###### RANSOMWARE DETECTION USING MACHINE LEARNING

Jaiwanth Vavaldas , Nithin Kodithyala , Aashish Aarukonda ,

Sandeep Jangapelli , Varun Sai Baddapuram

[{jaiwanthvavaldas@gmail.com](mailto:%7bjaiwanthvavaldas@gmail.com) , [kodithyalanithin153@gmail.com](mailto:kodithyalanithin153@gmail.com) , [aashisharukonda@gmail.com](mailto:aashisharukonda@gmail.com) , [varunsaibaddapuram@gmail.com](mailto:varunsaibaddapuram@gmail.com) , }

Department of CSE-AIML , Neil Gogte Institute Of Technology

**Abstract**: Ransomware is a type of malware that encrypts the victim's data and demands a ransom for its decryption. Ransomware attacks have become one of the most significant cybersecurity threats in recent years. The traditional signature-based detection methods are no longer effective due to the advanced techniques used by the attackers. In this paper, we propose a machine learning-based approach for detecting and eliminating ransomware. Our approach uses combination of supervised and unsupervised machine learning algorithms to detect ransomware in real-time. We also propose a novel method for eliminating ransomware by automatically reverting the encrypted files to their previous states.

**Introduction**: Ransomware attacks are a growing concern for individuals and organizations alike. The attackers use various techniques to infect the victim's system, such as social engineering, exploit kits, and phishing emails. Once the system is infected, the ransomware encrypts the files and demands a ransom in exchange for the decryption key. The attackers use advanced techniques such as polymorphism, encryption, and evasion to evade traditional signature-based detection methods.

**Background** **and** **related** **works**: Malicious software analysis is an important research topic This is caused by damage caused by malware with the youngest Advances in artificial intelligence (AI) are turning cybersecurity researchers' attention to machine learning (ML) and deep learning (DL) techniques to improve detection and classification of malicious files. if still Although the results are good, these methods still need improvement

**Perpose and outline:** The purpose of this study is to provide different approaches to this. Malware detection with a special focus on ransomware recognition. Our main contributions are algorithms such as: We concatenate two of her models, each optimized for malicious file and ransomware detection. This solution, called a "two-layer recognition model", serves as input. PE files report file maliciousness in the following format: The second step determines if this PE file can do this. Whether it is classified as ransomware. do my best As far as we know, we haven't found an equivalent approach Discussion of ransomware detection in existing literature. our The final model therefore consists of two special models. great because you can train independently Flexibility in optimization phase and production environment Surroundings**.**

**Dataset**

The main datasets are 3 types the are Ember, Bodmas and PEMachine Learning (for our project we used PEMachine Learning)

**1.Ember:** Ember is a dataset from Anderson et Al .Al Contains a total of 1.1 million extracted features 400,000 malicious files, 400,000 benign files, 300,000 files Files without labels. Anderson et al. It also provides the tools necessary to generate feature-based datasets.

**2.Bodmas:** Bodmas contributed datasets to our team Contains 134,435 binaries of the same format As an embers with pre-extracted features, 57,293 malicious files in RAW PE format. I have these files Collected within one year from August 2019 and September 2020, Caption: Author's Statement Category to which each file belongs. Table II shows the distribution of malware families within the Bodmas dataset. of The "Other" category does not contain about 10 other categories Represented by the Bodmas dataset. For example: "eyedropper", For example, “Downloader” or “Information Thief”.

**3.PE Machine Learning Dataset**: The third dataset used is PEMachineLearning, made available by M. Lester [21]. It contains 201, 549 binary files including 114,737 malicious files. These files have been gathered from different sources such as VirusShare1 , MalShare2 and TheZoo3.

“Our Dataset contains 96,724 binary files and 47,238 benign files and remaining ransom files.”

**Features**

Training a model for malware detection is done in several steps. procedure. The first step is to extract the information (that is, functionality) from the PE file. we are dependent on preprocessing algorithm. We Used **Random Forest Classifier Model** for training the model

Which was proposed by leo breinan in 2001.

**Random Forest**: Random forest is the collection of decision trees the main advantages of random forest are :

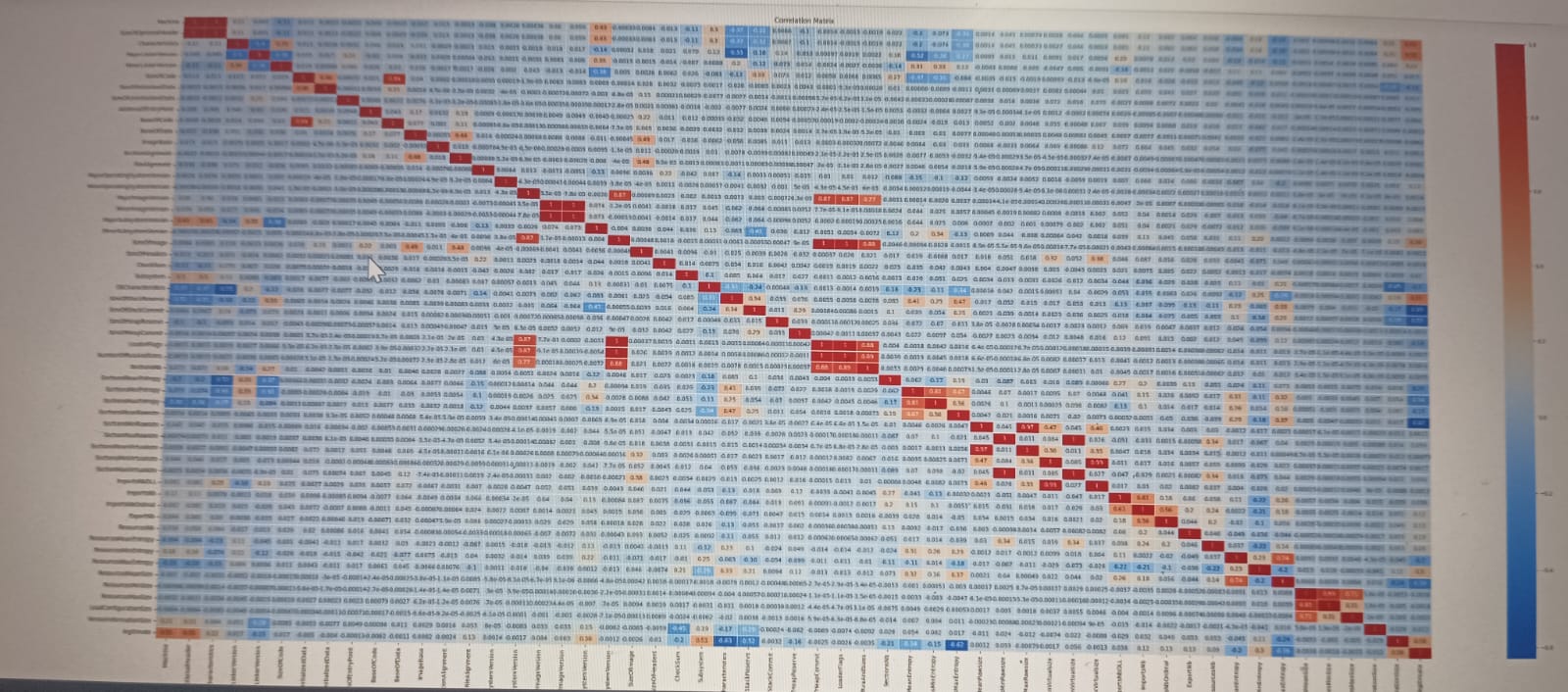
1.High Accuracy: Rand om Forest is known for its ability to achieve high accuracy in both classification and regression task.

2. Robustness to Overfitting: One of the major advantages of Random Forest is its robustness to overfitting, a common problem in machine learning.

3. Handles Large Datasets: Random Forest is well-suited for handling large datasets with many features.

4. Feature Importance: Another benefit of Random Forest is its ability to provide a measure of feature importance. It identifies which features contribute the most to the model's predictions, offering insights into the underlying relationships in the data.

5. Easy to Use: Random Forest is relatively easy to use, making it accessible to both beginners and experienced practitioners.



**Features Related to Ransomware Attack**

There are many Features which are related to Malware and Ransomware Attack and The Features which we have used for evaluation are :

1. 'SizeOfOptionalHeader': Indicates the size of the optional header in the PE file, which contains additional information about the file's characteristics.

2. 'MajorSubsystemVersion': Denotes the major version number of the subsystem required to run the executable.

3. 'DllCharacteristics': Represents characteristics of the executable, such as whether it is a DLL (Dynamic Link Library) or not.

4. 'SizeOfStackReserve': Refers to the amount of memory reserved for the stack during execution.

5. 'SectionsMeanEntropy': The mean value of entropies calculated for various sections of the PE file, which measures the level of randomness in those sections.

6. 'SectionsMaxEntropy': The maximum value of entropies among all sections, indicating the most random section in the PE file.

7. 'Subsystem': Specifies the required subsystem to run the executable (e.g., Windows GUI, console, or EFI application).

8. 'ResourcesMaxEntropy': An initial value of 6 is given, which may be replaced later based on entropy calculations for the resource section (if available).

**Model**

In machine learning, a model is a mathematical representation or algorithm that learns patterns and relationships from data to make predictions or decisions without being explicitly programmed. Models are the core building blocks of machine learning systems, and they are trained on historical data to generalize and make predictions on new, unseen data.

Here are some common types of machine learning models:

1. Linear Regression: A model used for regression tasks, where the goal is to predict a continuous output variable based on input features. It finds a linear relationship between the input features and the target variable.

2. Logistic Regression: A classification model used to predict binary outcomes (0 or 1) based on input features. It uses the logistic function to transform the output into probabilities.

3. Decision Trees: A non-linear model that uses a tree-like structure to make decisions based on feature values

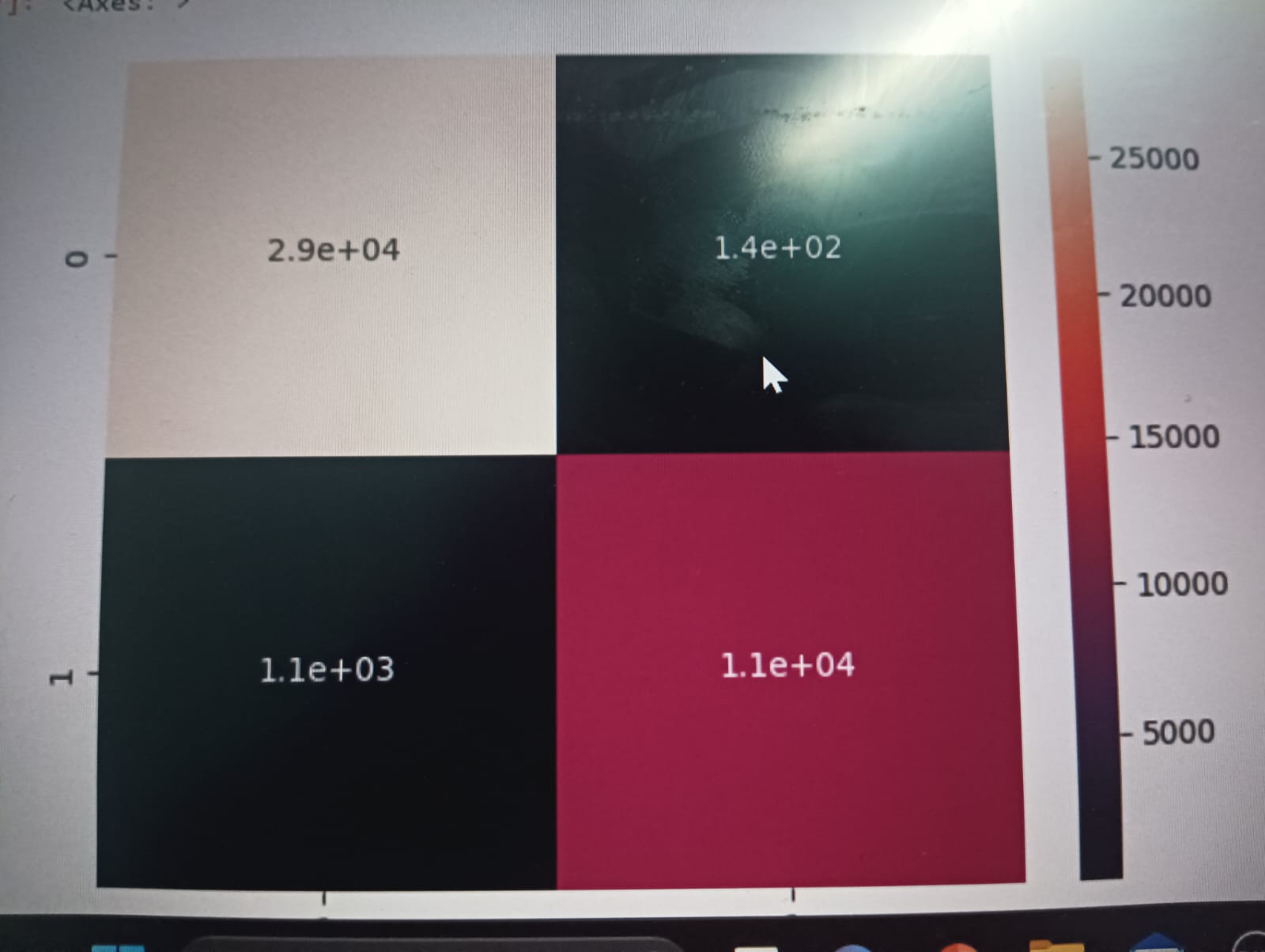
4. Random Forest: An ensemble learning method that combines multiple decision trees to improve accuracy and reduce overfitting. It uses averaging or voting to make predictions.

5. Support Vector Machines (SVM): A powerful classification model that finds the optimal hyperplane to separate different classes in a high-dimensional feature space.

**Correlation Matrix**

In machine learning, a correlation matrix is often used as a preliminary step in data analysis and feature selection. It helps to identify the relationships between different features (variables) in the dataset. The correlation matrix can be particularly helpful when dealing with tabular data, such as data stored in spreadsheets or database tables.

The output of our Correlation Matrix is given:



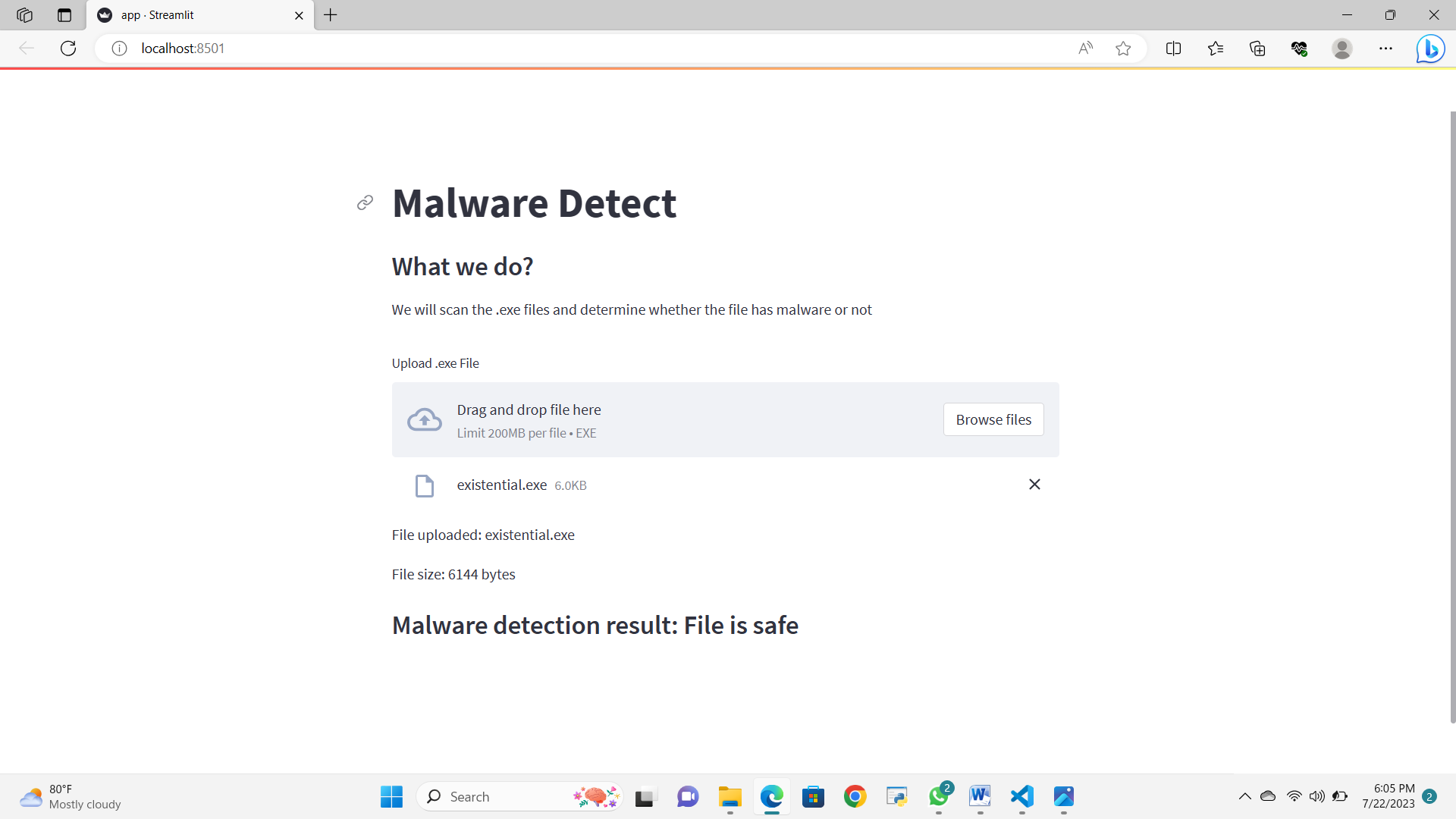
**Front-End**

There are many application or framework for deploying the output of our program:

**1.Flask**: Flask is a micro web framework for Python, designed to be simple and lightweight. It provides tools and libraries to build web applications easily and efficiently. Unlike full-stack web frameworks, Flask is not bundled with built-in components for database access, form validation, or other features. Instead, Flask follows a "minimalistic" philosophy, allowing developers to choose and integrate various libraries according to their specific project requirements.

**2.Streamlit** : Streamlit is a popular Python library that allows you to create interactive web applications for data science and machine learning projects with minimal effort. It enables data scientists and developers to quickly turn data scripts into shareable web applications, all in Python.

“For our project we used Streamlit Framework ”



**Conclusion**

* A ransomware detection project aims to protect organizations from ransomware attacks and their detrimental consequences. By implementing robust detection mechanisms and mitigation strategies, organizations can enhance their cybersecurity defenses
* The project focuses on early detection, incident response, user awareness, data backup, and continuous improvements.

**References:**

[1] Steve Morgan. Cybercrime to cost the world $10.5 trillion annually by 2025. https://cybersecurityventures.com/ cyberwarfare-report-intrusion/. Online; Post : 2020-11-13.

[2] Harjinder Singh Lallie, Lynsay A. Shepherd, Jason R.C. Nurse, Arnau Erola, Gregory Epiphaniou, Carsten Maple, and Xavier Bellekens. Cyber security in the age of COVID-19: A timeline and analysis of cyber-crime and cyber-attacks during the pandemic. Computers and Security, 105:102248, 2021.

[3] Av-test. https://www.av-test.org/fr/?r=1. Accessed: 2022-05-30.

[4] Todd Drew. Costa rica declares state of emergency after conti ransomware attack. https://www.secureworld.io/industry-news/ costa-rica-emergency-ransomware. Online; Post : 2022-05-12.

[5] Ross J. Anderson, Chris J. Barton, Rainer Bolme, Richard ¨ Clayton, Carlos Hernandez Gan˜an, Tommaso Grasso, Michael ´ Levi, Tyler W. Moore, and Marie Vasek. Measuring the changing cost of cybercrime. 2019.

[6] Daniele Ucci, Leonardo Aniello, and Roberto Baldoni. Survey of machine learning techniques for malware analysis. Computers and Security, 81:123–147, 2019.

[7] Edward Raff and Charles Nicholas. A Survey of Machine Learning Methods and Challenges for Windows Malware Classification. 2020