ISSN: 2395-7852

International Journal of Advanced Research in Arts, Science, Engineering & Management

Volume 10, Issue 3, May 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 6.551

+91 9940572462 +91 9940572462 ijarasem@gmail.com www.ijarasem.com

International Journal of Advanced Research in Arts, Science, Engineering & Management (IJARASEM)

| ISSN: 2395-7852 | www.ijarasem.com | Impact Factor: 6.551 |Bimonthly, Peer Reviewed & Referred Journal|

| Volume 10, Issue 3, May 2023 |

Bitcoin Attack Prediction UsingMachine Learning

Mr.A.Anbumani[1], Geeth Akshay Kumar M[2] , Sanjay M[3] , Thiyagarajan P G[4], Shaik Abdul Aleem[5]

[1]Assistant Professor, Department of Computer Science and Engineering,Velammal Institute of Technology, Panchetti, Chennai, India

[2]Department of Computer Science and Engineering, Velammal Institute ofTechnology, Panchetti, Chennai, India

[3]Department of Computer Science and Engineering, Velammal Institute ofTechnology, Panchetti, Chennai, India

[4]Department of Computer Science and Engineering, Velammal Institute ofTechnology, Panchetti, Chennai, India

[5]Department of Computer Science and Engineering, Velammal Institute ofTechnology, Panchetti, Chennai, India

ABSTRACT*:* Ransomware attacks are emerging as a major source of malware intrusion in recent times. While so far ransomware has affected general-purpose adequately resourceful computing systems. Many ransomware prediction techniques are proposed but there is a need for more suitable ransomware prediction techniques for machine learning techniques. This paper presents an attack of ransomware prediction technique that uses for extracting information features in Artificial Intelligence and Machine Learning algorithms for predicting ransomware attacks. The application of the data science process is applied for getting a better model for predicting the outcome. Variable identification and data understanding is the main process of building a successful model. Different machine learning algorithms are applied to the pre-processed data and the accuracy is compared to see which algorithm performed better other performance metrics like precision, recall, f1-score are also taken in consideration for evaluating the model. The machine learning model is used to predict the ransomware attack outcome.

I. INTRODUCTION

Cryptocurrencies have completely altered the digital transaction process all over the globe. Almost a decade after Satoshi Nakamoto generated the first Bitcoin block, many cryptocurrencies have been established. The Ransomware attack is a type of cybercrime and aclass of malware that encrypts the files and prevents users from accessing their data or systems and demands payment for decrypting and retrieving access to their files. Ransomware data classification using present data mining and machine learning methods is difficult because predictions aren't always correct. We aim to build two models that effectively address these challenges and can diagnose and classify Ransomware attacks accurately, then compare the performance of the models.

II. RELATED WORKS

A number of other approaches towards Intrusion detection and prediction have been made tilldate and previous studies, research, or publications that are relevant and closely related to the topic being discussed.
Ransomware is a type of malware that infects a victim's data and resources, and demands ransom to release them. In two main types, ransomware can lock access to resources or encrypt their content. In addition to computer systems, ransomware can also infect IoT and mobile devices [23]. Ransomware can be delivered via email attachments or web based vulnerabilities.More recently, ransomware have been delivered via mass exploits. For example, CryptoLockerused Gameover ZeuS botnet to spread through spam emails. Once the ransomware is installed, it communicates with a command and control center. Although earlier ransomware used hard-coded IPs and domain names, newer variants may use anonymity networks, such as TOR, to reach a hidden command and control server Once resources are locked or encrypted, the ransomware displays a message that asks a certain amount of bitcoins to be sent to a bitcoin address. This amount may depend on the number and size of the encrypted resources. After payment, a decryption tool is delivered to the victim. However, in some cases, such as with WannaCry, the ransomware contained a bug that made it impossible to identify who paid a ransomware amount.

Tracing cryptocurrencies payments due to malicious activity and criminal transactions is a complicated process. Therefore, the need to identify these transactions and label them is crucial to categorize them as legitimate digital currency trade and exchange or malicious activity operations. Machine learning techniques are utilized to train the machine to recognize specific transactions and trace them back to malicious transactions or benign ones. I propose to

work on the Bitcoin Heist data set to classify the different malicious transactions. The different transactions features are analyzed to predict a classifier label among the classifiers that have been identified as ransomware or associated with malicious activity. I use decision tree classifiers and ensemble learning to implement a random forest classifier.

Bitcoin might be a suburbanized form of payment system wherever the general public ledger is correctly supported in a very distributed manner. The unknown anonymous members referred to as miners, capital punishment a protocol that maintains and extends a distributed public ledger that records bitcoin transactions is termed a block chain. Block chain is enforced as a series of blocks. Bitcoin is that the known crypto-currency business. The transactions of bitcoin area unit utterly digital and unknown to a good extent. This case has crystal rectifier several cyber-crime perpetuators to use bitcoin as a secure haven for misbr transactions like Ransomware payments. Ransomware is malicious code that affects the payments entry reciprocally of ransom that should be paid. Machine Learning approaches could also be utilized to pour over the previous transactions as coaching information inorder to properly predict the people or teams to whom Ransomware payments area unit being created. This paper tries to explore the efficaciousness of various machine learning approaches in police work such payments. Ransomware may be a form of malware that infects a victim's information and resources, and demands ransom to unleash them.

Among all those ransomware attacks could be more impacting owing to attack methodology where victim systems become unusable until a ransom is paid, typically have attacker-defined timelines to respond, and can cause more monetary loss. Ransomware attacks, one of the malware attacks affect all types of security issues availability which causes monetary losses, and sensitive information loss . Crypto ransomware, locker ransomware, and hybrid ransomware are common types of ransomware.In crypto-ransomware attacks, data files are encrypted and the decryption key is provided only after paying the ransom. In locker ransomware attacks, the resources are blocked and are released only after paying the ransom. In hybrid ransomware attacks, both concepts of crypto ransomware and locker ransomware areused.

III. PROPOSED METHOD

The architecture of our system is illustrated in Figure 1. The major components of our system are Intrusion detection dataset, data pre-processing, data visualization, learning model using algorithm, choose best algorithm for accuracy and Deployment.

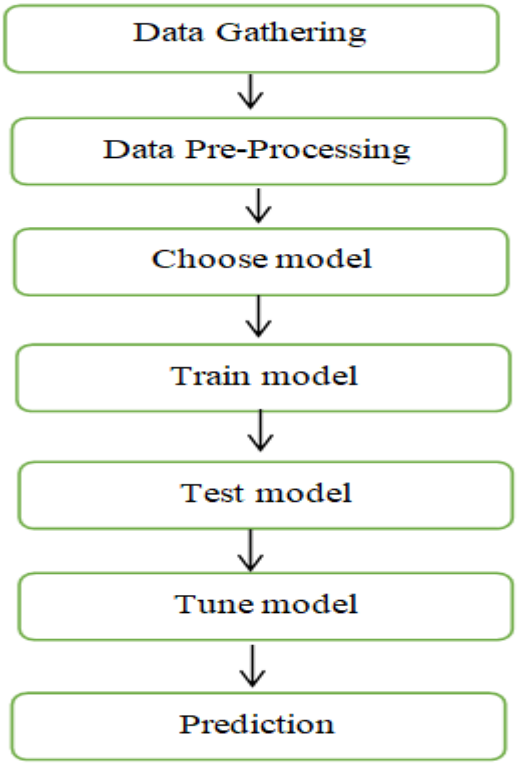


Fig 1: Architecture of proposed method

3.1 Data Description

The data set collected for predicting given data is split into Training set and Test set. Generally, 7:3 ratios are applied to split the Training set and Test set. The Data Model which was created using Random Forest, logistic, Decision tree algorithms, Support vector classifier(SVC), Multilayer Perceptron are applied on the Training set and based on the test result accuracy, Test set prediction is done.

3.2 Data Pre-Prepocessing

The data which was collected might contain missing values that may lead to inconsistency. To gain better results data need to be preprocessed so as to improve the efficiency of the algorithm. The outliers have to be removed and also variable conversion needto be done.
Validation techniques in machine learning are used to get the error rate of the MachineLearning (ML) model, which can be considered as close to the true error rate of the dataset. Ifthe data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that maynot be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuningmodel hyper parameters.

```
data.shape
```

```
(14514, 11)
```

```
df = data.dropna()
```

```
df.shape
```

```
(14514, 11)
```

```
df.isnull().sum()
```

```
Unnamed: 0    0
address       0
year          0
day           0
length        0
weight        0
count         0
looped        0
neighbors     0
income        0
label         0
dtype: int64
```

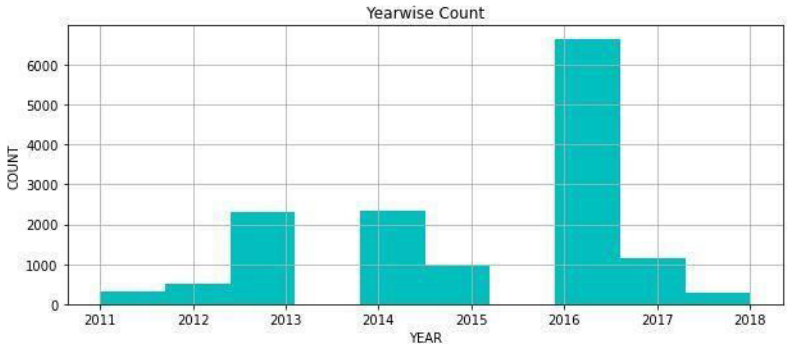
Fig:2 Pre-processing results show the whole dataset

3.3.Data Visualization

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral and stakeholders than measures of association or significance. Data visualization and exploratory data analysis are whole fields themselves and it will recommend a deeper dive into some the books mentioned at the end.

```
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')
```




```
# Heatmap plot diagram
fig, ax = plt.subplots(figsize=(15,5))
s.heatmap(df.corr(), ax=ax, annot=True, cmap='CMRmap')
```

```
<AxesSubplot:>
```

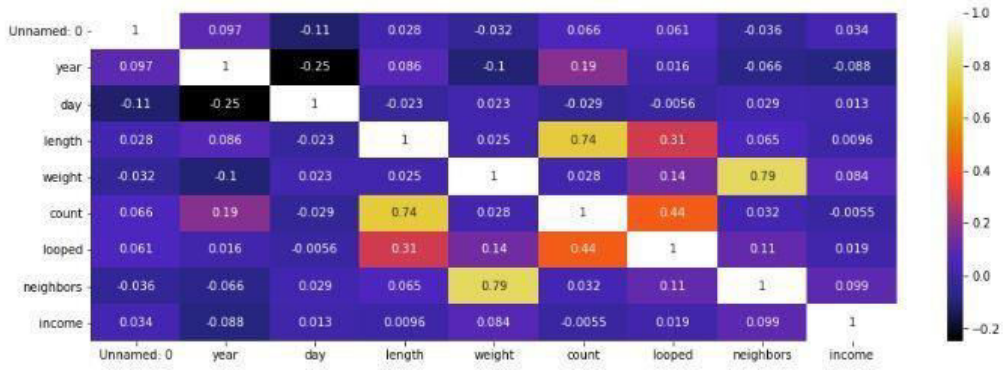


Fig 3: Types of Data visualization

3.4. Implementation of Algorithm:

It is important to compare the performance of multiple different machine learning algorithms consistently and it will discover to create a test harness to compare multiple different machine learning algorithms in Python with scikit-learn. It can use this test harness as a template on your own machine learning problems and add more and different algorithms to compare. Each model will have different performance characteristics. Using resampling methods like cross validation, you can get an estimate for how accurate each model may be onunseen data. It needs to be able to use these estimates to choose one or two best models from the suite of models that you have created. When have a new dataset, it is a good idea to visualizethe data using different techniques in order to look at the data from different perspectives. The same idea applies to model selection. You should use a number of different ways of looking at

the estimated accuracy of your machine learning algorithms in order to choose the one or twoto finalize. A way to do this is to use different visualization methods to show the average accuracy, variance and other properties of the distribution of model accuracies.

3.4.1 XG Boost classifier:

Its speed and performance are unparalleled and it consistently outperforms any other algorithms aimed at supervised learning tasks.The library is parallelizable which means the core algorithmcan run on clusters of GPUs or even across a network of computers. This makes it feasible to solve ML tasks by training on hundreds of millions of training examples with high performance.

```
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
```

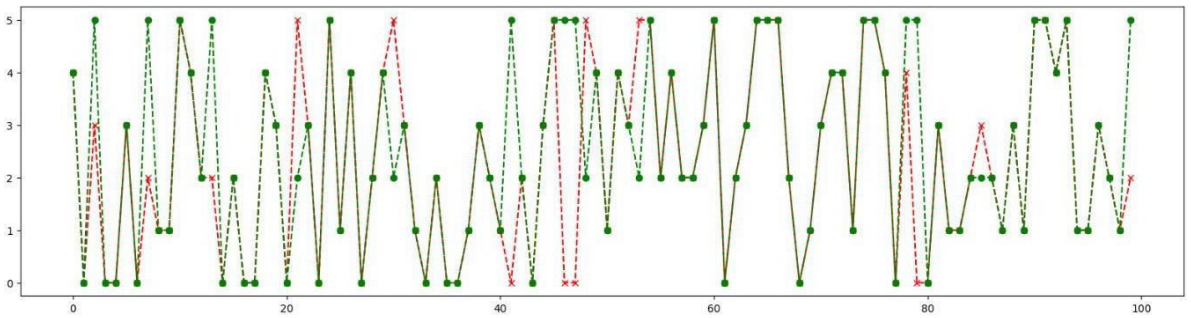


Fig 4: XG Boost classifier algorithm code and output

3.4.2 Voting Classifier:

A Voting Classifier is a machine learning model that trains on an ensemble of numerous

models and predicts an output(class) based on their highest probability of chosen class as the output.
It simply aggregates the findings of each classifier passed into Voting Classifier and predicts the output class based on the highest majority of voting. The idea is instead of creating

separate dedicated models and finding the accuracy for each them, we create a single model which trains by these models and predicts output based on their combined majority of voting for each output class.

```
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
```

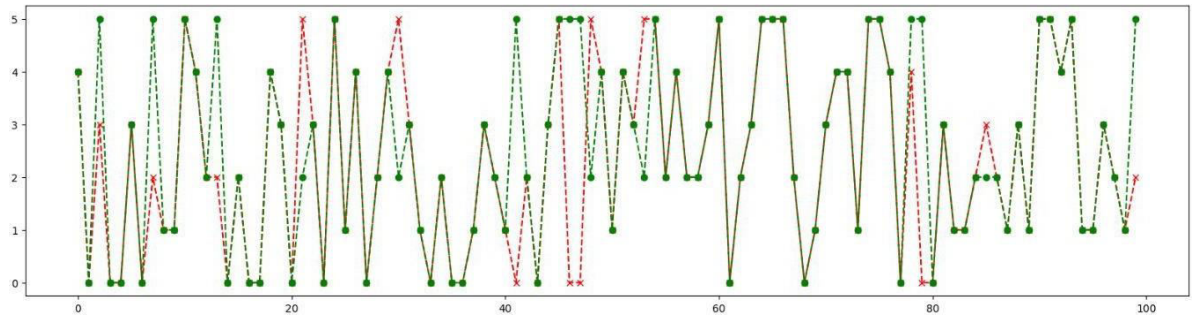


Fig 5: Voting classifier algorithm code and output

3.4.3 Random Forest Classifier

Random Forest is a popular machine learning algorithm that belongs to the supervised learningtechnique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of *combining multiple classifiers to solve a complex problem and to improve the performance of the model.*

As the name suggests, *"Random Forest is a classifier that contains a number of decision treeson 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, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

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

```
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
```

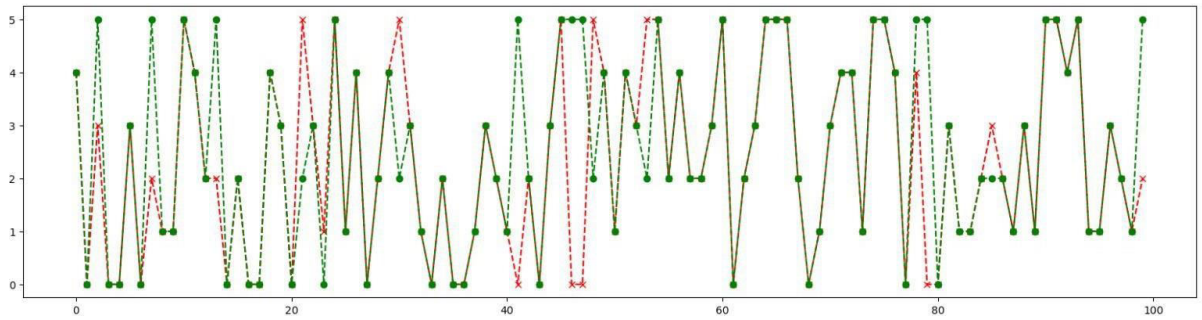


Fig 6: Random Forest algorithm code and output

3.4.4 Logistic regression

Logistic regression is another powerful supervised ML algorithm used for
binary classification problems (when target is categorical). The best way to think about logistic regression is that it is a linear regression but for classification problems. Logistic regression essentially uses a logistic function defined below to model a binary output variable (Tolles & Meurer, 2016). The primary difference between linear regression and logistic regression is that logistic regression's range is bounded between 0 and 1. In addition, as opposed to linear regression, logistic regression does not require a linear relationship betweeninputs and output variables.

```
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
```

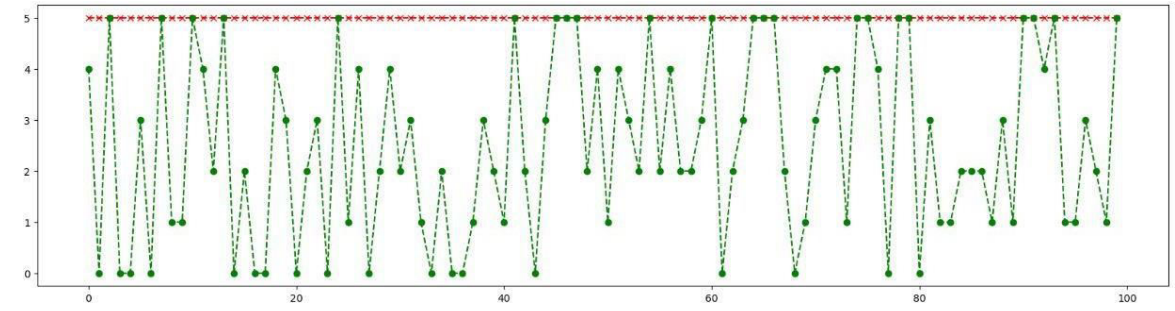


Fig 7: Logistic regression algorithm code and output

3.5 Deployment

In this module the trained machine learning model is converted into pickle data format file (.pkl file) which is then deployed for providing better user interface and predicting the outputof Human Stress and Deployment used here is Django Web Framework. Django is a micro web framework written in Python. It is classified as a micro-framework because it does not require particular tools or libraries

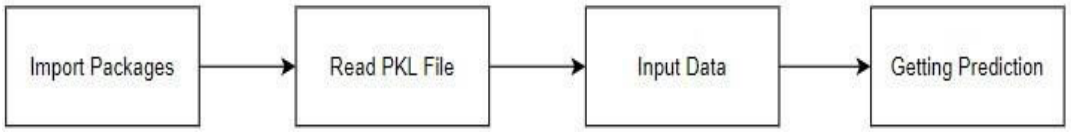


Fig 8: Deployment Module Diagram

IV. EXPERIMENTAL RESULTS

V. CONCLUSION

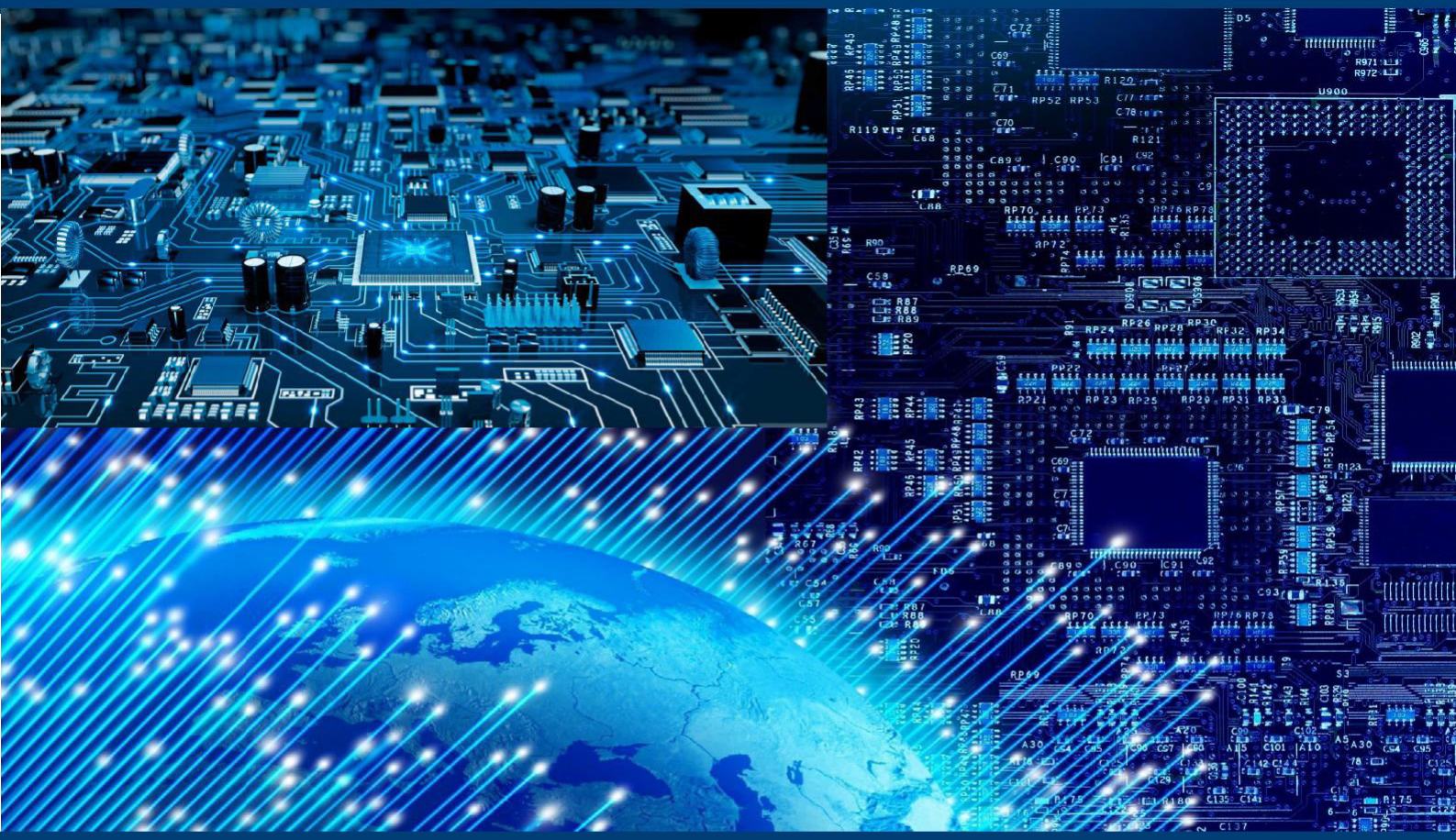
The analytical process started from data cleaning and processing, missing value,exploratory analysis and finally model building and evaluation. The best accuracy on public test set of higher accuracy score algorithm will be find out. The founded one is used in the application which can help to find the Bitcoin Heist randsomeware attack.

ACKNOWLEDGEMENTS

The work done by Geeth Akshay Kumar M, Sanjay M, Thiyagarajan P G and Shaik AbdulAleem is guided by Mr A.Anbumani and supported by Velammal Institute of Technology.

REFERENCES

1. Muniye, Temesgen & Rout, Minakhi & Mohanty, Lipika & Satapathy, Suresh. (2020). Bitcoin Price Prediction and Analysis Using Deep Learning Models. 10.1007/978-981-15-5397-4_63.
2. Al Harrack, Micheline, The BitcoinHeist: Classifications of Ransomware Crime Families (October 2021). International Journal of Computer Science & InformationTechnology (IJCSIT) Vol 13, No 5, October 2021.
3. Çağlar, Ersin & Kirikkaleli, Dervis. (2020). The Crypto-currency and Cyber-attack: Evidence from Causality Techniques. International Journal of Engineering Trends and Technology. 68. 1-4. 10.14445/22315381/IJETT-V68I9P201.
4. Mahajan, Rahul & Roychaudhary, Reema. (2021). Protective Mechanism form DDoS Attack for Cryptocoin. 12. 1421.
5. Shah, Naman & Dave, Sonal. (2022). Review on Types of Attacks on Bitcoin and Ethereum crypto currencies. International Journal of Engineering Research in Computer Science and Engineering. 9. 1-6. 10.36647/IJERCSE/09.10.Art001.



International Journal of Advanced Research in Arts, Science, Engineering & Management (IJARASEM)

| Mobile No: +91-9940572462 | Whatsapp: +91-9940572462 | ijarasem@gmail.com |