*3.1.1 Dataset One - 50 Features Dataset.* Utilized a 2000 sample dataset that distinguished between Ransomware and Good ware. Families of ransomware were not explicitly given and 50 features were extracted from a JSON file through a sandboxed based environment [1].

*3.1.2 Dataset Two - Collected Dataset for Ransomware, Malware, and Benign Software.* 257 sampled datasets were collected from the Cuckoo Sandbox™ environment. Fifty features were extracted and tested for Ransomware (57), Malware (120), and Benign software (80) samples accordingly. Families of Ransomware were specified in the data collection.

## Using Cuckoo Sandbox™

The initial dataset was collected by running software samples within the Cuckoo Sandbox™ environment. To initiate the cuckoo sandbox™ installation, an Ubuntu Linux machine was utilized as the foundation. Then VirtualBox was installed along with other required software for the Cuckoo Sandbox version 2.0.7. Next, we installed Cuckoo Sandbox™ and set up the virtual machine. For the testing environment, Windows 7 was utilized with modifications to make it more vulnerable. Python 2 was installed to run the files. Once the Cuckoo Sandbox™ was configured the files could be run in Python. The internet was scoured to find malicious and benign software. Several repositories were identified that aided our research. First was the Zoo [6], a popular malware and ransomware GitHub project that hosts both general malware and ransomware samples. The second repository was NTFS123’s Malware Database [3]. This repository is almost identical to the Zoo, but has some variations in samples. Using these two repositories we were able to procure enough samples to generate our dataset, Dataset 2.

## Feature Extraction

After running the cuckoo samples, python programming interface [1] was utilized to extract 50 features from each malware sample. These features were selected based on certain characteristics of a program that were affected in the infection process while the malware was running. The selected features were extracted from a JSON report contained within the cuckoo report that was generated after a sample was run. The python program generated a GUI tool that presented the 50 final features as a list of checkboxes. As a result, certain features could be filtered out that were related to specific processes such as network activity, registry /events, file processes, etc. After each extraction was complete, a .csv for each malware sample was generated displaying information on the features selected. The headers represented the features and the number associated with each header represented the number of occurrences of each process during the malware analysis. A total of 257 samples were run consisting of Malware, Ransomware, and good ware. Each sample representing a .csv file was generated. All .csv files were combined into one to represent a dataset set for machine learning analysis. This dataset represented Dataset Two. I also utilized a GitHub repository [1] to serve as a training set for both Benign Software versus Ransomware analysis which is represented as Dataset one.