UNIVERSITY OF CAMBRIDGE

Classification of Malware Involving Three Machine Learning Methods Sacrifices Efficiency to Attain Higher Accuracy

Runhan Yang

Pembroke

This dissertation is submitted
on July, 2022 for the Pembroke International Summer Programme

Abstract

Machine learning models have a wide use in almost every field involving computer science. They have a terrific performance on forecasting, classification clustering and decreasing the dimension, which can all be applied in cybersecurity to protect the safety of sending and downloaded data on personal computers. For example, machine learning algorithms on classification offer a new perspective to defend cybersecurity under the threat of ransomware attacks. The model uses previous data as input and predict whether the testing data is legitimate or not. There are some basic machine learning algorithms. Since they are fundamental methods of machine learning, it is extremely hard to have breakthrough on the logic of basic algorithms to achieve better performance. Meantime, the algorithm with the relatively highest accuracy takes more time to learn from the given data and finish the final classification.

In this paper, a method involving three different single machine learning algorithms are used to present a new way to balance the speed and accuracy of the detection of files in the malware attack. Three methods mentioned above are Logistic Regression, Random Forest, and Deep Neural Networks. Furthermore, three kinds of machine learning algorithms can be replaced by another methods and adding more algorithms is possible to improve the performance of the whole algorithm.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

Runhan Yang
May 11, 2023

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Kieren Nicôlas Lovell, who guided me throughout the whole project. Without the help of my supervisor, it will be impossible for me to accomplish the first research project in my life, let alone on a completely new topic for me. Mr Lovell is an enthusiastic and professional supervisor, who taught me the motivation and target to do the research.

I would also like to thank the technical and support staff in the Online Summer Research Programme and Pembroke College from Cambridge College for offering this platform for students like me to learn anything we are interested in and finish a project independently. I wish to acknowledge the special help for professors from my University, the Chinese University of Hong Kong, Shenzhen. They listened my concerns and idea about the project patiently and advanced their own proposal for me.

I would like to show my deep appreciation to my friends and my family. Thanks for Mengqi Su, who gave me her best support to debug my main code with me. Thanks for Kai Yu, who accompanied me online to finish the whole programme. Thanks for my mother and father, for the good care and encouragement to me during this period of time.

I wish to extend my special thanks to my bed, because I came up the main idea of my algorithm when I was dreaming.

Contents

Chapter 1

Introduction

Ransomware attack is the primary mean of cyber-attacks in the digital world. After the COVID-19 pandemic, online working and online teaching become an important part of people's life. Due to the increasing amount of information on the Internet and any other applications, malware attacks have been getting ever more frequent all over the world. As reported by Help Net Security, about 470 million ransomware attacks appeared in the first six months of 2021. Specifically, attackers used a vulnerability in the Zero-day Kaseya VSA software and affected approximately 1,500 businesses, covered retail business, medical institutions, education and so on. Finally, Kaseya afforded 70 million dollars for the decryption to bring all files back to normal, reported by the U.S. Office of the Director of National Intelligence. Meanwhile, this kind of ransomware attacks influenced a much broader range of industries than just medical institutions, retail business, like services from Kaseya, and public traffic[1, 2, 3]. Due to the huge harm to society, government pays great attention to cybersecurity all over the world.

To protect the information security from malware attack, machine learning has also been applied to detect the ransomware and these methods offers a completely new perspective to cybersecurity. They model the general rule from the infected and legitimate files and apply the generated model to decide whether the testing file is infected or not. However, the basic mathematical logic of these algorithms is fixed and new algorithm needs big breakthrough in mathematics or computer science to create. Thus, optimisation on the existing algorithms is more promising and possible to have better performance than finding new algorithms.

Furthermore, different basic classification algorithms have different performance in detecting malware attacks' pattern. For example, Random Forest spends a long time in finishing the task but has a relatively accuracy. On the contrary, the basic computation of logistic regression is fairly easy to run. It needs less time and consequently, makes more mistakes. A natural question is, how to balance from consuming time and classification accuracy in one machine learning method?

Chapter 2

Literature review

As Xin et al stated, a machine learning algorithm consists of four main steps [4]:
1. extract the features;
2. select the appropriate machine learning algorithm;
3. train the model and then select the model with the best performance by evaluating different algorithms and adjusting parameters;
4. classify or predict unknown data using the trained model.

Commonly used machine learning algorithms include Random Forest, Logistic Regression, Deep Neural Networks, etc. Different algorithms based on different functions, like classification and clustering, can be applied for different stage of defending malware's attack[5].

All machine learning algorithms can be divided into two categories, single-based and ensemble-based classifiers . As their name, single-based classifiers use the unique algorithm to do classification and ensemble-based classifiers combine two or more single algorithms to produce the final decision including Bagging and Random Forest. On the other aspect, according to the check point of detection, machine learning algorithms belong to delayed detection or early detection.

Support vector machines (SVM) plays an important role on detecting ransomware. The work conducted by Nasser R. Sabar, Xun Yi and Andy Song [6] models the SVM configuration as a bi-objective optimisation problem using the proposed hyper-heuristic frame work. Another research by Hafiz Abbas, Zahid Hussain et al [7] applied random forest algorithm for comprehensive estimations in achieving sustainable public service and enhanced its cybersecurity. More precisely, given 1,171 raw phishing emails and 1,718 legitimate emails researchers compared the error rates of six machine learning classifiers to check the performance of different algorithms [8]. These six algorithms are Logistic Regression, Classification and Regression Trees, SVM, Bayesian Additive Regression Trees Random Forest, and Neural Networks, where Random Forest has the lowest error rate and Neural Networks has the highest one when there are equally many legitimate and phishing emails. The error rate of Random Forest has an average error rate of 7.72%, followed by Classification and Regression Trees with 8.13%, followed by Logistic Regression with 8.58%, followed by Bayesian Additive Regression Trees Random Forest with 9.69%,then SVM with 9.90%, and finally Neural Networks with 10.73%. However, the above results depends on the weight of legitimate and phishing emails. If the proportion changes, the performance relying on the average rate of each machine learning algorithm will also change.

Table 2.1: Average error rate when there are equally many two kinds of emails for all classifiers

Logistic Regression	8.58%
Classification and Regression Trees	8.13%
SVM	9.90%
Neural Networks	10.73%
Bayesian Additive Regression Trees	10.73%
Random Forest	7.72%

More precisely, define false positives as legitimate emails that are classified as phishing by mistake and false negatives as phishing emails that are classified as legitimate by mistake according to machine learning algorithm. The following table shows the false positive rate and false negative rate of all classifiers mentioned above.

Table 2.2: False positive rate and false negative rate for all classifiers

Classifiers	False positive	False negative
Logistic Regression	4.89%	17.04%
Classification and Regression Trees	7.68%	12.93%
SVM	7.92%	17.26%
Neural Networks	5.85%	21.72%
Bayesian Additive Regression Trees	5.82%	18.92%
Random Forest	8.29%	11.12%

The survey points out that false positives and false negatives should be paid more attention that the average error rate since they show the proportions of two kinds of mistake. False positives are much more terrible than false negative in phishing detection for users may click the malware link which is misclassified as legitimate and get infected. Meanwhile, users do not want their legitimate emails, which can be really important, to be deleted or moved to the junk box.

Another survey conducted by Mariana Belgiu and Lucian Dra˘gut [9] shows that Random Forest classifier achieves better decision when multi-dimensional data are given and its processing speed is faster than the SVM classifier and other ensemble classifiers.

To predict data using machine learning algorithms, Jack V. Tu [10] made a comparison between the Neural Networks and Logistic Regression in prediction specific data. The paper points out the advantages and disadvantages of Neural Networks with comparison of Logistic Regression and gives suggestion for choosing the proper machine learning algorithm.

Table 2.3: Advantages and disadvantages of using neural networks for predicting medical outcomes

Advantages
1. Neural network models require less statistical training to develop
2. Neural network models can implicitly detect complex nonlinear relationships between independent and dependent variables
3. Neural network models have the ability to detect all possible interactions between predictor variables
4. Neural networks can be developed using multiple different training algorithms
Disadvantages
1. Neural networks are a ’black box’ and have limited ability to explicitly identify possible causal relationships
2. Neural networks models may be more difficult to use in the field
3. Neural networks modelling requires greater computational resources
4. Neural networks model development is empirical, and many methodological issues remain to be resolved

Chapter 3

Methodology

3.1 Motivation for the New Approach

With more details, there are four steps to implement a machine learning algorithm:
1. Choose and prepare a training data set.
Training data is information that is representative of the data which the machine learning application will ingest to tune model parameters. Training data is sometimes labelled, meaning it has been tagged to call out classifications or expected values the machine leaning model is asked to predict. For labelled data, they should be divided into a training subset and a testing subset. The former is used to train the model and the latter to evaluate the effectiveness of the model and try to optimise it. Other training data may be unlabelled, so the model will have to extract features and assign clusters autonomously.
2. Select an algorithm and apply it to the training data set.
The type of chosen machine learning algorithm will primarily depend on a few aspects. One is the size of the training data set. Another is the nature of the problem the model seeks to solve. For prediction or classification cases, regression algorithms like logistic regression are often used.
3. Train the algorithm to build the model.
Training the algorithm is the process of tuning model variables and parameters to more accurately predict the appropriate results. This process is usually iterative and uses a variety of optimisation methods depending upon the chosen model.
4. Use and improve the model.
The last step is to feed new data to the model as a means of improving its effectiveness and accuracy over time. Where the new information will come from depends on the nature of the problem to be solved. For instance, a machine learning model for self-driving cars will ingest real-world information on road conditions, objects on the road, and traffic laws.

As mentioned before, the basic mathematical theory for each machine learning algorithms is hard to optimise. Besides, performance of machine learning algorithm depends not only on the training and testing data but also the strategy it applies. Optimised algorithm has higher accuracy or faster processing speed only in the specific problem and thus cannot fit every situation. In order to improve the processing speed or the accuracy of detecting the legitimate files from the database, a more general method with better performance on at least one aspect is needed.

The idea of new algorithm is from the ensemble-based machine learning classifiers. The principle is quite easy: double check is more reliable than checking only once. First, choose two algorithms as the primary methods to finish the process separately. However, since two different models are applied to detect the same testing data, there must be some files with different outcomes from the different two methods. Here needs the third method, supervising method. Files with conflict will be tested again by the supervised algorithm. According to the new decision and two previous decisions, the model will return the most frequent outcome. Specifically, in detecting unsafe emails or link in the Internet, the outcome is either legitimate or normal. There are three decisions made by three different machine learning algorithm and two kinds of outcome. By the pigeonhole principle, it is clearly to claim that there must be one decision appearing two times, which means this decision has the higher possibility to be true. Then the algorithm will return the final decision, for files with and without conflict.

Considering the outstanding performance of Random Forest algorithm, this work focuses on picking Logistic Regression and Neural Networks as the two primary methods and Random Forest as the supervising method to make the final decision.

3.2 Random Forest

Random Forest is an ensemble machine learning model of decision tree. A decision trees is a hierarchical model for supervised learning whereby the local region is identified in s sequence of recursive splits. In a univariate tree, in each internal node, the test uses only one of the input dimensions. Tree induction (also known as learning or growing) is the construction of the tree given a training sample. The goal is. for a given training set, there exist many trees that node it with no error, and, for simplicity, the smallest one, where tree size is measured as the number of nodes in the tree and the complexity of the decision nodes, is more acceptable. However, finding the smallest tree is NP-complete, and people are focus to use local search procedures based on heuristics that give reasonable trees in reasonable time.

To implement this algorithm, first select an attribute and split the data into its children in a tree and continue splitting with available attributes until leaf nodes are pure (only one class remains), a maximum depth is reached and a performance metric is achieved. Clearly, the tree construction is conducted is a recursive way.

There are two main kinds of decision trees. One is classification tree and the other one is regression tree. In this situation, classification tree is needed to detect malware attack. As only one input variable is used at each step, which attribute and what split is the best for each step? There are some rules:
1. Random: an attribute chosen at random;
2. Least-Values: the attribute with the smallest number of possible values;
3. Most-Values: the attribute with the largest number of possible values;
4. Impurity Measure: the attribute that has the largest reduction of impurity.

For the node m, N_m is the number of training instances reaching node m. For the root node, it is N. N_m^i of N_m belong to class $C_i, i = 1, ..., K$ with $\sum_i N_m^i = N_m$. Given that an instance reaches node m, the estimate for the probability of class C_i is

$$\hat{P}(C_i|x, m) \equiv p_m^i = \frac{N_m^i}{N_m}.$$

Node m is pure if p_m^i for all i are either 0 or 1. It is 0 when none of the instance reaching node m are of class C_i, and it is 1 if all such instances are of $C_i, i = 1, ..., K$. If the split is pure, there is no need to split any further and add a leaf node labelled with the class for which p_m^i is 1.

3.3 Logistic Regression

The result of Logistic Regression represents the probability of an event occurring, depending on the given data set. Since it means a probability, the range of its value should be from zero to one. That is, an desired hypothesis function of Logistic Regression should be $f_{\boldsymbol{w},b}(\boldsymbol{x}) \in [0, 1]$. To this end, a novel function is introduced to accomplish the goal, as follows:

$$f_{\boldsymbol{w},b}(\boldsymbol{x}) = g(\boldsymbol{w}^\top \boldsymbol{x} + b) \in [0, 1], g(z) = \frac{1}{1 + exp(-z)},$$

where g is called sigmoid function or logistic function.

The value of $f_{\boldsymbol{w},b}(\boldsymbol{x})$ is the estimated probability that y = 1 of input $\boldsymbol{x}$. For example, consider a patient with tumor goes to the hospital to check whether the tumor is malignant or not. If $f_{\boldsymbol{w},b}(\boldsymbol{x}) = 0.8$, then it means that the patient with tumor size $\boldsymbol{x}$ has 80 percent chance of tumor being malignant tumor. In this circumstance, larger tumor size has a larger probability of being malignant tumor. Thus, the following result is clear

$$f_{\boldsymbol{w},b}(\boldsymbol{x}) = P(y = 1|\boldsymbol{x}; \boldsymbol{w}).$$

In conclusion, Logistic Regression has

$$f_{\boldsymbol{w},b}(\boldsymbol{x}) = g(\boldsymbol{w}^\top \boldsymbol{x} + b) = P(y = 1|\boldsymbol{x}; \boldsymbol{w}) \in [0, 1], g(z) = \frac{1}{1 + exp(-z)}.$$

Suppose that if $f_{\boldsymbol{w},b}(\boldsymbol{x}) \geq 0.5$,then the prediction is 1; if that if $f_{\boldsymbol{w},b}(\boldsymbol{x}) \leq 0.5$,then the prediction is 0. Correspondingly, if $\boldsymbol{w}^\top \boldsymbol{x} + b > 0$, the prediction is 1; if $\boldsymbol{w}^\top \boldsymbol{x} + b < 0$, the prediction is 0. It determines the decision boundary boundary, which is the curve/hyperplane corresponding to $f_{\boldsymbol{w},b}(\boldsymbol{x}) = 0.5$, or $\boldsymbol{w}^\top \boldsymbol{x} + b = 0$.

Given m training examples as $(\boldsymbol{x}_i, y_i)_{i=1}^m$, like linear regression, the cost function of Logistic Regression is residual sum of squares $J(\boldsymbol{w}) = \frac{1}{2m} \sum_{i=1}^m (g(\boldsymbol{w}^\top \boldsymbol{x}_i - y_i))^2$. However, it is non-convex with respect to $\boldsymbol{w}$. The real cost function for Logistic Regression is defined by

$$J(\boldsymbol{w}) = \frac{1}{m} \sum_{i=1}^{m} cost(y_i, f_{\boldsymbol{w},b}(\boldsymbol{x}_i)),$$

$$cost(y(\boldsymbol{x}), f_{\boldsymbol{w},b}(\boldsymbol{x})) = \begin{cases} -log(f_{\boldsymbol{w},b}(\boldsymbol{x}), & if \quad y(\boldsymbol{x}) = 1; \\ -log(1 - f_{\boldsymbol{w},b}(\boldsymbol{x}), & if \quad y(\boldsymbol{x}) = 0. \end{cases} \tag{3.1}$$

The next question is how to find $\boldsymbol{w}$ and b. Maximum likelihood estimate can help. Given a model with an unknown parameter θ and samples $X_1, X_2, ..., X_n$, the probability that the model generates the samples is called likelihood.

$$L(\theta) = P(X_1, X_2, ..., X_n|\theta).$$

To determine the best θ, choose $\hat{\theta}$ such that $L(\theta)$ is maximised at $\theta = \hat{\theta}$, which can be achieved by gradient descent method.

3.4 Neural Networks

In biological neuron model, the neurons are connected to others. There are positive and negative electric potentials in each neuron. If the number of positive electric potentials is over one threshold, then it is activated to send chemical substances to other connected neurons, leading to the change of the electric potentials of these neurons. Inspired by this phenomenon, in 1943, MaCulloch and Pitts [12] proposed the M-P neuron models.

In a model neuron, or unit, there are inputs $\boldsymbol{x}$, corresponding weights $\boldsymbol{w}$, bias b, activation function ϕ, and output y. Activation functions are a critical part of the design of a neural network. The choice of activation function will control how well the network model learns the training data set. Furthermore, the choice of the activation function in the output layer will define the type of predictions the model can make. The relation of the components is as follows: $y = \phi(\boldsymbol{w}^\top \boldsymbol{x} + b)$. Perceptron model is consist of one input layer to receive input signals, one output layer including one M-P neuron and hidden layers to connect the input layer to the output layer.

Similar with Logistic Regression, treat the neural networks as a black-box and use the gradient descent method and backpropagation. The reason of applying neural networks is that its samples belonging to same class may not from a convex set. Neural networks does not require the convexity but Logistic Regression does.

Chapter 4

Analysis and results

In previous algorithm, all training and testing data set will be executed only once and the final result is decided be by the classification algorithm. However, in the the new algorithm, the training data set will be used three times by two different primary methods and the supervised method and the testing data set may be detected two or three times, depending on the results offered by the two primary methods. If the two algorithms give the same answer, both 0 or 1, where 0 presents legitimate and 1 represents safe, the file will be classified as the corresponding status. Otherwise, one algorithm gives 0 and the other gives 1. Since the two methods make the different classification to the same file after learning from the same training data set, the model applies another algorithm, supervising method, to make the third decision. Consequently, files with decision conflict have 3 classification from 3 different learning methods in the end. To have a more accurate result, the model chooses the classification which appears twice as the class of the file.

Rigorously, it does not matter which machine learning algorithm is the supervised method. In this work, Choose Logistic Regression and Neural Network as the two primary methods and Random Forest as the supervising method due to the outstanding accuracy of Random Forest classifier.

Assume that the number of legitimate emails and phishing emails is equal, and the average error rates of Random Forest, Logistic Regression and Neural Networks are 7.72%, 8.58% and 10.73% [8]. If the email is misclassified by the algorithm, it is either false positive or false negative. The probability of making the wrong classification is probability of false positive error plus probability of false negative error. There are also two conditions when the algorithm makes mistake. The first one is that two primary methods misclassified at the same time and the second one is the supervising method makes the wrong decision.

$$\begin{aligned}P_{error} &= P_1 + P_2 \\ &= P_{error,1}P_{error,1} + P_{error,3}[P_{error,1}(1 - P_{error,2}) + (1 - P_{error,1})P_{error,2}] \\ &= 8.58\% \times 10.73\% + 7.72\% \times [8.58\% \times (1 - 10.73\%) + (1 - 8.58\%) \times 10.73\%] \\ &= 2.27\%\end{aligned} \tag{4.1}$$

According to the result from calculation, the theoretical average error rate of the new algorithm involving three different machine learning algorithms is only 2.27 % ,far less than the single average error rates of individual classifiers, which are 8.58 %, 10.73 % and 7.72 % respectively. To get the theoretical average false positive and false negative rate of the new algorithm, substitute the probability of error rate in the above formula.

$$\begin{aligned}P_{error,FP} &= P_{1,FP} + P_{2,FP} \\ &= P_{error,1,FP}P_{error,1,FP} + P_{error,3,FP}[P_{error,1,FP}(1 - P_{error,2,FP}) \\ &\quad + (1 - P_{error,1,FP})P_{error,2,FP}] \\ &= 4.89\% \times 5.85\% + 8.29\% \times [4.89\% \times (1 - 5.85\%) + (1 - 4.89\%) \times 8.29\%] \\ &= 1.13\%\end{aligned} \tag{4.2}$$

The following formula calculates the average false negative error rate of the new algorithm.

$$\begin{aligned}P_{error,FN} &= P_{1,FN} + P_{2,FN} \\ &= P_{error,1,FN}P_{error,1,FN} + P_{error,3,FN}[P_{error,1,FN}(1 - P_{error,2,FN}) \\ &\quad + (1 - P_{error,1,FN})P_{error,2,FN}] \\ &= 17.04\% \times 21.72\% + 11.12\% \times [17.04\% \times (1 - 21.72\%) + (1 - 17.04\%) \times 21.72\%] \\ &= 7.19\%\end{aligned} \tag{4.3}$$

Compared with single machine learning method, the average error rate, false negative rate and false positive rate of the new algorithm decrease dramatically. It is easy to conclude that this algorithm has much higher accuracy than single algorithm, including single-based, like Logistic Regression and ensemble-based classifiers, like Random Forest. The below table shows all the average rate of involving algorithms.

Table 4.1: Average error rate, false positive rate and false negative rate for all classifiers

Classifiers	Error rate	False positive	False negative
Logistic Regression	8.58 %	4.89%	17.04%
Neural Networks	10.73 %	5.85%	21.72%
Random Forest	7.72%	8.29%	11.12%
New Algorithm	2.27%	1.13%	7.19%

The main reason for its high accuracy is that it combines the results from three independent different machine learning algorithms. The algorithm gains much more reliable result, and at the same time, needs more time to do all classifications. Single machine learning classifier takes only two steps to finish detecting the phishing emails, learning from the training data set and according to the pattern from training data set giving the result of each file in the testing data set. All the procedure only does once. However, there are much more steps the new algorithm needs than previous single machine classifier. Firstly, the new algorithm requires three different classifiers learning from the training data set, while individual classifier only needs one of them. After implementing three algorithms, two primary methods make the classification for testing data set, which may take double time to finish compared with individual classifier. Then, the new algorithm has to compare results from the two primary methods, and pick files with different classes according to the result of two primary methods. Supervising methods does the final classification for files with conflicting decision. Finally, the new algorithm compares the three results and gives the ultimate outcome.

The new algorithm has many extra steps than single classifier. Besides, it spends more time in learning from the training data set and making classification of the testing data set. When the data set is huge enough, the efficiency of new algorithm may be quite low. If the data set is not so large and the classification is as accurate as possible, it is worthwhile to sacrifice time to have much more reliable result. From the above analysis, it is obvious to conclude that the new algorithm achieves remarkable breakthrough on decreasing average error rate, false positive rate, and false negative rate on detecting phishing emails. It means that users are less likely to click the dangerous link from emails and legitimate emails are less likely to be detected as phishing. If the target is high accuracy, this method will perform much better than any individual machine learning algorithm.

Notice that there is no limitation for primary and supervised methods, the model is much flexible to do classification for data set with different pattern for different target. Due to different circumstance and target, the two primary methods can be chosen from any two classification algorithms from all methods. Depending on the pattern of data set, choosing classifiers with the best performance and applying the algorithm will get much reliable classification result.

Chapter 5

Conclusion

Machine learning algorithm have been applied to cybersecurity and have outstanding outcome. To detect phishing emails, this work introduces a new algorithm which combines three independent machine learning methods, two primary method and one supervising method, to make the final decision. This paper assumes that there are equally many legitimate and phishing emails in the data set. The two primary algorithms are Logistic Regression and Neural Networks and the supervising algorithm is Random Forest. The result shows that this combination of three algorithms successfully improves the accuracy of classification. The average error rate, false positive rate, and false negative rate of the new algorithm decrease to 2.27%, 1.13%, and 7.19% respectively. The average rates of involved three machine learning algorithms are 9.01%, 6.34%, and 16.62%. The new algorithm attains more reliable classification result in phishing emails. At the same time, this algorithm needs much more steps to finish the whole process. Consequently, it is not suitable for too huge data set. It sacrifices efficiency to get accuracy. Furthermore, the chosen of three algorithms is arbitrary. This model can be flexibly adjusted to fit in different circumstances.

Bibliography

[1] A. Alqahtani and F. T. Sheldon, “A survey of crypto ransomware attack detection methodologies: An evolving outlook,” *Sensors*, vol. 22, no. 5, p. 1837, 2022.

[2] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, “Practical black-box attacks against machine learning,” in *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 2017, pp. 506–519.

[3] V. Ford and A. Siraj, “Applications of machine learning in cyber security.”

[4] Y. Xin, L. Kong, Z. Liu, Y. Chen, Y. Li, H. Zhu, M. Gao, H. Hou, and C. Wang, “Machine learning and deep learning methods for cybersecurity,” *Ieee access*, vol. 6, pp. 35 365–35 381, 2018.

[5] R. Islam, R. Tian, L. M. Batten, and S. Versteeg, “Classification of malware based on integrated static and dynamic features,” *Journal of Network and Computer Applications*, vol. 36, no. 2, pp. 646–656, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1084804512002214

[6] N. R. Sabar, X. Yi, and A. Song, “A bi-objective hyper-heuristic support vector machines for big data cyber-security,” *Ieee Access*, vol. 6, pp. 10 421–10 431, 2018.

[7] H. S. M. Abbas, Z. H. Qaisar, X. Xu, and C. Sun, “Nexus of e-government, cybersecurity and corruption on public service (pss) sustainability in asian economies using fixed-effect and random forest algorithm,” *Online Information Review*, 2021.

[8] S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair, “A comparison of machine learning techniques for phishing detection,” in *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit*, 2007, pp. 60–69.

[9] M. Belgiu and L. Drăguţ, “Random forest in remote sensing: A review of applications and future directions,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271616000265

[10] J. V. Tu, “Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes,” *Journal of Clinical Epidemiology*, vol. 49, no. 11, pp. 1225–1231, 1996. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0895435696000029