

tools like Cuckoo. Prediction of upcoming ransomware kinds that can be easily made utilising readily available toolkits like ADMMutate, Clet, and Phatbot follows. Considering the impact and threat they could pose as well as how challenging it would be to catch them once they deploy all of the aforementioned stealth techniques.

**III. PROPOSED METHODOLOGY**

The Bitcoin Heist Ransomware Attack Prediction includes six different stages in order to predict the type of attacks. Generally, the first step will be collecting input from the user who wants to know what attack it is. Parallely, the network will have several collection of features stored in dataset for prediction. The data will be preprocessed by removing all the cleaning the raw data and make it as a valuable dataset. Then, the next stage will be data visualization which is done to have a better glance of vast amount of data. It can be either in graphs or in charts. Now the dataset provided by the user will be compared with the four machine learning algorithms such as Logistic Regression, Random Forest, XG Boost and Voting Classifier. After the comparison is done, the system will predict the best algorithm which gives efficient accuracy and builds the model. Then the deployment will be carried out.

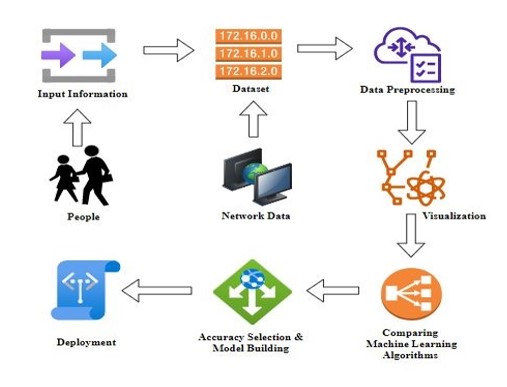


Fig 1.System Architecture

The goal of the system under consideration is to create a ransomware attack prediction model. The first step in the procedure is the identification of variables, such as dependent and independent variables, where we locate the target column. To deal with missing values, pre-processing procedures are then used. The pre-processed data is then utilised to create a model by splitting the dataset into 7:3 ratios, with 70% of the data being used for training purposes so that the model can learn the pattern and the remaining 30% being used for testing purposes so that our project can be evaluated for performance. The classification model can be employed to forecast the various ransomware attack types that target bitcoin.

Benefits:

* We are focusing our implementation on bitcoin ransomware assaults.
* We are putting the voting classifier into practise.
* Deployment can be done with necessary inputs and designed for the end users.
* The dataset involves more than 14k sample of features which could give the best accuracy for attack prediction.

*A. Data Collection and Preprocessing*

The error rate of the machine learning (ML) model is obtained using validation techniques, and is thought to be as close to the actual error rate of the dataset as possible. You might not require the validation techniques if the volume of the data is sufficient to be representative of the population. However, working with data samples that might not be a true representative of the population of a given dataset in real-world scenarios. Finding duplicate values, missing values, and information about the data type—whether a float variable or an integer—are all necessary. the subset of data used to assess a models fit to a training dataset while adjusting model hyperparameters. As skill from the validation dataset is incorporated into the model configuration, the evaluation becomes more skewed. A given model is evaluated using the validation set, but this is done frequently. This information is used by machine learning engineers to adjust the model hyperparameters. A time-consuming to-do list can result from the collection, analysis, and process of dealing with data content, quality, and structure. Understanding your data and its characteristics will help you choose which algorithm to use to construct your model during the data identification process.

Some of these sources contain merely careless errors. Sometimes there may be a more significant cause for missing data. Its critical from a statistical perspective to comprehend these various missing data types. The kind of missing data will affect how it is handled in terms of filling in the blanks, identifying missing values, basic imputation, and a thorough statistical approach. Before writing any code, its crucial to comprehend where the missing data is coming from. Here are a few typical explanations for missing data:

* User forgot to fill in a field.
* Data was lost while transferring manually from a legacy database.
* There was a programming error.
* Users chose not to fill out a field tied to their beliefs about how the results would be used or interpreted.

Variable identification with Uni-variate, Bi-variate and Multi-variate analysis:

* Import libraries for access and functional purpose and read the given dataset
* General Properties of Analyzing the given dataset
* Display the given dataset in the form of data frame
* Show columns
* Shape of the data frame
* To describe the data frame Checking data type and information about dataset
* Checking for duplicate data
* Checking Missing values of data frame
* Checking unique values of data frame
* Checking count values of data frame
* Rename and drop the given data frame
* To specify the type of values
* To create extra columns

*B. Data visualization*

Data visualisation is a key skill in applied statistics in ML. In actuality, the primary emphasis of statistics is on quantitative estimates and data descriptions. Data visualisation offers a crucial set of tools for gaining a qualitative understanding. This is useful when exploring and learning about a dataset .It will suggest looking more closely at some of the referenced books, as well as the fields of data visualisation and exploratory.

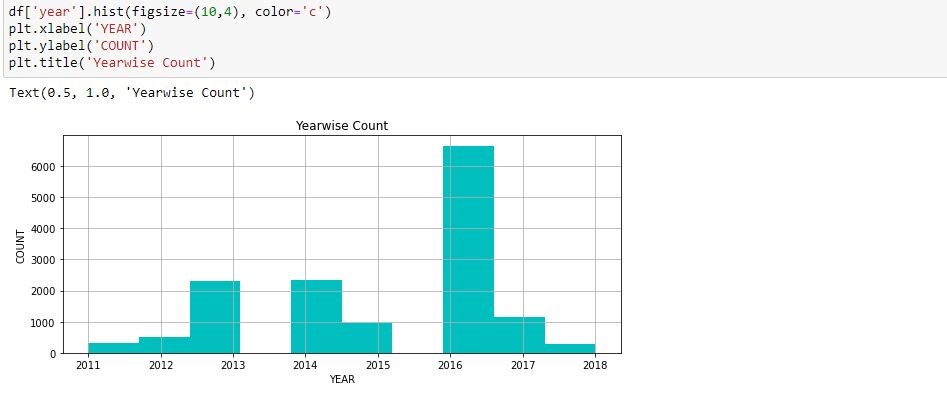


Fig 2.Visualization through bar chart

Data may not always make sense until it is presented visually. It's crucial to have the ability to visualise data samples and other objects quickly



Fig 3.Heatmap Plot Diagram

It will demonstrate how to use different plot types to comprehend your own data as well as the

different plot types need to be familiar with when visualising data in Python. It will

demonstrate how to visualise categorical data using bar charts and time series data using line

plots.

**ALGORITHM IMPLEMENTATION**

*C. XG Boost classifier*

It consistently outperforms all other algorithms designed for supervised learning tasks due to its unmatched speed and performance. The main algorithm can run on clusters of GPUs or even across a network of computers because the library is parallelizable. A machine learning algorithm called XG Boost classifier is used for structured and tabular data. A gradient boosted decision tree implementation made for speed and performance is called the XG Boost classifier. An ensemble modelling method that works with huge, complex datasets is XG Boost. XG Boost is a distributed gradient boosting library that has been developed to be very effective, adaptable, and portable.

This enables the high-performance training of ML tasks using hundreds of millions of training examples. Figure 6 represents the XG Boost Classifier.

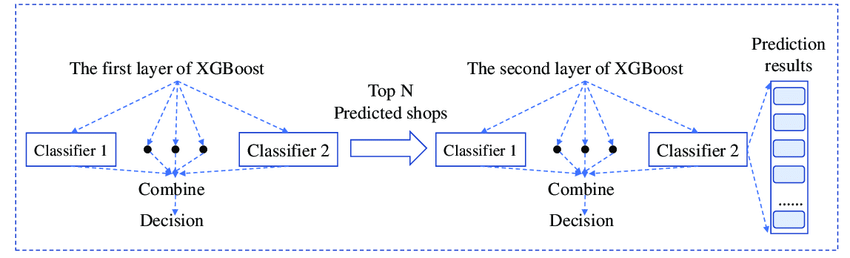


Fig 4. Diagrammatic representation of XG Boost Classifier Algorithm

*D. Random Forest Classifier*

The supervised learning method includes the well-known machine learning algorithm Random Forest. It can be applied to ML Classification and Regression issues. Its foundation is the idea of ensemble learning, which is the process of combining various classifiers to solve a challenging problem and enhance the performance of the model.

Random Forest, as the name implies, is a classifier that uses a number of decision trees on different subsets of the given dataset and averages them to increase the dataset's predictive accuracy. Instead of relying on a single decision tree, the random forest uses predictions from each tree and predicts the result based on the votes of the majority of predictions.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.



Fig 5. Diagrammatic Representation of Random Forest Classifier Algorithm

*E. Logistic Regression*

For binary classification problems, another potent supervised machine learning algorithm is logistic regression (when target is categorical). The best way to think of logistic regression is to consider it to be a linear regression applied to classification issues. In order to simulate a binary output variable, logistic regression essentially uses the logistic function defined below (Tolles & Meurer, 2016). The range of logistic regression is constrained to the range of 0 and 1, which is the main distinction between it and linear regression. Furthermore, logistic regression does not demand a linear relationship between the variables that make up the inputs and outputs, in contrast to linear regression.

Based on a set of independent factors and a dataset of independent variables, the machine learning classification technique known as logistic regression calculates the likelihood that an event will occur, such as voting or not voting.Finding the best-fitting model to explain the link between a set of independent (predictor or explanatory) factors and a dichotomous feature of interest (dependent variable = response or outcome variable) is the aim of logistic regression.

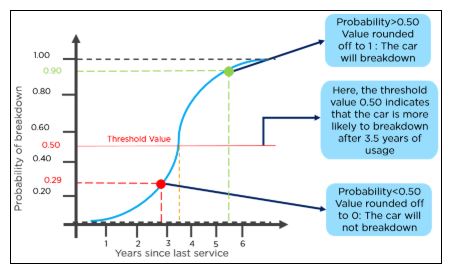


Fig 6. Diagrammatic Representation of Logistic Regression Algorithm

*F. Voting Classifier*

A voting classifier is a machine learning model that gains experience by training on a collection of various models and forecasts an output (class) based on the class with the highest likelihood of being the output. To predict the output class based on the highest majority of voting, it merely aggregates the results of each classifier that was passed into the voting classifier. The concept is to build a single model that learns from these models and predicts output based on their combined majority of voting for each output class, rather than building separate dedicated models and determining the accuracy for each of them.

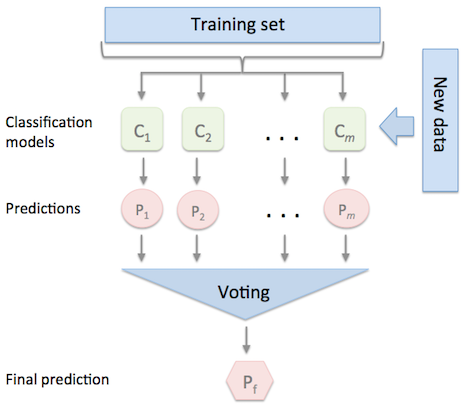


Fig 7. Diagrammatic Representation of Voting Classifier Algorithm

*G. Model Evaluation and Comparison*

We employ a number of conventional techniques to assess the overall performance of the applied classification models and contrast the results. We especially employ metrics like F-measure, recall, accuracy, and precision. Additionally, we employ the confusion matrix, a useful tool for debating the efficacy of the applied ML/DL models. In this matrix true positive and true negative indicate the number of samples that the model correctly classified, while false positive and false negative indicate the number of samples that were misclassified. Also as a measure of each model’s computing performance, we compute its speed.

*H. Limitations*

We faced (and had to work around) a number of data gathering and processing restrictions in our work. First even though we used a number of publicly accessible threat repositories(such as VirusShare and VirusTotal) to gather a large number of authentic ransomware instances from various families, the dataset we eventually obtained was imbalanced because each family had a different number of

ransomware samples. This is why we had used undersampling to build a balanced dataset by

1.Focusing on the ransomware families with the most recent attack occurrences and,

2.Randomly choosing a subset of malicious executables from the selected families.

**IV. EXPERIMENTAL RESULTS**

As described in Section III, we focus on 6 main ransomware families for our experimental analysis. Furthermore, as noted, given the imbalance dataset in terms of the sample distribution per family

This dataset contains 14514 records of features, which were then classified into 6 classes.

They are Cerber, Crypto, Crypto Locker, Crypto Wall, Locky and White.

Data pre-processing is performed to transform the raw data into a useful and efficient format by removing noise and inconsistent data.

So, it creates the reliable consistent data that improves the efficiency of the training data for analysis and also enables accurate decision – making.

1.Describing the data

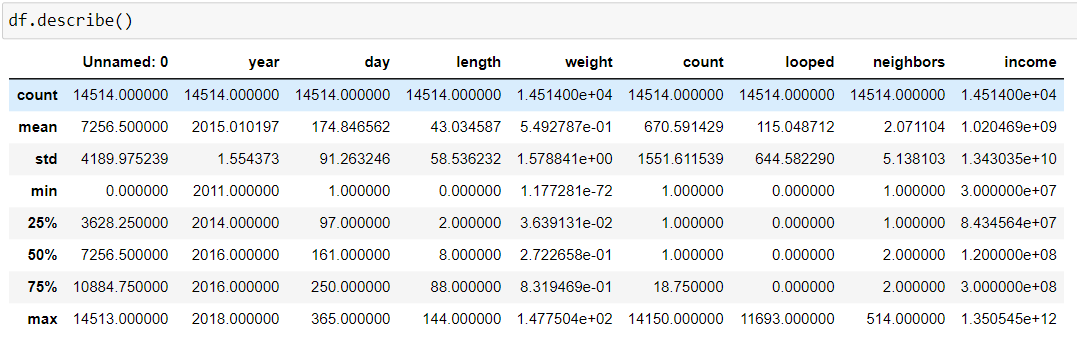


Fig 8. Description of data

2.Checking the unique length values

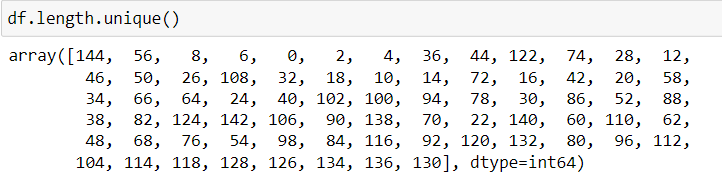


Fig 9. Unique length values

3.Identifying the label in accordance with year

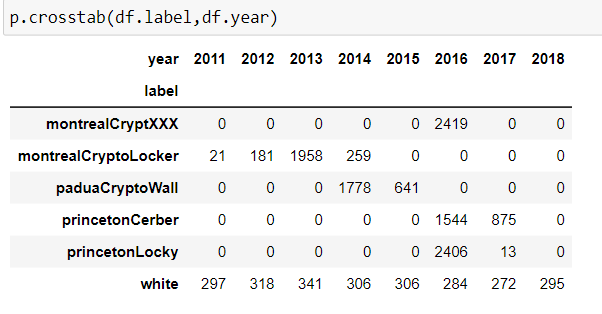
**

Fig 10. Attacks happened from 2011 to 2018

4.Identifying the label count and frequencies

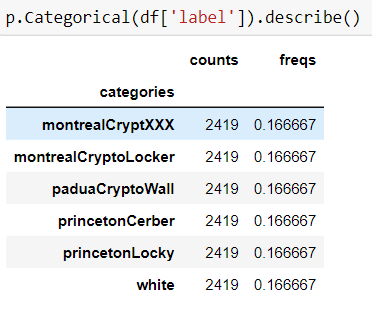


Fig 11. Counts and Frequencies of attacks

5.Identifying the duplicate data

6.Processing the completed series

Visualization plays a major role in organizing a vast amount of data into visible format. We have visualized it based on occurrence of attacks, based on year and the density plot for yearly attacks.

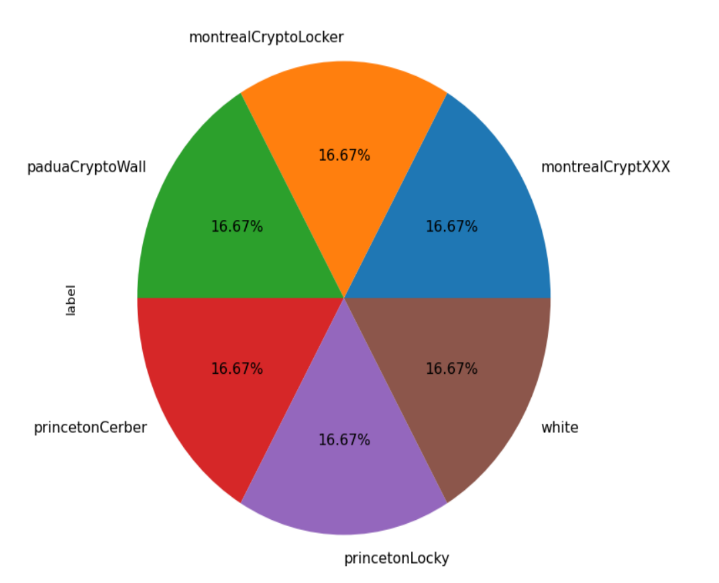


Fig 12.Pie Chart based on label

Several attacks were happened based on the years from 2011 to 2018

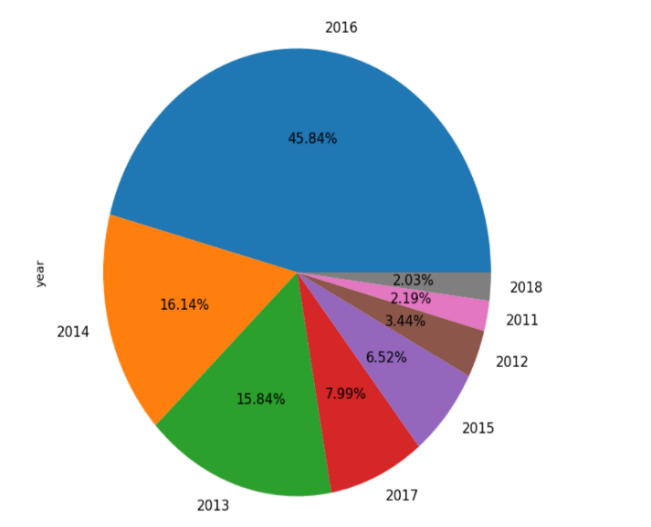


Fig 13.Pie Chart based on year

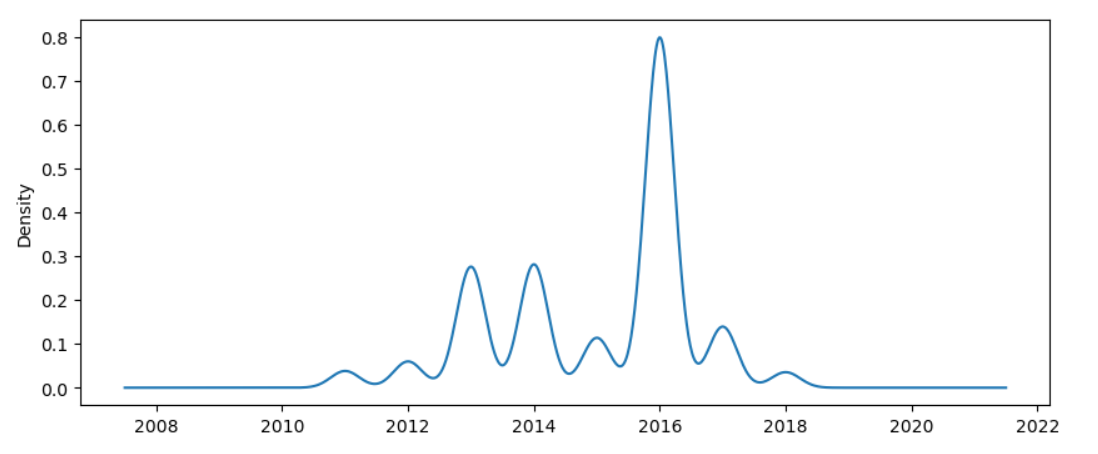


Fig 14.Density plot for yearly attacks

**PERFORMANCE METRICS**

The below tables will show the performance metrics of algorithms.

TABLE I

**Performance Metrics of XG Boost Algorithm**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1 score | Support |
| 0 | 0.90 | 0.95 | 0.93 | 726 |
| 1 | 0.97 | 0.97 | 0.97 | 726 |
| 2 | 0.86 | 0.92 | 0.89 | 725 |
| 3 | 0.91 | 0.98 | 0.94 | 726 |
| 4 | 0.97 | 0.97 | 0.97 | 726 |
| 5 | 0.85 | 0.69 | 0.76 | 726 |
| Accuracy |  |  | 0.91 | 4355 |
| Macro average | 0.91 | 0.91 | 0.91 | 4355 |
| Weighted average | 0.91 | 0.91 | 0.91 | 4355 |

TABLE II

**Performance Analysis of Random Forest Algorithm**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1 score | Support |
| 0 | 0.89 | 0.95 | 0.92 | 726 |
| 1 | 0.96 | 0.96 | 0.96 | 726 |
| 2 | 0.87 | 0.91 | 0.89 | 725 |
| 3 | 0.89 | 0.97 | 0.93 | 725 |
| 4 | 0.96 | 0.97 | 0.96 | 726 |
| 5 | 0.96 | 0.67 | 0.75 | 726 |
| Accuracy |  |  | 0.91 | 4355 |
| Macro average | 0.90 |  | 0.90 | 4355 |
| Weighted average | 0.90 |  | 0.90 | 4355 |

TABLE III

**Performance Analysis of Logistic Regression Algorithm**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1 score | Support |
| 0 | 0.00 | 0.00 | 0.00 | 726 |
| 1 | 0.00 | 0.00 | 0.00 | 726 |
| 2 | 0.00 | 0.00 | 0.00 | 725 |
| 3 | 0.00 | 0.00 | 0.00 | 726 |
| 4 | 0.00 | 0.00 | 0.00 | 726 |
| 5 | 0.17 | 1.00 | 0.29 | 726 |
| Accuracy |  |  | 0.17 | 4355 |
| Macro average | 0.03 | 0.17 | 0.05 | 4355 |
| Weighted average | 0.03 | 0.17 | 0.05 | 4355 |

TABLE IV

**Performance Analysis of Voting Classifier Algorithm**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1 score | Support |
| 0 | 0.89 | 0.95 | 0.92 | 726 |
| 1 | 0.97 | 0.96 | 0.97 | 726 |
| 2 | 0.88 | 0.90 | 0.89 | 725 |
| 3 | 0.92 | 0.96 | 0.94 | 726 |
| 4 | 0.98 | 0.96 | 0.97 | 726 |
| 5 | 0.83 | 0.74 | 0.78 | 726 |
| Accuracy |  |  | 0.91 | 4355 |
| Macro average | 0.91 | 0.91 | 0.91 | 4355 |
| Weighted average | 0.91 | 0.91 | 0.91 | 4355 |

*Evaluation and Comparison*

In the existing work, 3,432 samples had been drawn and the evaluations demonstrated the effectiveness of the implemented approach, with the Random Forest (RF) techniques producing an optimal classification accuracy (94.92%).

TABLE V

**Accuracy of Existing Work**

|  |  |
| --- | --- |
| Model | Accuracy |
| Random Forest Classifier | 94.92% |
| Bernoulli Naïve Bayes | 85.11% |
| k-Nearest Neighbour | 92.85% |
| Artificial Neural Network | 87.80% |

In the proposed work, 14,514 have been drawn and the evaluations demonstrated the effectiveness of the implemented approach, with the Extreme Gradient (XG) Boost producing an optimal classification accuracy (91.32%).

TABLE VI

**Accuracy of Proposed Work**

|  |  |
| --- | --- |
| Model | Accuracy |
| XG Boost Classifier | 16.67% |
| Random Forest Classifier | 90.53% |
| Logistic Regression | 91.34% |
| Voting Classifier | 91.32% |

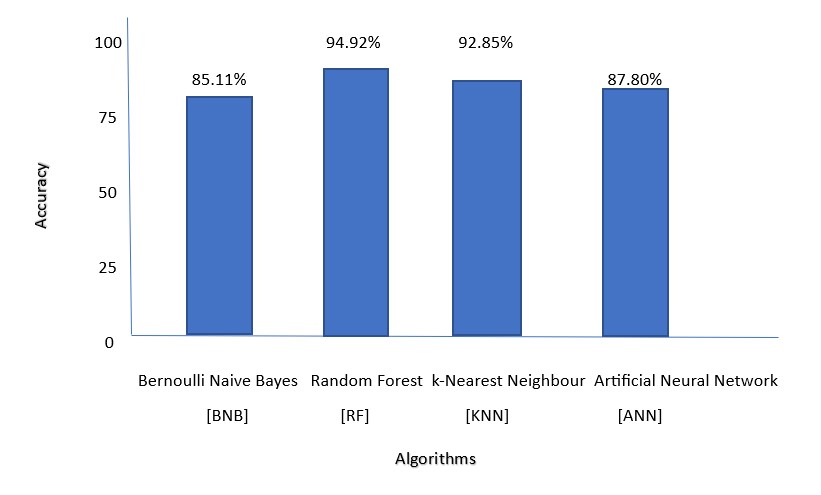


Fig 15. Accuracy of Existing Work With 3k Samples

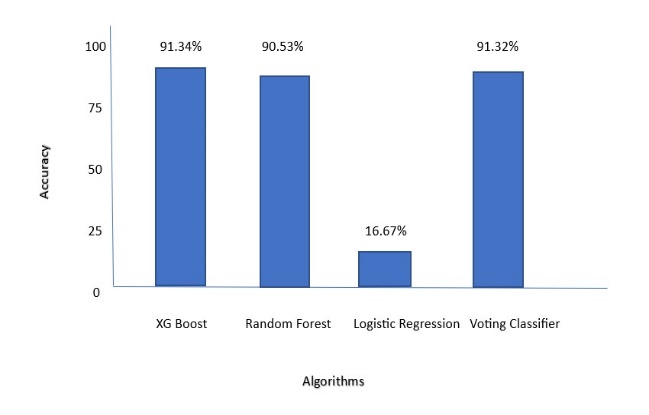


Fig 16.Accuracy of Proposed Work With 14k Samples

**Confusion Matrix**

Usually Confusion matrix is matrix that gives us the summarization of performance of the models.

It has used six classes such as Cerber, Crypto, Crypto Locker, Crypto Wall, Locky and White.

**Confusion Matrices Outcome for the Algorithms**

**[**Ce-Cerber ,Cr-Crypto ,Cr.L-Crypto Locker, Cr.W-Crypto Wall, Lo-Locker, W-White]

TABLE VII

**XG Boost Classifier**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Ce | Cr | Cr.L | Cr.W | Lo | W |
| Ce | 690 | 12 | 0 | 0 | 8 | 16 |
| Cr | 17 | 707 | 0 | 0 | 0 | 2 |
| Cr.L | 0 | 0 | 664 | 12 | 0 | 49 |
| Cr.W | 0 | 0 | 0 | 710 | 0 | 16 |
| Lo | 14 | 0 | 0 | 0 | 706 | 6 |
| W | 43 | 8 | 107 | 56 | 11 | 501 |

TABLE VIII

**Random Forest Classifier**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Ce | Cr | Cr.L | Cr.W | Lo | W |
| Ce | 691 | 15 | 0 | 0 | 9 | 11 |
| Cr | 22 | 696 | 0 | 0 | 6 | 2 |
| Cr.L | 0 | 0 | 663 | 15 | 0 | 47 |
| Cr.W | 0 | 0 | 1 | 705 | 0 | 20 |
| Lo | 18 | 3 | 0 | 0 | 703 | 2 |
| W | 45 | 8 | 97 | 73 | 18 | 485 |

TABLE IX

**Logistic Regression**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Ce | Cr | Cr.L | Cr.W | Lo | W |
| Ce | 0 | 0 | 0 | 0 | 0 | 726 |
| Cr | 0 | 0 | 0 | 0 | 0 | 726 |
| Cr.L | 0 | 0 | 0 | 0 | 0 | 725 |
| Cr.W | 0 | 0 | 0 | 0 | 0 | 726 |
| Lo | 0 | 0 | 0 | 0 | 0 | 726 |
| W | 0 | 0 | 0 | 0 | 0 | 726 |

TABLE X

**Voting Classifier**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Ce | Cr | Cr.L | Cr.W | Lo | W |
| Ce | 692 | 11 | 0 | 0 | 3 | 20 |
| Cr | 26 | 697 | 0 | 0 | 0 | 3 |
| Cr.L | 0 | 0 | 655 | 12 | 0 | 58 |
| Cr.W | 0 | 0 | 1 | 699 | 0 | 26 |
| Lo | 19 | 3 | 0 | 0 | 698 | 6 |
| W | 39 | 7 | 88 | 47 | 9 | 536 |

**V. CONCLUSION**

In this work we proposed an analysis for attributing ransomware samples based on their behavioral characteristic following their activities. The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set is higher accuracy score will be find out. This application can help to find the Prediction of Crypto Ransomware Attack. This way can help the organizations to know more about the attacks and also will be able to identify the attacks from spam, phishing emails, malicious and hacked websites.The future works can be implemented on Crypto Ransomware attack prediction which can be connected with the cloud model.

The optimization if this work can also be better if the system and the prediction model is connected with the embedded system.

**REFERENCES**

**[1]** Ricardo Misael Ayala Molina, Sadegh Torabi, Khaled Sarieddine, Elias Bou-harb, Nizar Bouguila, Chadi Assi. (2022**) ‘On Ransomware Family Attribution Using Pre-attack Paranoia Activities’**, IEEE Transactions On Network And Service Management, Vol. 19, No. 1, March 2022.

**[2]** Suleiman Ali Alsaif. (2022) **‘Machine Learning-based Ransomware Classification Of Bitcoin Transactions’**, Applied Computational Intelligence And Soft Computing, Vol. 2022; Article ID 6274260.

**[3]** J. Hernandez-castro, A. Cartwright and E. Cartwright. (2022) **‘An Economic Analysis Of Ransomware And Its Welfare Consequences’**, Royal Society Open Science, Vol. 07, Issue 3, March 2022; https://doi.org/10.1098/rsos.190023.

**[4]** Hardi Sabah, Talabani, Hezha M.TAREQ Abdulhadi. (2022) **‘Bitcoin Ransomware Detection Employing Rule Based Algorithms’**, Science Journal Of University Of Zakho, Vol. 10, No. 1: January - March Issue; https://doi.org/10.25271/sjuoz.2022.10.1.865.

[**5]** Abdullah Alqahtani, Frederick T. Sheldon. (2022) **‘A Survey Of Crypto Ransomware Attack Detection Methodologies: An Evolving Outlook’**, MDPI Open Access Journals, Vol. 22, Sensors, 22(5), 1837; https://doi.org/10.3390/s22051837.

**[6]** Sabira Karim, Shemitha PA. (2021) **‘A Survey On Detection And Classification Of Ransomware Bitcoin Transactions’**, IJARIIE-ISSN(O)-2395-4396, Vol. 07, Issue 3, 2021.

**[7]** Vytarani Mathane, P.V. Lakshmi. (2021) **‘Predictive Analysis Of Ransomware Attacks Using Context-aware AI In Iot Systems’**, (IJACSA) International Journal Of Advanced Computer Science And Applications, Vol. 12, No. 4, 2021.

**[8]** Micheline Al Harrack. (2021) **‘The Bitcoinheist: Classifications Of Ransomware Crime Families’**,International Journal Of Computer Science & Information Technology (IJCSIT), Vol. 13, No. 5, October 2021.

**[9]** Kirat Jadhav. (2020) **‘Investigating Machine Learning Approaches For Bitcoin Ransomware Payment Detection System’,** ISSN No:-2456-2165, Vol. 05, Issue 9, September – 2020.

**[10]** Jason E. Thomas. (2018) **‘Individual Cyber Security: Empowering Employees To Resist Spear Phishing To Prevent Identity Theft And Ransomware Attacks’**,International Journal Of Business Management, 12(3). 1-23; doi:10.5539/Ijbm.V13n6p1.

**[11]** Navneet Kaur Popli, Anup Girdhar. (2018) **‘Behavioural Analysis Of Recent Ransomwares And Prediction Of Future Attacks By Polymorphic And Metamorphic Ransomware’**, Computational Intelligence: Theories, Applications And Future Direction,Part Of The Advances In Intelligent Systems And Computing Book Series (AISC, Volume 799);10.1007/978-981-13-1135-2\_6.