A Framework for the Validation of Network Artifacts

Livinus Obiora Nweke[1], Stephen D. Wolthusen[1,2], and Luigi V. Mancini[3]

[1] Information Security and Communication Technology, Norwegian University of Science and Technology (NTNU), Norway
[2] School of Mathematics and Information Security, Royal Holloway, University of London, Egham, United Kingdom
[3] Dipartimento Di Informatica, La Sapienza Univerista Di Roma, Roma, Italy

Abstract

Digital forensics has been of growing interest over the past ten to fifteen years despite being a relatively new scientific field. Many technologies and forensics processes have been developed to meet the growing number of cases relying on digital artifacts. In this paper, we presents a framework for the validation of network artifacts in digital forensics investigations. Validation in the context of this work, refers to the overall probability of reaching the correct inferences about the artifacts, given a specific method and data. The main hypothesis of this work is that the validity of network artifacts can be determined based on probabilistic modelling of internal consistency of artifacts. The framework consists of three phases, namely: data collection, feature selection, and validation process. We demonstrate the functionality of the proposed framework using network artifacts obtained from Intrusion Detection Systems. Also, we assume that the initial acquisition of the network artifacts is forensically sound and steps are taken to ensure that the integrity of the artifacts is maintained during the data collection phase. A Monte Carlo Feature Selection and Interdependency Discovery algorithm is applied in selecting the informative features, while logistic regression is used as the probabilistic modelling methodology for the validation process. The experiment results show the validity of the network artifacts and can serve as a scientific methodology to support the initial assertions drawn from the network artifacts.

Contents

1 Introduction

Digital forensics has been of growing interest over the past ten to fifteen years despite being a relatively new scientific field [12]. This can be attributed to the large amount of data being generated by modern computer systems, which has become an important source of digital artifacts. The proliferation of modern computer systems and the influence of technology on the society as a whole have offered many opportunities not previously available. Unfortunately, this trend has also offered the same opportunities to criminals who aim to misuse the systems and as such leads to an increase in the number of recent cyber crimes [22].

Many technologies and forensics processes have been developed to meet the growing number of cases relying on digital artifacts. A digital artifact can be referred to as digital data that support or refute a hypothesis about digital events or the state of digital data [7]. This definition includes artifacts that are not only capable of entering into a court of law but may have investigative value. Network artifacts, on the other hand, are among the type of digital artifacts that have attracted a lot of attention in recent times. This is as a result of pervasive cyber crimes being witnessed nowadays. Network artifacts are digital artifacts which provide insight into network communications. As observed in [9], Dynamic Host Configuration Protocol servers, Domain Name System servers, Web Proxy Servers, Intrusion Detection Systems, and firewalls all can generate network artifacts which can be helpful in digital forensics investigations. Thus, there is a need to validate such artifacts to make them admissible in court proceedings.

Establishing the validity of network artifacts and digital artifacts, in general, is very challenging in digital forensics considering that the concept of validation has different meanings in the courtroom compared with research settings [15]. Validation, as applied in this paper, refers to the overall probability of reaching the correct inferences about the artifacts, given a specific method and data. It requires the verification of relevant aspects of the artifacts and estimating the error rate. The goal is to increase the confidence about the inferences drawn from the artifacts and also, to employ scientific methodology in doing so, as recommended by President's Council of Advisors on Science and Technology [16].

Eadaoin, Niamh, and Sue [15] noted that practitioners face great difficulty in meeting the standards of scientific criteria in courts. Investigators are required to estimate and describe the level of uncertainty underlying their conclusions to help the Judge or the Jury determine what weight to attach. Unfortunately, the field of digital forensics does not have formal mathematics or statistics to evaluate the level of uncertainty associated with digital artifacts [8]. There is currently lack of consistency in the way the reliability or accuracy of digital artifacts are assessed, partly because of the complexity and multiplicity of digital systems. Furthermore, the level of uncertainty investigators assigned to their findings is usually influenced by their experience.

Most of the existing research in digital forensics focuses on identification, collection, preservation, and analysis of digital artifacts. However, not much attention has been paid to the validation of digital artifacts and network artifacts in particular. Artifacts acquired during a digital forensics investigations could be invalidated if reasonable doubts are raised about the trustworthiness of the artifacts. The Daubert criteria are currently recognized as benchmarks for determining the reliability of digital artifacts [3]. It is common practice to follow the five Daubert tests in court proceedings for evaluating the admissibility of digital artifacts. However, these requirements are not exhaustive nor entirely conclusive, as artifacts may be accepted even when do not meet all the criteria. The requirements are generalized by Garrie and Morrissy [3]:

- Testing: can the scientific procedure be independently tested?
- Peer Review: Has the scientific procedure been published and subjected to peer review?

- Error rate: Is there a known error rate, or potential to know the error rate, associated with the use of the scientific procedure?
- Standards: Are there standards and protocols for the execution of the methodology of the scientific procedure?
- Acceptance: Is the scientific procedure generally accepted by the relevant scientific community?

The known or potential error rates associated with the use of the scientific procedure to which the Daubert requirements refer can include a number of parameters such as confidence interval, the statistical significance of a result, or the probability that a reported conclusion is misleading. For this work, statistical error is used. Statistical error as used in this paper, is the deviation between actual and predicted values, estimated by a measure of uncertainty in prediction. Also, selecting the appropriate statistical model is crucial in producing valid scientific methods with low estimated error rates and hence, it is important to show that the chosen model is actually a good fit.

In this paper, we propose a framework that can be used for the validation of network artifacts based on probabilistic modelling of the internal consistency of artifacts. The focus of this work is on the use of logistic regression analysis as the probabilistic modelling methodology in determining the validity of network artifacts. Network artifacts obtained from Intrusion Detection Systems are used as the domain example to demonstrate the working of the proposed framework. First, we assume the initial acquisition of the network artifacts is forensically sound and the integrity of the artifacts is maintained during the data collection phase of the proposed framework. Next, the selection of the subsets of the features of the artifacts for validation is done. Then, logistic regression analysis is applied in the validation stage of the proposed framework. Lastly, inferences are drawn from the results of the validation process as it relates to the validity of the network artifacts. The results of the validation can be used to support the initial assertions made about the network artifacts and can serve as a scientific methodology for supporting the validity of the network artifacts.

The rest of this paper is organised as follows. Section 2, provides definitions as well as background on digital forensics. The proposed framework is presented in section 3. Section 4 presents results of experiments carried out to demonstrate the functionality of the proposed framework. Section 5 offers insights into the results obtained and the limitations of the approach. Finally, section 6 concludes the paper and presents future works.

2 Background

This section provides definitions as well as background on digital forensics. The general concept of digital forensics and the goal of digital forensics are explored. It continues with an in-depth analysis of the admissibility of network artifacts to court proceedings and then concludes with a discussion on the importance of validation of network artifacts.

2.1 Digital Forensics

The term forensics comes from the Latin forum and the requirement to present both sides of a case before the judges (or jury) appointed by the praetor [7]. Forensics science is derived from diverse disciplines, such as geology, physics, chemistry, biology, computer science, and mathematics, in order to study artifacts related to crime. Digital forensics, on the other hand,

is a branch of forensics concerned with artifacts obtained from any digital devices. It can be defined as a science using repeatable process and logical deduction to identify, extract and preserve digital artifacts and can be referred back to as early as 1984 when the FBI began developing programs to examine computer artifacts [21]. Thus, digital forensics is an investigative technique used to uncover and gather artifacts relating to computer crimes.

During the last few years, there has been a pervasive incident of computer crimes which have led to a growing interest in digital forensics. The field of digital forensics has been primarily driven by vendors and applied technologies with very little consideration being given to establishing a sound theoretical foundation [4]. Although this may have been sufficient in the past, there has been a paradigm shift in recent times. The judiciary system has already begun to question the scientific validity of many of the ad-hoc procedures and methodologies and is demanding proof of some sort of a theoretical foundation and scientific rigor [16].

2.2 The Importance of Validation of Network Artifacts

Validation requires confidence about the inferences drawn from the artifacts. Court proceedings require that artifacts are validated and the reliability of those artifacts critically evaluated before presentation to the court. It has been observed in [15] that digital artifacts and in particular, network artifacts face difficulty in meeting the standards for scientific criteria for use in courts. Lack of trust in the digital forensics process and absence of an established set of rules for evaluation makes it possible to raise doubts about the reliability of digital artifacts presented in courts. Therefore, re-echoing the importance of the validation of the artifacts.

The validation of network artifacts involves ensuring the trustworthiness of the artifacts and ensuring that the reliability of the artifacts can be dependent upon in the court. Validation requires reporting not just the process used in the validation but also, the uncertainty in the process [16]. Reporting the error rate and the accuracy in the process used in the validation of the network artifacts will provide the judge or the jury the basis on which a decision can be reached to use or not to use the network artifacts in court proceedings [18]. Furthermore, it is pointed out in [3] that only scientifically proven methods that are verifiable and can be validated, should be used in evaluating digital artifacts to be used in courts.

The absence of a clear model for the validation of network artifacts and digital artifacts, in general, is one of the fundamental weaknesses confronting practitioners in the emerging disciple of digital forensics. If reasonable doubts can be raised about the validity of artifacts, the weight of those artifacts in a legal argument is diminished. It is then easy for the defense attorneys to challenge the use of the artifacts in the court. Thus, it is important that digital artifacts are validated using a scientifically proven methodology to increase the confidence in the inferences drawn from the artifacts and to show the certainty in the methodology used.

3 The Framework for the Validation of Network Artifacts

In this section, the proposed framework for the validation of network artifacts is described. The proposed framework for the validation of network artifacts comprises of three stages, namely: data collection, feature selection, and validation process. In the first stage of the proposed framework, network artifacts to be validated are collected, and it is assumed that the initial acquisition of the artifacts is forensically sound. The next stage involves selecting subsets of features of the network artifacts to be used for the validation process. Lastly, the actual validation process is performed using probabilistic modelling methodology. The following subsections

provide the description of each of the stages of the proposed framework for the validation of network artifacts.

3.1 Data Collection

Data collection is the first stage of the proposed framework, and it involves the collection of the network artifacts to be validated. There are requirements that the data collection process must meet, to ensure that the artifacts are forensically sound and can be used in court proceedings. To understand these requirements, it is important to understand what is meant by the term "forensically sound". Artifacts are said to be forensically sound if they acquired with two clear objectives set out in [20]:

- The acquisition and subsequent analysis of electronic data has been undertaken with all due regard to preserving the data in the state in which it was first discovered.
- The forensic process does not in any way diminish the probative value of the electronic data through technical, procedural or interpretive errors.

In order to meet these objectives, several processes and procedures needs to be adopted. The two widely used approaches as observed in [20] are the "Good Practice Guide for Computer Based Electronic Evidence" published by the Association of Chief Police Officers (United Kingdom) [17] and the International Organization on Computer Evidence (now Scientific Working Group on Digital Evidence(SWGDE)) [19]. The "Good Practice Guide for Computer Based Electronic Evidence" published by the Association of Chief Police Officers (United Kingdom) [17] lists four important principles related to the recovery of artifacts:

- No action taken by law enforcement agencies or their agents should change data held on a computer or storage media which may subsequently be relied upon in court.
- In exceptional circumstances, where a person finds it necessary to access original data held on a computer or on storage media, that person must be competent to do so and be able to give evidence explaining the relevance and the implications of their actions.
- An audit trail or other record of all processes applied to computer based electronic evidence should be created and preserved. An independent third party should be able to examine those processes and achieve the same result.
- The person in charge of the investigation (the case officer) has overall responsibility for ensuring that the law and these principles are adhered to.

Similarly, the SWGDE has the following guiding principle [19]:

"The guiding principle for computer forensic acquisitions is to minimize, to the fullest extent possible, changes to the source data. This is usually accomplished by the use of a hardware device, software configuration, or application intended to allow reading data from a storage device without allowing changes (writes) to be made to it."

For the purpose of this work and given the constraints under which the research is undertaken, it is assumed that the acquisition of the network artifacts to be used for the data collection stage of the framework is forensically sound and that chain of custody is maintained. Also, it is assumed that the data collection process follows a forensically sound methodology to ensure that the integrity of the network artifacts to be validated are preserved.

3.1.1 Features of the Domain Example

Network artifacts obtained from Intrusion Detection Systems are used as the domain example to demonstrate the functionality of the proposed framework. Network artifacts can be characterized using flow-based features. A flow is defined by a sequence of packets with the same values for Source IP, Destination IP, Source Port, Destination Port and Protocol (TCP or UDP). CICFlowMeter [2] is used to generate the flows and calculate all necessary parameters. It generates bidirectional flows, where the first packet determines the forward (source to destination) and backward (destination to source) directions, hence more than 80 statistical network traffic features such as Duration, Number of packets, Number of bytes, Length of packets, etc. can be calculated separately in the forward and backward directions.

Also, the CICFlowMeter has additional functionalities which include, selecting features from the list of existing features, adding new features, and controlling the duration of flow timeout. The output of the application is the CSV format file that has six columns labeled for each flow (FlowID, SourceIP, DestinationIP, SourcePort, DestinationPort, and Protocol) with more than 80 network traffic analysis features. It is important to note that TCP flows are usually terminated upon connection teardown (by FIN packet) while UDP flows are terminated by a flow timeout. The flow timeout value can be assigned arbitrarily by the individual scheme e.g., 600 seconds for both TCP and UDP [2]. Therefore, the output of the application forms the basis of the CICIDS2017 dataset [10], used as network artifacts for this work.

3.2 Feature Selection

The next step after the data collection stage is the selection of subsets of the features of the network artifacts to be used for the validation process. A typical nature of network artifacts is high-dimensionality of the features of the artifacts. It is only natural that after the collection of the network artifacts, subsets of the features of the network artifacts should be selected to remove redundant or non-informative features. This is because the successful removal of non-informative features aids both the speed of model training during the validation process and also, the performance and the interpretation of the results of the model.

The feature selection technique to be deployed in the feature selection stage of the proposed framework would depend on the nature of network artifacts to be validated. On one hand, it may be the case of simple network artifacts where the investigator is familiar with the features of the network artifacts and is able to select the subsets of the features that are most relevant from the network artifacts and then use in the validation process. On the other hand, it may be the case of complex network artifacts that requires the use of feature selection algorithm to select the subsets of the features of the network artifacts that are most relevant to be used for the validation process.

There are several approaches that can be deployed for selecting subsets of features where the network artifacts to be validated are complex. These approaches can be grouped into three, namely: filter methods, wrapper methods, and embedded methods [1]. Filter methods select subsets of features on the basis of their scores in various statistical tests for their correlation with the outcome variable. Some common filter methods are correlations metrics (Spearman, Pearson, Distance), Chi-Squared test, Anova, Fisher's Score, etc. In the case of wrapper methods, subsets of features are used to train a model, then based on the inferences that are drawn from the model, features are added or removed from the subset. Forward Selection, Backward Elimination are some of the examples of wrapper methods. Embedded methods are the algorithms that have their own built-in feature selection methods. An example of embedded methods is Least Absolute Shrinkage and Selection Operator (LASSO) regression.

Given the complexity of the network artifacts used for this work, several R packages of the feature selection algorithm are applied on the network artifacts to ascertain which one is best suited for the artifacts. An example of the R packages that is used is Boruta algorithm [14], which is a wrapper built around the random forest classification algorithm. The goal of the algorithm is to capture all the important features of the artifacts with respect to an outcome. It achieves this by first duplicating the artifacts, train a classifier such as a Random Forest Classifier on the artifacts, obtain the importance of each of the features in the classification, and using this importance to select the most relevant features. The drawback of Boruta algorithm is the difficulty in understanding the interdependencies of the selected subsets of features. Hence, Monte Carlo Feature Selection and Interdependency Discovery (MCFS-ID) algorithm [13], which does not only provide the ranking of features but also, includes a directed ID-Graph that presents interdependencies between informative features is used for feature selection stage of the proposed framework.

3.2.1 Monte Carlo Feature Selection (MCFS)

The MCFS algorithm is a feature selection algorithm that is based on intensive use of classification trees. The goal of the algorithm is to select s subsets of the original d features, each with a random selection of m features. It involves dividing each of the subsets into a training and test set with 2/3 and 1/3 of the objects respectively. This division is repeated t times, and a classification tree is trained on each training set. The main step involved in the MCFS algorithm is shown in the figure 1

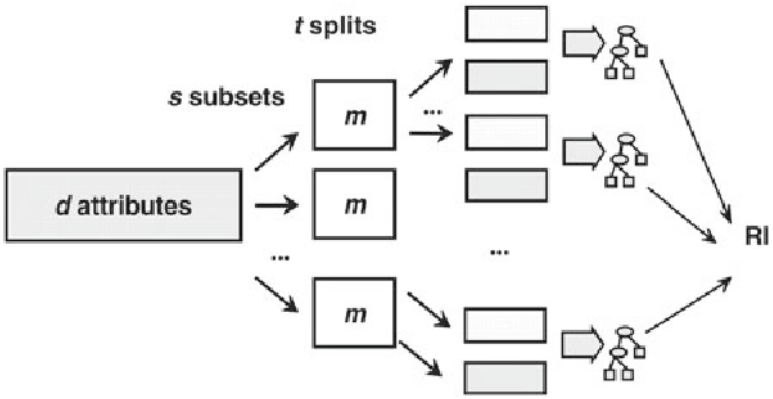

Figure 1: Overview of the MCFS Procedure. Reproduced from Draminski and Koronacki [13]

3.3 Validation Process

After the selection of subsets of the dataset, the next step is the validation of the network artifacts based on probabilistic modeling of internal consistency of the artifacts. The validation process involves using the selected subsets of the features obtained during the feature selection phase and then applying probabilistic modelling methodologies to observe the internal consistency of the artifacts and make inference about the validity of the network artifacts. Logistic regression analysis is used as the probabilistic methodology for the validation process.

The nature of network artifacts to be validated determines the choice of the probabilistic modeling methodology to be used in the validation process. There are cases where a linear relationship may exist between the dependent variable and the independent variables of the network artifacts. In such cases, it is only natural that linear regression model as the probabilistic modelling methodology is used. Linear regression is used to predict the value of an outcome variable (dependent variable) Y based on one or more input predictors variables (independent variables) X [5]. The goal is to establish a linear relationship between the predictor variables and the response variable in such a way that we can use this relationship to estimate the value of the response Y, when only the predictors X values are known.

However, in the case of the domain example used for this work, the network artifacts are such that the dependent variable is categorical and hence, linear regression model is inadequate. The plot of the such data appears to fall on parallel lines, each corresponding to a value of the outcome (1 = Benign, and 0 = PortScan). Thus, the variance of residuals at specific values of X equals p * (1-p), where p is the proportion of a particular outcome at specific X values. Also, the categorical nature of the outcome makes it impossible to satisfy either the normality assumption for residuals or the continuous, unbounded assumptions on Y. Hence, the significance tests performed on regression coefficients are not valid even though least squares estimates are unbiased. In addition, even if the categorical outcomes are calculated as probabilities, the predicted probabilities obtained from the linear regression model can exceed the logical range of 0 to 1. This is because of lack of provision in the model to restrict the predicted values.

Consequently, to overcome the inherent limitations of linear regression in handling the nature of the network artifacts under review, logistic regression model is used as the probabilistic modelling methodology. The justification for the choice of logistic regression is because it has been shown to be superior in dealing with categorical outcome variables when compared to alternative probabilistic modelling methodologies (e.g. discriminant function analysis, log-linear models, and linear probability models) for analyzing categorical outcome variables [6]. Also, in terms of classification and prediction, logistic regression has shown to produce fairly accurate results [11].

3.3.1 Logistic Regression

Logistic regression analysis is a type of regression analysis used in analyzing and explaining the relationship between independent variables and a categorical dependent variable and computing the probability of occurrence of an event by fitting data to a logistic curve [6]. The goal is to predict the outcome of a categorical dependent variable e.g whether the network traffic label is benign or malicious, based on the predictor variables e.g the selected subsets of the features of the network traffic. A typical regression model has the following general appearance:

$$y = b_0 + b_1x_1 + b_2x_2 + ... + b_px_p$$

where y is the estimated outcome variable value for the true Y (Benign or PortScan), b_0 is the constant of the equation, $b_1, ..., b_p$ are estimated parameters corresponding to predictor values $x_1, ..., x_p$ (selected subset of features); b_0 is alternatively called the Y-intercept; $b_1, ..., b_p$ are slopes, regression coefficients, or regression weights.

4 Experiment Results

In this section, the results of experiments carried out is presented to demonstrate the functionalities of the proposed framework. The section begins with a description of the experimental

setup and the dataset used for the experiments. Also, the results and analysis of the experiments are presented.

4.1 Experimental Setup

The experiments were conducted on 64-bit Microsoft Windows-based operating system, using R x64 3.5.0. The justification for the choice of R for this research work is based on the fact that the feature selection algorithm and the logistic regression model used for this work have already been implemented via R packages.

4.2 Dataset

The CICIDS2017 dataset [10] used for this work contains benign and PortScan attack. The dataset consists of labeled network flows, the corresponding profiles and the labeled flows (CSV), that is publicly available for researchers. In the dataset, there were 123789 benign sessions due to normal network traffic, and 153444 malicious network traffic due to PortScan attack. A summary of the dataset is shown in the Figure 2.

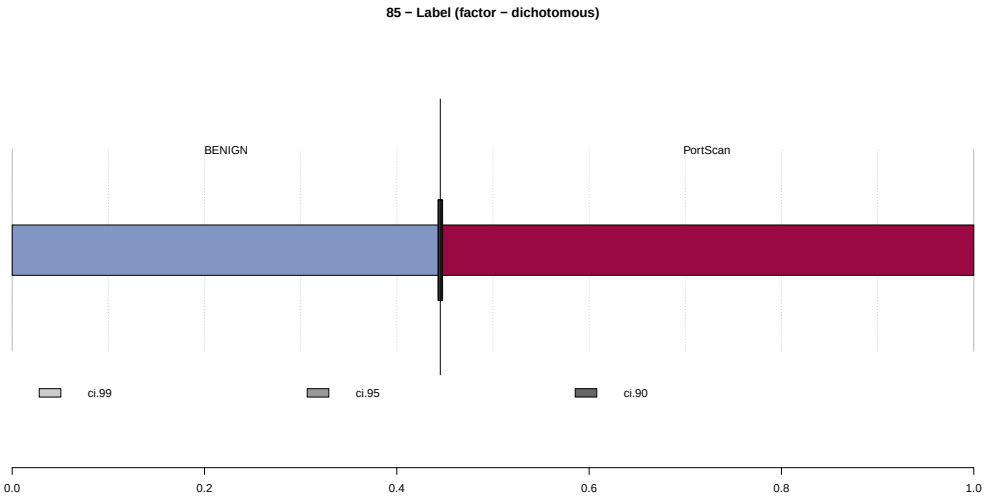

Figure 2: Dataset Summary

4.3 Feature Selection Experiment Results

The MCFS-ID algorithm is applied to the dataset described in 4.2. After successfully running the MCFS-ID algorithm, the next step is to check convergence of the algorithm. This is shown in Figure 3. The distance function shows the difference between two consecutive rankings; zero means no changes between two rankings (the left Y axis in figure 6.2). Common part gives the fraction of features that overlap for two different rankings (the right Y axis in figure 6.2). Ranking stabilizes over a number of iterations: distance tends to zero and common part tends to 1. Beta1 shows the slope of the tangent of a smoothed distance function. If Beta1 tends to 0 (the right Y axis) then the distance is given by a flat line.

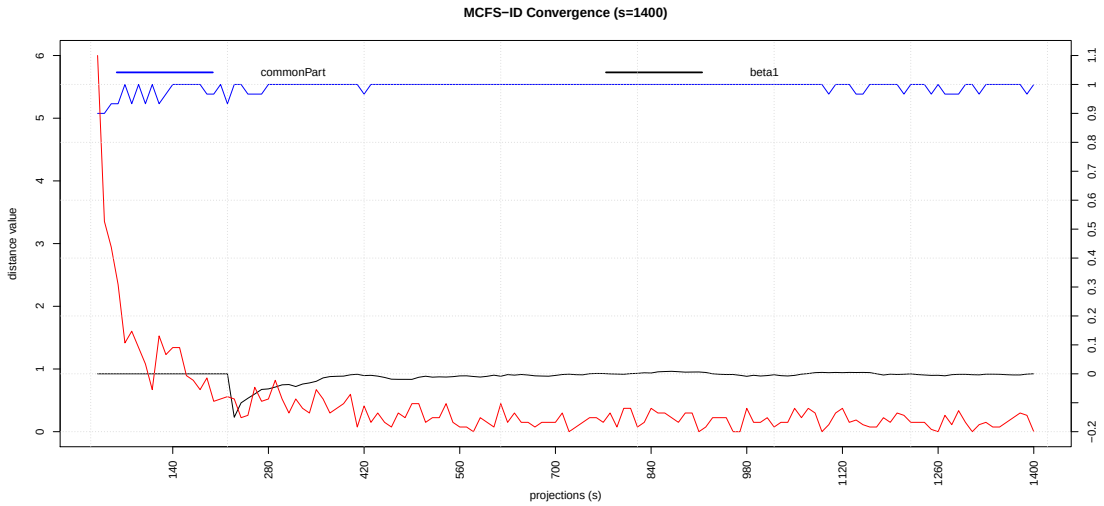

Figure 3: Distance Function

Next, it is important to see the relative importance (RI) values of the features. This is achieved by plotting the RI values in decreasing order from the top as shown in Figure 4. The line with red/gray dots gives RI values, the blue vertical barplot gives difference δ between consecutive RI values. Informative features are separated from non-informative ones by the cutoff value and are presented in the plot as red and gray dots, respectively.

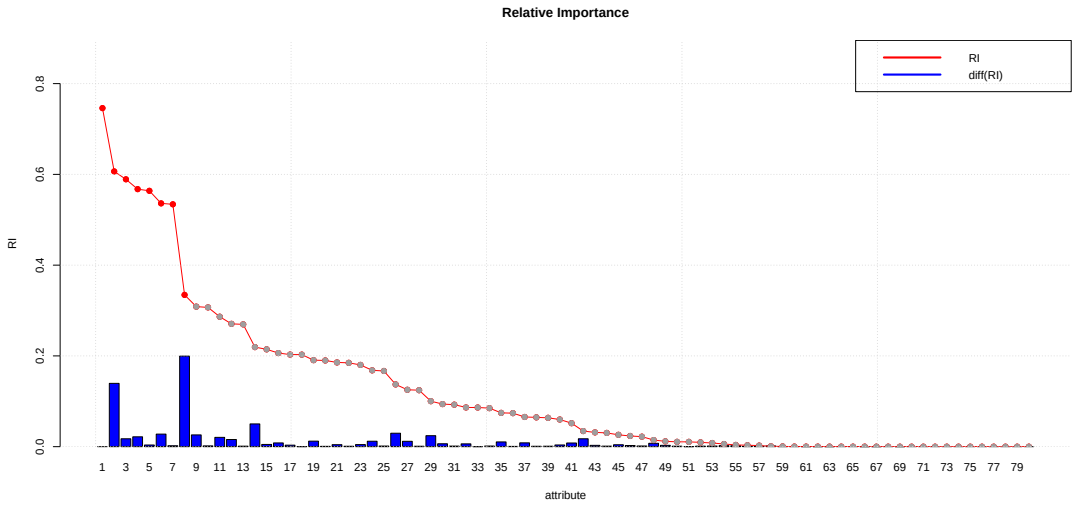

Figure 4: Relative Importance

Similarly, labels and RIs of the top features can be review. The resulting plot is presented

in Figure 5. It can be observed that all the eight features are highly important and their RIs are much higher than those of other features. The set of informative features is flagged in red in the plot.

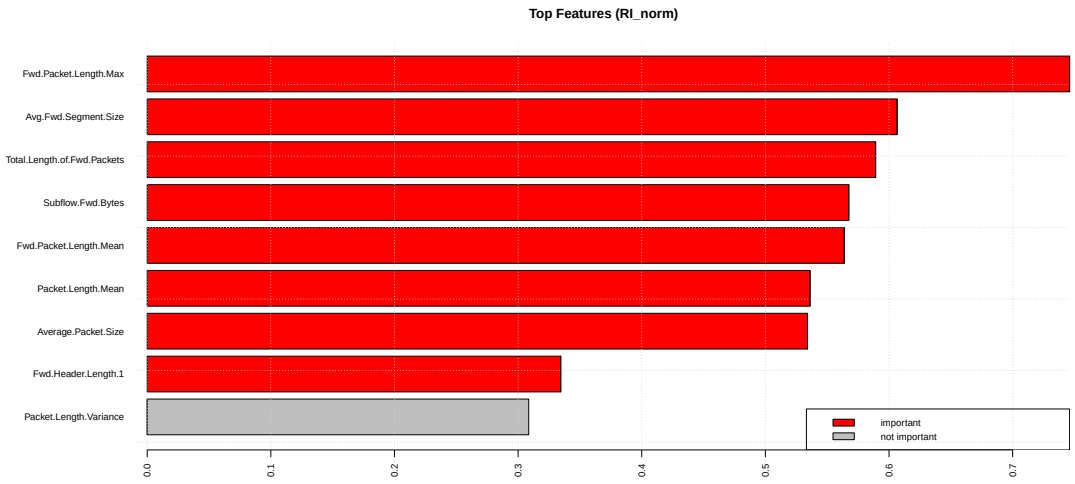

Figure 5: Features Selected

4.4 Logistic Regression Analysis Experiment Results

The values obtained after running the logistic regression model using the subset of features selected in the feature selection phase of the framework are listed in Table 1. The table reports the estimated coefficients, the statistical significance of each of the dependent variables, the number of observations, log likelihood, and Akaike Information Criteria (AIC). Also, Table 2 presents a final summary report of standardized coefficients. Standardized coefficients (or estimates) are usually used when the predictors are expressed in different units.

For the purpose of this research, the goal of the validation process is to support inferences drawn from the artifacts, that is, to provide empirical support for the classification of the artifacts as benign or PortScan (malicious). The method used for the experiment evaluates the ability of the logistic regression model to correctly predict the outcome category (Benign or PortScan) of the network artifacts.

Another important graph used to depict the reliability of the categorical outcome variables of the network artifacts is the ROC curve shown in Figure 6. In the ROC curve, the true positive rate (Sensitivity) is plotted as function of the false positive rate (Specificity) for different cut-off points of the parameter. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. The area under the ROC curve (AUC) is a measure of how well (accuracy) a parameter can distinguish between the two categorical outcome variables.

Table 1: Logistic Regression Model Estimation

	Dependent variable:
	Label
Fwd.Packet.Length.Max	0.073^{***}
	(0.006)
Avg.Fwd.Segment.Size	-1.117^{***}
	(0.007)
Total.Length.of.Fwd.Packets	-0.010^{*}
	(0.006)
Average.Packet.Size	0.031^{***}
	(0.001)
Fwd.Header.Length	-0.102^{***}
	(0.001)
Constant	6.966^{***}
	(0.041)
Observations	200,527
Log Likelihood	−31,584.900
Akaike Inf. Crit.	63,181.800
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 2: Final Summary Report

Statistic	N	Mean	St. Dev.	Min	Max
Standardized.Coeff	5	−13.944	24.250	−48.138	13.083

5 Discussion

The study of the validity of network artifacts using logistic regression as the probabilistic modelling methodology for modeling the internal consistency of artifacts demonstrates that inferences drawn from the artifacts can be supported using statistical results. Indeed, Table 1 depicts important statistics that support the validity of the artifacts used for the study. All the selected subsets of the features of the artifacts used for the validation process are highly significant in predicting the dependent variable. Also, the log likelihood test suggests that the logistic regression model used for the validation process is better than the null model. In the same way, the Akaike Information Criterion value indicates that the logistic regression model used for the validation process is a good fit.

Also, it is important to discuss the distribution of the network artifacts used for the exper-

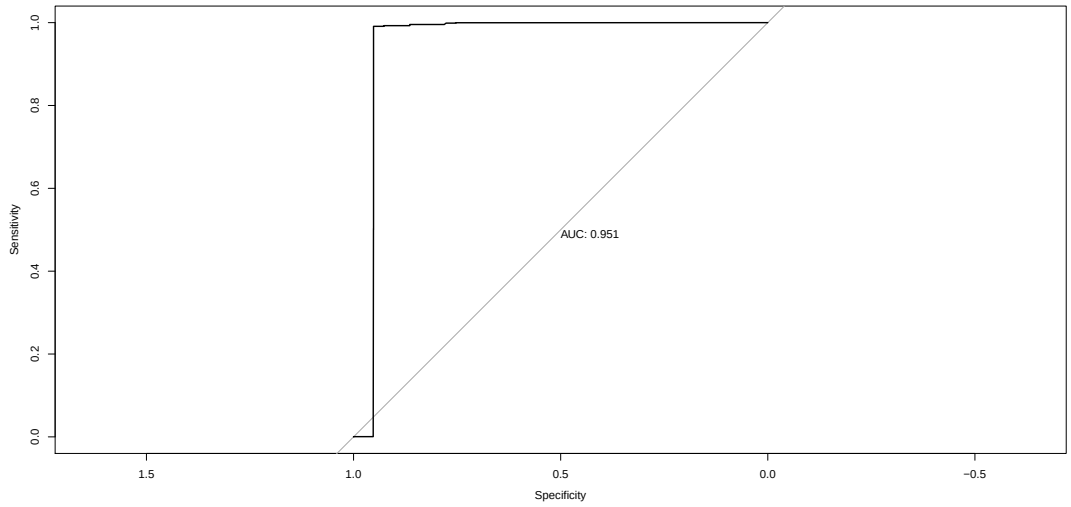

Figure 6: ROC Curve

iments. The summary of the statistical distribution of the network artifacts is given in Table 2. The standardized coefficients explains how increases in the independent variables affect the dependent variable. It aids in establishing a relationship between the independent variables and the dependent variable. Also, it can be inferred from the table that the nature of network artifacts used for the validation process follows a normal distribution and as such, provides a useful basis for interpreting the artifacts in terms of the true positive fraction (sensitivity) and the false positive fraction (specificity).

The ROC curve in Figure 6 graphically displays the trade-off between the true positive fraction and the false positive fraction and it is useful in describing how well a test discriminates between cases with and without a certain condition. An ROC curve is based on the notion of a separator scale, on which results for the Benign and PortScan form a pair of overlapping distributions. The complete separation of the two underlying distribution implies a perfectly discriminating test as in the case of the result from the experiment, while complete overlap implies no discrimination. The area under the curve (AUC) as shown in Figure 6 summarizes the entire location of the ROC curve rather than depending on a specific operating point. The AUC is an effective and combined measure of sensitivity and specificity that describes the inherent validity of the network artifacts.

However, the limitations of research method used have to do with the initial acquisition of the network artifacts and the data collection phase of the framework. It is assumed that the initial acquisition of the network artifacts is forensically sound and that the data collection phase of the framework ensured the integrity of the network artifacts is maintained. These are very strong assumptions that require rigorous processes and procedures to be achieved. This is because it is possible to raise doubts about the reliability of the process used in acquiring the network artifacts and also to claim that the tools used in the data collection phase of the framework may have altered the network artifacts in some way. In addition, if the initial classification of the artifacts as benign or malicious is achieved using probabilistic method, the use of probabilistic methodology for the validation process will not provide useful information

to support or refute the validity of the artifacts.

Notwithstanding the limitations of this study, the findings are very important in the validation of network artifacts. Logistic regression has been used in several fields for classification and predictions but there is little or no work in digital forensics where it has been applied. Its ability to show the significance of each of the independent variables in the classification of the dependent variable can be used in other areas of digital forensics. Also, measuring the contributions of the individual predictors can help in deciding which of the independent variables can be considered seriously as an artifact in proving or disproving the merit of a case.

6 Conclusions

In this paper, a framework for the validation of network artifacts is presented. The working of the proposed framework is demonstrated using a publicly available dataset as the network artifacts. It is assumed that the initial acquisition of the network artifacts is forensically sound and that the data collection stage of the proposed framework guaranteed the integrity of the network artifacts. The first experiment involves the use of Monte Carlos Feature Selection algorithm to select subsets of the features of the artifacts to be used for the validation process. Considering the nature of the network artifacts, logistic regression is then applied to the selected subsets of the features to check the internal consistency of the artifacts. Results from the experiments show the validity of the network artifacts and can be used as a scientific methodology to support inferences drawn from the network artifacts in court proceedings.

In future work, it is possible to extend the proposed framework to incorporate processes and procedures to ensure that the initial acquisition of the network artifacts is forensically sound and ensuring that the data collection stage of the proposed framework maintains the integrity of the network artifacts. The achievement this, requires setting up a lab to emulate the actual environment where the network artifacts are generated and collected. Such an enhanced solution will be able to address any doubts that could be raised on the reliability of the initial acquisition of the network artifacts and the integrity of the data collection process of the proposed framework.

References

[1] De Silva Anthony Mihirana and Leong Philip H. W. Grammar based feature generation. *Grammar-Based Feature Generation for Time-Series Prediction*, January 2015.

[2] Lashkari Arash Habibi, Draper-Gil Gerard, Mamun Mohammad Saiful Islam, and Ghorbani Ali A. Characterization of tor traffic using time based features. In Paolo Mori, Steven Furnell, and Olivier Camp, editors, *Proceedings of the 3rd International Conference on Information Systems Security and Privacy, ICISSP 2017, Porto, Portugal, February 19-21, 2017.*, pages 253–262. SciTePress, 2017.

[3] Garrie B. Daniel and Morrissy J. David. Digital Forensic Evidence in the Courtroom: Uderstanding Content and Quality. *Nw. J. TECH. & INTELL. PROP.*, 12(2), 2014.

[4] Lillis David, Becker Brett A., O'Sullivan Tadhg, and Scanlon Mark. Current challenges and future research areas for digital forensic investigation. *CoRR*, abs/1604.03850, 2016.

[5] Freedman David A. *Statistical Models: Theory and Practice.* Cambridge University Press, New York, NY, USA, 2010.

[6] Hosmer Jr. David W., Lemeshow Stanley, and Sturdivant Rodney X. *Applied Logistic Regression.* John Wiley & Sons, Inc., New Jersey, NJ, USA, 2013.

[7] Casey E. *Digital Evidence and Computer Crime: Forensic Science, Computers, and the Internet*. Academic Press, 2011.

[8] Vincze Eva A. Challenges in digital forensics. *Police Practice and Research*, 17(2):183–194, January 2016.

[9] Fraser Gordon. Building a Home Network Configured to Collect Artifacts for Supporting Network Forensic Incident Response. *SANS Institute InfoSec Reading Room*, 2016.

[10] Sharafaldin Iman, Lashkari Arash Habibi, and Ghorbani Ali A. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. *4th International Conference on Information Systems Security and Privacy (ICISSP)*, 2018.

[11] Holden Jocelyn E., Finch W. Holmes, and Kelley Ken. A comparison of Two-Group Classification Methods. *SAGE journals*, 71(5), May 2011.

[12] Conlan Kelvin, Baggili Ibrahim, and Breitinger Frank. Anti-forensics: Futhering digital forensic science through a new extended, granular taxonomy. In *Procedings of the 16th Annual USA Digital Forensics Research Conference*, pages S66–S75, USA, 2016. Elservier.

[13] Draminski Michal and Koronacki Jacek. rmcfs: An R Package for Monte Carlo Feature Selection and Interdependency Discovery. *Journal of Statistical Software*, 2018.

[14] Kursa Miron Bartosz. Package 'Boruta'. 2018.

[15] Eadaoin O'Brien, Niamh Nic Daeid, and Sue Black. Science in the court: pitfalls, challenges and solutions. *Phil. Trans. R. Soc. B*, May 2015.

[16] President's Council of Advisors on Science and Technology. *Report to the President Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*. Executive Office of the President, Washington, DC, 2016.

[17] Association of Chief Police Officers (United Kingdom). *ACPO Good Practice Guide for Digital Evidence*. Police Central e-crime Unit, United Kingdom, 2012.

[18] Scientific Working Group on Digital Evidence. *SWGDE establishing confidence in digital forensic results by error mitigation analysis*. Scientific Working Group on Digital Evidence, 2017.

[19] Scientific Working Group on Digital Evidence. *SWGDE Best Practices for Computer Forensic Acquisitions*. SWGDE, USA, 2018.

[20] McKemmish R. Advances in Digital Forensics. In *IFIP International Federation for Information Processing*, Boston, 2008. Springer.

[21] Garfinkel L. Simson. Digital forensics research: The next 10 years. *Digital Investigation*, 30:S64–S73, 2010.

[22] Morgan Steve. *2017 Cybercrime Report*. Cybersecurity Ventures, CA, USA, 2017.