SAPIENZA UNIVERSITA DI ROMA

Digital Forensics: Validation of Network Artifacts Based on Stochastic and Probabilistic Modeling of Internal Consistency of Artifacts

by

Livinus Obiora Nweke (1735405)

Supervisors: Prof. Luigi V. Mancini and Prof. Stephen D. Wolthusen (Royal Holloway, University of London)

A thesis submitted in partial fulfillment for the
degree of Master of Science in Computer Science

in the
Faculty of Information Engineering, Informatics, and Statistics
Department of Computer Science

July 2018

Declaration of Authorship

I, Livinus Obiora Nweke, declare that this thesis titled, 'Digital Forensics: Validation of Network Artifacts Based on Stochastic and Probabilistic Modeling of Internal Consistency of Artifacts' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a master degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

"When you want something, all the universe conspires in helping you to achieve it."

-Paulo Coelho

SAPIENZA UNIVERSITA DI ROMA

Abstract

Faculty of Information Engineering, Informatics, and Statistics
Department of Computer Science

Master of Science in Computer Science

by Livinus Obiora Nweke (1735405)

In this thesis, a framework for the validation of network artifacts in digital forensics investigations is presented. The main hypothesis of this thesis is that the validity of network artifacts can be determined based on stochastic and probabilistic modeling of internal consistency of artifacts. This framework consists of three phases, namely: data collection, feature selection, and validation process. The functionality of the proposed framework is demonstrated using network artifacts obtained from Intrusion Detection Systems. It is assumed that the initial acquisition of the network artifacts was forensically sound and steps were taken to ensure that the integrity of the artifacts was maintained during 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 stochastic and probabilistic modeling 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.

Acknowledgements

I would like to thank my supervisors, Prof. Luigi V. Mancini and Prof. Stephen D. Wolthusen (Royal Holloway, University of London), for the guidance, support and advice they provided throughout the period of writing this thesis. I have been extremely lucky to have supervisors who cared so much about my work.

Contents

Declaration of Authorship i

Abstract iii

Acknowledgements iv

List of Figures vii

List of Tables viii

1 Introduction 1
- 1.1 Background . . . 1
- 1.2 Research Question, Aim and Objectives . . . 4
  - 1.2.1 Research Question . . . 4
  - 1.2.2 Research Aim . . . 4
  - 1.2.3 Research Objectives . . . 4
- 1.3 Thesis Contributions . . . 5
- 1.4 Structure of the Thesis Report . . . 5

2 Literature Review 6
- 2.1 Digital Forensics . . . 6
- 2.2 Admissibility of Network Artifacts . . . 7
- 2.3 Digital Forensics Models . . . 8
- 2.4 Counter-forensics . . . 9
- 2.5 The Importance of Validation of Network Artifacts . . . 10
- 2.6 Related Works . . . 11
- 2.7 Summary . . . 11

3 Methodology 13
- 3.1 Research Design . . . 13
- 3.2 Implemented Research Methodology . . . 14
- 3.3 Summary . . . 16

4 Proposed Framework for the Validation of Network Artifacts 17
- 4.1 Data Collection . . . 17

4.1.1 Features of the Domain Example used for this Thesis 19
4.2 Feature Selection 19
4.2.1 Monte Carlo Feature Selection (MCFS) 20
4.2.2 Discovering Feature Interdependencies 22
4.3 Validation Process 22

5 Logistic Regression Analysis for the Validation of Network Artifacts 24
5.1 Justification for Using Logistic Regression Analysis 24
5.2 Logistic Regression 25
5.3 Logistic Regression Model Evaluation 28

6 Experiment Results 30
6.1 Experimental Setup 30
6.2 Dataset 30
6.3 Feature Selection Experiment Results 31
6.4 Logistic Regression Analysis Experiment Results 33

7 Discussion and Conclusions 37
7.1 Discussion 37
7.2 Conclusions 38

A An Appendix: Digital Forensics Report on the Validation of Network Artifacts 40
A.1 Overview/Case Summary 40
A.2 Objective 40
A.3 Forensic Acquisition & Exam Preparation 40
A.4 Findings and Report (Forensic Analysis) 41
A.5 Conclusion 42

B An Appendix: Glossary 43

C An Appendix: R Codes Used for the Experiments 46
C.1 R Codes Used for Feature Selection Experiment 46
C.2 R Codes Used for Logistic Regression Analysis Experiment 47

Bibliography 49

List of Figures

4.1 Overview of the MCFS procedure. Reproduced from Draminski and Koronacki in [1] . . . 21

5.1 Logistic Regression Curve . . . 27

6.1 Dataset Summary . . . 31
6.2 Distance Function . . . 31
6.3 Relative Importance . . . 32
6.4 Features Selected . . . 32
6.5 Interdependence Graph . . . 33
6.6 Cross Validation . . . 34
6.7 Lift Curve . . . 34
6.8 ROC Curve with AUC . . . 36
6.9 ROC Curve without AUC . . . 36

A.1 Features Selected . . . 41
A.2 ROC Curve . . . 42

List of Tables

6.1 Logistic Regression Model Estimation 35
6.2 Final Summary Report 35

This work is dedicated to my late father, who believed that I will accomplish great things but did not live to see them happen, and to my sweet mother who did not relent in spite of the demise of my father.

Chapter 1

Introduction

This chapter presents a background of the thesis work. The motivation for undertaking the research is opined. Next, the research question, aim and objectives are described. Then, the contributions of thesis work are presented. The chapter concludes with a description of the structure of the thesis report.

1.1 Background

Digital forensics has been of growing interest over the past ten to fifteen years despite being a relatively new scientific field [2]. 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 an increase in the number of recent cyber crimes [3].

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 [4]. This definition includes artifacts that are not only capable of entering into a court of law but may have investigative value. Altheide and Carvey in [5] make a distinction in terminology between evidence, which is a legal construct and is presented in a courtroom or other official proceedings, and an artifact, which is a piece of data that pertains to an alleged action or event and is of interest to an investigator.

Network artifacts, on the other hand, are among the type of digital artifacts that have attracted a lot of attention in recent times, and 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 [6], 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, it is imperative 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 [7]. Validation, as applied in this thesis, 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 use scientific methodology in doing so, as recommended in [8].

It was noted in [7] 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 [9]. There is currently lack of consistency in the way that the reliability or accuracy of digital artifacts are assessed, partly because of the complexity and multiplicity of digital systems. Furthermore, the level of uncertainty that investigators assigned to their findings is 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 [10]. 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 in [10]:

- 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 thesis, statistical error is used. Statistical error as used in this thesis, 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.

This thesis proposes a framework that can be used for the validation of network artifacts based on stochastic and probability modeling of the internal consistency of artifacts. The focus of this work is on the use of logistic regression analysis as the stochastic and probabilistic modeling methodology in determining the validity of network artifacts. Network artifacts obtained from Intrusion Detection Systems are used as the domain example to demonstrate the workings of the proposed framework. First, it is assumed that the initial acquisition of the network artifacts was forensically sound and that the integrity of the artifacts is maintained during the data collection phase of the proposed framework. The next step involves the selection of the subsets of the features of the artifacts for validation. 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.

1.2 Research Question, Aim and Objectives

The following subsections describe the research question, aim and objectives of this thesis work.

1.2.1 Research Question

The growing reliance on network artifacts in proving and disproving the merit of cases and the need to provide a scientific methodology that meets the requirements for presentation of network artifacts in court have led to the formulation of the following question in the context of network artifacts validation:

Can the validity of network artifacts be established based on stochastic and probabilistic modeling of internal consistency of artifacts?

The definition of validation used in this research is derived from a widely accepted standard for the presentation of digital artifacts in the court known as Daubert test [10]. Garrie and Morrissy noted in [10], that the reigning case in scientific evidence admission is *Daubert v. Merrell Dow Pharmaceuticals Inc.,* 509 U.S. 579, 595 (1993). The landmark decision describes the requirements for the admission of digital artifacts to court proceedings. These requirements include digital artifacts having a scientific basis, can be reviewed by others, known or potential error rates, and the degree to which the theory or technique is accepted by a relevant scientific community. Therefore, the principal themes of the research question are the concepts of network artifacts validation, stochastic and probabilistic modeling, and digital forensic investigation. These elements inform the underlying direction of this thesis and have been used to derive the research objectives in subsection 1.2.3.

1.2.2 Research Aim

The main aim of this research is:

- Validation of network artifacts based on stochastic and probabilistic modeling of internal consistency of artifacts.

1.2.3 Research Objectives

The achievement of the research aim depends on the completion of the following objectives:

- Development of a framework for the validation of network artifacts.
- Use network artifacts obtained from Intrusion Detection Systems as domain example to demonstrate the workings of the proposed framework.
- Use feature selection algorithm for selecting interesting features to be used for the validation process.
- Explore stochastic and probabilistic modeling methodologies to be used for the validation process.
- Perform experiments and analyze the results of the experiments.
- Make inferences based on the outcome of the experiments.

1.3 Thesis Contributions

The contribution of this thesis is to provide the following resources to the scientific community:

- A framework for the validation of network artifacts based on stochastic and probabilistic modeling of the internal consistency of artifacts.
- Logistic regression analysis as a stochastic and probabilistic modeling methodology for the validation of network artifacts.
- A scientific methodology that can be used to support the validity of network artifacts in court proceedings.

1.4 Structure of the Thesis Report

The rest of this thesis consist is organized as follows. In Chapter 2 a review of literature related to digital forensics is presented. Then, the research methodology adopted in this thesis to achieve the research objectives set out in section 1.2 is discussed in Chapter 3. After that, the proposed framework for the validation of network artifacts is described in Chapter 4 and Chapter 5 presents a description of the stochastic and probabilistic modeling methodology used for the validation process. Also, experiment results carried out to demonstrate the functionalities of the proposed framework are presented in Chapter 6. Lastly, the thesis presents a discussion of the experiments results in Chapter 7, with conclusions and future works.

Chapter 2

Literature Review

This chapter presents a review of literature related to digital forensics. The chapter begins by providing definitions as well as a background on digital forensics. The general concept of digital forensics and the goal of digital forensics are explored. The literature review then continues with an in-depth analysis of the admissibility of network artifacts to court proceedings. Furthermore, the literature review provides an overview of digital forensics models and counter-forensics. The importance of validation of network artifacts and related works are described, and the review concludes with a summary of the literature review.

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 [4]. 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 [11]. 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 [12]. 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 [8].

It was observed in [11] that the Digital Forensics Research Workshop in 2001, described the aims of digital forensics as law enforcement, information warfare and critical infrastructure protection (business and industry); while the primary goal of law enforcement is to prosecute those who break the law. Also, Altheide and Carvey observed in [5], that the goal of any digital forensics examination is to identify the facts relating to an alleged event and to create a timeline of these events that represents the truth. Thus, an investigator is required to strive to link these events to the identity of an individual but in most cases, this may not be possible and as such, the events may be described with unknown actors [13].

Another important goal in digital forensics is the attribution of assets and threat actors. Investigators are burdened with the responsibility of achieving good attribution of assets and threat actors using varying approaches. However, due to a large number of technical complexities associated with digital infrastructures, it is often impractical for investigators to determine fully the reliability of endpoints, servers, or network infrastructure devices and provide assurances to the court about the soundness of the processes involved and the complete attribution to a threat actor [14].

2.2 Admissibility of Network Artifacts

The admissibility of network artifacts to court proceedings is dependent on the weight and relevance of the artifacts to the issue at hand. Artifacts considered being vague or indefinite will have less weight compared to artifacts that can be proven. So, an artifact is deemed admissible if it goes to prove the fact at hand and if it provides implications and extrapolations that may assist in proving some key fact of the case. Such artifacts help legal teams, and the court develop reliable hypotheses or theories as to the threat actor [14]. Therefore, the reliability of network artifacts is vital to supporting or refuting any hypothesis put forward, including the attribution of threat actors.

If network artifacts are being contemplated for inclusion during legal hearings, the court must be satisfied that the network artifacts conform to established legal rules – the network artifacts must be scientifically relevant, authentic, reliable and must have been obtained legally [15]. If they fail any of these conditions, then they are likely to be

deemed by the court as inadmissible, preventing the judge or jury from examining and deliberating upon them.

Furthermore, the fragile nature of network artifacts pose additional challenges to the admissibility of the artifacts [16]. The constant evolution of technology, the fragility of the media on which electronic data is stored and the intangible nature of electronic data all make the artifacts vulnerable to claims of errors, accidental alteration, prejudicial interference and fabrication [17]. Thus, even when the artifacts have been admitted to court proceedings, these factors could still impact its weight in proving or disproving the issue at hand.

2.3 Digital Forensics Models

Over the past years, there have been a number of digital forensics models proposed by different authors. It is observed in this thesis work that some of the models tend to be applicable to very specific scenario while others applied to a wider scope. Some of the models tend to be quite detail and others too general. This may lead to difficulty for investigators to adopt the appropriate investigation model. A review of digital forensics models is presented in the following paragraphs.

As observed in [18], current digital forensics models can be categorized into three main types. The first type consists of general models that define the entire process of digital forensics investigation and include models that were proposed from 2000 to 2013. The second type focus on a particular step in an investigation process or a specific kind of investigative case. The last type defined new problems and/or explored new methods or tools to address specific issues. These models were identified and assessed using Daubert Test in [19] and they include:

- An Abstract Digital Forensic Model (ADFM) (Reith et al., 2002)
- The Integrated Digital Investigative Process (IDIP) (Carrier and Spafford, 2003)
- An Extended Model of Cybercrime Investigation (EMCI) (Ciardhuain, 2004)
- The Hierarchical, Objectives Based Framework for the Digital Investigation Process (HOBFDIP) (Beebe and Clark, 2005)
- The Four Step Forensic Process (FSFP) (Kent et al., 2006)
- The Computer Forensics Field Triage Process Model (CFFTPM) (Rogers et al., 2006)

- A Common Process Model for Incident Response and Computer Forensics (CP-MIRCF) (Freiling and Schwittay, 2007)
- The Two-Dimensional Evidence Reliability Amplification Process Model (TDER-APM) (Khatir et al., 2008)
- Mapping Process of Digital Forensic Investigation Framework (MPDFIF) (Selamat et al., 2008)
- The Systematic Digital Forensic Investigation Model (SDFIM) (Agarwal et al., 2011)
- The Integrated Digital Forensic Process Model (IDFPM). Kohn et al., 2013)

According to the review and assessment in [19], there is no comprehensive model encompassing the entire investigative process that is formal in that it synthesizes, harmonizes, and extends the previous models, and that is generic in that it can be applied in different fields of law enforcement, commerce and incident response. Also, it was noted from the assessment in [19] that Rogers et al's CFFTPM compared to other digital forensics models above, has taken the most scientific approach to develop its model.

Rogers et al's CFFTPM includes six phases: Planning, Triage, User Usage Profiles, Chronology Timeline, Internet and Case Specific. The goal of the model is to facilitate "onsite triage" to examine and analyze digital devices within hours as opposed to weeks or months. This model can only be applied in situation where a swift examination needs to be carried out at the crime scene. The most important contribution of the CFFTPM is that it took a different approach from the traditional digital forensic approach of seizing a digital device, transporting it to the lab, making a forensic image, and then searching the entire system for potential artifacts [19].

2.4 Counter-forensics

Counter-forensics can be seen as any methodology or technique that can be deployed to negatively affect forensics investigation. Several attempts have been made to formally define counter-forensics. According to [20], Rogers in 2005 defined counter-forensics as attempts to negatively affect the existence, amount and/or quality of artifacts from a crime scene, or make the analysis and examination of artifacts difficult or impossible to conduct. A more recent definition that was noted in [2] is that by Albano et al in 2012, which defined counter-forensics as methods undertaken in order to thwart the digital investigation process conducted by legitimate forensics investigators.

There have been attempts at the classification, identification, and characterization of counter-forensics tools and techniques. It was observed in [20] that Rogers in 2006 proposed a new approach for categorization of counter-forensics techniques. The taxonomy which is widely accepted in digital forensics research has four categories, namely: data hiding, artifact wiping, trail obfuscation and attacks against both the forensic process and forensic tools. Also, Dahbur and Mohammad in [21] presented a classification of counter-forensics mechanisms, tools and techniques and evaluated their effectiveness. Furthermore, they discussed the challenges of countermeasures against counter-forensics, along with a set of recommendations for future studies.

Another important work in counter-forensics worth noting are attempts at detection and indication of counter-forensics tool usage. Kyoung et al in [22] presented an anti-forensic trace detection in digital forensic triage investigations in an effort towards automatic detection using signature-based methods. Also, it was observed in [2] that Rekhis and Boudriga in 2012 described a theoretical approach to digital forensics investigations in which the investigation process is at all times aware of the possibility of counter-forensics attacks. The work created an investigated system context, the applied security solution, and a library of counter-forensics attacks that are used against the system, with the resulting artifacts being collected. Then, an inference system was proposed to mitigate the counter-forensics attacks. Potential scenarios were then generated from the counter-forensics traces to provide models of counter-forensics actions that can occur during digital forensics investigations.

2.5 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 [7] 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. Thereby, 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 [8]. 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 [15]. Furthermore, it was pointed out in [10] 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 imperative 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.

2.6 Related Works

Stochastic and probabilistic modeling techniques have been applied in the validation of digital artifacts. A review of issues in scientific validation of digital artifacts is presented in [23]. As opined in [24], Bayesian Networks have been used to facilitate the expression of opinions regarding the legal determinations on the credibility and relative weight of non-digital artifacts. Also, digital forensics researchers have used Bayesian Networks to reason about network artifacts in order to quantify their strengths in supporting hypotheses about the reliability of the digital artifacts [24].

Different from the above works, the focus of this thesis is explicitly on the validation of network artifacts based on stochastic and probability modeling of the internal consistency of artifacts. The solution is able to provide a sound scientific basis to support the validity of network artifacts. Although network artifacts obtained from Intrusion Detection Systems are used as the domain example to demonstrate the functionalities of the framework, the framework is more general in design and can be applied to other network artifacts.

2.7 Summary

This chapter reviewed existing literature surrounding the various facets of digital forensics as well as insights into how the presented work addressed the challenges identified within the field of digital forensics. It started with a background in digital forensics, identifying the origin of forensic science and the goals of digital forensics. A discussion

on the admissibility of network artifacts was also presented. Furthermore, different models of digital forensics processes and counter-forensics were explored. The importance of validation of network artifacts and related works were presented. It was observed that recent research works have focused on overcoming technical challenges relating to digital forensics and there is a need for further investigation into validation of digital artifacts to report the certainty in the inferences drawn from the artifacts. Further analysis of the validation process based on stochastic and probabilistic modeling is proposed in the subsequent chapters.

Chapter 3

Methodology

In this chapter, a discussion of the research methodology adopted in this thesis to achieve the research objectives set out in section 1.2 is presented. The research objectives focus on the development of a framework for the validation of network artifacts based on stochastic and probabilistic modeling of internal consistency of artifacts. These research objectives were the basis of the research methodology and the research process that was implemented to achieve them.

3.1 Research Design

Nicholas in [25] observed that research methodologies vary from qualitative to quantitative or both. The goal of each method is to aid the researcher to achieve the objectives of the research. According to [25], the three main requirements for a structured and well-designed research include: the knowledge claimed, the research strategy, and the method used for data collection. For this thesis research, the knowledge is digital forensics and the need to validate network artifacts. Through a comprehensive review of previous studies in the research domain and the gaps observed in the current methodologies for the validation of network artifacts; validation of network artifacts based on stochastic and probabilistic modeling of internal consistency of artifacts was established.

A methodological approach was adopted in achieving the proposed framework for the validation of network artifacts. The methodology used in this research involves performing a comprehensive study to identify the requirements for the validation of network artifacts. An extensive search of the literature relating to digital forensics, particularly with validation of network artifacts was carried out. The data collection method of this research includes the use of journals, conference proceedings, books, websites, workshops

and seminars understanding the issues on digital forensics and validation of network artifacts. Also, a publicly available dataset was collected and was used for the research.

3.2 Implemented Research Methodology

The aim of the study was to understand how network artifacts can be validated based on stochastic and probabilistic modeling of the internal consistency of artifacts. This type of evaluation requires an action inquiry, in which one improves practice by systematically oscillating between taking action in the field of practice and inquiring into it [26]. Hence, Action Research process was adopted as the research process for this inquiry. Action Research process is a research approach which involves diagnosing a problem, planning an intervention, carrying it out, analysis the results of the intervention, and reflecting on the lessons learned [26]. Although the main application of Action Research is in the field of social science and educational studies research, it has been used to validate security design methods [27].

The Action Research process used in this thesis is that proposed by Baskerville [28], who breaks the research process into five distinct phases:

- Diagnosis: Initial reflection on the problem that needs to be solved.
- Action Planned: Devising the planned process to be adopted in order to meet the intervention's objectives.
- Action Taken: Carrying out the planned steps.
- Evaluating: Evaluating the outcome of the intervention.
- Specifying Learning: Stating the actions which need to feed forward to future research.

The following paragraphs describe each of the steps in the Action Research process as was applied in this thesis.

As observed in the description of the diagnosing phase of Action research process above, the phase involves reflecting on the problem that needs to be solved. These problems usually arise from ambiguity in the situation researchers find themselves. In the case of this research, the problem that needs to be solved has to do with how network artifacts can be validated to increase the confidence on the inferences drawn from the artifacts. To address this problem, it needs to be made concrete by formulating a research question. This process was largely exploratory in nature, where the research question was

derived from gaps in the available literature. The problem was thought of as a reliability challenge and the research started to analyze what would be the best framework and methodology that can be developed and followed to address the problem. Furthermore, several features of network artifacts were considered at this stage but given the constraint of time and resources, validation of network artifacts based on stochastic and probabilistic modeling of internal consistency of artifacts was chosen.

In the action planning phase of Action Research, the goal is to develop a detailed plan of action that would lead to addressing the research question identified in the diagnosing process. For this thesis, this process was achieved by developing research objectives and realistic timeline for each of the research objectives. The first stage of the process was identifying the need for a framework that would be the basis for the validation of network artifacts. Next, was to decide the domain example that would be used to demonstrate the functionalities of the proposed framework. After the domain example had been chosen, a search was done for the suitable network artifacts that can be obtained and use for the research. The next step in the process was the selection of the subsets of the features of the network artifacts. Different methodologies were explored and the technique that was best suited for the network artifacts was selected. After the selection, several stochastic and probabilistic modeling methodologies were explored for the validation process. Then, experiments and analysis would be carried out. The last step of the action planned was making inferences based on the outcome of the experiments.

The action taking phase involves carrying out the planned steps identified in the action planning phase. The planned steps for this thesis have been defined by the research objectives. During this stage, the framework for the validation process was identified. The framework followed the same approach usually adopted in forensic investigation and anomaly detection. The framework consists of three steps: Data collection, feature selection, and validation process. Next, the processes that should be involved in each of the identified steps were explored. Upon concluding the exploration, the methods for the validation of network artifacts based on stochastic and probabilistic modeling of internal consistency of artifacts were investigated. Also, experiments and analysis were conducted. The outcome of this stage includes a description of data collection that meets the requirements for handling of artifacts during a forensic investigation, a feature selection algorithm for selecting the subsets of features of the artifacts and understanding the interdependence between the features, logistic regression analysis as the stochastic and probabilistic methodology for the validation process, and experiment results.

During the evaluation phase, the steps taken in the action taking phase were critically examined. The goal was to understand how those steps can provide a sound scientific basis for the presentation of network artifacts to the court. Further, the definition of scientific methodology was re-examined in the light of the observations of the actions taken. This was to ensure that the outcome of the action taking phase meets the requirements for the presentation of network artifacts to court and that they can be used to proving or disproving the merit of cases. All these, contribute to providing insights into the validation of network artifacts based on stochastic and probabilistic modeling of internal consistency of artifacts.

Lastly, the lessons learned from the research process were described. The goals of this phase are two folds: first, was to describe how the research process have addressed the problem identified in the diagnosing phase; second, was to identify actions that need to be fed forward to future research. These were achieved by first, reflecting on the findings of the experiments and analysis. These findings were used to draw conclusions on the significance of the contributions of this thesis to the validation of network artifacts within digital forensics context. Second, by identifying the works that still needs to be done but were not done due to time and resource constraints.

3.3 Summary

The major aspect covered in this chapter include the research design and the implemented research methodology. The research design identified the knowledge claimed for this thesis and the methodological approach adopted achieving the proposed framework for the validation of network artifacts. Also, it was observed that the research process for this thesis followed the Action Research process. The processes involved in the Action Research process were used for describing the implemented research methodology for this thesis.

Chapter 4

Proposed Framework for the Validation of Network Artifacts

In this chapter, 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 was 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 stochastic and probabilistic modeling methodology. The following sections provide the description of each of the stages of the proposed framework for the validation of network artifacts.

4.1 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 [29]:

- 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 [29] are the "Good Practice Guide for Computer Based Electronic Evidence" published by the Association of Chief Police Officers (United Kingdom) and the International Organization on Computer Evidence (now Scientific Working Group on Digital Evidence(SWGDE)). The "Good Practice Guide for Computer Based Electronic Evidence" published by the Association of Chief Police Officers (United Kingdom) [30] 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 [31]:

"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 thesis and given the constraints under which the research was undertaken, it is assumed that the acquisition of the network artifacts to be used for the data collection stage of the framework was forensically sound and that chain of custody was 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 were preserved.

4.1.1 Features of the Domain Example used for this Thesis

Network artifacts obtained from Intrusion Detection Systems were used as the domain example to demonstrate the functionalities 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 [32] was 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 CICFlowMater 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 [32]. Therefore, the output of the application forms the basis of the CICIDS2017 dataset [33], used as network artifacts for this thesis.

4.2 Feature Selection

The next step after the data collection stage of the proposed framework 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. 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 [34]. 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 thesis, several R packages of the feature selection algorithm were applied on the network artifacts to ascertain which one is best suited for the artifacts. An example of the R packages that was used is Boruta algorithm [35], 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 [1], which does not only provide the ranking of features but also, includes a directed ID-Graph that presents interdependencies between informative features was used for feature selection stage of the proposed framework.

4.2.1 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

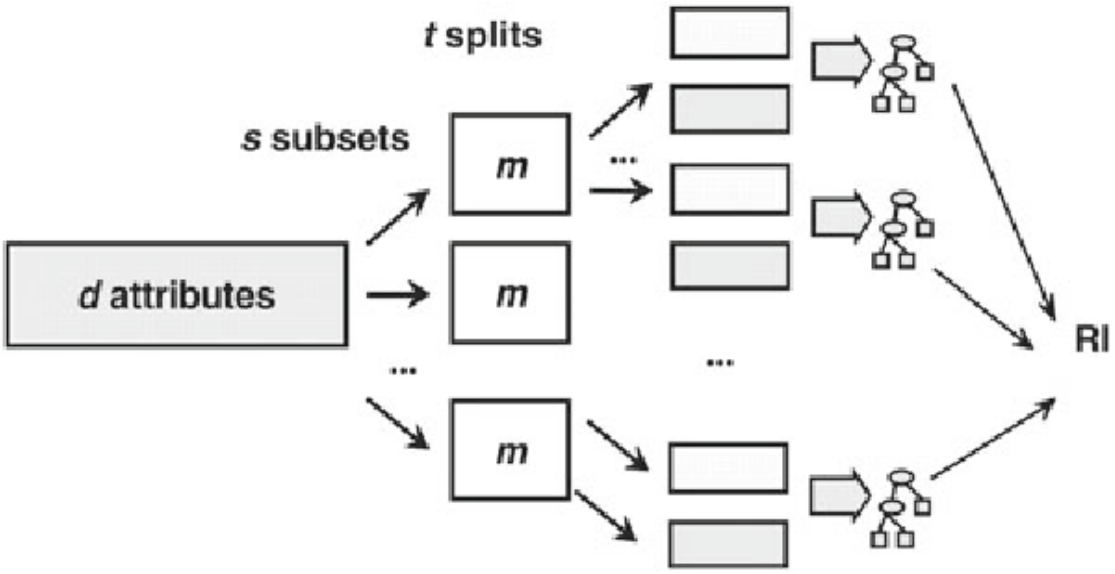

FIGURE 4.1: Overview of the MCFS procedure. Reproduced from Draminski and Koronacki in [1]

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 4.1

A particular feature is considered to be important based on its performance in classifying the samples into classes. The degree in which a feature takes part in the classification process is termed relative importance (RI). The RI of a feature g was defined by Draminski and Koronacki in [1] as:

$$RI_g = \frac{1}{M_g} \sum_{\tau=1}^{st} (wAcc)^u \sum_{n_g(\tau)} IG(n_g(\tau)) \left(\frac{no.in\, n_g(\tau)}{no.in\, \tau}\right)^v \tag{4.1}$$

where s is the number of subsets and t is the number of splits for each subset. M_g is the number of times the attribute g was present in the training set used to construct a decision tree. For each tree τ the weighted accurate $wAcc$ is calculated as the mean sensitivity over all decision classes, using

$$wAcc = \frac{1}{M_c} \sum_{i=1}^{c} \frac{n_{ii}}{n_{i1} + n_{i2} + ... + n_{ic}} \tag{4.2}$$

where c is the number of decision classes and n_{ij} is the number of objects from class i that were classified to class j.

Furthermore, for each $n_g(\tau)$ (a node n in decision tree τ that uses attribute g) the information gain (IG) of $n_g(\tau)$ and the fraction of the number of training set objects in (no.in) $n_g(\tau)$ compared to the number of objects in the tree root is computed. There

are two weighting factors u and v that determine the importance of the $w\mathrm{A}cc$ and the number of objects in the node.

4.2.2 Discovering Feature Interdependencies

Discovering interdependencies among features is an important step in understanding the subset of features selected by the algorithm. There are several approaches used in discovering interdependencies among features and the most widely adopted approach is finding correlations between features or finding groups of features that behave in some way similarly across samples. However, the approach to interdependence discovery as applied by MCFS-ID used for this thesis is significantly different from other approaches. MCFS-ID algorithm focus on identifying features that cooperate in determining that a sample belongs to a particular class. The idea is that given a training set of samples and a classification tree, for each class, a set of decision rules defining this class is provided. Each decision rule point to some interdependencies between the features appearing in the conditions.

The MCFS-ID algorithm uses an ID-graph, which is based on aggregating information provided by all the $s.t$ trees (shown in figure 1). In each of the $s.t$ trees the nodes represents features. For each path in the tree, a distance is defined between two nodes as the number of edges between these two nodes and the strength of the interdependencies between features g_i and g_j was defined by Draminski and Koronacki in [1] as

$$Dep(g_i, g_j) = \sum_{\tau=1}^{st} \sum_{\xi_\tau} \sum_{n_{gi}(\xi_\tau), n_{gj}(\xi_\tau)} \frac{1}{dist(n_{gi}(\xi_\tau), n_{gj}(\xi_\tau))} \tag{4.3}$$

where summation is over all the $s.t$ trees, within each τ-th tree over all paths ξ and, within each path ξ_τ, over all pair of nodes $(n_{gi}(\xi_\tau), n_{gj}(\xi_\tau))$ on which the splits are made respectively, on feature g_i and feature g_j; $Dep(g_i, g_j) = 0$ if in none of the trees there is a path along which there are two splits made respectively on g_i and g_j.

4.3 Validation Process

After the selection of subsets of the dataset, the next step is the validation of the network artifacts based on stochastic and 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 stochastic and probabilities modeling methodologies to observe the internal consistency of the artifacts and make

inference on the validity of the network artifacts. Logistic regression analysis was the stochastic and probabilistic methodology used for the purpose of this thesis research. The following chapter provides a complete description of logistic regression analysis used as it relates to the validation of the network artifacts.

Chapter 5

Logistic Regression Analysis for the Validation of Network Artifacts

The description of the stochastic and probabilistic modeling methodology used for the validation process is presented in this chapter. First, the justification for using logistic regression analysis for this thesis is offered with a brief discussion on an alternative method that could be applied in the validation process. This is because the choice of the stochastic and probabilistic modeling methodology to be used for the validation process depends on the nature of the network artifacts to be validated. Further, logistic regression as a stochastic and probabilistic modeling methodology is presented. In addition, the chapter describes metrics for evaluating logistic regression models as it relates to the validation of network artifacts.

5.1 Justification for Using Logistic Regression Analysis

The nature of network artifacts to be validated determines the choice of stochastic and 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 stochastic and probabilistic modeling 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 [36]. 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 thesis, the network artifacts were 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 was used as the stochastic and probabilistic modeling 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 stochastic and probabilistic modeling methodologies (e.g. discriminant function analysis, log-linear models, and linear probability models) for analyzing categorical outcome variables [37]. Also, in terms of classification and prediction, logistic regression has shown to produce fairly accurate results [38].

5.2 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 [37]. 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

$$y = b_0 + b_1x_1 + b_2x_2 + ... + b_px_p \tag{5.1}$$

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.

In the simplest case of one predictor X and one dichotomous outcome variable Y, the logistic regression model predicts the logit of Y from X. The logit is the natural logarithm (ln) of odds of $Y = 1$ (the outcome of interest). The simple logistic model has the form:

$$\text{logit}(y) = \ln(\frac{p}{1-p}) = \alpha + \beta x \tag{5.2}$$

Hence,

$$\text{Probability}(Y = \text{outcome of interest}|X = x) = P = \frac{\mathrm{e}^{\alpha+\beta x}}{1 + \mathrm{e}^{\alpha+\beta x}} = g\,(x), \tag{5.3}$$

where P is the probability of the outcome of interest under variable Y, α is the Y intercept, and β is the slope parameter. Both the Y intercept and the slope parameter are estimated by the maximum likelihood method. The maximum likelihood method aims to maximize the likelihood of obtaining the data given its parameter estimates. As can be inferred from Equation 5.3, the relationship is non-linear between parameters and the probability of observing a particular outcome in an observation (such as the label of the network traffic). However, the relationship between parameters and the logit is linear. Figure 5.1 shows a logistic regression curve with a single predictor X.

Considering the inferential context, the null hypothesis states that β equals zero in the population. Rejecting such a null hypothesis implies that a relationship exists between X and Y. If the predictor is binary, the expontentiated $\beta(= e^{\beta})$ is the odds ratio, or the ratio of two odds.

Also, it can be observed from the logistic function, that is, the g(x) in equation 5.3, has the following unique properties: unless $\beta = 0$, the binary logistic regression maps the regression line onto the interval (0,1) which is compatible with the logical range of probabilities; the regression line is monotonically increasing if $\beta > 0$, and monotonically decreasing if $\beta < 0$; and the function takes the value of 0.5 at $x = -\alpha/\beta$ and is symmetric to the point of $(-\alpha/\beta, 0.5)$. These properties illustrates that the logistic regression model ensures that the predicted probabilities will not fall outside the range of 0 to 1 and the logistic function has a point of inflection corresponding exactly to 50% probability.

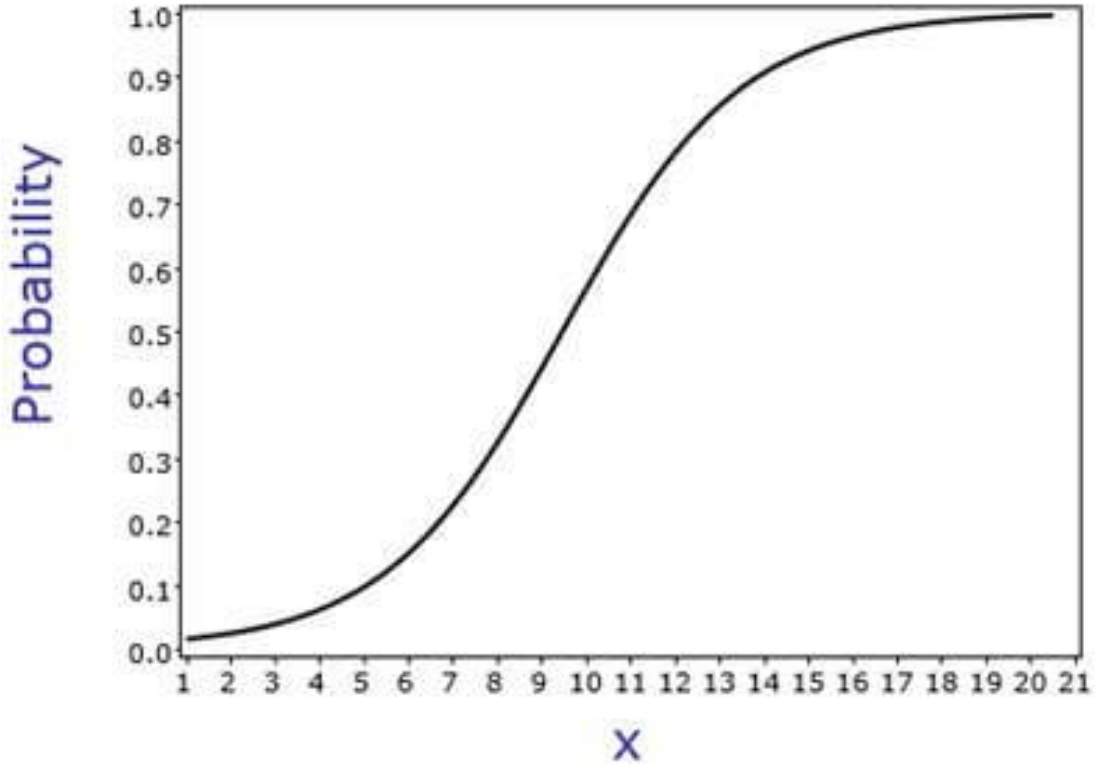

FIGURE 5.1: Logistic Regression Curve

The simple logistic regression model described so far can be extended to multiple predictors as was used for this thesis:

$$\text{logit}(y) = \ln(\frac{p}{1-p}) = \alpha + \beta_1 x_1 + \beta_2 x_2 + ...\beta_k x_k, \tag{5.4}$$

Therefore,

$$\begin{aligned} &\text{Probability}(Y = \text{outcome of interest} | X_1 = x_1, X_2 = x_2, ..., X_k = x_k) \\ &= P = \frac{e^{\alpha+\beta_1 X_1+\beta_2 X_2+...+\beta_k X_k}}{1 + e^{\alpha+\beta_1 X_1+\beta_2 X_2+...+\beta_k X_k}}. \end{aligned} \tag{5.5}$$

where P is the probability of the "event" under the outcome variable Y, α is the Y intercept parameter, βs are slope parameters, and Xs are a set of predictors. α and βs are estimated using the maximum likelihood method. The interpretation of βs is derived using the odds ratio. The null hypothesis states that all βs equal zero. A rejection of this null hypothesis suggests that at least one β does not equal to zero in the population.

5.3 Logistic Regression Model Evaluation

A logistic model is considered to provide a better fit to the data if it demonstrates an improvement over the intercept-only model also referred to as the null model. The null model serves as a good baseline because it contains no predictors. Thus, all observations would be predicted to belong in the largest outcome category. An improvement over this baseline is usually measured using the likelihood ratio, pseudo R^2 and Hosmer-Lemeshow test.

The likelihood ratio test uses the ratio of the maximized value of the likelihood function for the full model (L_1) over the maximized value of the likelihood function for the null model (L_0). It tests if the logistic coefficient for the dropped variable can be treated as zero, thereby justifying the dropping of the variable from the model. A non-significant likelihood ratio test indicates no difference between the full model and the null model, hence justifying dropping the given variable so as to have a model that works well. The likelihood ratio test statistic is given as:

$$-2\log(\frac{L_0}{L_1}) = -2[\log(L_0) - \log(L_1)] = -2(L_0 - L_1) \tag{5.6}$$

There are a number of pseudo R^2 metrics that could be used to determine the goodness of fit. However, McFadden's R^2 is used for the evaluation and it is defined as:

$$R^2 = 1 - \frac{\ln(\mathrm{L}_1)}{\ln(\mathrm{L}_0)} \tag{5.7}$$

where $\ln(L_1)$ is the log likelihood value for the fitted model and $\ln(L_0)$ is the log likelihood for the null model with only an intercept as a predictor. The measures ranges from 0 to just 1, with values closer to zero indicating that the model has no predictive power.

An additional approach for determining the goodness of fit is using the Homer-Lemeshow test. The test is computed on data after the observations have been divided into groups based on their similar predictive probabilities. It examines whether the observed proportions of events are similar to the predicted probabilities of occurrence in subgroups of the data using a Pearson Chi Square test. Small values with large p-values indicate a good fit to the data while large values with p-values below 0.05 indicate a poor fit.

Another important test in evaluating the logistic regression model is to measure the contributions of the individual predictors. The statistical significance of individual regression coefficients (i.e., βs) is tested using the Wald chi-square statistic. it is calculated by taking the ratio of the square of the regression coefficient to the square of the standard error of the coefficient. The goal is to test the hypothesis that the coefficient of an

independent variable in the model is significantly different from zero. If the test fails to reject the null hypothesis, this suggests that removing the variable from the model will not affect the fit of the model. The Wald test is given as:

$$\mathrm{W}_j = \frac{\beta_j^2}{\mathrm{SE}_j^2} \tag{5.8}$$

A critical question to answer when evaluating a logistic regression model has to do with its predictive accuracy. This is usually done using a classification table. A classification table involves the use of a table to cross-classify the observed values for the dependent outcome and that of the predicted values. The process involves using the logistic regression model estimates to predict values on the training set. Then, compare the predicted target variable and that of the observed values for each observation and presenting them using a table.

Also important in the validation of the predicted values is the receiving operating characteristic curve (ROC curve), which measures the classifier performance. The ROC curve uses the proportion of positive data points that are correctly considered as positive and the proportion of negative data points that are mistakenly considered as positive to generate a graph that shows the trade-off between the rates of correct prediction with the rate of incorrect prediction. The metrics ranges from 0.50 to 1.00, and values above 0.80 indicate that the model does a good job in discriminating between the two categories.

Chapter 6

Experiment Results

In this chapter, the experiments that were carried out to demonstrate the functionalities of the proposed framework are presented. The chapter starts with a description of the experimental setup and then the dataset that was used for the experiments. Also, the results and analysis of the experiments are presented. Furthermore, a discussion on the entire process is provided in Chapter 7 to offer insight into how it could support the proving or disprove the validity of network artifacts.

6.1 Experimental Setup

The experiments were conducted on 64-bit Microsoft Windows-based operating system, using R x64 3.5.0. R is a language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R provides a wide variety of statistical and graphical techniques for data manipulation, calculation, and graphical display. 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.

6.2 Dataset

The CICIDS2017 dataset [33] that was 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 6.1.

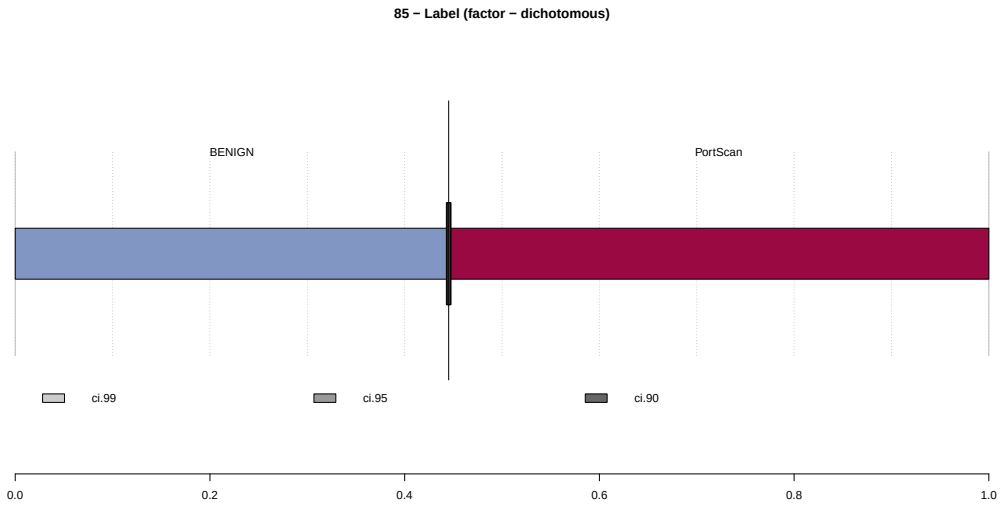

FIGURE 6.1: Dataset Summary

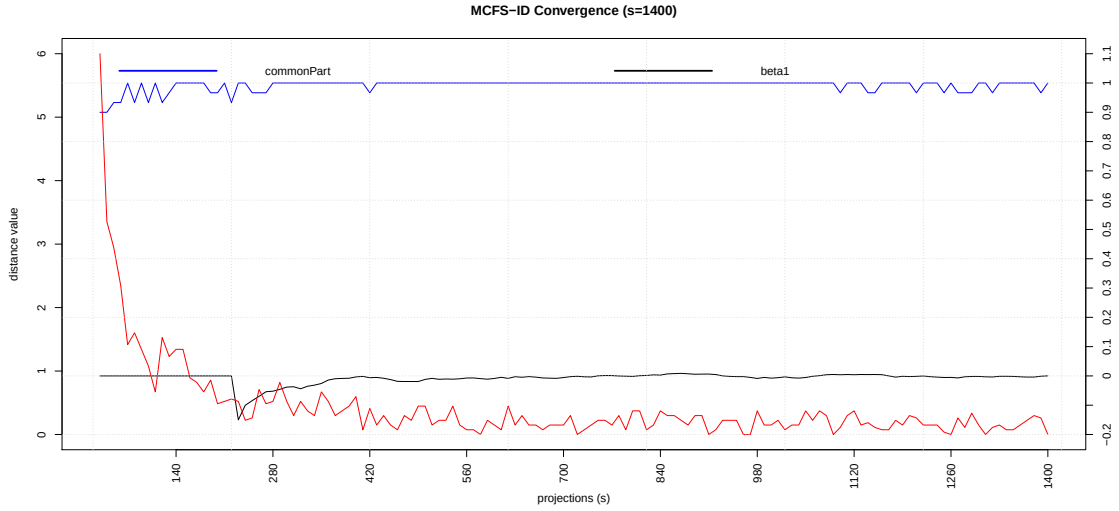

FIGURE 6.2: Distance Function

6.3 Feature Selection Experiment Results

The MCFS-ID algorithm was applied to the dataset described in 6.2. After successfully running the MCFS-ID algorithm, the next step is to check convergence of the algorithm. This is shown in Figure 6.2. 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.

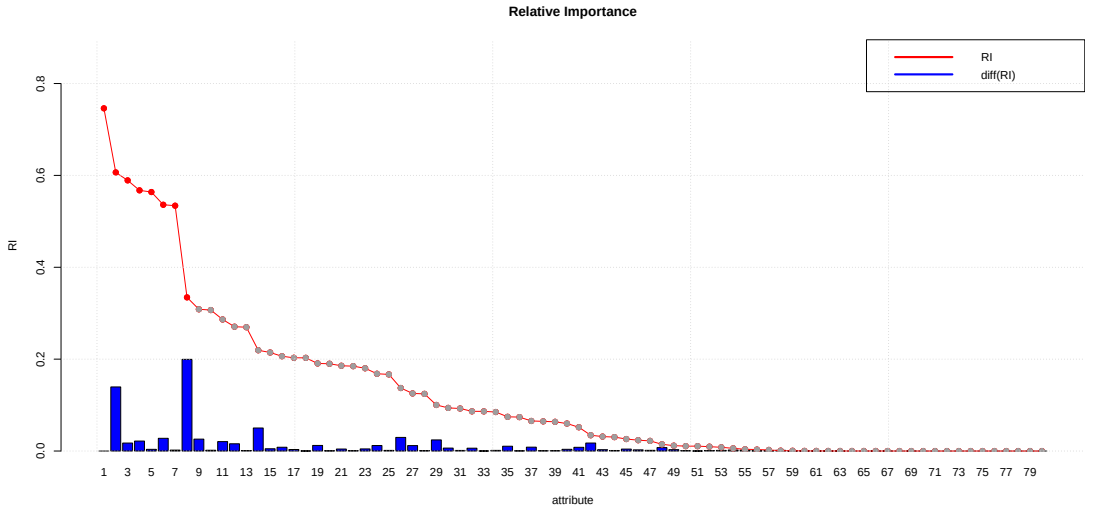

FIGURE 6.3: Relative Importance

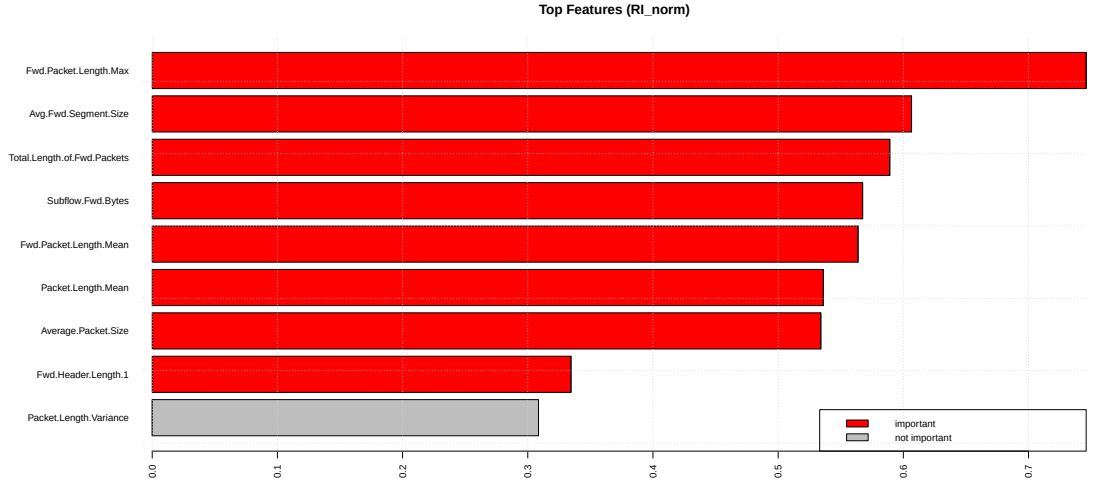

FIGURE 6.4: Features Selected

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 6.3. 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.

Similarly, labels and RIs of the top features can be review. The resulting plot is presented in Figure 6.4. 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.

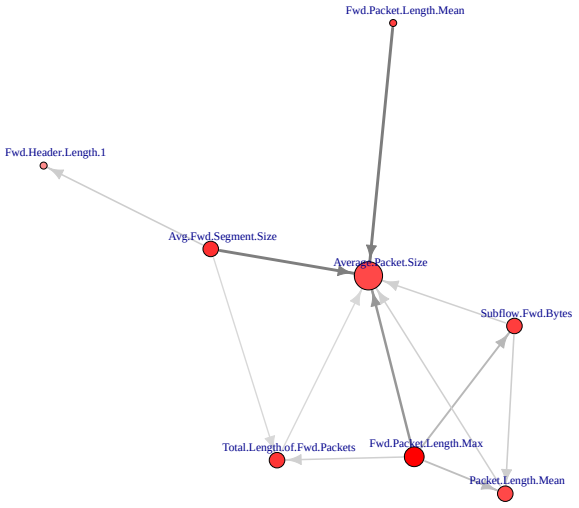

FIGURE 6.5: Interdependence Graph

Another important information useful in understanding the selected features is the interdependency between them. This can be visualized using the ID-Graph shown in Figure 6.5. The ID-Graph plot all the edges that connect the eight informative features as defined by the result cutoff value. The color intensity of a node is proportional to the corresponding feature's RI. The size of a node is proportional to the number of edges related to the node. The width and level of darkness of an edge is proportional to the ID weight of the edge. It can be observed from the figure that top 8 features along with top 12 ID weights are presented. However, three of the top features (Fwd.Header.Length.1, Packet.Length.Mean, and Subflow.Fwd.Bytes) do not cooperate in distinguishing bewteen classes and as such are not used in the logistic regression model.

Lastly, for top feature set, when the execution of MCFS-ID algorithm has finished, the procedure runs 10 fold cross validation (cv) on 6 different classifiers as shown in Figure 6.6. Each cv is repeated 3 times and the mean value of accuracy and weighted accuracy are gathered. Since the weighted accuracy is equal to the mean over all true positive, it is more meaningful for datasets with unbalanced classes. The accuracy of the top features depends on an algorithm and a given cv experiment. Th cv plot presents the result for the result cutoff value features (red label on X axis) and its multiple (0.25, 0.5, 0.75, 1, 1.25, 1.5, 2).

6.4 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 6.1.

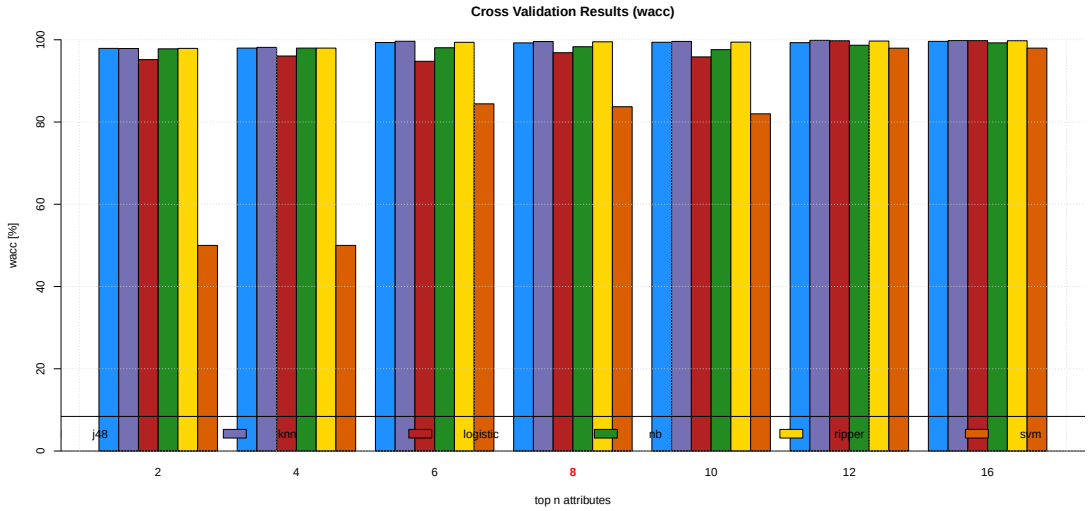

FIGURE 6.6: Cross Validation

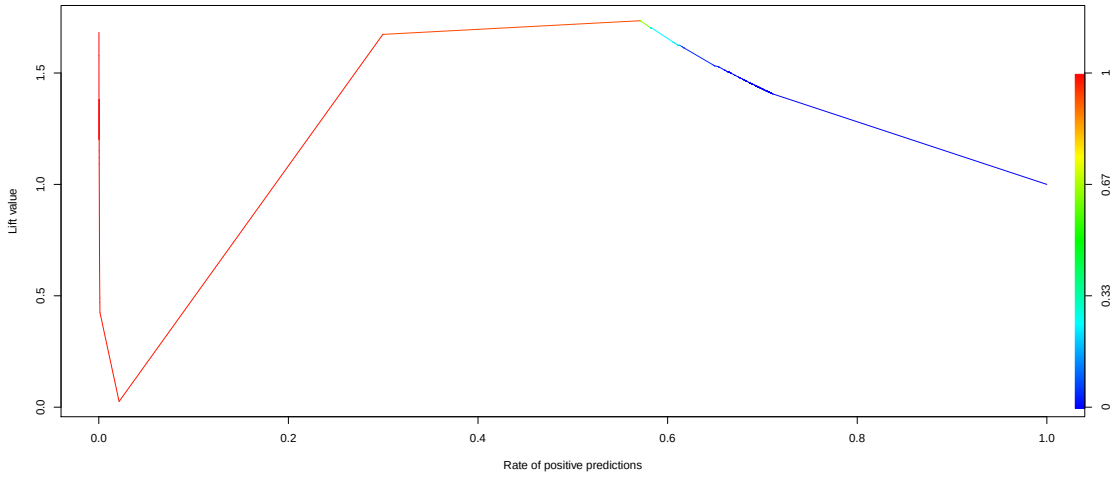

FIGURE 6.7: Lift Curve

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 6.2 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. Figure 6.7 shows a lift curve, which provide visual aid for measuring the logistic model performance.

TABLE 6.1: 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 6.2: Final Summary Report

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

Another important graph used to depict the reliability of the categorical outcome variables of the network artifacts is the ROC curve shown in Figure 6.8 and 6.9. 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.

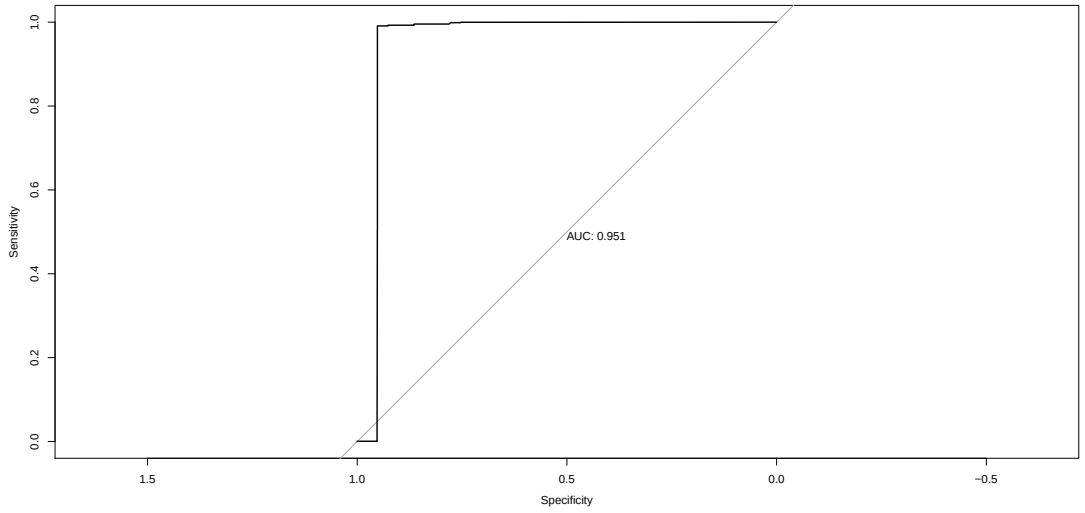

FIGURE 6.8: ROC Curve with AUC

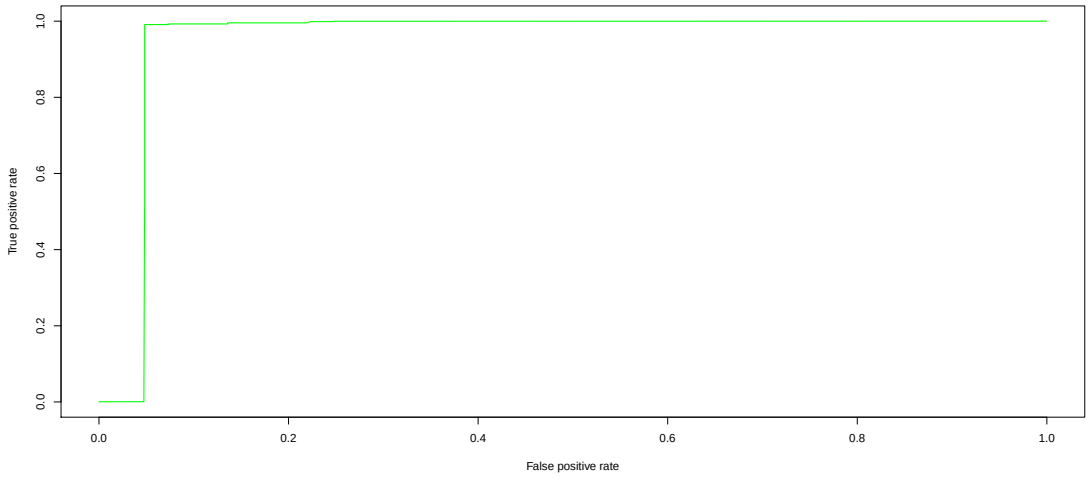

FIGURE 6.9: ROC Curve without AUC

Chapter 7

Discussion and Conclusions

This chapter presents a discussion of the experiment results. The goal is to highlight the major findings from the experiment results and interpret them as they relate to the validation of network artifacts. Furthermore, the limitation that resulted from the research methods is described. Lastly, the main conclusions from the research and suggestions for further research are presented.

7.1 Discussion

The study of the validity of network artifacts using logistic regression as the stochastic and probabilistic modeling methodology for modeling the internal consistency of artifacts demonstrates that inferences drawn from the artifacts can be supported using statistical results. Indeed, Table 6.1 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 experiments. The summary of the statistical distribution of the network artifacts is given in Table 6.2. 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 6.9 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 6.8 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 limitation that resulted from research methods has to with the initial acquisition of the network artifacts and the data collection phase of the framework. It was assumed that the initial acquisition of the network artifacts was forensically sound and that the data collection phase of the framework ensured that the integrity of the network artifacts was maintained. These were 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.

Notwithstanding the limitation of this research, 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.

7.2 Conclusions

In this thesis, a framework for the validation of network artifacts was proposed. The workings of the proposed framework were demonstrated using a publicly available dataset

as the network artifacts. It was assumed that the initial acquisition of the network artifacts was forensically sound and that the data collection stage of the proposed framework guaranteed the integrity of the network artifacts. Experiments were performed using the network artifacts. The first experiment involved the use of Monte Carlos Feature Selection algorithm to select subsets of the features of the artifacts to be used for the validation process. Given the nature of the network artifacts, logistic regression was then applied to the selected subsets of the features. Results from the experiments show the validity of the network artifacts and a can be used as a scientific methodology to support inferences drawn from the network artifacts in court proceedings.

In further 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. To achieve 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.

Appendix A

An Appendix: Digital Forensics Report on the Validation of Network Artifacts

A.1 Overview/Case Summary

On today's date, the security operation center (SOC) team acquired network artifacts from the Intrusion Detection Systems with regards to reported incident of attack (PortScan). The SOC team is requesting a forensic examination to see if the classification of the artifacts can be verified so as to support the initial assertions made by the Intrusion Detection Systems.

A.2 Objective

The goal of the forensic examination is to validate the network artifacts based on stochastic and probabilistic modeling of the internal consistency of artifacts.

A.3 Forensic Acquisition & Exam Preparation

- On today's date the forensic acquisition of the network artifacts was done while maintaining chain of custody.
- Using a sterile storage media (examination medium) that had been previously forensically wiped and verified by this examiner using Encase 8.06.1. The MD5

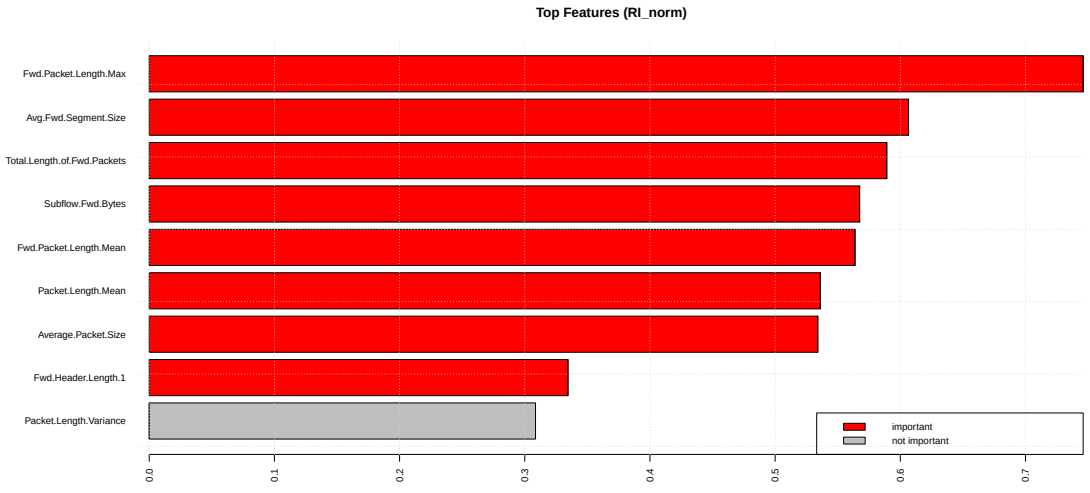

FIGURE A.1: Features Selected

hash value for the examination medium yielded the same MD5 hash value as previous forensic wipes to sterilize this media.

- At this point, the drive containing the artifacts was connected to hardware write-blocker, which is running the most recent firmware and has been verified by this examiner. After connecting the hardware write blocker to the drive, the hardware write blocker was connected via USB 2.0 to forensic machine to begin the acquisition.

A.4 Findings and Report (Forensic Analysis)

- After completing the forensic acquisition of the network artifacts, the analysis of the artifacts was carried out with Forensic Tools.
- The following tools were used for the forensic analysis:
  1. Encase 8.06.1
  2. R x64 3.5.0
- rmcfs R package was used to select subset of the features of the artifacts to be used for the validation process. The selected features are shown in Figure A.1
- Logistic regression was applied to the selected subset of features. The result from the regression analysis shows an accuracy of 95.1% of the initial assertions made by the intrusion detection systems using the selected subset of features as the independent variables. The result is shown in Figure A.2

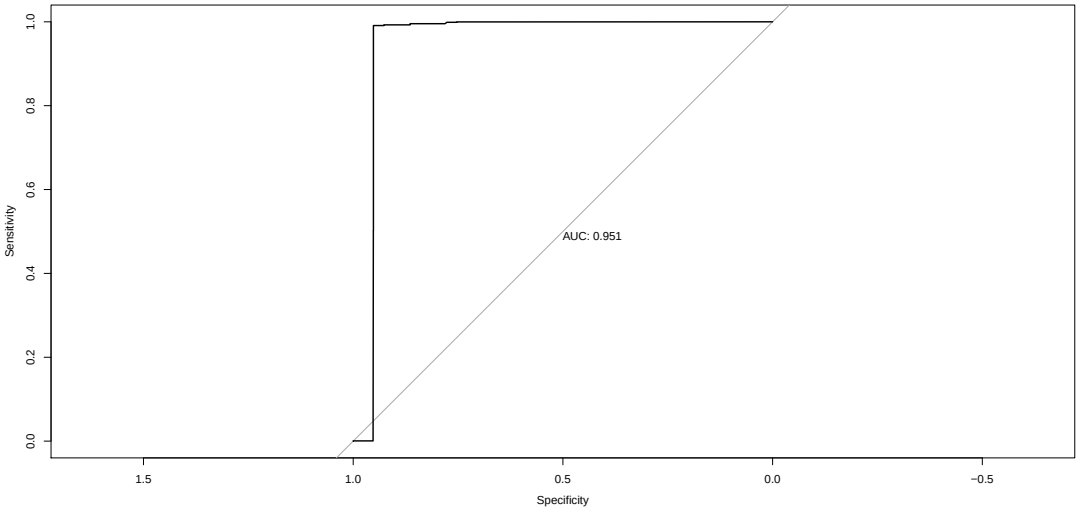

FIGURE A.2: ROC Curve

A.5 Conclusion

This forensics report verifies that the classification of the network artifacts by the intrusion detection systems is valid. Steps were taken during the data collection stage to ensure the integrity of the artifacts was maintained. The subset of features of the artifacts was selected using rmcfs R package because of the high-dimensionality of the artifacts. After the feature selection phase, logistic regression was applied in the validation process. As reported in the findings, 95.1% accuracy of the classification using the selected subset of the features of the artifacts supports the initial assertions made by the Intrusion Detection Systems.

Appendix B

An Appendix: Glossary

The following terms are included to assist the reader in understanding this thesis.

Acquisition: A process by which the digital artifacts are duplicated, copied, or imaged.

Action Inquiry: A method of improving practice by systematically oscillating between taking action in the field and inquiry into it.

Admissibility: Refers to admissible artifacts, artifacts which may be introduced in a court of law.

Analysis: To look at the results of an examination for its significance and probative value to the case.

Artifact Wiping: The process of permanently eliminating a particular artifact.

Attribution: The process of linking events to the identity of the actor(s) responsibility for the events.

Authentic: Legally valid because all necessary procedures have been followed correctly.

Boruta Algorithm: A wrapper type of feature selection algorithm which builds a wrapper around the random forest classification.

CICFlowMeter: A network traffic flow generator written in Java that generates Biflows from pcap files, and extracts features from the flows.

Chain of Custody: Refers to the chronological documentation or paper trail that records the sequence of custody, control, transfer, analysis, and disposition of digital artifacts.

Copy: An accurate reproduction of information contained on an original physical item, independent of the electronic storage device (e.g., logical file copy). Maintains contents, but attributes may change during the reproduction.

Counter-forensics: Any method undertaken in order to thwart the digital investigation process conducted by legitimate forensics investigators.

Data Hiding: The process of making data difficult to find while also keeping it accessible for future use.

Daubert Test: Generally accepted standard used by the court to make assessment of whether an expert's scientific testimony is based on reasoning or methodology that is scientifically valid and can be applied to the facts at issue.

Digital Forensics: A branch of forensic science encompassing the recovery and investigation of material found in digital devices, often in relation to computer crime.

Digital Artifact: Information stored or transmitted in binary form that may be relied on in court.

Error Rate: The frequency in which errors occur in a given time period.

Examination: Technical review that makes the digital artifact visible and suitable for analysis; tests performed on the evidence to determine the presence or absence of specific data.

Evaluation: A systematic determination of an artifact's merit, worth and significance using criteria governed by a set of standards (Daubert criteria).

Forensically Clean: Digital media that are completely wiped of nonessential and residual data, scanned for viruses, and verified before use.

Forensically Sound: The assurance that an artifact was not corrupted or destroyed during investigative processes whether on purpose or by accident.

Hashing: The process of using a mathematical algorithm against data to produce a numeric value that is representative of that data.

Hypothesis: A supposition or proposed explanation made on the basis of limited information as a starting point for further investigation.

Linear Regression: A stochastic and probabilistic modeling methodology which follows a linear approach in modeling the relationship between the dependent variable and the independent variables.

Logistic Regression: A stochastic and probabilistic modeling methodology which models the relationship between independent variables and a categorical dependent variable.

MCFS-ID Algorithm: Monte Carlo Feature Selection and Interdependence Discovery Algorithm is a type of feature selection algorithm that is based on intensive use of classification and a feature is considered important or informative if it is likely to take part in the classification. Also, the algorithm provides a mechanism for discovery interdependencies among selected features.

Network: A group of computers connected to one another to share information and resources.

Network Artifacts: Digital artifacts that provide insight into network communications.

Network Flow: A sequence of packets with the same values for Source IP, Destination IP, Source Port, Destination Port and Protocol (TCP or UDP).

Peer Review: The process of subjecting a research work to the scrutiny of others who are experts in the same field.

Reliability: The degree to which the result of a measurement can be depended on to be accurate.

Scientific Methodology: An approach to seeking knowledge that involves forming and testing a hypothesis.

Testing: An investigation conducted to provide information about the reliability of the artifacts under test.

Trail Obfuscation: Techniques used to confuse, disorient, and divert the forensic examination process.

Validation - The process of checking or proving the validity or the reliability of the digital artifact.

Write Protection: Hardware or software methods of preventing data from being written to a disk or other medium.

Appendix C

An Appendix: R Codes Used for the Experiments

C.1 R Codes Used for Feature Selection Experiment

```
#Read Data File
dataset = read.csv("dataset.csv")

#Load Monte Carlos Feature Selection Package
library(rmcfs)

#Run the feature selection algorithm on the dataset
result = mcfs(Label ~., dataset, cuttoffPermutations = 0)

#Plot the distance function to check convergence of the algorithm
plot(result, type="distances")

#Check the final cutoff value for selected features
result$cutoff_value

#Plot the Relative Importance values of features
plot(result, type = "ri")

#Displays the Top features selected
plot(result, type = "features")

#Builds and displays the ID-Graph
gid = build.idgraph(result)
plot(gid, label_dist = 1)
```

```
#Displays Cross Validation results for top features
plot(result, type = "cv", measure = "wacc")
```

C.2 R Codes Used for Logistic Regression Analysis Experiment

```
#Read Data File
dataset = read.csv("dataset.csv")

#Summary
summary(dataset)

#Split data into training (70%) and validation (30%)
dt = sort(sample(nrow(dataset), nrow(dataset)*.7))
train = dataset[dt,]
val = dataset[-dt,]

#Check number of rows in training and validation data sets
nrow(train)
nrow(val)

#Run Logistic Regression
mylogistic = glm(Label ~ Fwd.Packet.Length.Max + Avg.Fwd.Segment.Size + To-
tal.Length.of.Fwd.Packets + Subflow.Fwd.Bytes + Fwd.Packet.Length.Mean + Packet.Length.Mean
+ Average.Packet.Size + Fwd.Header.Length.1, data = train, family = "binomial")
summary(mylogistic)$coefficient

#Stepwise Logistic Regression
mylogit = step(mylogistic)

#Logistic Regression Coefficient
summary.coeff0 = summary(mylogit)$coefficient

#Calculating Odd Ratios
OddRatio = exp(coef(mylogit))
summary.coeff = cbind(Variable = row.names(summary.coeff0), OddRatio, summary.coeff0)
row.names(summary.coeff) = NULL

#R Function : Standardized Coefficients
stdz.coff = function (regmodel)
{ b = summary(regmodel)$coef[-1,1]
```

```
sx = sapply(regmodel$model[-1], sd)
beta = (3^(1/2))/pi * sx * b
return(beta)
}

std.Coeff = data.frame(Standardized.Coeff = stdz.coff(mylogit))
std.Coeff = cbind(Variable = row.names(std.Coeff), std.Coeff)
row.names(std.Coeff) = NULL

#Final Summary Report
final = merge(summary.coeff, std.Coeff, by = "Variable", all.x = TRUE)

#Prediction
pred = predict(mylogit,val, type = "response")
finaldata = cbind(val, pred)

#Storing Model Performance Scores
library(ROCR)
pred_val = prediction(pred, finaldata$Label)

#Maximum Accuracy and prob. cutoff against it
acc.perf = performance(pred_val, "acc")
ind = which.max( slot(acc.perf, "y.values")[[1]])
acc = slot(acc.perf, "y.values")[[1]][ind]
cutoff = slot(acc.perf, "x.values")[[1]][ind]

#Print Results
print(c(accuracy= acc, cutoff = cutoff))

#Calculating Area under Curve
perf_val = performance(pred_val,"auc")
perf_val

#Plotting Lift curve
plot(performance(pred_val, measure="lift", x.measure="rpp"), colorize=TRUE)

#Plot the ROC curve
perf_val2 = performance(pred_val, "tpr", "fpr")
plot(perf_val2, col = "green", lwd = 1.5)

test_roc = roc(finaldata$Label  pred, plot=TRUE,print.auc=TRUE,colorize=TRUE)
```

Bibliography

[1] Michal Draminski and Jacek Koronacki. rmcfs: An R Package for Monte Carlo Feature Selection and Interdependency Discovery. *Journal of Statistical Software*, 2018. doi: https://cran.r-project.org/web/packages/rmcfs/vignettes/jss2621.pdf.

[2] 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. doi: https://doi.org/10.1016/j.diin.2016.04.006.

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

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

[5] Altheide C. and Carvey H. *Digital Forensics with Open Source Tools.* Syngress, 2011.

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

[7] Eadaoin O'Brien, Niamh Nic Daeid, and Sue Black. Science in the court: pitfalls, challenges and solutions. *Phil. Trans. R. Soc. B*, May 2015. doi: http://dx.doi.org/10.1098/rstb.2015.0062.

[8] 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.

[9] Eva A. Vincze. Challenges in digital forensics. *Police Practice and Research*, 17(2): 183–194, January 2016. doi: https://doi.org/10.1080/15614263.2015.1128163.

[10] Daniel B. Garrie and J. David Morrissy. Digital Forensic Evidence in the Courtroom: Uderstanding Content and Quality. *Nw. J. TECH. & INTELL. PROP.*, 12 (2), 2014. doi: http://scholarlycommons.law.northwestern.edu/njtip/vol12/iss2/5.

[11] Simson L. Garfinkel. Digital forensics research: The next 10 years. *Digital Investigation*, 30:S64–S73, 2010. doi: https://doi.org/10.1016/j.diin.2010.05.009.

[12] David Lillis et al. Current Challenges and Future Research Areas for Digital Forensic Investigation. *11th Annual ADFSL Conference on Digital Forensics, Security and Law*, 2016. doi: https://arxiv.org/pdf/1604.03850v1.pdf.

[13] Jawwad A. Shamsi et al. Attribution in cyberspace: techniques and legal implications. *Security and Communication Neworks*, 2016. doi: https://onlinelibrary.wiley.com/doi/epdf/10.1002/sec.1485.

[14] Santos O. and Muniz J. *CCNA Cyber Ops SECOPS 210-255 Official Cert Guide*. Pearson Education, Inc., 2017.

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

[16] Shahzad Saleem. *Protecting the Integrity of Digital Evidence and Basic Human Rights During the Process of Digital Forensics*. PhD thesis, Computer and Systems Sciences, Stockholm University, San Diego, CA, USA, 2015. URL `https://www.diva-portal.org/smash/get/diva2:806849/FULLTEXT02.pdf`.

[17] Tobias Kuhn and Michel Dumontier. Making Digital Artifacts on the Web Verifiable and Reliable. *IEEE Transactions on Knowledge and Data Engineering*, 27(7):2390–2400, 2015. doi: https://doi.org/10.1109/TKDE.2015.2419657.

[18] Xiaoyu Du, Nhien-An Le-Khac, and Mark Scanlon. Evaluation of digital forensic process models with respect to digital forensics as a service. *CoRR*, abs/1708.01730, 2017. doi: https://arxiv.org/ftp/arxiv/papers/1708/1708.01730.pdf.

[19] Reza Montasari. Review and Assessment of the Existing Digital Forensic Investigation Process Models. *International Journal of Computer Application*, 147(7):0975–8887, 2016. doi: https://pdfs.semanticscholar.org/7542/aabfa6f69d7c219c85c6b951702367cc0053.pdf.

[20] Murat Gul and Emin Kugu. A survey on anti-forensics techniques. In *Artificial Intelligence and Data Processing Symposium (IDAP), 2017 International*, Malatya, Turkey, 2017. IEEE. doi: https://doi.org/10.1109/IDAP.2017.8090341.

[21] Dahbur K and Mohammad B. Toward understanding the challenges and countermeasures in computer anti-forensics. *Cloud Comput. Adv Des Implement Tech*, 176, 20122.

[22] Kyoung et al. Anti-Forensic Trace Detection in Digital Forensic Triage Investigations. *Digital Forensics Security and Law*, 12(1), 2017. doi: https://doi.org/10.15394/jdfsl.2017.1421.

[23] Humaira Arshad, Aman Bin Jantan, and Oludare Isaac Abiodun. Digital Forensics: Review of Issues in Scientifc Validation of Digital Evidence. *J Inf Process Syst*, 14 (2):346–376, 2018. doi: https://doi.org/10.3745/JIPS.03.0095.

[24] Liu C., Singhal A., and Wijesekera D. A Probabilistic Network Forensic Model for Evidence Analysis. In *Peterson G., Shenoi S. (eds) Advances in Digital Forensics XII. DigitalForensics 2016. IFIP Advances in Information and Communication Technology*. Springer, Cham, 2016. doi: https://doi.org/10.1007/978-3-319-46279-0_10.

[25] Nicholas Walliman. *Research Methods: the basics*. Routledge, New York, NY, USA, 2011.

[26] Peter Reason and Hilary Bradbury. *Handbook of Action Research: Participative Inquiry and Practice*. SAGE, London, United Kingdom, 2001.

[27] Ivan Flechais, Cecilia Mascolo, and M. Angela Sasse. Integrating security and usability into the requirements and design process. *Int. J. Electronic Security and Digital Forensics*, 1(1), 2007. doi: https://pdfs.semanticscholar.org/7a5a/b8d42efa881b366d7f9676dc70f3434ca558.pdf.

[28] Richard L. Baskerville. Investigating Information Systems with Action Research. *Communications of the Association for Information Systems*, 2(19), 1999. doi: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.137.2608&rep=rep1&type=pdf.

[29] McKemmish R. Advances in Digital Forensics. In *IFIP International Federation for Information Processing*, Boston, 2008. Springer. doi: https://doi.org/10.1007/978-3-319-46279-0_10.

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

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

[32] Arash Habibi Lashkari et al. Characterization of Tor Traffic Using Time Based Features. In *The proceedings of the 3rd International Conference on Information System Security and Privacy*, Porto, Poturgal, 2017. SCITEPRESS. doi: https://doi.org/10.5220/0006105602530262.

[33] Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. *4th International Conference on Information Systems Security and Privacy (ICISSP)*, 2018. doi: http://www.scitepress.org/Papers/2018/66398/66398.pdf.

[34] De Silvia et al. *Grammer-Based Feature Generation.* Springer, Singapore, 2015.

[35] Miron Bartosz Kursa. Package 'Boruta'. 2018. doi: https://cran.r-project.org/web/packages/Boruta/Boruta.pdff.

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

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

[38] Jocelyn E. Holden, W. Holmes Finch, and Ken Kelley. A comparison of Two-Group Classification Methods. *SAGE journals*, 71(5), May 2011. doi: https://doi.org/10.1177/0013164411398357.