Date of publication 04, 2018.

Dynamic and System Agnostic Malware Detection Via Machine Learning

Michael Sgroi[1], Doug Jacobson[2]
[1]Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50010 USA
[2] Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50010 USA

Corresponding author: Michael Sgroi (e-mail: mcsgroi@iastate.edu, mcsgroi321@gmail.com).

ABSTRACT This paper discusses malware detection in personal computers. Current malware detection solutions are static. Antiviruses rely on lists of malicious signatures that are then used in file scanning. These antiviruses are also very dependent on the operating system, requiring different solutions for different systems. This paper presents a solution that detects malware based on runtime attributes. It also emphasizes that these attributes are easily accessible and fairly generic meaning that it functions across systems and without specialized information. The attributes are used in a machine learning system that makes it flexible for retraining if necessary, but capable of handling new variants without needing to modify the solution. It can also be run quickly which allows for detection to be achieved before the malware gets too far.

INDEX TERMS Malware, Machine Learning

I. INTRODUCTION

Malware is a large problem in modern technology. It causes many issues for people individually, as well as companies. This becomes more of an issue when you take into account the fact that malware is constantly evolving. As can be imagined, this makes it an incredibly difficult problem to solve. Antivirus hasn't changed much at all over the past 20 years for this reason. The solutions we employ are still fairly static. They rely on the antivirus publisher collecting samples continuously. These samples have to be analyzed to generate a signature that can then be used in detection. This is challenging for antivirus developers because they have to find ways of obtaining these samples and they have to invest resources in analyzing them. The signatures obtained have to be added to a list that is pushed to clients. From a client's perspective this means constant updates and slow response to new malware variants.

When new malware strains are introduced or the malware is obfuscated the antivirus becomes completely ineffective. This leaves clients vulnerable to attacks, maybe even more so than without the antivirus because they assume it will keep them safe.

This is where the solution I am proposing comes into play. It seems self-evident that malware should be detectable based on runtime attributes. These would be aspects of malware that on some high level would never change.

The other issue that this paper aims to solve is that antivirus is effectively in itself malware that requires itself to be tightly coupled with the machine and incredibly specialized. This means that it needs to be designed specifically for each operating system and requires a large amount of information to function.

What this paper proposes is a dynamic model that utilizes easily accessible runtime attributes in a generalizable way such that it can be extended between operating systems. These attributes are correlated in a statistically meaningful way by using machine learning.

In this paper, I will outline what previous research has been done in this area. I will then detail the proposed solution after which the testing implementation will be laid out. There will then be discussion on the results of these tests. Lastly, the accomplishments of this paper and ideas for future work in this area will be summarized.

II. PREVIOUS RESEARCH

It is quickly becoming common knowledge that existing antivirus solutions are inadequate. There are even articles appearing in common technical magazines outlining the idea of changing from static analysis methods to dynamic methods [1]. The technical ideas supporting this change of thought are slightly sparser and this is due to the technical challenge involved in implementation. This is due to the fact that antivirus must be incredibly accurate and minimize false positives. This works in favor of static analysis, which will only positively flag malware if it is an exact match for known malware. Dynamic systems will always have more false positives since they are dependent on behavior that cannot be hard coded.

There are a few dynamic solutions that have been proposed, but none of them match the criteria I have outlined here.

Liu et al. [2] proposed an algorithm that takes into account malware behavior features and outputs a judgment based on these features. This doesn't utilize a machine learning model as they created a custom predictor. The solution they proposed is also tied into Windows and requires low level information from the operating system.

Wijnands [5] also proposed a very similar algorithm taking into account malware behavior features such as filesystem, registry, process creation/exiting, and thread creation/exiting. This compared feature sets by utilizing a matrix to calculate distance between nodes. This was also tied in with Windows.

Aubrey-Jones [3] suggests intercepting API calls or using a virtualized environment to capture low level calls. Unfortunately, this only suggests a concept and provides no implementation or proof of concept.

Tobiyama et al. [4] builds on this concept of intercepting API calls and adds the idea of using a Markov chain to construct behavior patterns for processes. These behavior patterns can then be labeled as malicious or benign. This also, only works on Windows, however. Xie et al. [6] also proposes using a Markov chain detection method, but this implementation is based on user behavior/interaction so that it can determine anomalous behavior. This implementation is specific to Android systems though.

Shahzad et al. [7] uses low level process information such as page frames and context switches along with more general information like launcher size. This implementation is specific to Linux.

Ferrante et al. [8] suggests using system calls as well as CPU and memory usage. This is more similar to the attribute set that is used in the solution proposed in this paper, but still requires low level attributes, has a fairly limited number of features and is specific to Android.

Gheorghe et al. [9] is very similar in that it also utilizes CPU and memory usage, but instead of system calls, it uses system settings such as WiFi enabling/disabling and Bluetooth enabling/disabling. As can be surmised, this couples it to the operating system again - Android in this case.

Milosevic et al. [10] is effectively the same attribute set that is used in this paper and does, in fact, use much of the same analysis process. The notable difference is that their solution is tied to Android.

It is quite noticeable that the implementations in existence currently are very low level and require a tight knit coupling with the specific operating system in use. The solution proposed here is similar to most of these solutions, but significantly more generalized.

III. SOLUTION

Based on the problem statement outlined in the introduction, the solution that is being proposed here is a machine learning model that utilizes process statistics to flag malicious programs. The process statistics that are being utilized are similar to what would come from a “top” or “ps” command on a Unix based system.

Since the goal of this system is to be cross platform, it is important that the method of obtaining these process statistics is easily portable. With this in mind, a program call SIGAR [11] was selected. This is a Java library that captures process information using a DLL or shared object library file. The list of operating systems this supports is shown in Figure 1. While it is not completely universal, it is close and could be extended to support other operating systems as needed.

File	Language	Description	Required
sigar.jar	Java	Java API	Yes (for Java only)
log4j.jar	Java	Java logging API	No
libsigar-x86-linux.so	C	Linux AMD/Intel 32-bit	*
libsigar-amd64-linux.so	C	Linux AMD/Intel 64-bit	*
libsigar-ppc-linux.so	C	Linux PowerPC 32-bit	*
libsigar-ppc64-linux.so	C	Linux PowerPC 64-bit	*
libsigar-ia64-linux.so	C	Linux Itanium 64-bit	*
libsigar-s390x-linux.so	C	Linux zSeries 64-bit	*
sigar-x86-winnt.dll	C	Windows AMD/Intel 32-bit	*
sigar-amd64-winnt.dll	C	Windows AMD/Intel 64-bit	*
libsigar-ppc-aix-5.so	C	AIX PowerPC 32-bit	*
libsigar-ppc64-aix-5.so	C	AIX PowerPC 64-bit	*
libsigar-pa-hpux-11.sl	C	HP-UX PA-RISC 32-bit	*
libsigar-ia64-hpux-11.sl	C	HP-UX Itanium 64-bt	*
libsigar-sparc-solaris.so	C	Solaris Sparc 32-bit	*
libsigar-sparc64-solaris.so	C	Solaris Sparc 64-bit	*
libsigar-x86-solaris.so	C	Solaris AMD/Intel 32-bit	*
libsigar-amd64-solaris.so	C	Solaris AMD/Intel 64-bit	*
libsigar-universal-macosx.dylib	C	Mac OS X PowerPC/Intel 32-bit	*
libsigar-universal64-macosx.dylib	C	Mac OS X PowerPC/Intel 64-bit	*
libsigar-x86-freebsd-5.so	C	FreeBSD 5.x AMD/Intel 32-bit	*
libsigar-x86-freebsd-6.so	C	FreeBSD 6.x AMD/Intel 64-bit	*
libsigar-amd64-freebsd-6.so	C	FreeBSD 6.x AMD/Intel 64-bit	*

FIGURE 1. List of possible SIGAR library files by operating system.

This library is used in a script that outputs to either the terminal or a text file about all process information it can capture as frequently as possible. Note that this means that benign and malicious processes are both logged to the same file since all processes are captured. The features that are captured are shown in Table 1.

These features and the classification are organized in a flat text CSV file in entries like the following:

29736960, 5009408, -1, -1, -1, 1635, -1, -1, -1, '-1', 117, 8, 'R', -1, 'WmiPrvSE', 'clean'
67579904, 5136384, -1, -1, -1, 2236, 0.0, 187, 156, 'C:\Users\michael\AppData\Roaming\sktoeys.exe', 57, 2, 'R', 187, 'C:\Users\michael\AppData\Roaming\sktoeys.exe', 'infected'
1441792, 233472, -1, -1, -1, 12491, -1, -1, -1, '-1', 328, 65, 'R', -1, 'System', 'clean'

These CSV files are then mapped to ARFF files and the malicious data labeled using the identified malicious EXE with string attributes removed. The reason that string attributes are removed is that they limit the number of model types that can be used and don't really provide any meaningful data unless parsed for specific pieces of content. For the purposes of this paper, it was unnecessary to keep this information, but could potentially be used in future implementations. ARFF files are the proprietary data format of the machine learning library WEKA [12]. This was chosen here due to its simplicity of implementation, vast feature selection, and data visualization tools.

TABLE I

SIGAR PROCESS ATTRIBUTE LIST [11]

Attribute	Type	Description
pid	STRING	Process ID
mem_size	NUMERIC	Total process virtual memory
mem_resident	NUMERIC	Total process resident memory
mem_share	NUMERIC	Total process shared memory
mem_minor_faults	NUMERIC	Non I/O page faults
mem_major_faults	NUMERIC	I/O page faults
mem_page_faults	NUMERIC	Total number of page faults
cpu_percent	NUMERIC	Process cpu usage
cpu_total	NUMERIC	Process cpu time (sum of user and kernel time)
cpu_system	NUMERIC	Process cpu time (kernel time)
proc_name	STRING	Name of process executable
proc_file_descriptors	NUMERIC	Total number of open file descriptors
proc_threads	NUMERIC	Number of active threads
proc_state	STRING	Process state (Running, Zombie, etc.)
proc_time	NUMERIC	Process cpu time (sum of user and kernel)

Once a model is generated, it can be used in correlation with the script that captures data to classify processes as malicious or not.

IV. TESTING IMPLEMENTATION

There were three steps in setting up the testing for this. These were the selection of datasets, features, and machine learning model types.

A. DATASETS

For testing purposes there were a few malware instances from theZoo malware database [13] whose runtime attributes were sampled. Note that these datasets include both benign and malicious data even though they are the dataset for a specific malware, but that they are labeled benign/malicious appropriately. There was also a large dataset of just clean data for false positive testing. These are all listed in Table 2.

Once the data was collected it was segregated into 4 training and 4 testing sets.

TABLE 2

DATASETS

Malware Name	Malicious EXE	Malware Type	Number of Data Entries
Waski.Upatre	utilview.exe	Trojan	23,150 malicious, 1,523,816 clean
Win32.Alina.3.4.B	jucheck.exe	Trojan	13,047 malicious, 881,478 clean
EquationDrug	EquationDrug_4556CE5EB007AF1DE5BD3B457F0B216D.exe	Trojan	769 malicious, 10,936,625 clean
ZeusVM	dwm.exe	Botnet	11,473 malicious, 1,203,780 clean
IllusionBot	BOTBINARY.EXE	Botnet	249,050 malicious, 14,292,470 clean
Teslacrypt	sktoeys.exe	Ransomware	53 malicious, 2,247 clean
Jigsaw	drpbx.exe	Ransomware	114 malicious, 4,562 clean
Locky	svchost.exe	Ransomware	80 malicious, 4,525 clean
Clean		Clean Data	12,093,240 clean

The first set was for Trojan testing. For this, the Waski.Upatre and Win32.Alina.3.4.B datasets were used for training and the EquationDrug dataset was used for testing.

The second set was for botnet testing. For this, the IllusionBot dataset was used for training and the ZeusVM dataset was used for testing.

The third set was for ransomware testing. For this, the Jigsaw and Locky datasets were used for training and the Teslacrypt dataset was used for testing.

The last set was an aggregation of all of these malware variants and used combined training and testing sets. In other words, the training dataset was Waski.Upatre, Win32.Alina.3.4.B, IllusionBot, Jigsaw, and Locky. The testing dataset consisted of EquationDrug, ZeusVM, Teslacrypt, and the purely clean data.

B. FEATURE SELECTION

Three feature selection algorithms in Weka were used to determine which of the acquired process attributes should be used in model training and testing. The algorithms used were CfsSubsetEval, CorrelationAttributeEval, and InfoGainAttributeEval.

1) CFSSUBSETEVAL

This is a means of evaluating the value of a subset of attributes by comparing the value of an attribute with how redundant it is with other attributes in the subset. It utilized BestFirst which searches via greedy hillclimbing with backtracking.

2) CORRELATIONATTRIBUTEEVAL

This picks the most relevant attributes based on how likely a class is for that specific variable. This utilized Ranker which simply organizes by the highest values achieved by attribute evaluators such as entropy.

3) INFOGAINATTRIBUTEEVAL

This evaluates an attribute based on how much class information is gained from it. This also utilized Ranker.

4) FINAL FEATURES

After running the above feature selection algorithms, the attribute rankings and what they represented were used to construct a list of the most valuable attributes for each of the 4 test datasets.

The attributes chosen were as follows:

- **Trojan datasets:**
  - mem_size
  - mem_resident
  - proc_file_descriptors
  - proc_threads
- **Botnet datasets:**
  - mem_page_faults
  - mem_size
  - proc_file_descriptors
  - proc_threads
- **Ransomware datasets:**
  - proc_file_descriptors
  - mem_resident
  - mem_size

- **Aggregate datasets:**
    - proc_file_descriptors
    - mem_size
    - mem_resident
    - mem_page_faults

C. MODELS

There were six Weka machine learning models chosen for testing. These are as follows:

- **Decision Table**
    - This is a simple Decision Table majority classifier. It utilizes a grid to map features to the likeliest classification.
- **Logistic**
    - This is a Logistic Regression model which includes a ridge estimator.
- **NaiveBayes**
    - This is a NaiveBayes implementation using estimator classes. The estimator uses a precision that is based on the input data.
- **PART**
    - This is a decision list based on tree data. Effectively it constructs partial C4.5 decision trees and makes the best leaf from each into a rule in the list.
- **REPTree**
    - This is a fast regression tree that uses information gain for tree derivation and is pruned. It sorts the attributes once and if anything needs to be added splits existing instances.
- **Voted Perceptron**
    - This is a voting system where weight vectors are used with a set number of nodes to vote on data. This is supposed to be similar to SVM except faster.

For all of these models, the default parameters specified in Weka were used, except for Voted Perceptron where the number of nodes was changed from 10000 to 3000.

V. RESULTS

First each training set was used to create each of the 6 classifiers. Each of these classifiers was evaluated in two ways, using 10 fold cross validation and via the test dataset outlined previously.

A. 10 FOLD CROSS VALIDATION RESULTS

When the classifier was being made, 10 fold cross validation was performed. This means that the data is split into 10 pieces and for each of those pieces one piece is used for testing while the other 9 are used for training. This generated the results outlined in Figures 2 - 25.

1) TROJAN

As can be seen in Figures 2 - 7, the Decision Table, NaiveBayes, PART, and REPTree perform about equally and have near perfect accuracy.

2) BOTNET

As can be seen in Figures 8 - 13, all of the classifiers have near perfect accuracy with the exception of Voted Perceptron.

3) RANSOMWARE

According to Figures 14 - 19, the Decision Table, PART, and REPTree have near perfect accuracy and the NaiveBayes and Logistic have moderate performance.

4) COMBINED

As can be seen in Figures 20 - 25, the Decision Table, PART, and REPTree perform extremely well. The NaiveBayes also performs fairly well, but has an increased false positive rate.

5) EVALUATION

This shows that the classifiers would work for moderately similar data, but are at least fairly extensible. The only consistently bad classifier was the Voted Perceptron which consistently missed identification of malware.

B. TEST RESULTS

The next step then was to analyze completely unseen malware samples' runtime attributes. This was where the classifiers just generated were then tested using the test data outlined in the previous section. The results of this are shown in Figures 26 - 49.

1) TROJAN

As can be seen in Figures 26- 31, none of the classifiers correctly identify a single malware sample.

2) BOTNET

This demonstrates that the logistic classifier at least starts to identify the malicious samples as shown in Figure 33. Even so, it only classifies a small portion of the samples and all of the other classifiers fail completely in malicious identification as shown in Figures 32 - 37.

```
=== Summary ===

Correctly Classified Instances     2441343               99.9939 %
Incorrectly Classified Instances       148                0.0061 %
Kappa statistic                          0.9979
Mean absolute error                      0.0005
Root mean squared error                  0.008
Relative absolute error                  1.6499 %
Root relative squared error              6.6327 %
Total Number of Instances          2441491

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.999    0.000    0.997      0.999    0.998      0.998    1.000     1.000     infected
                 1.000    0.001    1.000      1.000    1.000      0.998    1.000     1.000     clean
Weighted Avg.    1.000    0.001    1.000      1.000    1.000      0.998    1.000     1.000

=== Confusion Matrix ===

       a       b   <-- classified as
   36146      51 |       a = infected
      97 2405197 |       b = clean
```

FIGURE 2. **Trojan, Decision Table, 10 Fold Cross Validation Results.**

```
=== Summary ===

Correctly Classified Instances     2405294               98.5174 %
Incorrectly Classified Instances     36197                1.4826 %
Kappa statistic                          0
Mean absolute error                      0.029
Root mean squared error                  0.1207
Relative absolute error                 99.3472 %
Root relative squared error             99.8631 %
Total Number of Instances          2441491

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    0.000      0.000    0.000      0.000    0.741     0.028     infected
                 1.000    1.000    0.985      1.000    0.993      0.000    0.741     0.996     clean
Weighted Avg.    0.985    0.985    0.971      0.985    0.978      0.000    0.741     0.981

=== Confusion Matrix ===

       a       b   <-- classified as
       0   36197 |       a = infected
       0 2405294 |       b = clean
```

FIGURE 3. **Trojan, Logistic, 10 Fold Cross Validation Results.**

```
=== Summary ===

Correctly Classified Instances     2427613               99.4316 %
Incorrectly Classified Instances     13878                0.5684 %
Kappa statistic                          0.8347
Mean absolute error                      0.0058
Root mean squared error                  0.0755
Relative absolute error                 20.0218 %
Root relative squared error             62.5116 %
Total Number of Instances          2441491

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.988    0.006    0.727      0.988    0.837      0.845    0.999     0.976     infected
                 0.994    0.012    1.000      0.994    0.997      0.845    0.999     1.000     clean
Weighted Avg.    0.994    0.012    0.996      0.994    0.995      0.845    0.999     1.000

=== Confusion Matrix ===

       a       b   <-- classified as
   35757     440 |       a = infected
   13438 2391856 |       b = clean
```

FIGURE 4. **Trojan, NaiveBayes, 10 Fold Cross Validation Results.**

```
=== Summary ===

Correctly Classified Instances     2441491             100      %
Incorrectly Classified Instances         0               0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances          2441491

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     infected
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     clean
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

       a       b   <-- classified as
   36197       0 |       a = infected
       0 2405294 |       b = clean
```

FIGURE 5. Trojan, PART, 10 Fold Cross Validation Results.

```
=== Summary ===

Correctly Classified Instances     2441489              99.9999 %
Incorrectly Classified Instances         2               0.0001 %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0.0009
Relative absolute error                  0.0042 %
Root relative squared error              0.7489 %
Total Number of Instances          2441491

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     infected
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     clean
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

       a       b   <-- classified as
   36195       2 |       a = infected
       0 2405294 |       b = clean
```

FIGURE 6. Trojan, REPTree, 10 Fold Cross Validation Results.

```
=== Summary ===

Correctly Classified Instances     2405294              98.5174 %
Incorrectly Classified Instances     36197               1.4826 %
Kappa statistic                          0
Mean absolute error                      0.0148
Root mean squared error                  0.1218
Relative absolute error                 50.7517 %
Root relative squared error            100.7496 %
Total Number of Instances          2441491

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.015     infected
                 1.000    1.000    0.985      1.000    0.993      0.000    0.500     0.985     clean
Weighted Avg.    0.985    0.985    0.971      0.985    0.978      0.000    0.500     0.971

=== Confusion Matrix ===

       a       b   <-- classified as
       0   36197 |       a = infected
       0 2405294 |       b = clean
```

FIGURE 7. Trojan, Voted Perceptron, 10 Fold Cross Validation Results.

```
=== Summary ===

Correctly Classified Instances     14541519          100      %
Incorrectly Classified Instances          1            0      %
Kappa statistic                           1
Mean absolute error                       0
Root mean squared error                   0.0003
Relative absolute error                   0.0022 %
Root relative squared error               0.2025 %
Total Number of Instances          14541520

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     infected
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     clean
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

        a        b   <-- classified as
   249050        0 |        a = infected
        1 14292469 |        b = clean
```

FIGURE 8. Botnet, Decision Table, 10 Fold Cross Validation Results.

```
=== Summary ===

Correctly Classified Instances     14531485           99.931  %
Incorrectly Classified Instances      10035            0.069  %
Kappa statistic                           0.9799
Mean absolute error                       0.0047
Root mean squared error                   0.0315
Relative absolute error                  13.9226 %
Root relative squared error              24.2486 %
Total Number of Instances          14541520

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.001    0.961      1.000    0.980      0.980    0.999     0.901     infected
                 0.999    0.000    1.000      0.999    1.000      0.980    0.999     1.000     clean
Weighted Avg.    0.999    0.000    0.999      0.999    0.999      0.980    0.999     0.998

=== Confusion Matrix ===

        a        b   <-- classified as
   249050        0 |        a = infected
    10035 14282435 |        b = clean
```

FIGURE 9. Botnet, Logistic, 10 Fold Cross Validation Results.

```
=== Summary ===

Correctly Classified Instances     14537350           99.9713 %
Incorrectly Classified Instances       4170            0.0287 %
Kappa statistic                           0.9914
Mean absolute error                       0.0003
Root mean squared error                   0.0169
Relative absolute error                   0.8514 %
Root relative squared error              13.0466 %
Total Number of Instances          14541520

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.983    0.000    1.000      0.983    0.992      0.991    1.000     1.000     infected
                 1.000    0.017    1.000      1.000    1.000      0.991    0.992     1.000     clean
Weighted Avg.    1.000    0.016    1.000      1.000    1.000      0.991    0.992     1.000

=== Confusion Matrix ===

        a        b   <-- classified as
   244880     4170 |        a = infected
        0 14292470 |        b = clean
```

FIGURE 10. Botnet, NaiveBayes, 10 Fold Cross Validation Results.

```
=== Summary ===

Correctly Classified Instances     14541520         100      %
Incorrectly Classified Instances          0               0      %
Kappa statistic                           1
Mean absolute error                       0
Root mean squared error                   0
Relative absolute error                   0      %
Root relative squared error               0      %
Total Number of Instances          14541520

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     infected
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     clean
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

        a        b   <-- classified as
   249050        0 |        a = infected
        0 14292470 |        b = clean
```

FIGURE 11. Botnet, PART, 10 Fold Cross Validation Results.

```
=== Summary ===

Correctly Classified Instances     14541520         100      %
Incorrectly Classified Instances          0               0      %
Kappa statistic                           1
Mean absolute error                       0
Root mean squared error                   0
Relative absolute error                   0      %
Root relative squared error               0      %
Total Number of Instances          14541520

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     infected
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     clean
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

        a        b   <-- classified as
   249050        0 |        a = infected
        0 14292470 |        b = clean
```

FIGURE 12. Botnet, REPTree, 10 Fold Cross Validation Results.

```
=== Summary ===

Correctly Classified Instances     14292470          98.2873 %
Incorrectly Classified Instances     249050           1.7127 %
Kappa statistic                           0
Mean absolute error                       0.0171
Root mean squared error                   0.1309
Relative absolute error                  50.8712 %
Root relative squared error             100.8675 %
Total Number of Instances          14541520

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.017     infected
                 1.000    1.000    0.983      1.000    0.991      ?        0.500     0.983     clean
Weighted Avg.    0.983    0.983    ?          0.983    ?          ?        0.500     0.966

=== Confusion Matrix ===

        a        b   <-- classified as
        0   249050 |        a = infected
        0 14292470 |        b = clean
```

FIGURE 13. Botnet, Voted Perceptron, 10 Fold Cross Validation Results.

```
=== Summary ===

Correctly Classified Instances        9278               99.9677 %
Incorrectly Classified Instances         3                0.0323 %
Kappa statistic                          0.992
Mean absolute error                      0.0014
Root mean squared error                  0.0185
Relative absolute error                  3.342  %
Root relative squared error             12.9352 %
Total Number of Instances             9281

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.985    0.000    1.000      0.985    0.992      0.992    1.000     0.992     infected
                 1.000    0.015    1.000      1.000    1.000      0.992    1.000     1.000     clean
Weighted Avg.    1.000    0.015    1.000      1.000    1.000      0.992    1.000     1.000

=== Confusion Matrix ===

    a    b   <-- classified as
  191    3 |    a = infected
    0 9087 |    b = clean
```

FIGURE 14. Ransomware, Decision Table, 10 Fold Cross Validation Results.

```
=== Summary ===

Correctly Classified Instances        9201               99.138  %
Incorrectly Classified Instances        80                0.862  %
Kappa statistic                          0.7362
Mean absolute error                      0.0154
Root mean squared error                  0.0894
Relative absolute error                 37.5522 %
Root relative squared error             62.5259 %
Total Number of Instances             9281

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.588    0.000    1.000      0.588    0.740      0.763    0.989     0.764     infected
                 1.000    0.412    0.991      1.000    0.996      0.763    0.989     1.000     clean
Weighted Avg.    0.991    0.404    0.991      0.991    0.990      0.763    0.989     0.995

=== Confusion Matrix ===

    a    b   <-- classified as
  114   80 |    a = infected
    0 9087 |    b = clean
```

FIGURE 15. Ransomware, Logistic, 10 Fold Cross Validation Results.

```
=== Summary ===

Correctly Classified Instances        9201               99.138  %
Incorrectly Classified Instances        80                0.862  %
Kappa statistic                          0.7362
Mean absolute error                      0.0291
Root mean squared error                  0.1058
Relative absolute error                 70.8308 %
Root relative squared error             73.9598 %
Total Number of Instances             9281

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.588    0.000    1.000      0.588    0.740      0.763    0.902     0.616     infected
                 1.000    0.412    0.991      1.000    0.996      0.763    0.902     0.998     clean
Weighted Avg.    0.991    0.404    0.991      0.991    0.990      0.763    0.902     0.990

=== Confusion Matrix ===

    a    b   <-- classified as
  114   80 |    a = infected
    0 9087 |    b = clean
```

FIGURE 16. Ransomware, NaiveBayes, 10 Fold Cross Validation Results.

```
=== Summary ===

Correctly Classified Instances        9280               99.9892 %
Incorrectly Classified Instances         1                0.0108 %
Kappa statistic                          0.9974
Mean absolute error                      0.0001
Root mean squared error                  0.0104
Relative absolute error                  0.2625 %
Root relative squared error              7.2558 %
Total Number of Instances             9281

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    0.995      1.000    0.997      0.997    1.000     0.995     infected
                 1.000    0.000    1.000      1.000    1.000      0.997    1.000     1.000     clean
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      0.997    1.000     1.000

=== Confusion Matrix ===

    a    b   <-- classified as
  194    0 |    a = infected
    1 9086 |    b = clean
```

FIGURE 17. Ransomware, PART, 10 Fold Cross Validation Results.

```
=== Summary ===

Correctly Classified Instances        9279               99.9785 %
Incorrectly Classified Instances         2                0.0215 %
Kappa statistic                          0.9947
Mean absolute error                      0.0003
Root mean squared error                  0.0146
Relative absolute error                  0.8424 %
Root relative squared error             10.2338 %
Total Number of Instances             9281

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.995    0.000    0.995      0.995    0.995      0.995    1.000     0.994     infected
                 1.000    0.005    1.000      1.000    1.000      0.995    1.000     1.000     clean
Weighted Avg.    1.000    0.005    1.000      1.000    1.000      0.995    1.000     1.000

=== Confusion Matrix ===

    a    b   <-- classified as
  193    1 |    a = infected
    1 9086 |    b = clean
```

FIGURE 18. Ransomware, REPTree, 10 Fold Cross Validation Results.

```
=== Summary ===

Correctly Classified Instances        9087               97.9097 %
Incorrectly Classified Instances       194                2.0903 %
Kappa statistic                          0
Mean absolute error                      0.0209
Root mean squared error                  0.1446
Relative absolute error                 50.9307 %
Root relative squared error            101.0616 %
Total Number of Instances             9281

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.021     infected
                 1.000    1.000    0.979      1.000    0.989      0.000    0.500     0.979     clean
Weighted Avg.    0.979    0.979    0.959      0.979    0.969      0.000    0.500     0.959

=== Confusion Matrix ===

    a    b   <-- classified as
    0  194 |    a = infected
    0 9087 |    b = clean
```

FIGURE 19. Ransomware, Voted Perceptron, 10 Fold Cross Validation Results.

```
=== Summary ===

Correctly Classified Instances     16991939          99.9979 %
Incorrectly Classified Instances       353            0.0021 %
Kappa statistic                          0.9994
Mean absolute error                      0.0002
Root mean squared error                  0.0053
Relative absolute error                  0.4903 %
Root relative squared error              4.1519 %
Total Number of Instances          16992292

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.999    0.000    1.000      0.999    0.999      0.999    1.000     1.000     infected
                 1.000    0.001    1.000      1.000    1.000      0.999    1.000     1.000     clean
Weighted Avg.    1.000    0.001    1.000      1.000    1.000      0.999    1.000     1.000

=== Confusion Matrix ===

       a        b   <-- classified as
  285171      270 |        a = infected
      83 16706768 |        b = clean
```

FIGURE 20. Combined, Decision Table, 10 Fold Cross Validation Results.

```
=== Summary ===

Correctly Classified Instances     16706851          98.3202 %
Incorrectly Classified Instances    285441            1.6798 %
Kappa statistic                          0
Mean absolute error                      0.032
Root mean squared error                  0.1276
Relative absolute error                 96.9109 %
Root relative squared error             99.2994 %
Total Number of Instances          16992292

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.899     0.084     infected
                 1.000    1.000    0.983      1.000    0.992      ?        0.899     0.998     clean
Weighted Avg.    0.983    0.983    ?          0.983    ?          ?        0.899     0.983

=== Confusion Matrix ===

       a        b   <-- classified as
       0   285441 |        a = infected
       0 16706851 |        b = clean
```

FIGURE 21. Combined, Logistic, 10 Fold Cross Validation Results.

```
=== Summary ===

Correctly Classified Instances     14755044          86.8337 %
Incorrectly Classified Instances   2237248           13.1663 %
Kappa statistic                          0.1731
Mean absolute error                      0.1423
Root mean squared error                  0.3562
Relative absolute error                430.9311 %
Root relative squared error            277.1719 %
Total Number of Instances          16992292

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.968    0.133    0.110      0.968    0.198      0.303    0.981     0.888     infected
                 0.867    0.032    0.999      0.867    0.928      0.303    0.980     1.000     clean
Weighted Avg.    0.868    0.034    0.984      0.868    0.916      0.303    0.980     0.998

=== Confusion Matrix ===

       a        b   <-- classified as
  276304     9137 |        a = infected
 2228111 14478740 |        b = clean
```

FIGURE 22. Combined, NaiveBayes, 10 Fold Cross Validation Results.

```
=== Summary ===

Correctly Classified Instances     16992287          100      %
Incorrectly Classified Instances          5                0      %
Kappa statistic                           1
Mean absolute error                       0
Root mean squared error                   0.0005
Relative absolute error                   0.0011 %
Root relative squared error               0.4189 %
Total Number of Instances          16992292

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     infected
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     clean
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

        a        b   <-- classified as
   285439        2 |        a = infected
        3 16706848 |        b = clean
```

FIGURE 23. Combined, PART, 10 Fold Cross Validation Results.

```
=== Summary ===

Correctly Classified Instances     16992273          99.9999 %
Incorrectly Classified Instances         19               0.0001 %
Kappa statistic                           1
Mean absolute error                       0
Root mean squared error                   0.001
Relative absolute error                   0.0047 %
Root relative squared error               0.8071 %
Total Number of Instances          16992292

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     infected
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     clean
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

        a        b   <-- classified as
   285430       11 |        a = infected
        8 16706843 |        b = clean
```

FIGURE 24. Combined, REPTree, 10 Fold Cross Validation Results.

```
=== Summary ===

Correctly Classified Instances     16706851          98.3202 %
Incorrectly Classified Instances     285441           1.6798 %
Kappa statistic                           0
Mean absolute error                       0.0168
Root mean squared error                   0.1296
Relative absolute error                  50.8542 %
Root relative squared error             100.8506 %
Total Number of Instances          16992292

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.017     infected
                 1.000    1.000    0.983      1.000    0.992      ?        0.500     0.983     clean
Weighted Avg.    0.983    0.983    ?          0.983    ?          ?        0.500     0.967

=== Confusion Matrix ===

        a        b   <-- classified as
        0   285441 |        a = infected
        0 16706851 |        b = clean
```

FIGURE 25. Combined, Voted Perceptron, 10 Fold Cross Validation Results.

```
=== Summary ===

Correctly Classified Instances     10930507          99.937  %
Incorrectly Classified Instances       6887               0.063  %
Kappa statistic                         -0.0001
Mean absolute error                      0.0029
Root mean squared error                  0.0203
Total Number of Instances          10937394

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.001    0.000      0.000    0.000      -0.000   0.281     0.000     infected
                 0.999    1.000    1.000      0.999    1.000      -0.000   0.281     1.000     clean
Weighted Avg.    0.999    1.000    1.000      0.999    1.000      -0.000   0.281     1.000

=== Confusion Matrix ===

        a        b   <-- classified as
        0      769 |        a = infected
     6118 10930507 |        b = clean
```

FIGURE 26. Trojan, Decision Table, Testing Results.

```
=== Summary ===

Correctly Classified Instances     10936625          99.993  %
Incorrectly Classified Instances        769               0.007  %
Kappa statistic                          0
Mean absolute error                      0.0134
Root mean squared error                  0.0191
Total Number of Instances          10937394

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    0.000      0.000    0.000      0.000    0.978     0.006     infected
                 1.000    1.000    1.000      1.000    1.000      0.000    0.978     1.000     clean
Weighted Avg.    1.000    1.000    1.000      1.000    1.000      0.000    0.978     1.000

=== Confusion Matrix ===

        a        b   <-- classified as
        0      769 |        a = infected
        0 10936625 |        b = clean
```

FIGURE 27. Trojan, Logistic, Testing Results.

```
=== Summary ===

Correctly Classified Instances     10908645          99.7371 %
Incorrectly Classified Instances     28749                0.2629 %
Kappa statistic                         -0.0001
Mean absolute error                      0.0032
Root mean squared error                  0.044
Total Number of Instances          10937394

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.003    0.000      0.000    0.000      -0.000   0.437     0.000     infected
                 0.997    1.000    1.000      0.997    0.999      -0.000   0.470     1.000     clean
Weighted Avg.    0.997    1.000    1.000      0.997    0.999      -0.000   0.470     1.000

=== Confusion Matrix ===

        a        b   <-- classified as
        0      769 |        a = infected
    27980 10908645 |        b = clean
```

FIGURE 28. Trojan, NaiveBayes, Testing Results.

```
=== Summary ===

Correctly Classified Instances     10760758          98.385  %
Incorrectly Classified Instances     176636             1.615  %
Kappa statistic                         -0.0001
Mean absolute error                      0.0161
Root mean squared error                  0.1271
Total Number of Instances          10937394

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.016    0.000      0.000    0.000      -0.001   0.492     0.000     infected
                 0.984    1.000    1.000      0.984    0.992      -0.001   0.492     1.000     clean
Weighted Avg.    0.984    1.000    1.000      0.984    0.992      -0.001   0.492     1.000

=== Confusion Matrix ===

        a        b   <-- classified as
        0      769 |        a = infected
   175867 10760758 |        b = clean
```

FIGURE 29. Trojan, PART, Testing Results.

```
=== Summary ===

Correctly Classified Instances     10762624          98.4021 %
Incorrectly Classified Instances     174770             1.5979 %
Kappa statistic                         -0.0001
Mean absolute error                      0.016
Root mean squared error                  0.1264
Total Number of Instances          10937394

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.016    0.000      0.000    0.000      -0.001   0.492     0.000     infected
                 0.984    1.000    1.000      0.984    0.992      -0.001   0.492     1.000     clean
Weighted Avg.    0.984    1.000    1.000      0.984    0.992      -0.001   0.492     1.000

=== Confusion Matrix ===

        a        b   <-- classified as
        0      769 |        a = infected
   174001 10762624 |        b = clean
```

FIGURE 30. Trojan, REPTree, Testing Results.

```
=== Summary ===

Correctly Classified Instances     10936625          99.993  %
Incorrectly Classified Instances        769             0.007  %
Kappa statistic                          0
Mean absolute error                      0.0001
Root mean squared error                  0.0084
Total Number of Instances          10937394

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.000     infected
                 1.000    1.000    1.000      1.000    1.000      0.000    0.500     1.000     clean
Weighted Avg.    1.000    1.000    1.000      1.000    1.000      0.000    0.500     1.000

=== Confusion Matrix ===

        a        b   <-- classified as
        0      769 |        a = infected
        0 10936625 |        b = clean
```

FIGURE 31. Trojan, Voted Perceptron, Testing Results.

```
=== Summary ===

Correctly Classified Instances     1203780              99.0559 %
Incorrectly Classified Instances     11473               0.9441 %
Kappa statistic                          0
Mean absolute error                      0.0094
Root mean squared error                  0.0972
Total Number of Instances          1215253

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    0.000      0.000    0.000      0.000    0.367     0.009     infected
                 1.000    1.000    0.991      1.000    0.995      0.000    0.367     0.988     clean
Weighted Avg.    0.991    0.991    0.981      0.991    0.986      0.000    0.367     0.979

=== Confusion Matrix ===

       a       b   <-- classified as
       0   11473 |       a = infected
       0 1203780 |       b = clean
```

FIGURE 32. Botnet, Decision Table, Testing Results.

```
=== Summary ===

Correctly Classified Instances     1176783              96.8344 %
Incorrectly Classified Instances     38470               3.1656 %
Kappa statistic                         -0.0119
Mean absolute error                      0.0351
Root mean squared error                  0.1769
Total Number of Instances          1215253

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.003    0.022    0.001      0.003    0.002      -0.013   0.161     0.010     infected
                 0.978    0.997    0.990      0.978    0.984      -0.013   0.355     0.987     clean
Weighted Avg.    0.968    0.988    0.981      0.968    0.975      -0.013   0.353     0.978

=== Confusion Matrix ===

       a       b   <-- classified as
      30   11443 |       a = infected
   27027 1176753 |       b = clean
```

FIGURE 33. Botnet, Logistic, Testing Results.

```
=== Summary ===

Correctly Classified Instances     1203780              99.0559 %
Incorrectly Classified Instances     11473               0.9441 %
Kappa statistic                          0
Mean absolute error                      0.0094
Root mean squared error                  0.0972
Total Number of Instances          1215253

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    0.000      0.000    0.000      0.000    0.372     0.007     infected
                 1.000    1.000    0.991      1.000    0.995      0.000    0.500     0.991     clean
Weighted Avg.    0.991    0.991    0.981      0.991    0.986      0.000    0.499     0.981

=== Confusion Matrix ===

       a       b   <-- classified as
       0   11473 |       a = infected
       0 1203780 |       b = clean
```

FIGURE 34. Botnet, NaiveBayes, Testing Results.

```
=== Summary ===

Correctly Classified Instances     1203780              99.0559 %
Incorrectly Classified Instances     11473               0.9441 %
Kappa statistic                          0
Mean absolute error                      0.0094
Root mean squared error                  0.0972
Total Number of Instances          1215253

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.009     infected
                 1.000    1.000    0.991      1.000    0.995      0.000    0.500     0.991     clean
Weighted Avg.    0.991    0.991    0.981      0.991    0.986      0.000    0.500     0.981

=== Confusion Matrix ===

       a       b   <-- classified as
       0   11473 |       a = infected
       0 1203780 |       b = clean
```

FIGURE 35. Botnet, PART, Testing Results.

```
=== Summary ===

Correctly Classified Instances     1203780              99.0559 %
Incorrectly Classified Instances     11473               0.9441 %
Kappa statistic                          0
Mean absolute error                      0.0094
Root mean squared error                  0.0972
Total Number of Instances          1215253

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.009     infected
                 1.000    1.000    0.991      1.000    0.995      0.000    0.500     0.991     clean
Weighted Avg.    0.991    0.991    0.981      0.991    0.986      0.000    0.500     0.981

=== Confusion Matrix ===

       a       b   <-- classified as
       0   11473 |       a = infected
       0 1203780 |       b = clean
```

FIGURE 36. Botnet, REPTree, Testing Results.

```
=== Summary ===

Correctly Classified Instances     1203780              99.0559 %
Incorrectly Classified Instances     11473               0.9441 %
Kappa statistic                          0
Mean absolute error                      0.0094
Root mean squared error                  0.0972
Total Number of Instances          1215253

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.009     infected
                 1.000    1.000    0.991      1.000    0.995      0.000    0.500     0.991     clean
Weighted Avg.    0.991    0.991    0.981      0.991    0.986      0.000    0.500     0.981

=== Confusion Matrix ===

       a       b   <-- classified as
       0   11473 |       a = infected
       0 1203780 |       b = clean
```

FIGURE 37. Botnet, Voted Perceptron, Testing Results.

```
=== Summary ===

Correctly Classified Instances        2247               97.6957 %
Incorrectly Classified Instances        53                2.3043 %
Kappa statistic                          0
Mean absolute error                      0.0235
Root mean squared error                  0.1518
Total Number of Instances             2300

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    0.000      0.000    0.000      0.000    0.347     0.023     infected
                 1.000    1.000    0.977      1.000    0.988      0.000    0.347     0.969     clean
Weighted Avg.    0.977    0.977    0.954      0.977    0.966      0.000    0.347     0.948

=== Confusion Matrix ===

    a    b   <-- classified as
    0   53 |    a = infected
    0 2247 |    b = clean
```

FIGURE 38. Ransomware, Decision Table, Testing Results.

```
=== Summary ===

Correctly Classified Instances        2247               97.6957 %
Incorrectly Classified Instances        53                2.3043 %
Kappa statistic                          0
Mean absolute error                      0.0282
Root mean squared error                  0.153
Total Number of Instances             2300

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    0.000      0.000    0.000      0.000    0.653     0.048     infected
                 1.000    1.000    0.977      1.000    0.988      0.000    0.653     0.990     clean
Weighted Avg.    0.977    0.977    0.954      0.977    0.966      0.000    0.653     0.969

=== Confusion Matrix ===

    a    b   <-- classified as
    0   53 |    a = infected
    0 2247 |    b = clean
```

FIGURE 39. Ransomware, Logistic, Testing Results.

```
=== Summary ===

Correctly Classified Instances        2247               97.6957 %
Incorrectly Classified Instances        53                2.3043 %
Kappa statistic                          0
Mean absolute error                      0.0364
Root mean squared error                  0.1513
Total Number of Instances             2300

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    0.000      0.000    0.000      0.000    0.660     0.114     infected
                 1.000    1.000    0.977      1.000    0.988      0.000    0.660     0.990     clean
Weighted Avg.    0.977    0.977    0.954      0.977    0.966      0.000    0.660     0.970

=== Confusion Matrix ===

    a    b   <-- classified as
    0   53 |    a = infected
    0 2247 |    b = clean
```

FIGURE 40. Ransomware, NaiveBayes, Testing Results.

```
=== Summary ===

Correctly Classified Instances        2247               97.6957 %
Incorrectly Classified Instances        53                2.3043 %
Kappa statistic                          0
Mean absolute error                      0.023
Root mean squared error                  0.1518
Total Number of Instances             2300

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.023     infected
                 1.000    1.000    0.977      1.000    0.988      0.000    0.500     0.977     clean
Weighted Avg.    0.977    0.977    0.954      0.977    0.966      0.000    0.500     0.955

=== Confusion Matrix ===

    a    b   <-- classified as
    0   53 |    a = infected
    0 2247 |    b = clean
```

FIGURE 41. Ransomware, PART, Testing Results.

```
=== Summary ===

Correctly Classified Instances        2247               97.6957 %
Incorrectly Classified Instances        53                2.3043 %
Kappa statistic                          0
Mean absolute error                      0.023
Root mean squared error                  0.1518
Total Number of Instances             2300

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.023     infected
                 1.000    1.000    0.977      1.000    0.988      0.000    0.500     0.977     clean
Weighted Avg.    0.977    0.977    0.954      0.977    0.966      0.000    0.500     0.955

=== Confusion Matrix ===

    a    b   <-- classified as
    0   53 |    a = infected
    0 2247 |    b = clean
```

FIGURE 42. Ransomware, REPTree, Testing Results.

```
=== Summary ===

Correctly Classified Instances        2247               97.6957 %
Incorrectly Classified Instances        53                2.3043 %
Kappa statistic                          0
Mean absolute error                      0.023
Root mean squared error                  0.1518
Total Number of Instances             2300

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.023     infected
                 1.000    1.000    0.977      1.000    0.988      0.000    0.500     0.977     clean
Weighted Avg.    0.977    0.977    0.954      0.977    0.966      0.000    0.500     0.955

=== Confusion Matrix ===

    a    b   <-- classified as
    0   53 |    a = infected
    0 2247 |    b = clean
```

FIGURE 43. Ransomware, Voted Perceptron, Testing Results.

```
=== Summary ===

Correctly Classified Instances     24222148          99.8926 %
Incorrectly Classified Instances     26039            0.1074 %
Kappa statistic                          0.3499
Mean absolute error                      0.004
Root mean squared error                  0.0264
Total Number of Instances          24248187

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.571    0.001    0.253      0.571    0.350      0.379    0.964     0.520     infected
                 0.999    0.429    1.000      0.999    0.999      0.379    0.964     1.000     clean
Weighted Avg.    0.999    0.429    0.999      0.999    0.999      0.379    0.964     1.000

=== Confusion Matrix ===

        a        b   <-- classified as
     7021     5274 |        a = infected
    20765 24215127 |        b = clean
```

FIGURE 44. Combined, Decision Table, Testing Results.

```
=== Summary ===

Correctly Classified Instances     24218030          99.8756 %
Incorrectly Classified Instances     30157            0.1244 %
Kappa statistic                          0.0026
Mean absolute error                      0.0129
Root mean squared error                  0.0432
Total Number of Instances          24248187

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.004    0.001    0.003      0.004    0.003      0.003    0.836     0.002     infected
                 0.999    0.996    0.999      0.999    0.999      0.003    0.836     1.000     clean
Weighted Avg.    0.999    0.996    0.999      0.999    0.999      0.003    0.836     0.999

=== Confusion Matrix ===

        a        b   <-- classified as
       49    12246 |        a = infected
    17911 24217981 |        b = clean
```

FIGURE 45. Combined, Logistic, Testing Results.

```
=== Summary ===

Correctly Classified Instances     20487840          84.4923 %
Incorrectly Classified Instances   3760347           15.5077 %
Kappa statistic                          0.0011
Mean absolute error                      0.1549
Root mean squared error                  0.3766
Total Number of Instances          24248187

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.321    0.155    0.001      0.321    0.002      0.010    0.790     0.001     infected
                 0.845    0.679    1.000      0.845    0.916      0.010    0.780     1.000     clean
Weighted Avg.    0.845    0.679    0.999      0.845    0.915      0.010    0.780     0.999

=== Confusion Matrix ===

        a        b   <-- classified as
     3949     8346 |        a = infected
  3752001 20483891 |        b = clean
```

FIGURE 46. Combined, NaiveBayes, Testing Results.

```
=== Summary ===

Correctly Classified Instances     24189517         99.758  %
Incorrectly Classified Instances      58670              0.242  %
Kappa statistic                          0.28
Mean absolute error                      0.0024
Root mean squared error                  0.0492
Total Number of Instances          24248187

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.931    0.002    0.165      0.931    0.281      0.392    0.964     0.154     infected
                 0.998    0.069    1.000      0.998    0.999      0.392    0.964     1.000     clean
Weighted Avg.    0.998    0.069    1.000      0.998    0.998      0.392    0.964     1.000

=== Confusion Matrix ===

        a        b   <-- classified as
    11443      852 |        a = infected
    57818 24178074 |        b = clean
```

FIGURE 47. Combined, PART, Testing Results.

```
=== Summary ===

Correctly Classified Instances     23985950         98.9185 %
Incorrectly Classified Instances     262237              1.0815 %
Kappa statistic                          0.0794
Mean absolute error                      0.0108
Root mean squared error                  0.104
Total Number of Instances          24248187

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.931    0.011    0.042      0.931    0.080      0.196    0.960     0.039     infected
                 0.989    0.069    1.000      0.989    0.995      0.196    0.960     1.000     clean
Weighted Avg.    0.989    0.069    0.999      0.989    0.994      0.196    0.960     0.999

=== Confusion Matrix ===

        a        b   <-- classified as
    11443      852 |        a = infected
   261385 23974507 |        b = clean
```

FIGURE 48. Combined, REPTree, Testing Results.

```
=== Summary ===

Correctly Classified Instances     24235892         99.9493 %
Incorrectly Classified Instances      12295              0.0507 %
Kappa statistic                          0
Mean absolute error                      0.0005
Root mean squared error                  0.0225
Total Number of Instances          24248187

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.001     infected
                 1.000    1.000    0.999      1.000    1.000      ?        0.500     0.999     clean
Weighted Avg.    0.999    0.999    ?          0.999    ?          ?        0.500     0.999

=== Confusion Matrix ===

        a        b   <-- classified as
        0    12295 |        a = infected
        0 24235892 |        b = clean
```

FIGURE 49. Combined, Voted Perceptron, Testing Results.

3) RANSOMWARE

As can be seen in Figures 38 - 43, none of the classifiers correctly identify a single malware sample.

4) COMBINED

As shown in Figures 44 - 49, the combined data models start to correctly classify samples. In particular the PART and REPTree models perform very well with a relatively small number of false positives. The NaïveBayes and Decision Table perform reasonably as well.

5) EVALUATION

The test datasets provide an interesting outcome. You'll note that these performed incredibly poorly except for the combined dataset which succeeded with certain classifiers. This seems to be due to the lower amounts of data in the training sets. Since there is less data, there is less process diversity. This process diversity does not seem to be related to malware types, e.g. botnets, ransomware, or trojans, either. It seems that malware tends to share traits across families and variants and training across these spreads provides a level of robustness to the system that is demonstrated in the combined data testing.

Note that the important factor here is low false positives. The malicious samples are repeatedly taken which means that even if a malware process is missed the first time, it can be caught in the future. This means that even if the rate of flagging malware correctly is low, it doesn't mean that it wouldn't perform well in practice as long as its false positive rate is low.

Another point to make is that over the course of the testing there were two models that performed consistently better than the other models. These were the PART and REPTree classifiers. It is worth noting that these are both tree based classifiers which shows that trees can be used for simple identification of malicious process attributes. The other win here is that trees are fairly efficient meaning that in an identification system, they would not add much overhead.

C. PERFORMANCE

The last point to address here is the speed performance. The process monitor script was run on an Intel Core i7-6700k 4GHz processor. The machine was running 385 processes and the average time to iterate over all of these processes in the script was 24.748 seconds. This means that the overhead to run the process capture script is roughly 64 ms per process running on a given machine. This would of course be added to the amount of time necessary to classify a given instance. This would be dependent on the machine learning algorithm chosen, but would be fairly insignificant. This means that this system could be run repeatedly and quite frequently to ensure that malware is caught almost immediately upon entering a system.

In addition, the models themselves are anywhere from 2 to 10 kilobytes meaning that the memory needed to use them for classification is fairly low. This means that it could be run on low memory systems as well.

VI. SUMMARY

This section provides a brief summary of what was accomplished in this paper. First, a system was proposed that allows for cross platform evaluation of malicious process behavior. While this was tested on a Windows system, the solution only relies on a system's ability to support the process statistics gathering library SIGAR and Java, both of which are widely supported. The malware flagging system is also fairly low power and can be run as often as needed so it can be run more frequently to catch malware more quickly or can be run less frequently for better performance which would allow it to work on IOT and Android devices as well as personal computers.

The second contribution was that it evaluated multiple machine learning models on 4 different datasets and showed which models performed the best. This demonstration showed that tree based models seem to provide the most accurate classification method for this data.

It also showed that this identification could be performed quickly and efficiently. Due to the statistics being gathered being fairly accessible and the simplicity of the solution, it doesn't cause a large amount of overhead on the system. This means that the classification can be done as often as needed.

Lastly, it showed that having malicious data spread across different variants and families provides a robust system capable of flagging a wide variety of malware. This was shown based on the inaccuracy of the classifiers when used on individual malware families when compared with the success of the classifiers when trained on cross family malware datasets and evaluated on diverse variants. It also had fairly low false positive rates indicating that the attributes chosen in the data were fairly indicative of malicious processes and had little overfitting.

VII. FUTURE WORK

A few ideas will now be proposed for how to increase the effectiveness of this malware detection system.

The first is taking into account some standard telltale signs of malware. For instance, malware is typically installed to the “AppData” or “tmp” folder and provides a decent estimator of malicious behavior. It was not included in this implementation since it is operating system specific. That said it could be used in future implementations by simply checking a variety of different common malware installation folder paths. Another example would be utilizing processes’ tree structures. Malware typically spins off multiple processes to accomplish its goals so using the tree structure as part of the input to the classifier may help. Both these and other signs could be used as part of the dataset that the model is trained on to increase accuracy.

The second improvement that could be made is the usage of behavior over time. This system simply checks a process’ statistics at a given time. This could be made significantly more robust by taking multiple samples for regression classification. This modifies the machine learning to account for time. Another possible implementation of time based behavior identification that wouldn’t require modification of the existing classifier would be to check for sequential flags on a process. In other words, if a process is flagged as malicious by the model multiple times in succession, there is a high likelihood of it being malicious. This modification of the system would reduce the likelihood of false positives. This could be used to balance false positives with malware identification.

Lastly, this paper was designed on the prospect of making adaptable dynamic systems that are cross platform compatible. The reality, though, is that most platforms already have static systems in place. It might be beneficial to tie into these systems and leverage their abilities with the strengths of a dynamic system. The dynamic system could also be used to find potentially malicious processes to send samples of to the antivirus manufacturer for addition to the malware signature list. This would be a means of obtaining a large number of malicious samples that would take less time to process due to the high malicious classification accuracy of the dynamic system.

REFERENCES

[1] Cloonan, J. (2018). *Advanced Malware Detection - Signatures vs. Behavior Analysis*. [online] Infosecurity Magazine. Available at: https://www.infosecurity-magazine.com/opinions/malware-detection-signatures/ [Accessed 10 Apr. 2018].

[2] Ieeexplore.ieee.org. (2018). *Behavior-Based Malware Analysis and Detection - IEEE Conference Publication*. [online] Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6128413&tag=1 [Accessed 10 Apr. 2018].

[3] Pdfs.semanticscholar.org. (2018). *Behaviour Based Malware Detection*. [online] Available at: https://pdfs.semanticscholar.org/08ec/24106e9218c3a65bc3e16dd88dea2693e933.pdf [Accessed 10 Apr. 2018].

[4] Ieeexplore.ieee.org. (2018). *Malware Detection with Deep Neural Network Using Process Behavior - IEEE Conference Publication*. [online] Available at: https://ieeexplore.ieee.org/document/7552276/ [Accessed 10 Apr. 2018].

[5] TUDelft. (2018). *Using endpoints process information for malicious behavior detection*. [online] Available at: https://repository.tudelft.nl/islandora/object/uuid:e1678077-9056-47ac-82e6-2762bfb40a63?collection=education [Accessed 10 Apr. 2018].

[6] Cse.psu.edu. (2018). *pBMDS: A Behavior-based Malware Detection System for Cellphone Devices*. [online] Available at: http://www.cse.psu.edu/~sxz16/papers/pBMDS.pdf [Accessed 10 Apr. 2018].

[7] Pdfs.semanticscholar.org. (2018). *In-execution dynamic malware analysis and detection by mining information in process control blocks of Linux OS*. [online] Available at: https://pdfs.semanticscholar.org/6c0e/9f196a82098804ed8f95a6fbe1a3886f15cb.pdf [Accessed 10 Apr. 2018].

[8] Arts.units.it. (2018). *Spotting the Malicious Moment: Characterizing Malware Behavior Using Dynamic Features*. [online] Available at: https://arts.units.it/retrieve/handle/11368/2889183/137832/2016-IWSMA-SpottingMaliciousMoment.pdf [Accessed 10 Apr. 2018].

[9] Onlinelibrary.wiley.com. (2018). *Smart malware detection on Android*. [online] Available at: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sec.1340 [Accessed 10 Apr. 2018].

[10] Milosevic, J., Malek, M. and Ferrante, A. (2018). *A Friend or a Foe? Detecting Malware using Memory and CPU Features*. [online] Scitepress.org. Available at: http://www.scitepress.org/DigitalLibrary/PublicationsDetail.aspx?ID=+yi+YAt4Z8o=&t=1 [Accessed 10 Apr. 2018].

[11] Hyperic SIGAR. (2018). *SIGAR*. [online] Available at: https://github.com/AlexYaruki/sigar [Accessed 10 Apr. 2018].

[12] Cs.waikato.ac.nz. (2018). *Weka 3 - Data Mining with Open Source Machine Learning Software in Java*. [online] Available at: https://www.cs.waikato.ac.nz/ml/weka/ [Accessed 10 Apr. 2018].

[13] GitHub. (2018). *ytisf/theZoo*. [online] Available at: https://github.com/ytisf/theZoo [Accessed 11 Apr. 2018].