**Malware Detection Using Machine Learning**

**M.Bilal Atiq – 1712278**

**Areeb Khemani – 1712260**

### Abstract:

#### Malware detection is an important factor in the security of the computer systems. However, currently utilized methods cannot provide accurate detection of zero-day attacks and polymorphic viruses. That is why the need for machine learning-based detection arises.

This work presents recommended methods for machine learning based malware classification and detection.

Introduction:

With the rapid development of the Internet, malware became one of the major cyber threats nowadays. Any software performing malicious actions, including information stealing, espionage, etc. can be referred to as malware.

Therefore, malware protection of computer systems is one of the most important cybersecurity tasks for single users and businesses, since even a single attack can result in compromised data and sufficient losses. Massive losses and frequent attacks dictate the need for accurate and timely detection methods. Current static and dynamic methods do not provide efficient detection, especially when dealing with zero-day attacks. For this reason, machine learning-based techniques can be used.

Objective:

The goal of this project is to develop the proof of concept for the machine learning based malware classification based on VirusShare. VirusShare will be utilized for the extraction of the behavior of the malware samples, which will be used as an input to the machine learning algorithms. The goal is to determine the best feature representation method and how the features should be extracted, the most accurate algorithm that can distinguish the malware families with the lowest error rate. The accuracy will be measured both for the case of detection of wheher the file is malicious and for the case of classification of the file to the malware family. The accuracy of the obtained results will also be assessed in relation to current scoring implemented in VirusShare, and the decision of which method performs better will be made.

Algorithms Used:

Random Forest

Random Forest is one of the most popular machine learning algorithms. It requires almost no data preparation and modeling but usually results in accurate results. Random Forests are based on the decision trees described in the previous section. More specifically, Random Forests are the collections of decision trees, producing a better prediction accuracy. That is why it is called a ’forest’ – it is basically a set of decision trees.

Gradient Boost

Gradient Boost is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

Dataset:

The Dataset used in our Malware Detection project consists of 137,000 samples. 41,000 binaries (exe, dll files) are not a malware file. 96,000 samples which are virus infected are taken form virusshare.com.

Procedure:

* Firstly we will perform feature extraction and drop all the unnessary coloumns from our dataset
* We then split our data into legit and virus files.
* We used ExtraTreeClassifier to prevent our model from over fitting.
* We used two algorithms Random Forrest and Gradient Boost and compared their performances.

Conclusion:

We got 99.40239043824701% accuracy with Random Forest Classifier and 98.75407461064832 % accuracy with Gradient Boost Classifier therefore Random Forrest Classifier is more preferred due to its high accuracy.