APPLYING SUPERVISED MACHINE LEARNING TO DETECT MALWAR

APPLYING SUPERVISED MACHINE LEARNING TO DETECT MALWAR

A MS course project report submitted in partial
fulfillment of the requirements for the degree of
Master of Computer Science

By

Satya Teja Nimmakayala
MS CS, University of Colorado - Denver

Fall 22
University of Colorado Denver

ABSTRACT

In this project, malware detection in the system is accomplished via supervised machine learning. This study has made use of primary data. Malware is executed in the malware analysis lab and then subjected to dynamic malware analysis.

In this project the gathered log files will be used to gather information about malware and train the machine learning model. The sandbox runs on a Windows 10 installation using Flare Virtual machine. Running malware and alternative source on Flare virtual machine will yield the log files. Using the NLP method known as bag of words, data is retrieved from these log files and tagged thereafter. A harmful software developed or built with the main objective of obtaining access, interfering with services, or stealing data. Code, as well as executables, library files, or other kinds of programs. Training and assessment of the model come next. The model was trained using four distinct algorithms: Logistic Regression, SVC, Random Forest, and Decision Tree

This Project Report is approved for recommendation to the Graduate Committee.

Project Advisor:

Dr. Haadi Jafarian

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION

This main focuses on dynamic malware analysis, which involves running malware files through an analysis process in a controlled setting. The gathered log files will be used to gather information about malware and train the machine learning model. It is difficult for normal programming approaches to successfully address the increasing number of assaults due to the sheer volume and variety of malware. Malware on the system may be more precisely detected using ML. Any cyber security breech can negatively and immediately impact the company. Cybersecurity must be a part of a country's overall security. The limitations of conventional programming approaches are outclassed by machine learning (ML) when it comes to processing enormous amounts of data and avoiding cyberattacks.

1.1 Purpose of Project

Simple pre-execution rules that had been manually set in the early days of the cyber age were frequently sufficient to identify dangers because there weren't many malware threats at the time.

Manually constructed detection criteria were no longer practicable due to the Internet's quick development and the resulting surge in malware, necessitating the development of new, cutting-edge security solutions. To improve their malware detection and categorization, anti-malware organizations turned to machine learning, an area of computer science that has previously been employed effectively in picture recognition, searching, and decision-making. Utilizing multiple types of data on hosts, networks, and

cloud-based anti-malware components, machine learning today improves malware detection.

1.2 Project Statement

The malware market has exploded during the last ten years. Due of the sheer number and variety of malware, it is challenging for traditional programming techniques to effectively combat growing cyberattacks. Malware in the system can be more accurately detected using Machine Learning. Detecting if a particular file or piece of software is malware is the key to defending a device against malware assault.

1.3 Approach

The overall mission of this model's ultimate objective was trained using four separate algorithms: Random Forest, Decision Tree, Logistic Regression, and SVM. We start evaluating the methods using machine learning after extracting the Train and Test split data from the data. Each branch of a decision tree has a test result and a leaf node in a class label, resembling the structure of a flow chart. Additionally, there is another supervised learning method called random forest that builds a forest using a collection of decision trees and has a hyperparameter modification that produces excellent results.

1.4 Organization of this Project Report

The remainder of this paper will be divided into the following sections: The background is covered in Chapter 2. Architecture is discussed in Chapter 3. The technique, results, and analysis are included in Chapter 4. Conclusions and recommendations for further research are included in Chapter 5.

2. BACKGROUND

2.1 Key Concepts

I'll give a quick overview of my project's structure in this chapter. In Supervised Machine Learning there are many algorithm's techniques used but for the detection of malware attacks we used some of the algorithms called SVM and logistic Regression and confusion matrix. The detail explanation will be in the below key words

2.1.1 Support Vector Machine (SVM Model)

One of the traditional machine learning methods that may still be used to aid with large data categorization issues is support vector machine. It can be especially helpful for multidomain applications in a huge data setting. The support vector machine is nevertheless, computationally, and theoretically demanding. One such section's major goal is to make this strategy simpler using process and data flow diagrams so that readers may better comprehend the theory and put it into practice. Each section is split into three sections to accomplish this goal: A linear support vector machine, a nonlinear support vector machine, and the Lagrangian support vector machine method and its applications are the first three.

2.1.2 Logistic Regression

Among the most often used Machine Learning algorithms, within the category of Supervised Learning, is logistic regression. Using a predetermined set of selected variables, it is used to predict the categorical dependent variable. In a categorical dependent variable, the output is predicted via logistic regression. As a result, the result

must be a discrete or categorical value. Rather of providing the precise values of 0 and 1, it provides the probabilistic values that fall between 0 and 1. It can be either Yes or No, 0 or 1, true or false, etc.

Except for how they are applied, logistic regression and linear regression are very similar. While logistic regression is used to solve classification difficulties, linear regression is used to solve regression problems.

2.1.3 Confusion Matrix

A confusion matrix is a matrix of dimension n*n that is used to evaluate the effectiveness of a classification model, where n represents the number of class labels. The confusion matrix displays the numbers of the expected values and the actual values. The observed class is shown by the confusion matrix's row, while the anticipated class is indicated by the confusion matrix's columns, or inversely. The terms True Plus, True Negatives, False Plus, and False Negatives are used in the confusion matrix.

2.2 Related Work

In TF-IDF as a unique approach to identify the key N-grams elements during the training of a machine learning algorithm. The study then evaluates the proposed strategy utilizing a range of supervised machine learning techniques [1]. Along with an accessible and thorough introduction to this class of methods. The first part of the review is devoted to an intuitive but thorough description of decision tree-based methods and a discussion of their advantages.[2] The second part of the review is to provide a survey of the application in a context of computational and systems biology.[3] approach was demonstrated along with an algorithm for logistic classification. Additionally, they

reported actual results on two data sets for tasks involving sentence categorization and examined how our techniques behaved. The case studies in SVM with semi-supervised learning can categorize many short texts and enhance the traditional technique, allowing for the extraction of meaningful information from the texts.[4]. The goal of this research is to see how Co-Forest a random forest semi-supervised technique is for classifying huge biological data.[5]. This study's main goal was to evaluate the feasibility of utilizing a machine vision system to infer a fish’s diet swiftly, affordably, and unobtrusively from a photo of its skin. No studies that we are aware of have looked at how different diets affect the skin of live fish using image-based criteria. The evaluation of the nutrition and welfare of fish study will benefit greatly from this method. This study also analyzed the performance of several machine learning algorithms for classifying rainbow trout according to their diets to identify the best accurate image processing methods.[6]. It discovered through studying malware that security analysts, too, need to regularly update their protection measures. One crucial element is antivirus software. Endpoint protection provides a variety of security technologies, including firewalls, URL filters, email security, anti-spam software, and sandboxing.[7]

3. ARCHITECTURE

3.1 High Level Design

The implementation plan was to extract the log files analyzed using the "bag of words" NLP technique before being labeled. The model's training and assessment follow next. Using the machine learning algorithms Random Forest, Decision Tree, Logistic Regression, and SVM were used to train the model. We extract Train and Test split data from the data, and then we begin utilizing machine learning to test the methods. Next ,we need to use every algorithm we can start finding the confusion matrix and we get a True label and predicted label from that if we classification the report we get accuracy and macro average and weighted average, and then we find the results of the algorithm and see the data results of every algorithm which we use to supervise by using machine learning and compare all the results of the supervised algorithms.

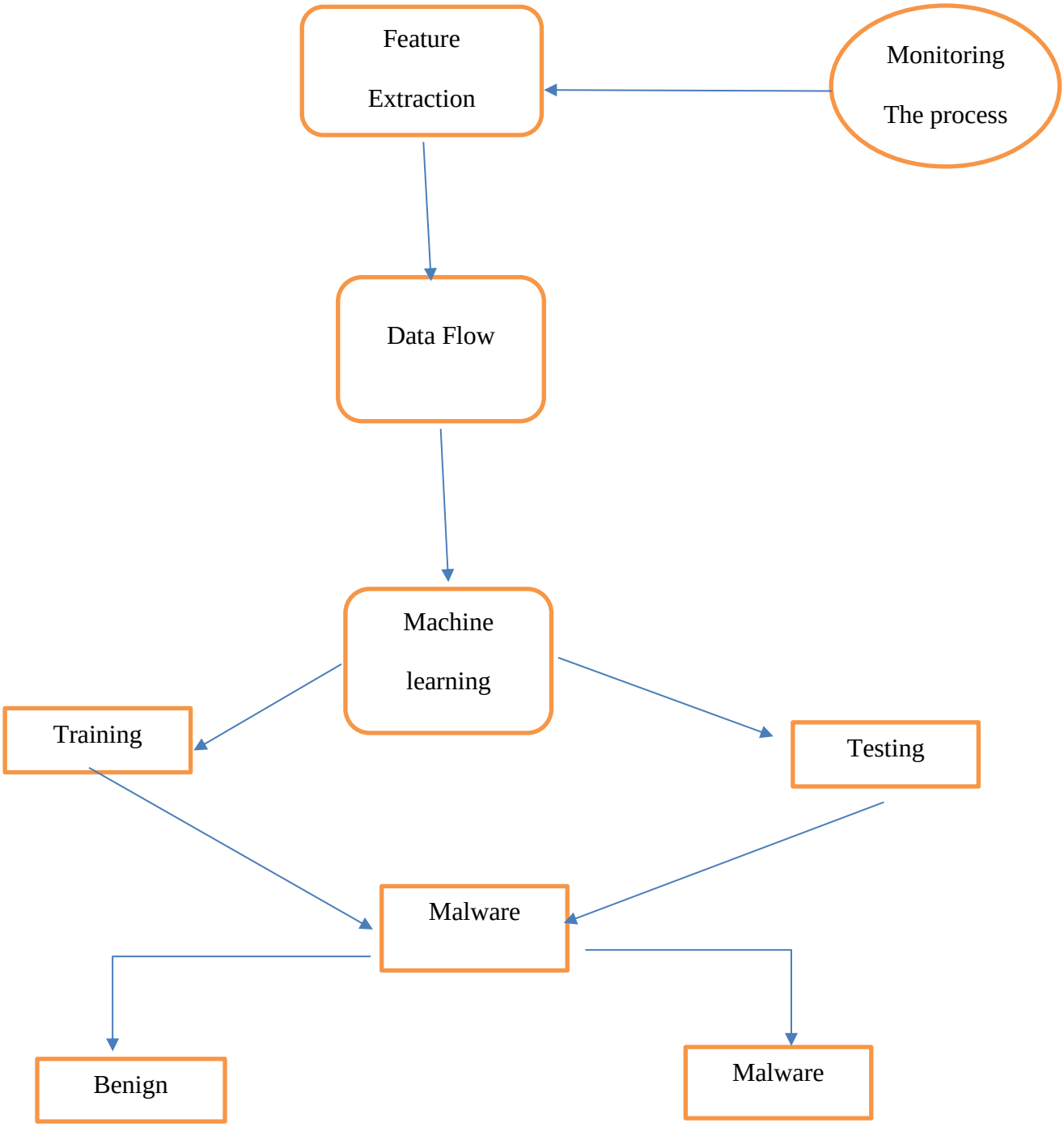
Feature
Extraction
Monitoring
The process
Data Flow
Machine
learning
Training
Testing
Malware
Benign
Malware

Figure 1

3.2 Implementation

This project has been implemented using the Python language. The library NumPy, pandas and Matplotlib has been used for plotting the prediction results for further analysis

3.2.1 Log file generation

The creation of log files establishing a lab for malware analysis and capturing a photo. Then gathering samples of both good and bad software. After the execute malware samples, then mark the log files as malicious malwarematrixdata.csv revert the lab to the most recent snapshot. We can then start executing samples of Data ware and store the log filesasData.csv.

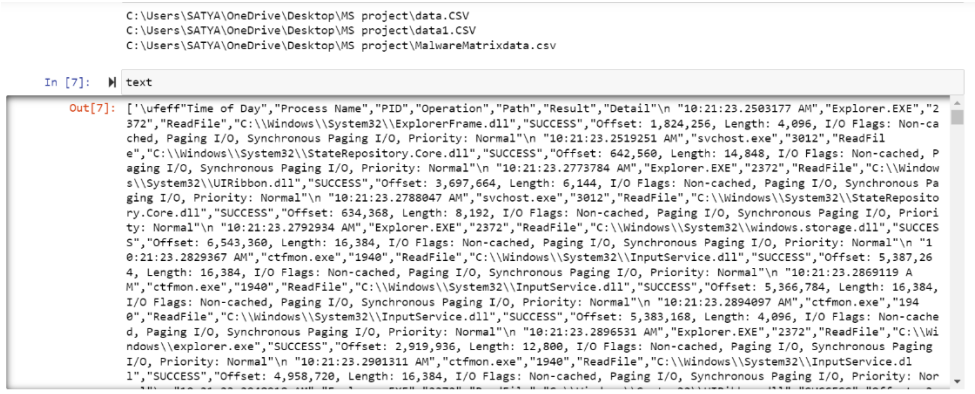

Figure (2)

3.2.2. Data Extraction from log files

The data from log files applying the count vectorizer from sklearn together with the bag of words NLP approach to extract data Additionally, the process of categorizing the dataset which resulting final dataset is in CSV format. It comprises 34,371 rows and 501 columns.

3.2.3 Using the Machine learning algorithm

A decision tree has a structure like a flow chart in internal nodes and test on attributes, and each branch has a test result and a leaf node in a class label. There is also another supervised learning method called random forest that constructs a forest using a collection of decision trees and has a hyperparameter modification that produces excellent results. When the data is very noisy or overlaps the class, we utilize SVM to construct the hyperplane data, which shouldn't be used otherwise. We begin by determining the confusion matrix for each algorithm, from which we obtain a True label and a predicted label. If we then classify the report, we obtain accuracy, a macro average, and a weighted average. Finally, we determine the algorithm's score, review the data results for each algorithm that we used to supervise using machine learning, and compare all the outcomes of the supervised algorithms. The aim of this research is to detect viruses in the system using machine learning.

4. METHODOLOGY, RESULTS AND ANALYSIS

In this chapter we will talk about the technique, findings, and analysis in this chapter. We'll talk about testing in the methodology section.

4.1 Methodology

Logistic regression gathers data, which is referred to as input, and takes constants in a classifier, which produces supervised classification. We utilize SVM to create the hyperplane data, which overlaps the class or shouldn't be used if the data is too noisy. We begin by determining the confusion matrix for each algorithm, from which we obtain a True label and a predicted label. If we then classify the report, we obtain accuracy, a macro average, and a weighted average. Finally, we determine the algorithm's score, review the data results for each algorithm that we used to supervise using machine learning, and compare all the outcomes of the supervised algorithms.

4.1.1 Algorithms Used

4.1.2 Logistic regression

Essentially, supervised categorization is what logistic regression does. For a certain collection of characteristics (or inputs), X, the target variable (or output), y, can only take discrete values in a classification issue. Because of the binary nature of this categorization, logistic regression can be used.

4.1.3 SVM

A line or a hyperplane that divides the data into classes is produced by the algorithm. When there are overlapping classes or when the data is excessively noisy, it shouldn't be used. This was deemed appropriate due to the numerous characteristics.

4.1.4 Decision Trees

A decision tree is a type of tree that looks like a flowchart, where each internal node represents a test on a property, each branch a test result, and each leaf node (terminal node) a class label. Decision trees frequently produce excellent outcomes.

4.1.5 Random Forest Tree

A supervised learning algorithm is random forest. It constructs a forest using a collection of decision trees. It is a simple machine learning technique that, even without hyperparameter modification, frequently yields excellent results. It improves accuracy over the decision tree by combining the findings of several decision trees.

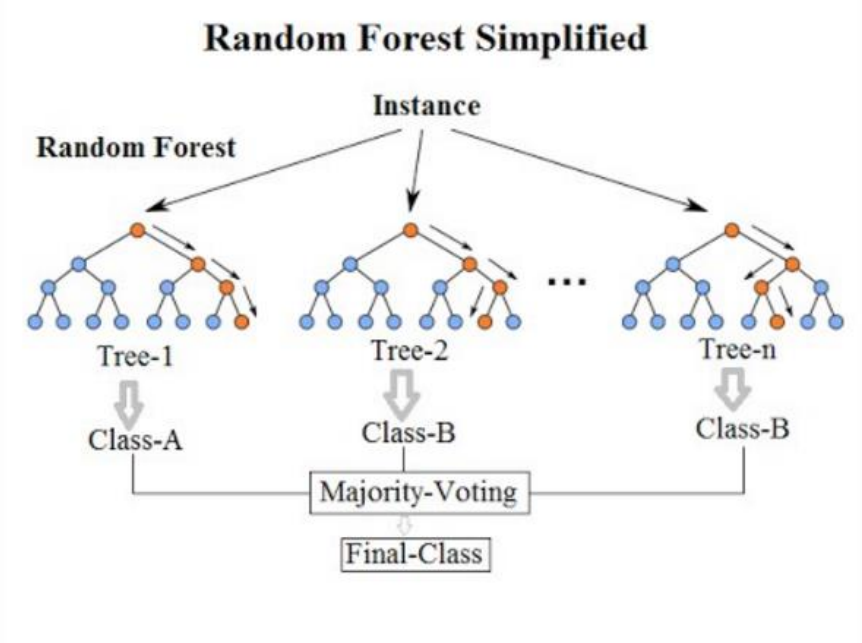

Figure (3)

source:Wikipedia

pro:

1. By averaging or merging the outcomes of many decision trees, it solves the overfitting issue.

2. Compared to a single decision tree, random forests perform better over a wide variety of data items.

3. Compared to a single decision tree, the random forest has lower variance.

4. Random forests are extremely adaptable and highly accurate.

5. A random forest approach is not necessary for data scaling.

6. Random Forest algorithms remain accurate despite a large amount of missing data.

Cons

1. The model requires a lot of storage space.

2. It required a lot more time and processing power.

4.2 Results

4.2.1 Model Accuracy

Model	Accuracy
Logistic Regression	99.976725
SVC	99.511230
Random Forest	99.959269
Decision Tree	99.72652

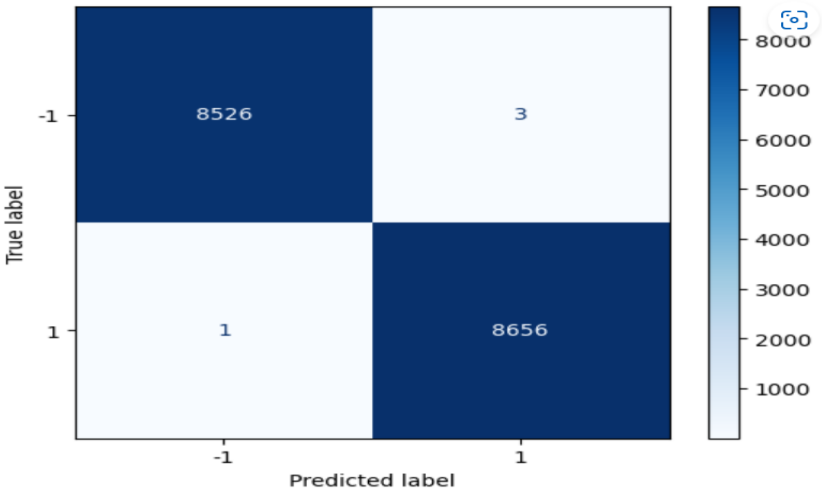

Figure (4)

Clf_results	precision	recall	f1-score	support
-1	1.00	1.00	1.00	8529
1	1.00	1.00	1.00	8657
accuracy			1.00	17186
macro avg	1.00	1.00	1.00	17186
weighted avg	1.00	1.00	1.00	17186

Figure (5)

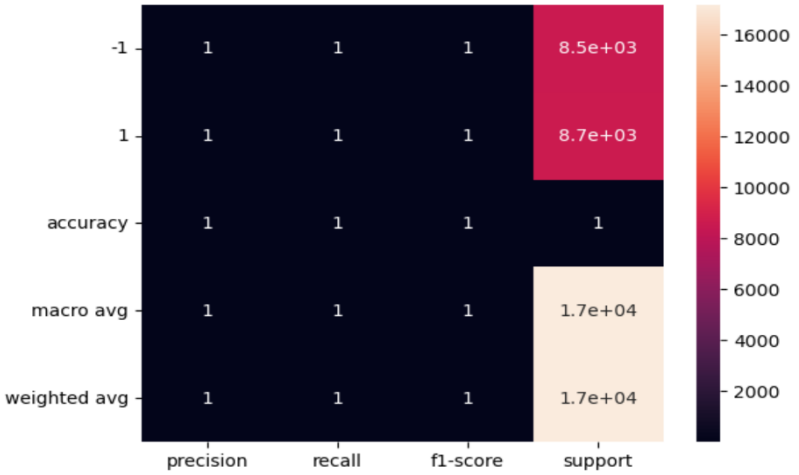

Figure (6)

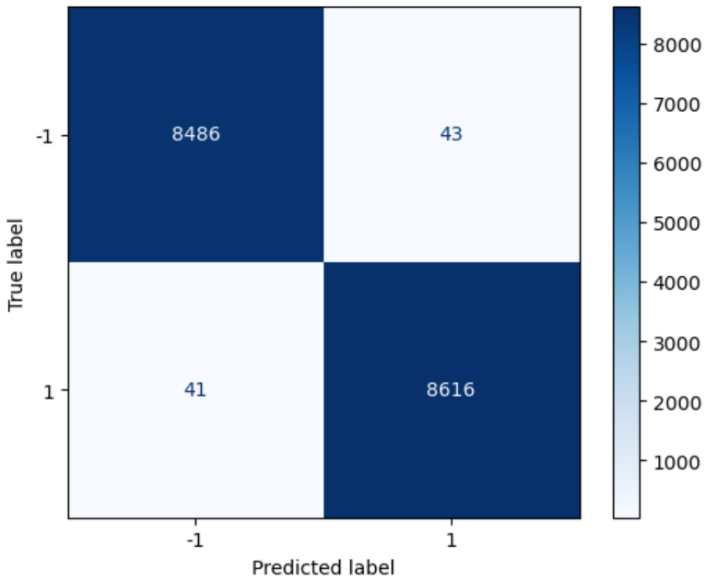

Figure (7)

Clf_results		precision	recall	f1-score	support
-1	1.00	1.00	1.00	8529	
1	1.00	1.00	1.00	8657	
accuracy			1.00	17186	
macro avg	1.00	1.00	1.00	17186	
weighted avg	1.00	1.00	1.00	17186	

Figure (8)

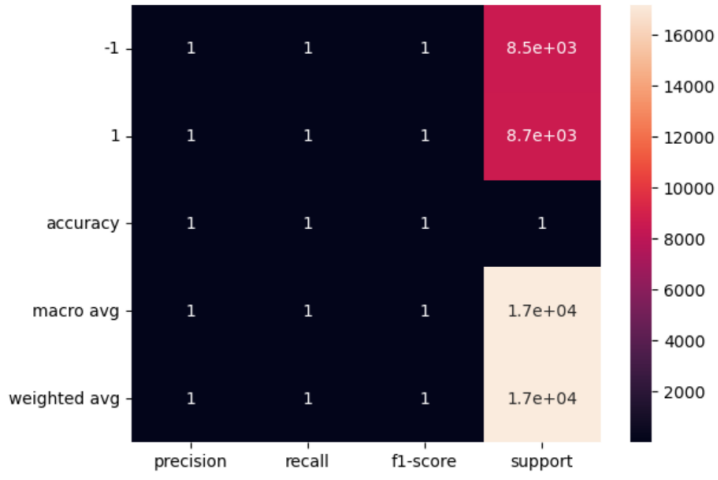

Figure (9)

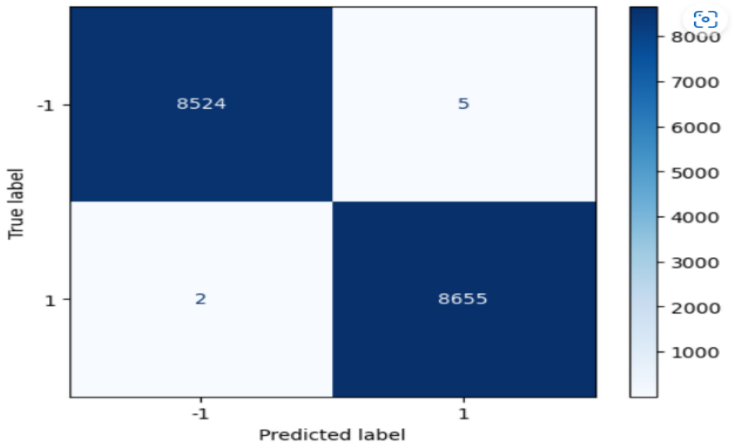

Figure (10)

Clf_results	precision	recall	f1-score	support
-1	1.00	1.00	1.00	8529
1	1.00	1.00	1.00	8657
accuracy			1.00	17186
macro avg	1.00	1.00	1.00	17186
weighted avg	1.00	1.00	1.00	17186

Figure (11)

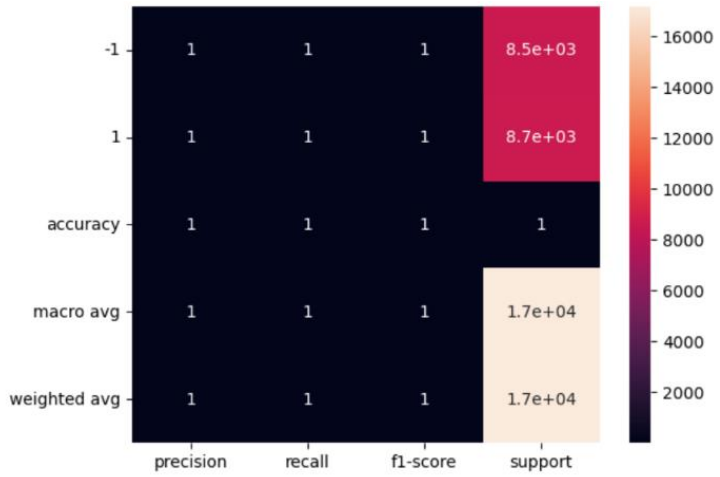

Figure (12)

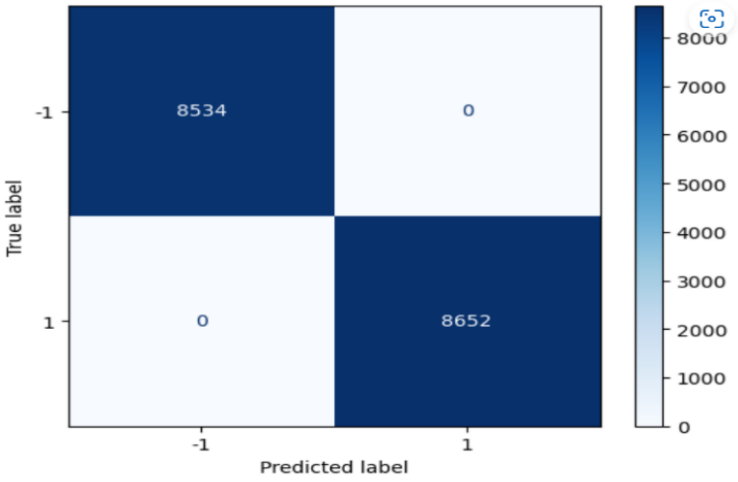

Figure (13)

Clf_results	precision	recall	f1-score	support
-1	1.00	1.00	1.00	8529
1	1.00	1.00	1.00	8657
accuracy			1.00	17186
macro avg	1.00	1.00	1.00	17186
weighted avg	1.00	1.00	1.00	17186

Figure (14)

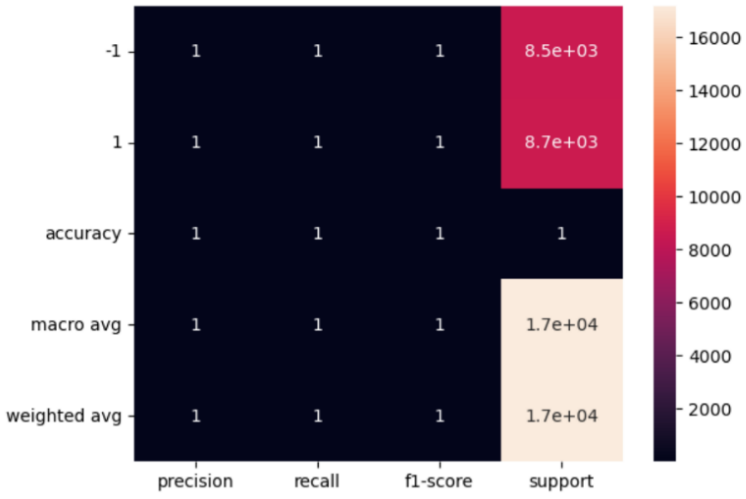

Figure (15)

4.3 Analysis

Four alternative machine learning models were acquired, and all four models exhibit accuracy of 99% with very little variations. The accuracy of the Logistic Regression model is the greatest, followed by that of the Random Forest, Decision Tree, and SVM models.

5. CONCLUSIONS

5.1 Summary

In this project Applying Supervised Machine Learning to Detect Malware. The data must be representative, relevant to the current malware landscape and correctly labeled when needed. We became experts in extracting and preparing data and training our algorithms. We made an efficient collection with billions of file samples to empower machine learning. Understand theoretical machine learning and how to apply it to cybersecurity. We understand how machine learning works in general and keep track of state-of the-art approaches emerging in the field. On the other hand, we are also experts in cybersecurity, and we recognize the value each innovative theoretical approach brings to cybersecurity practices. Understand user needs and be an expert at implementing machine learning into products that help users with their practical needs. We make machine learning work effectively and safely. We build innovative solutions that the cybersecurity market needs and by using the algorithms model we had the opportunity to bring accuracy rate.

5.2 Contributions

This project proves that we can guess it depends on produced log files, develop a machine learning model to find malware in the system. Today we can see the entire security of a nation must include cyber security. When it comes to handling cyberattacks, ML outperforms traditional programming techniques in terms of handling massive volumes of data. So, we started using some of the algorithm techniques to improve the

malware attacks accuracy by using supervised machine learning techniques to identify compromised systems and increase personal or corporate cyber security.

5.3 Future Work

Further malware development models are available. merging it into a system to enable live operation. There are ways to mix several models to increase accuracy. constructed analogous models for several operating systems. built a variation of this model without supervision. Even additional static properties, such as control flow graphs, may be included, and more methods can be used for feature selection like chi-square distribution etc. To identify even more sophisticated malware kinds, we can combine dynamic traits with static characteristics. For real-time examination of files stored in the cloud, we may create a site and host it online.

REFERENCES

[1] Ali, M.; Shiaeles, S.; Bendiab, G.; Ghita, B. MALGRA: “Machine Learning and N-Gram Malware Feature Extraction and Detection System”. Electronics October 2020, 9, 1777.

[2] Geurts, Pierre, Alexandre Irrthum, and Louis Wehenkel. "Supervised learning with decision tree-based methods in computational and systems biology". Molecular Biosystems 5.12 October (2009): 1593-1605.

[3] Amini, Massih-Reza, and Patrick Gallinari. "Semi-supervised logistic regression." ECAI. Vol. 2. No. 4. July 2002.

[4] C. Yin, J. Xiang, H. Zhang, J. Wang, Z. Yin and J. -U. Kim, "A New SVM Method for Short Text Classification Based on Semi-Supervised Learning," 2015 *4th International Conference on Advanced Information Technology and Sensor Application* (AITS), August 2015, pp. 100-103.

[5] N. Settouti, M. El Habib Daho, M. E. Amine Lazouni and M. A. Chikh, "Random Forest in semi-supervised learning (Co-Forest)," *2013 8th International Workshop on Systems, Signal Processing and their Applications* (WoSSPA), Decmber2013, pp. 326-329.

[6] Saberioon, M., Císař, P., Labbé, L., Souček, P., Pelissier, P., & Kerneis, T. March (2018). “Comparative Performance Analysis of Support Vector Machine, Random Forest, Logistic Regression and k-Nearest Neighbours in Rainbow Trout” (Oncorhynchus Mykiss) Classification Using Image-Based Features. Sensors, , 1027.

[7] Pranckevičius, T., & Marcinkevičius, V. “Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification. Baltic Journal of Modern Computing, January.2017

This is the final page of a Project Report and should be a blank page