Shape

Description automatically generated with medium confidence

**Report of Use of the Neural networks and Machine learning for malware detection.**

Course Name: [Cyber Threats, Vulnerabilities & Countermeasures](https://teams.microsoft.com/l/team/19%3aa4de569672a94477a4aba7f3b614f1bc%40thread.tacv2/conversations?groupId=caab0eea-9a46-4758-9897-9bbe60da93ed&tenantId=5f2063d3-1d5b-4e54-b7e9-a79687383110)

Course Code: CSI\_7\_SYS

**Submitted: Ioannis Iatropoulos**

Email: iatropoi@lsbu.ac.uk

**Submitted By**

|  |  |
| --- | --- |
|  | **Name: Asim Gul**  ID: 4003245​  Email:Gula2@lsbu.ac.uk |

Contents

[1. Abstract 3](#_Toc71845982)

[2. Introduction 3](#_Toc71845983)

[3. Data understanding 3](#_Toc71845984)

[4. Data preprocessing 5](#_Toc71845985)

[4.1. Feature selection 5](#_Toc71845986)

[4.2. Encoding 6](#_Toc71845987)

[5. Algorithms 6](#_Toc71845988)

[5.1. Random Forest 6](#_Toc71845989)

[5.2. Logistic regression 7](#_Toc71845990)

[5.3. Neural networks 8](#_Toc71845991)

[6. Experimental evaluation 8](#_Toc71845992)

[6.1. Methodology 8](#_Toc71845993)

[6.1.1. Accuracy 8](#_Toc71845994)

[6.1.2. Confusion matrix 9](#_Toc71845995)

[6.1.3. F1 Score 10](#_Toc71845996)

[**6.2.** **Results** 10](#_Toc71845997)

[6.2.1. Accuracy & F1 score 10](#_Toc71845998)

[6.2.2. Confusion matrix 11](#_Toc71845999)

[7. Conclusion 13](#_Toc71846000)

[8. Reference 14](#_Toc71846001)

**Use of the Neural networks and Machine learning for malware detection.**

# Abstract

Malware is considered a malicious software which are used to harm the user’s computers. There are many types of malwares, some make computer slow down and some steal sensitive information of the users such as bank details etc. Detecting malware is a challenging task for the researchers, because everyday malwares are growing in numbers and making themselves so advanced that it is difficult for the typical detecting systems to detect new advance malwares. Signature based technique was popular but that can only detect static malwares. Nowadays, researchers are developing machine learning & Deep learning-based techniques for the detection. In this project, I have built two machine learning: Logistic regression, Random forest and one deep neural network malware detection model, evaluated them and compared the accuracy of all models. I have found out that Random forest performed better than other two algorithms.

# Introduction

Malwares are also called malicious software, which are used by the hackers to slow down, damage or hack the sensitive information of the single user or organizations computers. There are many types of malware such as worm, virus and trojan. They work differently, some reproduce itself and spread, some slow down the systems and some steal important information such as bank details or documents [9]. Detecting the malware is the big challenge for researcher. In the past, signature-based detection technique was so popular, it was only effective on static old malware dataset whereas malware is changing and advancing itself every day to become more complex and powerful.

Current state-of-the-arts shows that researchers are developing machine learning and deep learning technique to detect the malware files. The research paper [10] proposed a hybrid machine learning malware detecting technique called OPEM for both static and dynamic features. In [11], the researcher represented the reviews on many machine learning techniques for windows system. The author said that the random forest algorithm gives better accuracy as compared to other algorithms such as SVM, Decision tree and naïve bays. Furthermore, the paper [12] presented deep learning and machine learning model based on the opcode frequency as a feature and achieved better results on the static features dataset.

In this project, I have built three malware detector model, one with deep learning and 2 with machine learning. I used Random forest and logistic regression techniques from machine learning, whereas I used Neural networks on tabular dataset. At the end, I made comparison of all three model with multiple evaluation metrics and presented the performance of the models on the static malware dataset. I have only targeted to the windows system malware detection, not any other such as android etc.

# Data understanding

Malware datasets in many different forms, the most common one is Portable Executable (PE) format. It is defined as “The Portable Executable (PE) format is a [file format](https://en.wikipedia.org/wiki/File_format) for [executables](https://en.wikipedia.org/wiki/Executable), [object code](https://en.wikipedia.org/wiki/Object_file), [DLLs](https://en.wikipedia.org/wiki/Dynamic-link_library) and others used in 32-bit and 64-bit versions of [Windows](https://en.wikipedia.org/wiki/Microsoft_Windows) [operating systems](https://en.wikipedia.org/wiki/Operating_system).” (Wikipedia, 2021).

The dataset [8] that is used in this project is in the form of csv files which is extracted from the PE files. The total number of files are 138047 with 56 features. These features incudes name, hash of the file etc.

Graphical user interface

Description automatically generated

Fig (5). Malware dataset data frame in python

In this dataset [8], the total number of legitimate files are 41323 whereas the total number of malware files are 96724. The last column Legitimate is the target class variable which contains only 0 and 1 binary digits. The 0 means the file malware file whereas 1 means the file is not malware but legitimate file. The remaining variables except Legitimate, will be used as an input variable after applying preprocessing techniques on them.

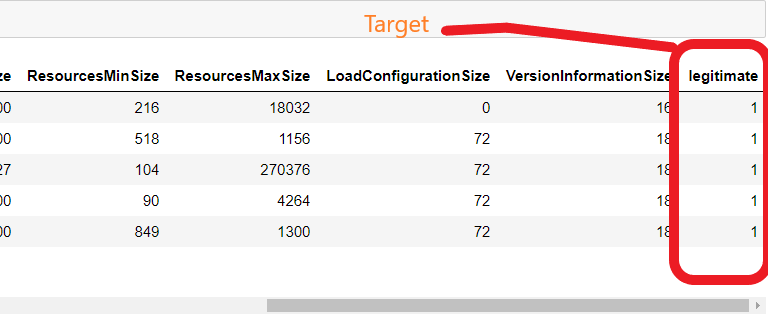


Fig (6). Malware dataset data with target variable

Chart

Description automatically generated

Details of the data variables

As the malware files are more than legitimate files, it clearly shows that the dataset in completely imbalance. In the fig. 7, the 0 bar represent the malwares, which are double of the legitimate files.

Chart

Description automatically generated

Fig (7). Legitimate variable bar graph

# Data preprocessing

## Feature selection

For training good model, features selection plays an important role in removing unwanted and biased variables. The unwanted variable may result into poor performance of model.

In this project, the Name and md4 two variables are removed, as every file has its own unique name or hash, so it does not play an important role in model prediction.

Text

Description automatically generated

Fig (8). Feature selection in malware dataset [8].

## Encoding

As machine learning does not understand any language except numerical. So before training we need to convert all categorial values of structured data in into numerical form.

If we look in the malware dataset [8], they all are already in the form of numeric. So, we do not need to perform any kind of encoding on it.

# Algorithms

I have used three algorithms to detect the malware from the dataset.

## Random Forest

Random forest algorithm is an ensemble-based learning model which is made by merging many decision trees and give well balanced predictions. It can be used for regression and classification problems. As it takes input dataset and give output in the form of target class (classification) or average of the predictions (regression) [1].

Diagram, radar chart

Description automatically generated

Fig (1). Random forest architecture [3]

Random forest is very simple to implement and understand. In addition, it saves from the overfitting problem and give best performance as compared to other algorithms [2].

## Logistic regression

Logistic regression is a machine learning classification algorithm in which the target variable is categorical variable. For example, if the problem is to classify the spam emails, then spam and not spam will be the target classes [4].

Diagram

Description automatically generated

Fig (2). Logistic regression architecture [4]

There are many types of logistic regression model. Such as binary, multinomial, and ordinal logistic regression model [4].

In my project, I have used binary logistic regression model for the classification of the malwares.

## Neural networks

Neural networks are the deep learning model, which works like the human brain to recognize patterns. One neuron in the model is the mathematical function which performs processing on the input data and produces some outputs [5].

It is used for regression, classification, and clustering problems. Furthermore, Neural networks architecture is made of layers. It takes input data through input layers; it has many hidden layers which process the given input and then produce the output from the output layer [5].

Diagram, schematic

Description automatically generated

Fig (3). Neural network architecture (Sabrina Jiang, 2020)

# Experimental evaluation

## Methodology

This section will describe the methods or metrics to evaluate the performance of the Malware detection models.

## Accuracy

Classification accuracy is the metric which is used to evaluate the performance of the balanced targeted classes. It is the ratio of total number of classes predicted by total number of classes in the dataset [7].

A picture containing table

Description automatically generated

Accuracy metric does not provide accurate results on the imbalanced dataset. That why we use f1 score instead of accuracy [7].

## Confusion matrix

Confusion matrix is the table which is used to evaluate the performance of any machine or deep learning classifier. In that table, each section provides an important information about the model performance [6].

Chart

Description automatically generated

Fig (4). Confusion matrix for Binary classification [6]

Each section in the matrix describes as following:

**True Positive**: True positive is when a model successfully predicted positive classes as positive.

**False Positive**: when the actual classes were negative, but model predicted positive.

**False negative**: When a model predicted positive but that were negative.

**True negative**: When a model predicted all classes as negative and there were also negative.

False negative and false positive are the dangerous section for the model. If these are more as compared to true negative and positive, it means that is the performance of the model is poor.

## F1 Score

F1 score is also called F measure which is used to measure the performance of the imbalance dataset classification model. If the targeted classes of the dataset are imbalanced then accuracy metric does not give us accurate results of the model. But F1 provides us the accurate performance of the model [7].

F1 score calculates the performance between 0 and 1. If the values are 1 or near to 1, it means that the performance of the model is high and vice versa [7].

## **Results**

## Accuracy & F1 score

I have measured the accuracy of my three models on train and test dataset. Results are shown below in the table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Models** | **Train Data Accuracy** | **Test Data**  **Accuracy** | **F1- score** | **Position** |
| **Random Forest** | 0.982 | 0.983 | 0.973 | 1st |
| **Logistic Regression** | 0.70 | 0.69 | 0.0 | 3rd |
| **Neural Networks** | 0.951 | 0.953 | 0.91 | 2nd |

Table 1. Accuracy & F1 score of the model.

The above table presents the comparison between all three models based on accuracy and f1 score evaluation parameters.

The Random forest model performed efficiently on the Malware detection dataset with the accuracy of 98 % whereas Neural network is on the second number with 95% accuracy. At the end, logistic regression performed poor than all models with the accuracy of 96%.

As dataset was imbalance, so I am considering F1 score as my evaluation metric instead of accuracy. If we compare the F1 scores of all the model, the random forest model achieved high scores.

The table 1 results shows that the performance of the Random forest is better than the other two models.

## Confusion matrix

|  |  |  |
| --- | --- | --- |
|  | Predicted  Yes | Predicted  No |
| Actual Yes | TP= 19080 | FP=170 |
| Actual NO | FN=277 | TN=8083 |

Random forest

Chart

Description automatically generated

Table 2. Confusion matrix of Random forest.

Logistic Regression

|  |  |  |
| --- | --- | --- |
|  | Predicted  Yes | Predicted  No |
| Actual Yes | TP= 19250 | FP=0 |
| Actual NO | FN=8360 | TN=0 |

Chart

Description automatically generated

Table 3. Confusion matrix of logistic regression.

Neural Network

|  |  |  |
| --- | --- | --- |
|  | Predicted  Yes | Predicted  No |
| Actual Yes | TP= 19030 | FP=220 |
| Actual NO | FN=1055 | TN=7305 |

Table 4. Confusion matrix of Neural Networks.

These confusion matrixes show the total number of correct and wrong predictions predicted by the models. In table 2, the True positive and true negative predictions are much higher than as compared to False parameters. It shows that the model has greater capability to differentiate between malware and original files. But there are few files which model cannot classify correctly. It is the most dangerous situation for the system.

The table 3 and 4 are also showing some valuable performance. The logistic regression model performed poor in detecting malware and original files whereas, the Neural network performed quite better but it the number of False negatives is greater in the confusion matrix, which means the file was Malware, but model considered that file No malware file.

As a result, the Random forest model performance is much better than the logistic and Neural network model.

# Challenges

Malware detection models and techniques are facing few challenges, that is why they are not performing well on the live streaming dataset. Firstly, lack of enough accurate dataset. Companies does not share the data that contains malware because of data protection act or that data contain sensitive information. Secondly, malware detecting models are trained on old static dataset, which have no information of new and advance viruses. It makes systems to become vulnerable to the attacks. Lastly, neural networks do not work well on tabular dataset whereas few algorithms of machine learning perform better than neural network. So choosing best algorithms will allow to make an efficient model for the harmful computer viruses.

# Conclusion

Machine learning and deep neural networks play important role in the protection of the systems form the unwanted attacks and highly dangerous malwares. In the past there were many techniques to safe systems from these attacks, but with the advancement in the viruses, those techniques seem useless. In this project, I evaluated three artificial intelligence model such as random forest, logistic regression, and neural networks on the malware datasets. That datasets [8] contain original and malware files. The evaluation is done on accuracy, F1 score and confusion matrix. After comparing the accuracy and f1 score, I found out that the Random forest performed much better than other two algorithms with accuracy of 98% on the test data and 97% f1 score. Whereas neural network also performs such as 95% on test data and 91% f1 score. But it is less as compared to Random forest.

# Reference

1. Wikipedia (2014) Random forest. Available from: <https://en.wikipedia.org/wiki/Random_forest> [Accessed 9 May 2021].
2. Niklas Donges (2021) A COMPLETE GUIDE TO THE RANDOM FOREST ALGORITHM. Available from: <https://builtin.com/data-science/random-forest-algorithm> [Accessed 9 May 2021].
3. [ABHISHEK SHARMA](https://www.analyticsvidhya.com/blog/author/abhishek-shrm/)**,**MAY 12, 2020 Decision Tree vs. Random Forest – Which Algorithm Should You Use? Available from: <https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/> [Accessed 9 May 2021].
4. Saishruthi Swaminathan, MAY 15, 2018 Logistic Regression — Detailed Overview. Available from: https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc [Accessed 9 May 2021].
5. James Chen (2020) Neural Network. Available from: <https://www.investopedia.com/terms/n/neuralnetwork.asp#:~:text=A%20neural%20network%20is%20a,organic%20or%20artificial%20in%20nature>. [Accessed 9 May 2021].
6. Towardsdatascience (2020) Confusion matrix for your multi class machine learning model. Available from: <https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826> [Accessed 10 May 2021]
7. Tavish Srivastava (2019) 11 important model evaluation metrics for machine learning everyone should know. Available from: <https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/> [Accessed 10 may 2021]
8. Chihebchebbi (2018) MalwareData. Available from: <https://github.com/PacktPublishing/Mastering-Machine-Learning-for-Penetration-Testing/blob/master/Chapter03/MalwareData.csv.gz> [Accessed 1 may 2021]
9. Fruhlinger (2019) Malware explained: How to prevent, detect and recover from it. Available from: <https://www.csoonline.com/article/3295877/what-is-malware-viruses-worms-trojans-and-beyond.html> [Accessed 12 May 2021]
10. Santos, I., Devesa, J., Brezo, F., Nieves, J., & Bringas, P. G. (2013). OPEM: A Static-Dynamic Approach for Machine-Learning-Based Malware Detection. In Á. Herrero, V. Snášel, A. Abraham, I. Zelinka, B. Baruque, H. Quintián, J. L. Calvo, J. Sedano, & E. Corchado (Eds.), *International Joint Conference CISIS’12-ICEUTE{\textasciiacute}12-SOCO{\textasciiacute}12 Special Sessions* (pp. 271–280). Springer Berlin Heidelberg.
11. S. Naz and D. K. Singh, "Review of Machine Learning Methods for Windows Malware Detection," 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2019, pp. 1-6, doi: 10.1109/ICCCNT45670.2019.8944796.
12. Rathore, H., Agarwal, S., Sahay, S. K., & Sewak, M. (2018). Malware Detection Using Machine Learning and Deep Learning. In A. Mondal, H. Gupta, J. Srivastava, P. K. Reddy, & D. V. L. N. Somayajulu (Eds.), *Big Data Analytics* (pp. 402–411). Springer International Publishing.