**Detection of Malware Using Machine Learning Algorithms**

**Shubhanshu Kumar1, Satyam Singh2, Yusuf Ahmad3, Yash Anjana4**

1,2,3,4Student

Department of CSE, HKBKCE, Bangalore, India

**\*Corresponding Author**

**E-mail Id**: - shubhanshubb@gmail.com

**ABSTRACT:**

The issue of malware is a significant problem faced by internet users, and polymorphic malware is a newer, more adaptable form of malicious software that can evade traditional signature-based detection models. To combat this, machine learning techniques were used to identify malware threats, and the algorithm with the highest accuracy was selected for use in the system. The confusion matrix provided additional information about false positives and false negatives, allowing for a more comprehensive evaluation of the system's effectiveness. The research demonstrated that machine learning algorithms could be used to detect harmful traffic on computer systems, improving the security of computer networks. The results showed that DT, RANDOM FOREST, and GRADIENT BOOSTING had high detection accuracy rates compared to other classifiers, indicating that these methods could be effective in detecting increasingly complex and common forms of malware. These findings have important implications for computer security and highlight the potential of machine learning in addressing the challenges of modern cybersecurity threats.

1. **INTRODUCTION:**

The increasing prevalence and complexity of cyberattacks are a significant concern in modern technology. Malware is a type of cyberattack that can take many forms and is designed to harm computers, users, businesses, or computer systems. Malicious software is typically installed and run without the user's knowledge or consent, and can include viruses, Trojan horses, ransomware, spyware, adware, rogue software, wipers, scareware, and other types of threats.

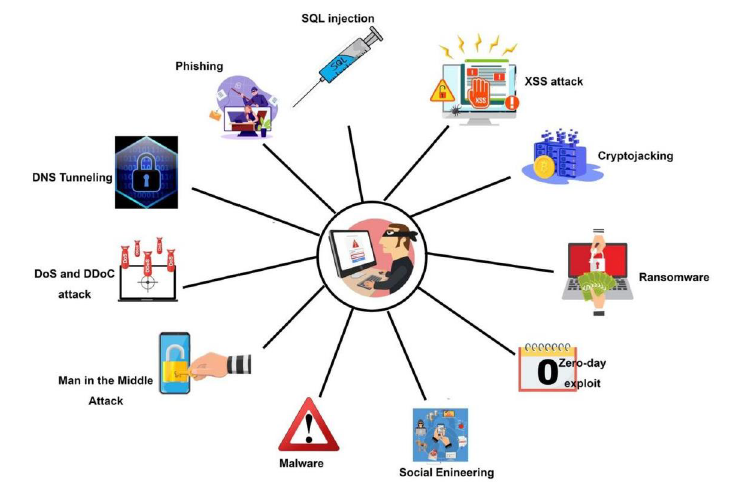
This study highlights the potential of machine learning algorithms to detect harmful traffic on computer systems and improve the security of computer networks. The proposed approach involved using malware analysis and detection with machine learning algorithms to compute the difference in correlation symmetry integrals. Several classification algorithms, including Random Forest, AdaBoost, GNB, Gradient Boosting, DT, and the proposed approach, were evaluated in terms of their detection accuracy rates.

These findings have important implications for improving the effectiveness of cybersecurity measures and protecting computer systems from malware and other cyber threats. By leveraging the power of machine learning and data analysis, organizations can better identify and respond to potential threats and reduce the risk of data breaches and other cyber-attacks.

To evaluate whether or not a particular piece of software or network connection constitutes a security issue, malware detection modules are responsible for analysing data they have acquired and been educated with. Consider a machine learning system that can articulate the concepts behind the patterns it has discovered clearly. Machine learning systems that have trained algorithms can help them anticipate better by giving them feedback on how well they handled earlier tasks and using that information to adjust their behaviour.

Through the use of malicious software and the theft of private data, cybercriminals around the world pose a serious threat to organisations, institutions of higher learning, governments, and individuals. Every day, thousands of con artists use malicious software to access networks, steal data, or send money. As a result, safeguarding confidential information has become a pressing issue in the scientific community. This study used data mining and machine learning classification techniques to provide a comprehensive framework for identifying harmful software and safeguarding personal data from hackers. In this research, we examine anomaly-based and signature-based characteristics to build a strong and efficient malware classification and detection method. Experiments have shown that the suggested method is better than alternatives.

Modern malware is becoming more prevalent and sophisticated, posing a serious danger to the security of contemporary websites. Cyberattack kinds are shown in Figure 1 in the digital or cyberspace realm. Malware is software designed specifically to harm a computer or network, for as by tracking its users or stealing their money. Attacks by malware are getting more frequent and now even impact IoT devices, medical equipment, and industrial and environmental control systems. Modern spyware frequently modifies its code and behaviour, making it infamously difficult to detect.



Malware has become more prevalent, making conventional signature-based defences ineffectual.

Instead, a wider variety of defensive measures must be taken.

The use of machine learning algorithms for malware detection is becoming increasingly important as cybercriminals continue to develop more sophisticated and complex malware. Dynamic analysis, which takes into account the behaviour of dangerous files by tracking data flows and function calls, is particularly useful in identifying these types of malware. By leveraging both static and behavioural features, machine learning algorithms can accurately describe the structure of contemporary malware and identify attacks that could otherwise avoid detection by traditional signature-based techniques. Deep learning algorithms are particularly effective in this regard as they can perform feature engineering on their own, resulting in more accurate feature representation.

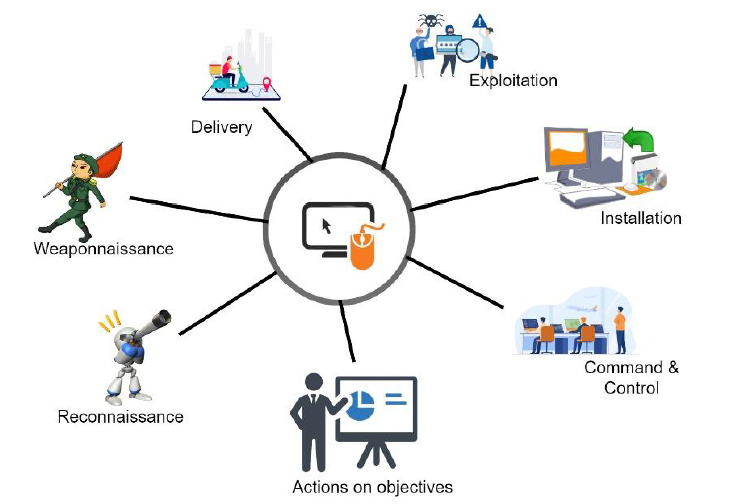


Figure 2 illustrates the Martin (2018) Cyber Kill Chain, which is used for cyberattack protection and as a security measure to protect networks. Unfortunately, despite such measures, large-scale cyberattacks continue to occur. For example, in February of 2020, AWS was the target of a massive distributed denial of service (DDoS) attack, which the organization withstood. In July of 2020, three hackers gained access to Twitter and took over several prominent users' accounts, generating over $100,000 in profits from Bitcoin scams uploaded from the stolen accounts. In 2018, a cyberattack at Marriott's Starwood Hotels exposed the personal information of over 500 million customers. And in 2017, the WannaCry ransomware attack affected over 300,000 systems in 150 countries and cost billions to fix. These examples demonstrate the urgent need for effective cybersecurity measures, including the use of machine learning algorithms for malware detection and prevention.

Russia attempted a cyberattack on the Ukrainian energy grid in 2017 as part of its ongoing effort to destabilise its neighbours. For the first time, this attack demonstrated Russia's capability for extensive cyberwarfare. This intricate effort was the first successful cyberattack on a power facility, despite taking place a full year after Russia's invasion of Crimea, which is widely recognised as the official start of Russia's conflict with the Ukraine. The command centre was attacked by the Russian cybermilitary unit Sandworm, and because of the command center's weakness, the hackers were able to take over the computer systems of the substation and shut it down. Other substations were soon the target of attacks after that. Sandworm, a Russian military unit, was allegedly responsible for the attack on the power grid in Ukraine. The attack was carried out by using a type of malware called Black Energy to gain access to the power grid's control systems. Once in, the hackers were able to remotely switch off power to over 200,000 people. The attack caused widespread disruption and raised concerns about the vulnerability of critical infrastructure to cyberattacks. The attack is considered to be one of the first instances of a cyberattack causing a power outage.

1. **LITERATURE REVIEW:**

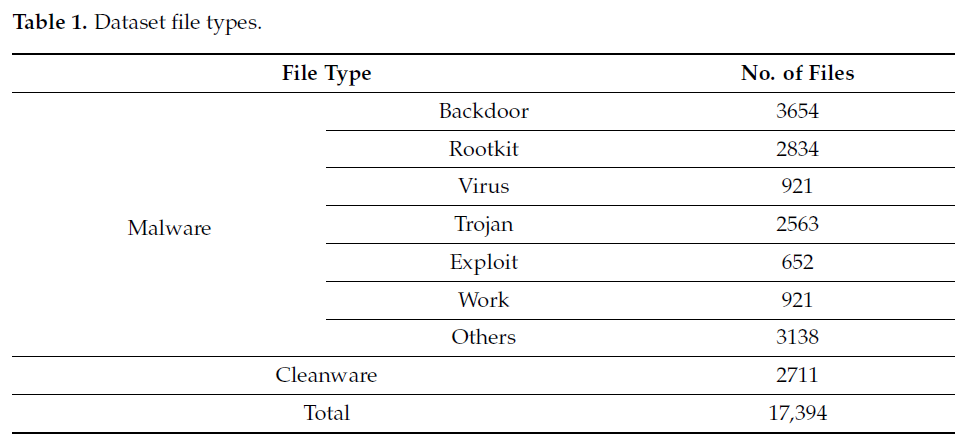
The world is vulnerable to cyberattacks because there are so many computers, smartphones, and other devices that can connect to the Internet. The surge in malware activity has led to the emergence of a plethora of malware detection techniques. When attempting to determine researchers employ a number of big data technologies and machine learning techniques to detect malicious code.

Traditional machine learning-based malware detection techniques take a long time to process, but they may successfully spot recently discovered malware. Due to the widespread use of contemporary machine learning algorithms, such as deep learning, feature engineering may become outdated. In this study, we looked at various methods for identifying and categorising malware. Researchers have developed methods to examine samples for malicious intent using machine learning and deep learning.

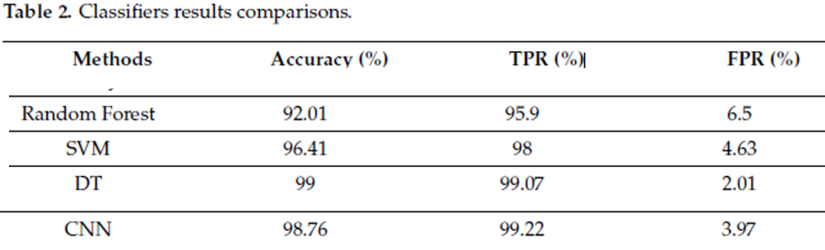
It is crucial to understand the significance of data security while implementing digital applications. As highlighted by Armaan (2021), models cannot function accurately without reliable data. Hence, precautions must be taken to safeguard data from potential cyber risks.

Machine learning is a cutting-edge approach that enables precise prediction by selecting relevant features. However, feature selection can be a challenging task, particularly when dealing with non-standard data. Therefore, it is essential to develop adaptable workarounds to handle such data. To effectively prevent future attacks, malware analysis must be performed to identify new rules and patterns. IT security professionals can use malware analysis tools to find patterns and create new rules to counter emerging malware types, as shown in Table 1. Overall, data security and analysis are critical components in developing effective digital applications that can accurately predict outcomes and prevent potential cyber risks.

The availability of technologies that analyze malware samples and determine their level of malignancy is a significant boon to the cybersecurity sector. These tools can help monitor security alerts and prevent malware attacks before they cause significant damage. Malware analysis tools are essential in identifying and categorizing malware, determining its capabilities, and assessing its potential impact. If malware is deemed dangerous, it must be eliminated before it can propagate further and cause additional harm. Moreover, as the number and complexity of malware threats continue to grow, businesses must find effective ways to lessen their effects. Malware analysis is becoming increasingly popular as it provides valuable insights into how malware operates, enabling IT security professionals to develop appropriate mitigation strategies. The availability of malware analysis tools is essential for businesses to safeguard their data and systems against cyber threats. These tools enable effective monitoring and prevention of malware attacks, helping businesses to reduce the risks associated with malware threats.



Chowdhury's (2018) proposed malware detection approach utilizing machine learning classification techniques, specifically incorporating N-gram and API call capabilities, has proven to be effective and dependable. However, further experiments were conducted to determine if adjusting certain parameters could increase the accuracy of malware classification. The experimental evaluation has confirmed the efficacy and dependability of the proposed technique, indicating that it outperformed other competing approaches. The results of this evaluation are presented in Table 2, where the Chowdhury approach was clearly superior. Future work in this area will focus on merging a large number of features to increase detection precision while decreasing false positives. This would further enhance the accuracy and effectiveness of the malware detection approach proposed by Chowdhury. The proposed approach utilizing machine learning classification techniques, N-gram, and API call capabilities, provides a reliable and effective means of detecting malware. The results of the experiments confirm the validity of this approach, and future work will continue to build upon its strengths to further improve malware detection accuracy.



The proliferation of malicious software poses a significant threat to global stability, with the prevalence of malware increasing as the number of interconnected computers exploded in the 1990s. In response to this phenomenon, multiple protective measures have been created. However, the current safeguards are not sufficient to keep up with modern threats created by malware authors to thwart security programs. In recent years, researchers have focused on using machine learning (ML) algorithm strategies for malware detection. In this research paper, a protective mechanism is presented that evaluates four ML algorithm approaches to malware detection and selects the most appropriate one. Using ML algorithm strategies for malware detection provides a promising avenue for improving current protective measures against malware. The results of this research suggest that the machine learning algorithm is a highly effective method for detecting malware and could be further developed and implemented in future protective mechanisms.

We collectively taught machine learning algorithms to distinguish between harmful and beneficial information. The DT machine learning approach was the most effective classifier we looked at, with 99% accuracy, as shown in Table 2. In order to achieve the maximum detection accuracy and the most accurate representation of malware, this experiment showed the possibility of static analysis based on PE information and selected important data elements. As the Internet evolved, malicious programmes and associated threats, or "malware," increased in prevalence and sophistication. Because of its quick spread across the Internet, malware writers now have access to a wide range of malware production tools.

The study focused on analyzing and measuring classifier performance to better understand how machine learning works in the detection of malware. Features were extracted from the recovered PE file and library information using latent analysis, and six classifiers based on machine learning techniques were evaluated. The experimental outcomes revealed that the random forest method is preferable for data categorization, with an accuracy of 99.4 percent. This indicates that the PE library is compatible with static analysis and that focusing on only a few properties could improve malware detection and characterization. The main benefit of this approach is that it reduces the risk of accidental installation of malicious software, as users can check the validity of a file before opening it. It is recommended that machine learning systems be trained and tested to determine whether a file is harmful or not.

1. **PROPOSED SYSTEM:**

Malicious components of malware can be detected through static analysis, which involves analyzing the malware binaries to identify harmful strings, or dynamic analysis, which involves monitoring the software as it operates in a controlled environment. While both methods have their pros and cons, it's best to use both when analyzing malware. To improve the accuracy of malware detection, it may be helpful to reduce the number of dangerous features and focus on more robust characteristics. This process begins with identifying potential methods or algorithms for feature selection. Ideally, solutions should be able to detect previously unseen malware while reducing the number of required characteristics.

1. **METHODOLOGY:**

This paper introduces the different steps and elements of a typical machine learning workflow for malware detection and classification, investigates the difficulties and constraints of such a workflow, and evaluates the most recent advancements and trends in the field, with a focus on deep learning methods. Below is a description of the planned research approach for this study.

Figures 3 and 4 show the workflow process from beginning to end to give a more thorough understanding of the proposed machine learning method for malware detection.



**Figure 3.** Proposed ML malware detection method.

DATASET:

The Canadian Institute for Cybersecurity's data were the sole source of information for this study. Numerous data files in the collection contain log data for various forms of malware. A wide range of models may be trained using these recovered log features. The samples contained roughly 51 different malware families. The dataset comprised 279 columns and 17,394 rows with over 17,394 data points from various locations.

PRE-PROCESSING:

The file system kept data as binary code, and the files themselves were raw executables. We prepared them before we began our investigation. It was necessary to use a protected environment, or virtual machine (VM), to unpack the executables. Automated unpacking of compressed executables by PEiD software.

 **Figure 4.** Workflow process illustration.

FEATURES EXTRACTION:

Tens of thousands of features are typically included in twentieth-century datasets. It has become obvious in recent years as feature counts have increased that the resulting machine learning model has been overfit. In order to solve this issue, we created a smaller set of features from a bigger set; this method is frequently used to preserve accuracy while utilising fewer characteristics. The objective of this study was to improve the dynamic and static feature dataset currently available by preserving the most beneficial features and removing the least beneficial ones for data analysis.

FEATURES SELECTION:

Feature extraction required the discovery of additional features; feature selection follows feature extraction. A key step in improving accuracy, streamlining the model, and minimising overfitting was feature selection, which entailed picking features from a pool of recently detected traits. In the past, numerous feature classification techniques have been employed by researchers in an effort to locate malicious software code. The feature rank technique was heavily used in this work since it is particularly good at selecting the appropriate features for creating malware detection models.

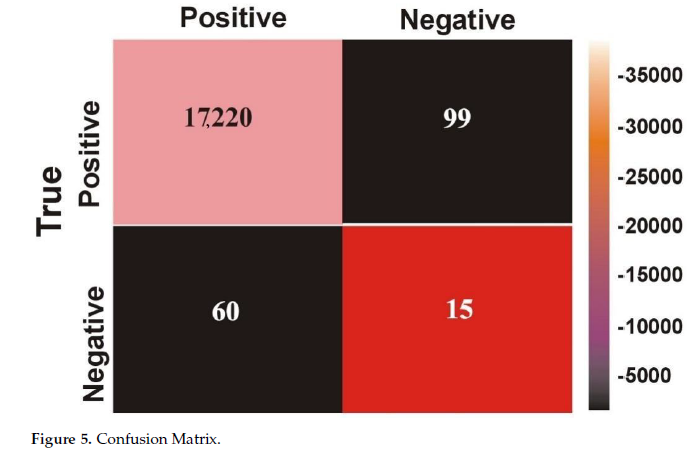
TRAINING OF THE CLASSIFIERS:

After the feature selection, next step is to find the best classifier for the detection of advanced malware. Next step is to compare different classifiers using top 20 features.

1. **RESULTS:**

Training and testing were the two key stages of the classification procedure. Both dangerous and safe files were supplied to a system to teach it. A learning algorithm was used to teach automated classifiers. With each batch of data it annotated, each classifier (RANDOM FOREST, AdaBoost, GNB, GRADIENT BOOSTING, or DT) got more intelligent. A classifier was provided a set of fresh files during testing, some of which were hazardous and some of which were not. The classifier evaluated whether the files were malicious or not.

Figure 5 shows that DT had the maximum accuracy (99%) and TPR (99.07%), while FPR had the lowest accuracy (0.43%). The confusion matrix clearly shows that DT had a greater accuracy than the other machine learning algorithms or classifiers (RANDOM FOREST, and GRADIENT BOOSTING).



Using the collected malware and cleanware, an experimental evaluation of our recommended method for malware categorization and detection was conducted. To investigate and categorise malware, we employed supervised machine learning techniques or classifiers (RANDOM FOREST, Gradient Boosting, and DT).

Based on the statistical analysis of the results presented in Table 2, it can be concluded that the decision tree (DT) classifier is the most optimal model for malware detection, with an accuracy of 99.2%. The second-best model for malware detection is the convolutional neural network (RANDOM FOREST) with an accuracy of 98.76%, followed by support vector machine (GRADIENT BOOSTING) with an accuracy of 98.41%. In terms of true positive rate (TPR), RANDOM FOREST had the highest rate of 99.22%, followed by DT with a rate of 99.07%, and GRADIENT BOOSTING with a rate of 98%. Among the classifiers, the RANDOM FOREST had the lowest false positive rate (FPR) of 0.43%, making it the best choice for malware detection. While other classifiers like KNN, Naïve Byes, and Random Forest had comparable accuracy and performance, it is recommended to use the three most optimal algorithms (DT, GRADIENT BOOSTING, and RANDOM FOREST) for identifying malware, with DT being the best choice. However, it should be noted that the dataset used in this study may not cover all possible types of malware, and further testing on a wider range of malware types is recommended for more accurate evaluation of the classifiers.

1. **FUTURE WORK:**

Use a wider dataset: Although the dataset utilised in this study is sizable and includes the majority of malware kinds relevant to the present world, it does not include every conceivable variety. The arduous process of gathering a malware dataset takes a lot of time and effort. It is recommended to test the models on every conceivable sort of malware, including spyware, adware, rootkits, backdoors, banking malware, etc., for a more accurate evaluation of the predictors. Additionally, it's critical to realise that the model can only forecast samples of the families that it has already observed. Prior to project launch for real-world contexts, the most number of families possible should be used in a real-world application.

1. **CONCLUSION:**

The study presented a protective mechanism that evaluated four machine learning (ML) algorithm approaches to malware detection and selected the most appropriate one. The results showed that compared to other classifiers, decision tree (DT) (99.1%), Random Forest (99.42%), Gradient Boosting (98.8%), AdaBoost (98.56%) and (GNB) (70.54%) performed well in terms of detection accuracy. The performance of DT, Random Forest, and Gradient boosting algorithms in detecting malware on a small false positive rate (FPR) (DT = 2.01%, Random Forest = 0.43%, and Gradient Boosting = 4.63%) in a given dataset was compared. The study evaluated and quantified the detection accuracy of an ML classifier that used static analysis to extract features based on PE data and compared it with two other ML classifiers.

The Random Forest machine learning method had the highest accuracy (99.42%) among all the classifiers evaluated. In addition to potentially providing the highest detection accuracy and accurately characterizing malware, static analysis based on PE information and carefully selected data showed promise in experimental findings. The significant benefit of this approach is that it does not require executing anything to determine if data are malicious. Overall, the study highlights the importance of using machine learning techniques in malware detection and characterisation. The results demonstrate that DT, Random Forest, and Gradient Boosting algorithms can effectively identify dangerous versus benign data, and static analysis based on PE information and carefully selected data can improve malware detection accuracy.

**REFERENCES:**

1. Nikam, U.V.; Deshmukh V.M. Performance evaluation of machine learning classifiers in malware detection. In Proceedings of the 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), Ballari, India, 23–24 April 2022; pp. 1–5.

2. Akhtar, M.S.; Feng, T. IOTA based anomaly detection machine learning in mobile sensing. EAI Endorsed Trans. Create. Tech. 2022, 9, 172814.

3. Sethi, K.; Kumar, R.; Sethi, L.; Bera, P.; Patra, P.K. A novel machine learning based malware detection and classification framework. In Proceedings of the 2019 International Conference on Cyber Security and Protection of Digital Services (Cyber Security), Oxford, UK, 3–4 June 2019; pp. 1–13.

4. Abdulbasit, A.; Darem, F.A.G.; Al-Hashmi, A.A.; Abawajy, J.H.; Alanazi, S.M.; Al-Rezami, A.Y. An adaptive behavioral-based increamental batch learning malware variants detection model using concept drift detection and sequential deep learning. IEEE Access 2021, 9, 97180–97196.

5. Feng, T.; Akhtar, M.S.; Zhang, J. The future of artificial intelligence in cybersecurity: A comprehensive survey. EAI Endorsed Trans. Create. Tech. 2021, 8, 170285.

6. Sharma, S.; Krishna, C.R.; Sahay, S.K. Detection of advanced malware by machine learning techniques. In Proceedings of the SoCTA 2017, Jhansi, India, 22–24 December 2017.

7. Chandrakala, D.; Sait, A.; Kiruthika, J.; Nivetha, R. Detection and classification of malware. In Proceedings of the 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), Coimbatore, India, 8–9 October 2021; pp. 1–3.

8. Zhao, K.; Zhang, D.; Su, X.; Li, W. Fest: A feature extraction and selection tool for android malware detection. In Proceedings of the 2015 IEEE Symposium on Computers and Communication (ISCC), Larnaca, Cyprus, 6–9 July 2015; pp. 714–720.

9. Akhtar, M.S.; Feng, T. Detection of sleep paralysis by using IoT based device and its relationship between sleep paralysis and sleep quality. EAI Endorsed Trans. Internet Things 2022, 8, e4.

10. Gibert, D.; Mateu, C.; Planes, J.; Vicens, R. Using convolutional neural networks for classification of malware represented as images. J. Comput. Virol. Hacking Tech. 2019, 15, 15–28.

11. Firdaus, A.; Anuar, N.B.; Karim, A.; Faizal, M.; Razak, A. Discovering optimal features using static analysis and a genetic search

based method for Android malware detection. *Front. Inf. Technol. Electron. Eng.* **2018**, *19*, 712–736.

12. Dahl, G.E.; Stokes, J.W.; Deng, L.; Yu, D.; Research, M. Large-scale Malware Classification Using Random Projections And Neural Networks. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing-1988, Vancouver, BC,

Canada, 26–31 May 2013; pp. 3422–3426.

13. Akhtar, M.S.; Feng, T. An overview of the applications of artificial intelligence in cybersecurity. *EAI Endorsed Trans. Create. Tech.* **2021**, *8*, e4.

14. Akhtar, M.S.; Feng, T. A systemic security and privacy review: Attacks and prevention mechanisms over IOT layers. *EAI Endorsed*

*Trans. Secur. Saf.* **2022**, *8*, e5.

15. Anderson, B.; Storlie, C.; Lane, T. "Improving Malware Classification: Bridging the Static/Dynamic Gap. In Proceedings of the

5th ACM Workshop on Security and Artificial Intelligence (AISec), Raleigh, NC, USA, 19 October 2012; pp. 3–14.

16. Varma, P.R.K.; Raj, K.P.; Raju, K.V.S. Android mobile security by detecting and classification of malware based on permissions

using machine learning algorithms. In Proceedings of the 2017 International Conference on I-SMAC (IoT in Social, Mobile,

Analytics and Cloud) (I-SMAC), Palladam, India, 10–11 February 2017; pp. 294–299.

17. Akhtar, M.S.; Feng, T. Comparison of classification model for the detection of cyber-attack using ensemble learning models. *EAI*

*Endorsed Trans. Scalable Inf. Syst.* **2022**, *9*, 17329.

18. Rosmansyah, W.Y.; Dabarsyah, B. Malware detection on Android smartphones using API class and machine learning. In

Proceedings of the 2015 International Conference on Electrical Engineering and Informatics (ICEEI), Denpasar, Indonesia, 10–11 August 2015; pp. 294–297.

19. Tahtaci, B.; Canbay, B. Android Malware Detection Using Machine Learning. In Proceedings of the 2020 Innovations in Intelligent

Systems and Applications Conference (ASYU), Istanbul, Turkey, 15–17 October 2020; pp. 1–6.

20. Baset, M. Machine Learning for Malware Detection. Master’s Dissertation, HeriotWatt University, Edinburg, Scotland, December

2016.