Ransomware Threat Analysis Using Machine Learning

Dr.Anitha T.N[1], and Ashish Kumar Mahto[2], Harsh Pandey[2], Sakshi Pandey[2]
[1]Professor, Department of CSE, Sir M. Visvesvaraya Institute Technology, Bangalore, India
1 anitharesddytn72@gmail.com
[2]Students, CSE, Sir M. Visvesvaraya Institute of Technology, Bangalore, India
2{ashishhero737,harshpandey3883,sakshipandeyiffco09}@gmail.com

ABSTRACT:

With the advancement and easy accessibility of computer and internet technology, network security has become vulnerable to hacker threats, including ransomware attacks targeting smartphone operating systems (e.g., Android) and applications. Despite efforts to detect malicious URLs, lexical features sometimes need to be improved. This document provides an overview of ransomware, a timeline of assaults, and comprehensive research on methods for identifying, avoiding, minimising, and recovering from ransomware attacks. Analysing studies between 2017 and 2024 offers up-to- date knowledge on developments in ransomware detection and advancements in combating ransomware attacks, highlighting unanswered concerns and potential research challenges.

KEYWORDS:

Machine Learning; Ransomware, Cybers-attacks; Android; Malware; Hacker; Ransomware detection.

INTRODUCTION:

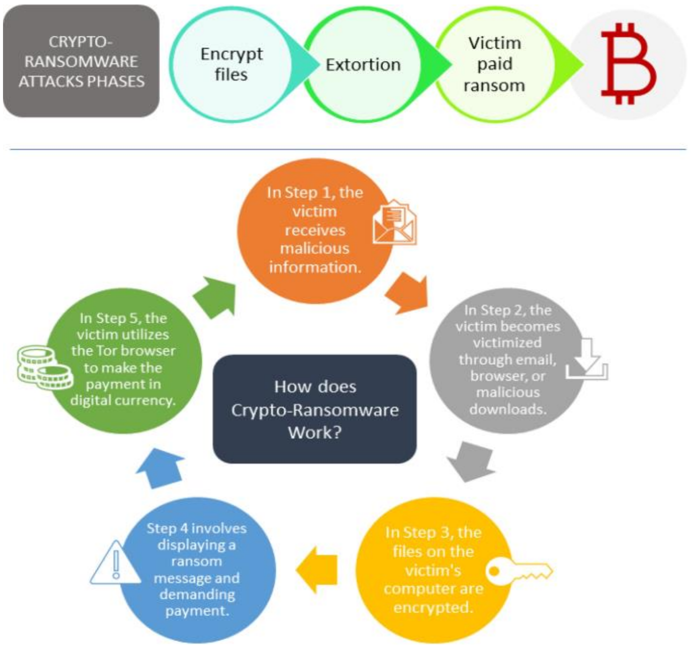
Android is the most popular operating system for smartphones and other smart devices, making it a prime target for cyber-attacks. With more people using smartphones for everything from studying to shopping, the risk of attacks has increased. Unlike some other operating systems, Android is open and doesn't restrict what apps you can download, making it easier for cybercriminals to create malicious apps. One of the most dangerous types of attack is called ransomware. This is when hackers take control of your device and either lock you out of it or encrypt your data, demanding money to give it back. In 2020, ransomware caused the death of a woman in Germany when it hit a hospital's computer systems, preventing doctors from treating her. The number of mobile apps being used is growing fast, and this means more traffic on the internet, especially from apps that are always connected, like social media apps. As the internet gets faster with 5G, more devices like self-driving cars and smart home gadgets will be connected. This means even more internet traffic. To protect against malware on Android devices, antivirus software uses standard methods, but these are well-known to hackers. A better solution is to use machine learning to detect unusual patterns in network traffic, which can help identify ransomware attacks before they cause damage.

RELATED WORK:

In the realm of ransomware detection, researchers have explored various approaches over the years. In 2017, Chen et al. introduced a system employing data mining techniques and dynamic analysis of Application Programming Interfaces (APIs) to detect ransomware. By monitoring API calls and mapping them into a feature space, they achieved high accuracy using the SL algorithm. Following this, in 2018, Al-rimy et al. Provided a detailed review of ransomware, outlining various detection and prevention technologies. Around the same time, Cusack et al. Utilised a programmable forwarding engine to monitor network flows between command and control servers and infected computers, achieving an 86% detection rate. In a different approach, Zhang et al. in 2018 used static feature analysis, transforming opcode sequences into N-gram sequences and achieving 91% accuracy in ransomware classification. Moreover, Alhawi et al. proposed NetConverse, a machine learning method utilising network traffic conversations, achieving 97% accuracy. In 2019, Kaiiali et al. focused on crypto ransomware network activities, achieving high detection accuracy at packet and flow levels. Additionally, Noorbehbahani et al. Analysed six machine learning techniques for ransomware detection, with Random Forest proving the most effective. In subsequent years, research diversified, with studies focusing on Android malware detection, utilising machine learning methods such as Random Forest and Deep Neural Networks. Finally, in 2021, Lee et al. emphasised the growing threat of Advanced Persistent Threats (APTs) and proposed an open-source framework for ransomware detection at system and network levels, aiming to quickly extract and respond to ransomware attack features.

BACKGROUND:

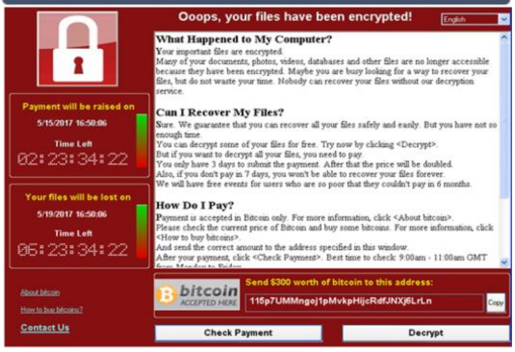
Crypto-ransomware is a type of malicious software that targets computer systems and networks. It works by encrypting files and data using both symmetric and asymmetric encryption algorithms. Once the files are encrypted, they become inaccessible to the user. Even if the ransomware is removed from the infected computer or the compromised storage device is introduced into another system, the encrypted data remains unusable. However, the malware often leaves essential system files uncorrupted, allowing the compromised device to still be used to pay the ransom. This visual representation illustrates how crypto-ransomware operates and why it has become increasingly prevalent in cyberattacks.


CRYPTO-RANSOMWARE ATTACKS PHASES
Encrypt files
Extortion
Victim paid ransom
In Step 1, the victim receives malicious information.
In Step 2, the victim becomes victimized through email, browser, or malicious downloads.
In Step 3, the files on the victim's computer are encrypted.
Step 4 involves displaying a ransom message and demanding payment.
In Step 5, the victim utilizes the Tor browser to make the payment in digital currency.
How does Crypto-Ransomware Work?

Crypto-Ransomware Facts

1. It doesn't steal but rather renders it impossible for users to access information.
2. It spreads through targeted email-based phishing campaigns.
3. Detection is not the solution as it won't restore lost data.
4. It is the favored attack tool for hackers because it is easy to produce and there are a number of well-documented cryptographic libraries available.

A screenshot shows a WannaCry ransomware executable.



Ooops, your files have been encrypted!

Payment will be raised on
5/15/2017 16:50:06
Time Left
02:23:34:22

Your files will be lost on
5/19/2017 16:50:06
Time Left
05:23:34:22

What Happened to My Computer?

Your important files are encrypted.
Many of your documents, photos, videos, databases and other files are no longer accessible because they have been encrypted. Maybe you are busy looking for a way to recover your files, but do not waste your time. Nobody can recover your files without our decryption service.

Can I Recover My Files?

Sure. We guarantee that you can recover all your files safely and easily. But you have not so enough time.
You can decrypt some of your files for free. Try now by clicking <Decrypt>.
But if you want to decrypt all your files, you need to pay.
You only have 3 days to submit the payment. After that the price will be doubled.
Also, if you don't pay in 7 days, you won't be able to recover your files forever.
We will have free events for users who are so poor that they couldn't pay in 6 months.

How Do I Pay?

Payment is accepted in Bitcoin only. For more information, click <About bitcoin>.
Please check the current price of Bitcoin and buy some bitcoins. For more information, click <How to buy bitcoins>.
And send the correct amount to the address specified in this window.
After your payment, click <Check Payment>. Best time to check: 9:00am - 11:00am GMT from Monday to Friday.

bitcoin ACCEPTED HERE
Send $300 worth of bitcoin to this address:
115p7UMMngoj1pMvkpHijcRdfJNXj6LrLn

Figure I: Crypto-Ransomware

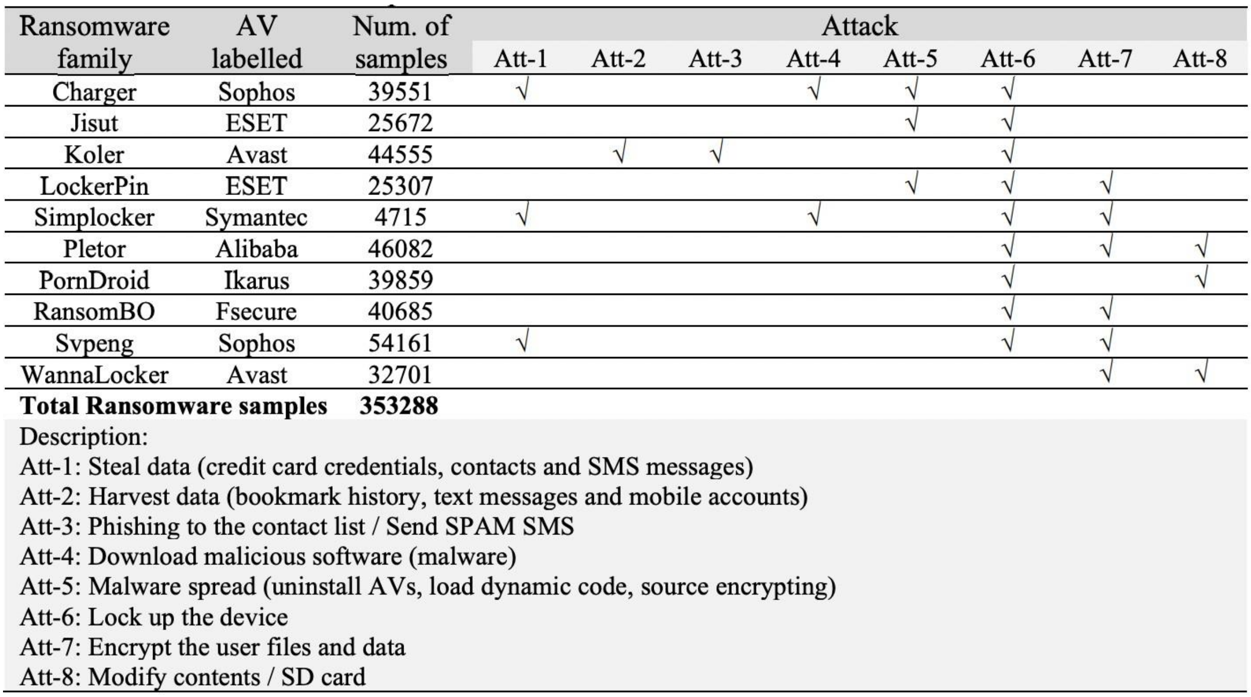
THE NETWORK TRAFFIC AND DATA SET USED:

Analyzing network traffic is a crucial way to detect malware. By looking at the data moving through a network, including what it contains and where it's going, we can spot suspicious activity. When malware infects a device, it often connects to other servers to do bad things like stealing information or getting updates. So, by keeping an eye on both incoming and outgoing network traffic, as well as what's happening within a network, we can find malware. In a recent study, researchers used a dataset called And Mal2017, which has both good and bad samples, including 42 different types of malware. They focused on detecting ransomware, a type of malware that holds your files hostage until you pay money. By looking at the data from over 600,000 samples, they found that certain features in the network traffic can help us spot ransomware effectively.

- **Ransomware Dataset:**

In this study, a dataset of 353,288 ransomware samples was utilised, comprising 85 features collected from 10 prevalent ransomware families. Table I presents the behaviour and characteristics of ransomware along with the number of samples used for each family.

TABLE 1: Description of behaviour and characteristics of ransomware dataset



Ransomware family	AV labelled	Num. of samples	Attack							
			Att-1	Att-2	Att-3	Att-4	Att-5	Att-6	Att-7	Att-8
Charger	Sophos	39551	√			√	√	√		
Jisut	ESET	25672					√	√		
Koler	Avast	44555		√	√			√		
LockerPin	ESET	25307					√	√	√	
Simplocker	Symantec	4715	√			√		√	√	
Pletor	Alibaba	46082						√	√	√
PornDroid	Ikarus	39859						√		√
RansomBO	Fsecure	40685						√	√	
Svpeng	Sophos	54161	√					√	√	
WannaLocker	Avast	32701							√	√
Total Ransomware samples		**353288**								

Description:
Att-1: Steal data (credit card credentials, contacts and SMS messages)
Att-2: Harvest data (bookmark history, text messages and mobile accounts)
Att-3: Phishing to the contact list / Send SPAM SMS
Att-4: Download malicious software (malware)
Att-5: Malware spread (uninstall AVs, load dynamic code, source encrypting)
Att-6: Lock up the device
Att-7: Encrypt the user files and data
Att-8: Modify contents / SD card

- **Benign Dataset:**

The research used over 6,000 benign applications from the Google Play Store, published between 2015 and 2017. These were grouped based on popularity. 92 samples with 85 network traffic features were extracted and categorized into classes such as Flow-ID, Packet-based, Byte-based, Flow-based, and Time-based. The applications were checked using the Virustotal Web Service with two Antivirus Products. The findings were published in the Journal of Education and Science, Volume 30, Number 5, 2021, pages 86-102.

THE PROPOSED SYSTEM:

The proposed system in this research consists of five steps as shown in Figure II.

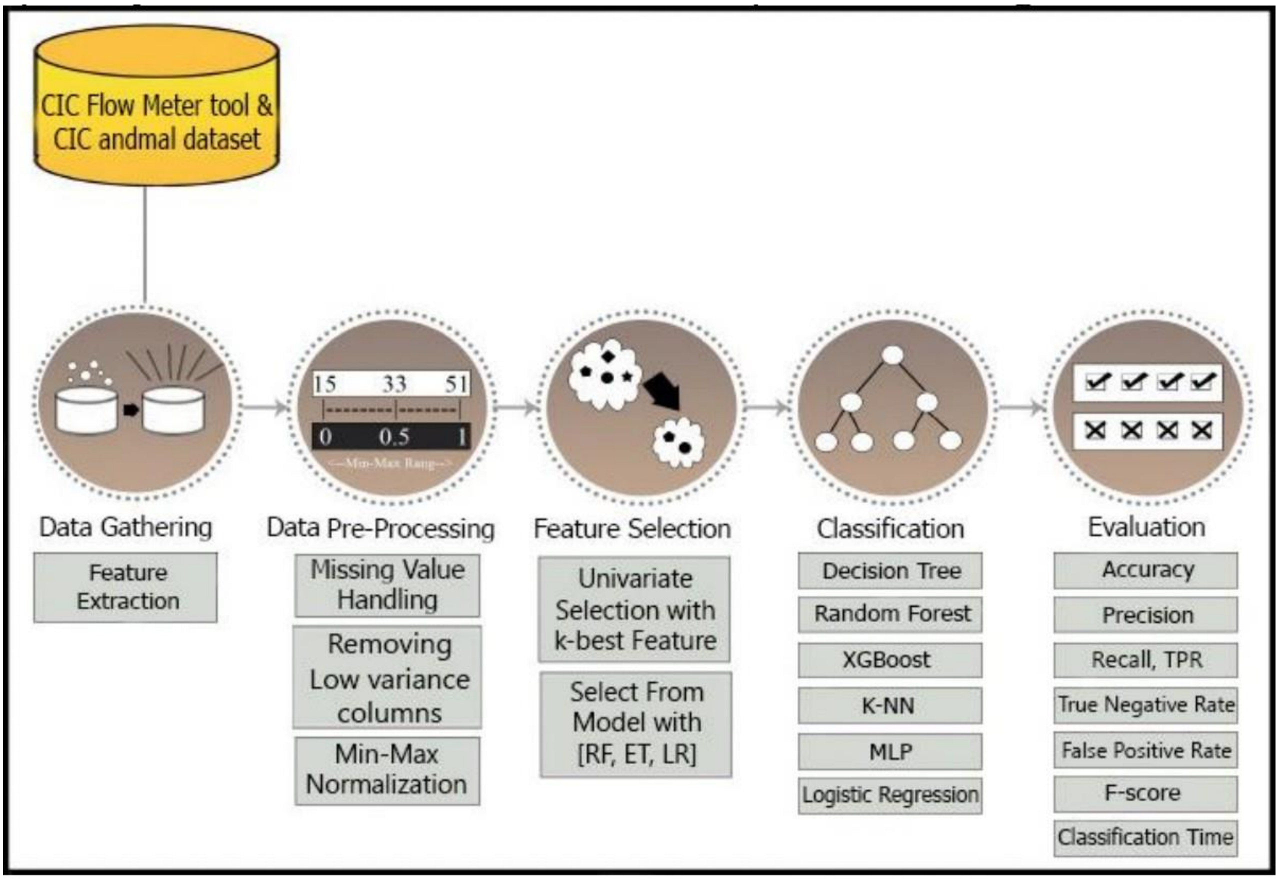


Figure II: The methodology of the proposed system

A. Collect Data: First, we gather network data using tools like CICFlowMeter or from reliable websites.

B. Prepare Data: Next, we clean up the data by removing any missing values and features with low variance. We also scale the data using a method called normalization, which puts all the data on a scale from 0 to 1.

C. Select Features: We then analyze the data to choose the most important features for our analysis.

D. Train Algorithms: After selecting the features, we use six different machine learning algorithms to train our model. We use 80% of our data for training and the remaining 20% for testing.

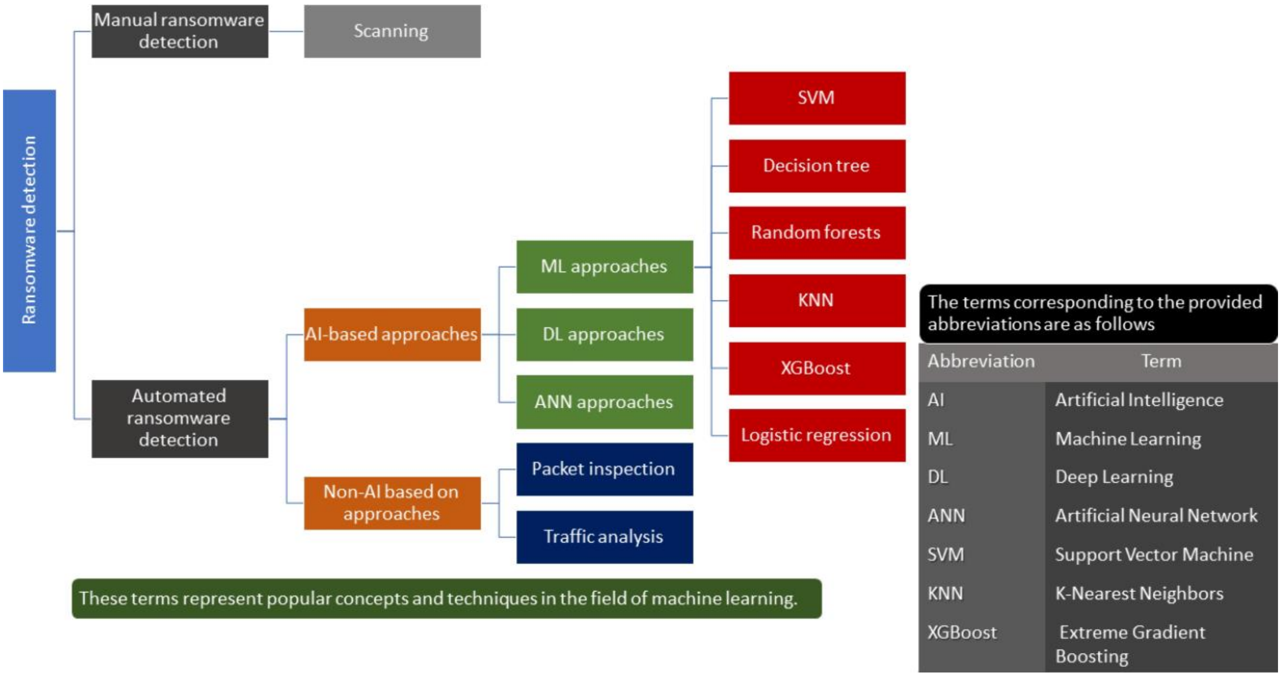
E. Test and Evaluate: Finally, we test our machine learning model using the 20% of data we set aside for testing. This helps us evaluate how well our model performs.

RANSOMWARE DETECTION:

There are two main ways to detect ransomware: automated and manual.

- Automated methods use software tools to find and report ransomware attacks. These tools can potentially stop attacks before they cause damage.
- Manual detection involves regularly checking data and devices for signs of attacks. This includes looking for changes to file extensions, seeing if authorized users can still access their devices and files, and checking for any other unusual changes.

You can see the flow of this process in Figure III.



These terms represent popular concepts and techniques in the field of machine learning.

The terms corresponding to the provided abbreviations are as follows

Abbreviation	Term
AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
ANN	Artificial Neural Network
SVM	Support Vector Machine
KNN	K-Nearest Neighbors
XGBoost	Extreme Gradient Boosting

Figure III: Ransomware detection taxonomy

ALGORITHMS:

Machine learning, a type of artificial intelligence, allows computer systems to improve performance without explicit programming. Ransomware, malicious software, encrypts files and demands payment for decryption. To combat this, various machine learning algorithms such as support vector machines, decision trees, random forests, k-nearest neighbours, XGBoost, and logistic regression are employed. Each method has its advantages and disadvantages, and the choice depends on the situation and data.

Table 2. Machine learning algorithms

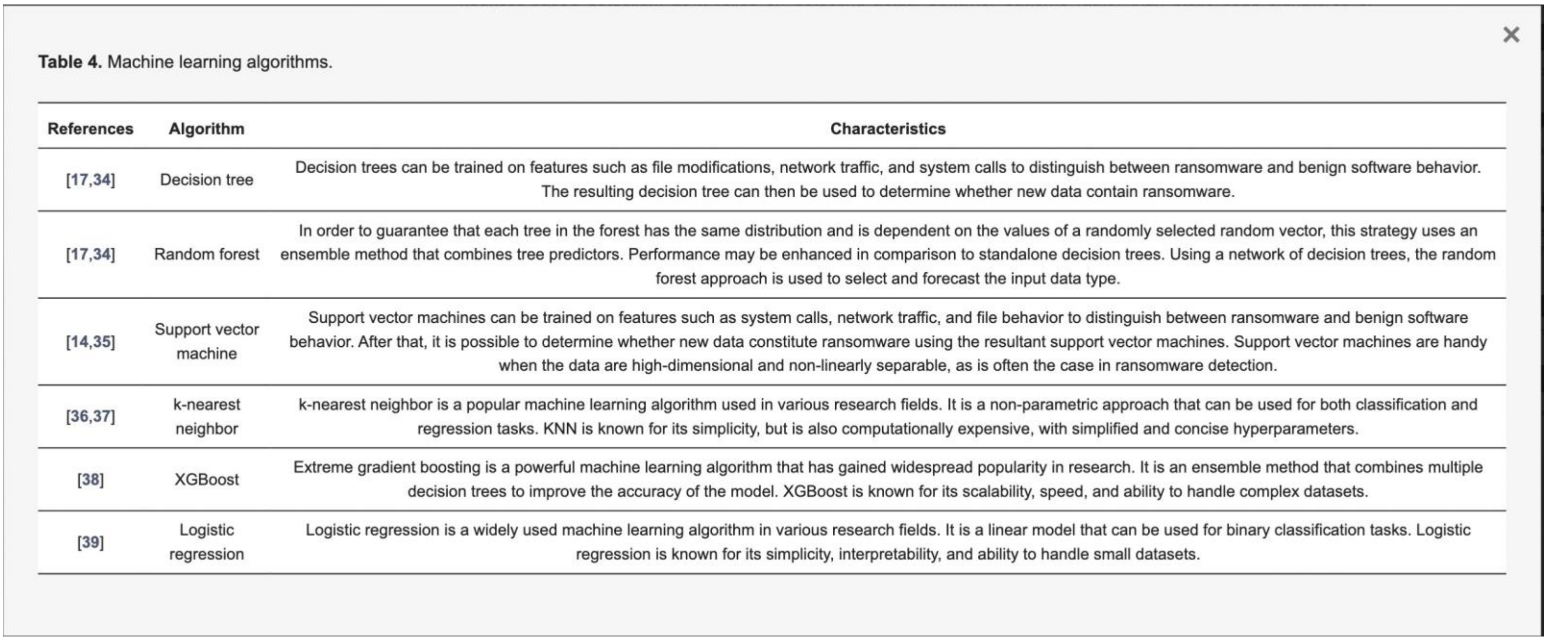


Table 4. Machine learning algorithms.

References	Algorithm	Characteristics
[17,34]	Decision tree	Decision trees can be trained on features such as file modifications, network traffic, and system calls to distinguish between ransomware and benign software behavior. The resulting decision tree can then be used to determine whether new data contain ransomware.
[17,34]	Random forest	In order to guarantee that each tree in the forest has the same distribution and is dependent on the values of a randomly selected random vector, this strategy uses an ensemble method that combines tree predictors. Performance may be enhanced in comparison to standalone decision trees. Using a network of decision trees, the random forest approach is used to select and forecast the input data type.
[14,35]	Support vector machine	Support vector machines can be trained on features such as system calls, network traffic, and file behavior to distinguish between ransomware and benign software behavior. After that, it is possible to determine whether new data constitute ransomware using the resultant support vector machines. Support vector machines are handy when the data are high-dimensional and non-linearly separable, as is often the case in ransomware detection.
[36,37]	k-nearest neighbor	k-nearest neighbor is a popular machine learning algorithm used in various research fields. It is a non-parametric approach that can be used for both classification and regression tasks. KNN is known for its simplicity, but is also computationally expensive, with simplified and concise hyperparameters.
[38]	XGBoost	Extreme gradient boosting is a powerful machine learning algorithm that has gained widespread popularity in research. It is an ensemble method that combines multiple decision trees to improve the accuracy of the model. XGBoost is known for its scalability, speed, and ability to handle complex datasets.
[39]	Logistic regression	Logistic regression is a widely used machine learning algorithm in various research fields. It is a linear model that can be used for binary classification tasks. Logistic regression is known for its simplicity, interpretability, and ability to handle small datasets.

CONCLUSION & FUTURE WORK:

Some attackers try to mimic normal network traffic to evade detection systems by changing certain characteristics like flow duration or using fake IP addresses. This research focuses on selecting the best features from a range of network traffic characteristics to detect ransomware. The experiments showed that these network traffic features are very effective in detecting ransomware, especially when the data is taken from online network traffic. Various techniques were used to select the best features, and six machine-learning algorithms were tested for detecting Android ransomware. Seven performance metrics were used to evaluate these algorithms. The results showed that decision trees (DT) and extreme gradient boosting (XGB) had an average detection accuracy of over 99%, with very low false positive rates (0.016% for DT and 0.029% for XGB). For future research, the suggestion is to use network traffic features to detect and classify other types of Android malware. Additionally, the aim is to incorporate other types of features, such as memory dump, permissions, logs, or API calls, to create a more comprehensive Android ransomware detection framework.

ACKNOWLEDGEMENT:

The authors of this research would like to thank the University of VTU SIR Mvit, Bangalore for their facilities, support, and guidance throughout the project, which have helped to enhance the quality of this work. The authors would also like to thank their friends and families for believing in them and continuously supporting them.

REFERENCES:

[1] A. Bhattacharya, R. T. Goswami, "Community-Based Feature Selection Method for Detection of Android Malware," Journal of Global Information Management (JGIM), 2018.

[2] Cyber Security Services in Washington, DC. The 2021 CYBER SECURITY STATISTICS,

DATA, & TRENDS. Available: https://purplesec.us/cyber-security-trends-2021/

[3] CISCO. The Cisco Annual Internet Report (2018–2023) White Paper. March 9, 2020. Available: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internetreport/ white-paper-c11-741490.html

[4] F. Noorbehbahani, F. Rasouli, M. Saberi, "Analysis of machine learning techniques for ransomware detection," 2019 16th International ISC (Iranian Society of Cryptology) Conference on Information Security and Cryptology (ISCISC), IEEE, 2019.

[6] B. A. Al-rimy, M. A. Maarof, S. Z. Shaid, "Ransomware threat success factors, taxonomy, and countermeasures: A survey and research directions," Computers & Security, 2018.

[7] Jessica Ellis. December 15, 2020. Year In Review: Ransomware. Available: https://securityboulevard.com/2020/12/year-in-review-ransomware/

[8] Canadian Institute for Cybersecurity. Canada's research leader in cybersecurity. accessed by Jan 2020. Available: https://www.unb.ca/cic/

[9] The CICFlowMeter packet capture tool. Available: https://www.unb.ca/cic/research/applications.html#CICFlowMeter

[10] Virustotal website for security services. (2020). Available: https://www.virustotal.com/en Abdullah, Mohammed Hamid Abdulraheem. Designing Deep Learning Based Network Intrusion Detection System for Software Defined Network. Diss. University of Mosul, 2020.

[11] Scikit learn. Variance threshold technique for feature selection, by sci-kit learning. Available: https://scikitlearn.org/stable/modules/generated/sklearn.feature_selection.VarianceThreshold.html

[12] Jason Brownlee. Machine Learning Mastery. The feature selection methods. November 27, 2019. Available: https://machinelearningmastery.com/feature-selection-with-real-and-categorical- data/

[13] Scikit learn. f_classif technique for feature selection, by sci-kit learning. Available: https://scikitlearn.org/stable/modules/generated/sklearn.feature_selection.f_classif.html

[14] Scikit learn. k_best technique for feature selection. Available: https://scikitlearn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html

[15] Scikit learn. Select from the model technique for feature selection, by sci-kit learning. Available: https://scikit-learn.org/stable/modules/generated/ sklearn.feature_selection.SelectFromModel.html