**Alma Mater Studiorum – Universitá di Bologna**

in cotutela con University of Luxembourg - Universitè du Luxembourg

**DOTTORATO DI RICERCA IN**

LAW, SCIENCE AND TECHNOLOGY

**Ciclo 34°**

**Settore Concorsuale:** 01/B1 - INFORMATICA

**Settore Scientifico Disciplinare:** INF/01 - INFORMATICA

# DIGITAL FORENSICS AI: ON PRACTICALITY, OPTIMALITY, AND INTERPRETABILITY OF DIGITAL EVIDENCE MINING TECHNIQUES

**Presentata da: SOLANKE, Abiodun Abdullahi**

**Coordinatore Dottorato**

Monica Palmirani

**Supervisore**

Maria Angela Biasiotti

**Supervisore**

Sjouke Mauw

**Esame Finale Anno 2022**

**Alma Mater Studiorum – University of Bologna**

In Collaboration with the University of Luxembourg - Universitè du Luxembourg

**Ph.D. PROGRAMME IN**

LAW, SCIENCE AND TECHNOLOGY

**Cycle 34°**

**Competition Sector:** 01 / B1 - COMPUTER SCIENCE

**Scientific Disciplinary Sector:** INF / 01 - COMPUTER SCIENCE

# DIGITAL FORENSICS AI: ON PRACTICALITY, OPTIMALITY, AND INTERPRETABILITY OF DIGITAL EVIDENCE MINING TECHNIQUES

**Submitted By: SOLANKE, Abiodun Abdullahi**

**Ph.D. Coordinator**

Monica Palmirani

**Supervisor**

Maria Angela Biasiotti

**Supervisor**

Sjouke Mauw

**Final Exam Year 2022**

UNIVERSITÉ DU LUXEMBOURG

PhD-FSTM-2022-076
The Faculty of Science, Technology and Medicine

Department of Legal Studies

**DISSERTATION**

Defence held on 17/06/2022 in Bologna, Italy

to obtain the degree of

**DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG EN INFORMATIQUE**

**AND**

**DOTTORE DI RICERCA**

*in Law, Science and Technology*

By

**SOLANKE, ABIODUN ABDULLAHI**

Born on 2nd January, 1983 in Kaduna (Nigeria)

# DIGITAL FORENSICS AI: ON PRACTICALITY, OPTIMALITY, AND INTERPRETABILITY OF DIGITAL EVIDENCE MINING TECHNIQUES

**Dissertation Defence Committee**

**Pierluigi Perri**, *Chairman*
*Professor, Universitá di Milano*
**Stefano Ferretti**, *Vice-Chairman*
*Professor, Universitá di Urbino*
**Stefano Pietropaoli**, *Member*
*Professor, Universitá di Firenze*
**Dr Maria Angela Biasiotti**, *Dissertation Supervisor*
*Ricercatore, ITTIG-CNR*
**Sjouke Mauw**, *Dissertation Supervisor*
*Professor, Université du Luxembourg*

# Declaration of Authorship

I, Abiodun Abdullahi SOLANKE, declare that this thesis titled, “Digital Forensics AI: on Practicality, Optimality, and Interpretability of Digital Evidence Mining Techniques” and the work presented in it are my own. I confirm that:

- This work was completed in whole or mainly while in candidature for doctoral degree at this university.
- Where any part of this thesis has been presented before for a degree or other qualification at this University or another institution, this has been explicitly stated.
- Where I have consulted the published work of others, proper citation is always provided.
- Where I have quoted the work of others, I always provide the source. Except for such quotations, this thesis is entirely my own work.
- I have acknowledged all major sources of help.
- Where the thesis is based on work that I did in collaboration with others, I have made it clear exactly what they did and what I contributed.

Signed:

_____

Date:

_____

# Abstract

SOLANKE Abiodun Abdullahi

*Digital Forensics AI: on Practicality, Optimality, and Interpretability of Digital Forensics Techniques*

Digital forensics as a field has progressed alongside technological advancements over the years, just as digital devices have gotten more robust and sophisticated. However, criminals and attackers have devised means for exploiting the vulnerabilities or sophistication of these devices to carry out malicious activities in unprecedented ways. Their belief is that electronic crimes can be committed without identities being revealed or trails being established. Several applications of artificial intelligence (AI) have demonstrated interesting and promising solutions to seemingly intractable societal challenges. This thesis aims to advance the concept of applying AI techniques in digital forensic investigation. Our approach involves experimenting with a complex case scenario in which suspects corresponded by e-mail and deleted, suspiciously, certain communications, presumably to conceal evidence. The purpose is to demonstrate the efficacy of Artificial Neural Networks (ANN) in learning and detecting communication patterns over time, and then predicting the possibility of missing communication(s) along with potential topics of discussion. To do this, we developed a novel approach and included other existing models. The accuracy of our results is evaluated, and their performance on previously unseen data is measured. Second, we proposed conceptualizing the term "Digital Forensics AI" (DFAI) to formalize the application of AI in digital forensics. The objective is to highlight the instruments that facilitate the best evidential outcomes and presentation mechanisms that are adaptable to the probabilistic output of AI models. Finally, we enhanced our notion in support of the application of AI in digital forensics by recommending methodologies and approaches for bridging trust gaps through the development of interpretable models that facilitate the admissibility of digital evidence in legal proceedings.

**Subject**: Legal Informatics

**Keywords**: Digital Forensics AI, Evidence Mining, Digital Forensics, Digital Evidence, ANN, DNN, DL, ML, CNN, VAE, VGAE, GRU, Optimization, Evaluation, Natural Language Processing, Explainable AI, Interpretable AI, Topic Modelling, E-mail Artifacts, LDA, Latent Dirichlet Allocation, NMF, Non-Matrix Factorization

Il Riassunto

La Digital forensics come campo ha progredito insieme ai progressi tecnologici nel corso degli anni, proprio come i dispositivi digitali sono diventati più robusti e sofisticati. Tuttavia, criminali e aggressori hanno escogitato mezzi per sfruttare le vulnerabilità o la sofisticazione di questi dispositivi per svolgere attività dannose in modi senza precedenti. La loro convinzione è che i crimini elettronici possano essere commessi senza che le identità vengano rivelate o che vengano stabiliti percorsi. Diverse applicazioni dell'intelligenza artificiale (AI) hanno dimostrato soluzioni interessanti e promettenti a sfide sociali apparentemente intrattabili. Questa tesi mira a far avanzare il concetto di applicazione delle tecniche di IA nell'indagine forense digitale. Il nostro approccio prevede la sperimentazione di un caso complesso in cui i sospetti corrispondevano via e-mail e cancellavano, sospettosamente, alcune comunicazioni, presumibilmente per nascondere prove. Lo scopo è dimostrare l'efficacia delle reti neurali artificiali (ANN) nell'apprendimento e nel rilevamento dei modelli di comunicazione nel tempo, e quindi prevedere la possibilità di comunicazioni mancanti insieme a potenziali argomenti di discussione. Per fare questo, abbiamo sviluppato un nuovo approccio e incluso altri modelli esistenti. Viene valutata l'accuratezza dei nostri risultati e viene misurata la loro performance su dati precedentemente non visti. In secondo luogo, abbiamo proposto di concettualizzare il termine “Digital Forensics AI” (DFAI) per formalizzare l'applicazione dell'IA in digital forensics. L'obiettivo è quello di evidenziare gli strumenti che facilitano i migliori risultati evidenziali e meccanismi di presentazione che sono adattabili all'output probabilistico dei modelli AI. Infine, abbiamo migliorato la nostra nozione a sostegno dell'applicazione dell'IA nella digital forensics raccomandando metodologie e approcci per colmare le lacune di fiducia attraverso lo sviluppo di modelli interpretabili che facilitano l'ammissibilità delle prove digitali nei procedimenti legali.

**Oggeto**: Informatica legale

**Parole chiave:** Digital Forensics AI, Evidence Mining, Digital Forensics, Prove digitali, ANN, DNN, DL, ML, CNN, VAE, VGAE, GRU, Ottimizzazione, Valutazione, Elaborazione del linguaggio naturale, AI spiegabile, AI interpretabile, Modellazione argomento, Artefatti e-mail, LDA, allocazione di Dirichlet latente, NMF, fattorizzazione non matrice

# Acknowledgements

All praise is due to the Almighty God, The Most Beneficent and Most Merciful who has made the beginning and conclusion of this journey possible. I would like to express my profound gratitude to the coordinator, Prof. Monica Palmirani, and the entire Board and staff members of the Last-JD programme, for finding me suitable for this doctoral programme.

My supervisors, Dr. Maria Angela Biasiotti, and Prof. Sjouke Mauw, deserve my deepest gratitude for their constant support, guidance, mentorship, and counsel throughout my doctoral research. Indeed, I am appreciative of their compassion and attentiveness. Thank you! Prof. Leone Van Der Torre and Dr. Reká Markovich deserve a substantial portion of this compliments for their contributions to our remarkable research experience at the University of Luxembourg. Most especially, I would like to thank Dr. Xihui Chen and Dr. Ramrez-Cruz Yunior for their contributions to the basis upon which the direction of this research is founded. I heartily appreciate and commend your efforts, the constant push, and the demand for improvement.

To my spouse, Aminat, even though we are miles apart, you are the reason why bad days did not turn into doomsday. Thank you for our shared love, affection, and understanding. The completion of this thesis is possible because you provided the motivation to persevere. My gratitude goes to my mother for her prayers and words of encouragements, and my siblings; Olukayode, Olatunde, Oluwayemis, Temitope, Oluwapelumi, Tomiloba, Omotolani, and Ibukunoluwa for their unconditional love, supports, and absolute understanding when I am unwilling.

To my friends and relatives; I value your affection and concerns. Special thanks to Dr. Babatunde Giwa, Oyeneye Tunde, Odunuga Olaide, and everyone else whose name I cannot put here; you have made this journey worthwhile. God bless you all.

To all my colleagues and friends that I met during the course of my research work, you have been fantastic throughout, and I am grateful for our shared professional relationship and the way in which we inspired one other to succeed. Grazie tutti!

To my late father, whose DNA of resilience I inherited, may you continue to rest in peace.

# Contents

# List of Figures

# List of Tables

# Part I

# General Introduction

# Chapter 1
# INTRODUCTION

## 1.1 Introduction

The history of forensics dates back thousands of years — the Babylonians in 200 BC used fingerprints to sign contracts, but the concept got popularized in 1892, when Sir Francis Galton conducted research by categorizing fingerprint patterns to determine the likelihood of two persons having the same sets of fingerprints. Galton's research eventually resulted in the development of what we now refer to as forensics.

Digital forensics (DF) origin dates all the way back to more than five decades, when two data recovery engineers successfully restored the lone copy of a database file that had been deleted mistakenly (Garfinkel, 2010). The discipline of computer forensics first appeared in a 1992 publication by (Collier and Spaul, 1992) and has progressively evolved into forensic science since then. Now, the discipline's scope has been expanded to encompass the presentation of forensic investigation findings in a court of law. Hence, the terms computer forensics, forensic computing, and digital forensics are all used interchangeably in the context of the acquisition, investigation, analysis, and presentation of digital evidence in a legal proceeding (Schatz, 2007).

While DF as a domain has evolved over time to keep up with technological advancements, as digital devices have become more robust and sophisticated in their ability to solve seemingly intractable societal problems, the intent and eventual use of these technologies to commit crimes has become more complex and widespread. Attackers and criminals have discovered methods for exploring the vulnerabilities or sophistication of these devices to carry out malicious operations in previously unthinkable ways. The primary motivation for electronic crime, as opposed to traditional physical crime, is the hope that evidence can be hidden; identities can be masked; actions can go undetected; and trails can be obfuscated.

What typically necessitates digital forensic investigation is cybercrime[1], i.e., the commission of crime using electronic means, which prompted the development of a system of rules, processes, and standards for systematically investigating such crimes. Among the early classifications of what was considered cybercrime were fraud; theft; use of unsolicited software; violation of privacy; hacking; malware or virus attack, and so on (Furnell, 2003). However, nowadays, these crimes have evolved in sophistication and operational complexities to include: Cyberextortion[2], Cryptojacking[3], Cyberespionage[4], ransomware attacks, Distributed-Denial-of-Service[5], phishing[6] and so on. Cybercrime has advanced to the point where it is expected to cost the global economy ten trillion dollars yearly by 2025, an almost 300 percent increase over 2015 (Steve, 2021). It was estimated that, if cybercrime were a country, it would rank third in terms of GDP after the United States and China. Every half-minute, an attack, attempted attack, intent to commit — or actual commission of e-crime occurs, according to reports (Marija, 2021).

When paired with the activities of cyber criminals, these staggering figures create a new type of challenge for the fields of law, criminology, law enforcement, and cyber security, among others. The whole nature of crime has shifted dramatically, with security and justice systems scrambling to devise measures and redefine laws to combat electronic-based criminal activity. Despite increased resources for cyber defence by governments and organizations, constant amendments to laws, and improvements on the skills and capacities of law enforcement and

[1] Cybercrime is referred to as an illegal action directed against or involving a computer, a network, or an interconnected devices. Individuals or organizations commit cybercrime for a variety of motives, including economic, personal, or political. Description available at https://www.kaspersky.com/resource-center/threats/what-is-cybercrime

[2] "Cyberextortion is a crime involving an attack or threat of an attack coupled with a demand for money or some other response in return for stopping or remediating the attack." Definition available at https://searchsecurity.techtarget.com/definition/cyberextortion

[3] "Cryptojacking is a threat that embeds itself within a computer or mobile device and then uses its resources to mine cryptocurrency. Cryptocurrency is digital or virtual money, which takes the form of tokens or coins." Definition available at https://www.kaspersky.com/resource-center/definitions/what-is-cryptojacking

[4] "Cyber Espionage, or cyber spying, is a type of cyberattack in which an unauthorized user attempts to access sensitive or classified data or intellectual property (IP) for economic gain, competitive advantage or political reason." Definition available at https://www.crowdstrike.com/cybersecurity-101/cyberattacks/cyber-espionage/

[5] "A distributed denial-of-service (DDoS) attacks target websites and online services. The aim is to overwhelm them with more traffic than the server or network can accommodate. The goal is to render the website or service inoperable." Definition available at https://us.norton.com/internetsecurity-emerging-threats-what-is-a-ddos-attack-30sectech-by-norton.html

[6] "Phishing is a cybercrime in which a target or targets are contacted by e-mail, telephone or text message by someone posing as a legitimate institution to lure individuals into providing sensitive data such as personally identifiable information, banking and credit card details, and password." Definition available at https://www.phishing.org/what-is-phishing

security experts, prevention of attacks remain elusive. This is especially expected given that technological advancements have usually outpaced legislations, or the development of detective/preventive mechanisms.

The interconnectedness of digital devices, particularly with the concept of the Internet of Things (IoTs[7]) (Ashton, 2009; Atzori, Lera, and Morabito, 2010; Xia, et al., 2012), which has become indispensably pervasive, has facilitated unprecedented access to and sharing of information (a commodity in and of itself). Oftentimes, this information overload (also referred to as infobesity), which is associated with excessive exposure to (mainly superfluous) information, has resulted in the unlawful storage of personal data for a variety of reasons, one of which is malicious.

Equally, developers/producers of technological systems are constantly enhancing the security of their systems to avoid illicit data exfiltration or unauthorized access by third parties, by incorporating data encryption techniques. Strong encryption-protected devices have significantly harmed the capacity of DF experts and LEAs to conduct adequate investigations in a short amount of time or without violating certain rules.

Computer/Digital forensics developed out of the need to accurately identify, trace, analyse, and report on the activities (or potential) of cybercriminals, in most situations, so that a court of competent jurisdiction can decide on the degree of the commission (or non-commission) of crime and adjudicate accordingly. Computer forensics is defined in (Kuchta, 2000) as "the science concerned with the relation and application of computers and legal issues." According to (Casey, 2004), DF process entails identifying investigative activity (including determining relevant digital sources), gathering information, safeguarding it against inadvertent alterations, analyzing, and reporting the examination's findings. In (Kerr, 2011), DF integrates computer science principles, such as computer architecture (operating and file systems), software engineering, and computer networking, with legal procedures defining criminal, civil, cyber, and evidence laws.

Over the last decade, the area of digital forensics has seen remarkable developments. Several improvements in the testing and validation of forensic tools; open-source tools with verifiable codes; the development of interoperable, community-oriented, ontology-based, investigative

[7] IoT connect "things and people – all of which collect and share data about the way they are used and about the environment around them." Definition available at: https://www.ibm.com/blogs/internet-of-things/what-is-the-iot/

information exchange tools, such as CASE[8] (Casey et al., 2017a; Casey et al., 2017b), EVIDENCE2-eCODEX[9], and 'The Evidence Project' (Biasiotti et al., 2018a; Biasiotti et al., 2018b); cross-border collaborative efforts; distributed or interconnected devices handing techniques, etc. Additionally, there are a number of challenges (Farid, 2019; Montasari and Hill, 2019; Casey, 2019) that the field is still facing — some are a result of recurring issues that have not been adequately addressed, while others are a result of the rapid growth in popularity of digital devices, which is putting forensic analysis' efficiency and capabilities to the test. (Garfinkel, 2010) highlighted several of these issues over the last decade and advocated a new research direction. This study will discuss the highlighted challenges and assess the current state of research and techniques implementation.

## 1.2 Digital Evidence

When crime is committed (particularly through electronic means), most often unintentional, often invisibly, footprints are left behind, and the goal of Law Enforcement Agencies (LEAs) or Forensic Investigators (FIs) is to follow this trail and reconstruct events that may be of potential probative value in the detection and subsequent prosecution of the perpetrator. What is most noteworthy in digital forensics is the evidence (referred to as digital evidence in this context) that is extracted, analyzed, and presented to prove culpability, complicity, intent, or guilt (or lack thereof) of a suspect, in a criminal or civil case. Digital evidence is the main resultant data of value in a forensic investigation. Michelle Theer was convicted for the first time using digital evidence in 2000. E-mails extracted from her computer revealed her complicity in a conspiracy to assassinate her spouse (Bryan, 2017). The significance of this (usually hidden) value in proving or disproving the commission of a crime, which may eventually result in a subject's conviction or acquittal, emphasizes its fragility, from the initial point of identification through the presentation in court. Numerous scientific methods have been used to extract evidence from digital devices, with many more being developed or proposed. However, because to the rapid and ongoing improvements in technology, many extraction procedures have become challenging, if not impossible.

[8] Cyber-investigation Analysis Standard Expression (CASE) is an ontology-based, community-exchanged digital evidence analysis system." Available at https://caseontology.org/

[9] The 'EVIDENCE2-eCODEX' project, among other things, implements existing European legislation concerning collection, preservation, and exchange of e-Evidence amongst member states, with best practices and guidelines integrated into a comprehensive framework. Available at: http:// http://www.evidenceproject.eu/

While forensic science has applications in a variety of science disciplines, such as biology, toxicology, and physics, these fields often view evidence in a physical context. In contrast, digital evidence is complex due to its volatility, volume, changeability, and sophisticated architecture of digital devices. These complexities have constantly cast doubt on the authenticity of digital evidence, as well as its admissibility in court. This means that the procedure used to get the evidence, its handling, and the chain of custody should guarantee that the evidence was not tampered with inadvertently and was obtained in accordance with all applicable fundamental rights. This can be seen as the legality of the procedure and the evidence. In Chapter 2 of this thesis, we discuss digital evidence in detail.

## 1.3 Motivation

The concept of evidence as described above is intended to emphasize the critical role of digital evidence in supporting or refuting hypothesis in a criminal or civil case. The narrated components stress the uniqueness of each of the points, in part or in whole, as a standard requirement, the absence of which could render the offered exhibit inadmissible as evidence. This indicates that the methodologies for extracting (Soltani and Seno, 2017; Horsman, 2019) digital evidence will be complex, massive, and time consuming. Numerous researchers have conceptualized and/or codified several approaches (Gaby and Benjamin, 2013; Novak, Grier, and Gonzalez, 2018; Kwon and Jeong, 2021) to reduce these complexities, but we are still a long way from developing an all-in-one efficient solution. The majority of forensic methods are either inadequate or incapable of discovering substantial evidence in artifacts, particularly when the items are "*out-of-the-ordinary, out of place, or subtly modified*" (Garfinkel, 2010).

While digital forensic artifacts from which evidence can be inferred provide critical validity for proof of facts, such as attribution of evidence to a suspect (Chaski, 2005; Himal, 2010; Kumar, et al., 2012; Sarunas and Jevgenijus, 2020), determination of intents (Mohammad, 2021), source identification, confirmation/corroboration of alibi or statement, etc., they are not without inherent complexities. Artifacts are similar to footprints in that they exist and are difficult for end users to access or manipulate. However, if the investigation is not conducted correctly, it is very possible for these artifacts to be missed. Textual data is a key source of artifacts, and it can take on a variety of forms — most frequently, conversations. Textual data includes e-mails, documents, tweets, and text messages, as well as log files and the registry. Complex scientific, linguistic, and philosophical techniques can be utilized to analyse this data

in order to elicit critical traces establishing or pointing the way to substantial evidence. A small number of existing forensic tools provide the capability to examine complex traces. While qualified experts can sift through these clues, it is possible that critical footprints may be ignored or missed.

Inspired by contemporary breakthroughs in Artificial Intelligence (AI) (Turin, 1950; McCarthy, 2004) — which is penetrating the entire landscape of technology and services at the moment. Digital devices, tasks, and major business services, for example, are leveraging machines' intelligence, predictive/forecasting capabilities, and pattern recognition capabilities (as well as their ability to mimic the human mind) to boost productivity and support critical organizational/governmental decision making. For example, the Big Data[10] field has evolved significantly over the years and demonstrated innovative approaches to handling and transforming data, as have 'Data Mining' (Agrawal and Psaila, 1995; Chung and Gray, 1999; David, Padhraic and Heikki, 2001) and Machine Learning (ML) (Carbonell, Michalski and Mitchell, 1983; Jordan and Mitchell, 2015) – both of which are subfields of AI. Statistical and computational methods can be useful in uncovering hidden patterns in huge datasets via data mining. This establishes a promising link between data mining techniques and the analysis of digital evidence. Additionally, ML is concerned with data and computer algorithms; it employs statistical and probabilistic methods to categorize, predict, and deduce important insights about the structure of data. (Farid and Rahman, 2010) introduced a Bayesian algorithm-based solution for detecting anomalous network intrusions. Their approach correctly classified various types of attacks with a low false positive rates. (Panagiotis, Theodoros and Petros, 2018) investigated the efficacy of deep computational methods for classifying and predicting crimes using open data from police reports. (Costantin, Giovanni and Olivieri., 2019a) investigated the potential for Answer Set Programming (ASP) to automate evidence analysis. The objective was to provide possible hypotheses as evidence in court. (Khan, Hanif and Muhammad, 2021) recently conducted a survey on the application of ML in the acquisition of digital evidence. The survey examined a variety of applications with the goal of promoting ML's potential in digital forensics.

The most recent, yet most advanced branch of AI to date is the Artificial Neural Network (ANN) (Jure, 1994; Wang, 2003; Dongare, Kharde and Kachare, 2012), a subfield of Deep Learning (DL) (LeCun, Yoshua and Geoffrey, 2015) or Deep Learning Neural Networks

[10] Big data is a field of AI that deals with systematic extraction or analysis of large volume and complex data sets.

(DNN). DL and ML are frequently used interchangeably, and except for the fact that DLs are designed with several layers, DL is a subdivision of ML. ANNs, also known as Neural Networks (NNs), are a form of computing system design that is inspired by the way neurons interconnect in the biological human brain. Although human brains appear to interpret real-world situations differently than computers or machines, the purpose of NNs is to attempt to simulate a network of neurons (also called nodes) connected (by edges) in layers, by sending signals (of real numbers) among each other, in order for computers to learn and make human-like decisions. ANNs can learn and model complex and non-linear relationships in data by identifying meaningful or unusual patterns in the data during training process. Whatever is learned during this training phase is expected to be generalizable and capable of being utilized to make predictions on previously unseen data. Due to the heterogeneity, volatility, and variation of data, ANNs are an ideal candidate for learning hidden relationships because they do not impose a fixed relationship on the data; rather, they compute the error rate during their learning phase and backpropagate by penalizing these errors for accurate predictions.

After a brief description of the complexities of digital forensics artifacts, the functionalities of AI models, and the potentially promising interconnections between them, one of the goals of this thesis – which forms the majority of Part II — is to develop an AI-based technique for detecting evidential patterns in textual communications. Although e-mails were employed in this research, the approach is extensible to other types of textual communications, including documents, tweets, and text/instant messages. The practicality of this part of the research is demonstrated in a project using Deep Neural Networks that is modelled on a real-world use case. The project uses Deep Neural Networks to reconstruct events based on prior learnable inferences in order to predict the likelihood of the occurrence of some prior events. Specifically, we deduced that a suspect may have deleted certain suspicious e-mails in order to conceal a suspected fraud. The fundamental objective of this paper is to promote and support and advance the concept and proposal for the (foundational) usage of AI in DF, particularly in detecting complex latent patterns that are difficult to infer manually.

In Part III of this thesis, we discussed the recent scepticism expressed by prosecutors and courts in general about AI-based evidence extraction methods/processes. Understandably, many have questioned the suitability of AI — commonly referred to as a closed-box – as a tool to be deployed in evidence mining[11], or more broadly, in digital forensics. While the operations of

[11] See Section 3.2

AI models are firmly steeped in pure mathematics, statistics, computational theories, and perhaps philosophy and psychology, they argued the interpretability and comprehensibility of these models, particularly in terms of how they arrive at a certain decision. Additionally, this sparked a lengthy argument and triggered a surge in research interest in computational science and law. The results of the majority of AI tasks are stochastic in nature - they are not deterministic. The challenge is that DF, as it relates to the presentation of evidence, should not be subject to probability — it must be well-established; scientifically verifiable; generally uncontroversial; and legally admissible, otherwise, the evidence will be excluded from proceedings or become inadmissible. We examined scepticism in AI-based evidence acquisition in terms of explainability and interpretability in this section. This article presents a framework defining the concepts and criteria that can be used to mitigate some of these mistrusts. In another dimension, we examined the many machine-driven methodologies currently being used for evidence mining, as well as the associated challenges and multiple attempts or proposals to solve complex problems. We demonstrated that different models, standards, and optimization strategies are appropriate for different tasks and hence have a significant impact on the result.

Overall, this thesis establishes, via practical examples, the notion of "***Digital Forensics AI***[12]" as a concept that encompasses the models, methods, acceptable standards, evaluation, and optimization techniques associated with AI models deployed for digital forensics purposes. In Chapter 5 of this thesis, we discuss this concept in detail.

## 1.4 Problem Statement

The 'Explosion of Complexity' (Caviglione, Wendzel and Mazurczyk, 2017) and 'Scale' (Casey, 2019) are two of the most difficult components of forensic investigation. The complexity of data – exacerbated by the development of novel technologies, and the rate at which this data is generated in huge quantities — amplified in particular by the 'Internet of Everything (IoE)' (Mahdi et al., 2015; Hussain, 2017), means that potential evidence may be dispersed over several physical or virtual hosts and geographical locations. The most significant issue that arises as a result of this is that forensic analysts become overwhelmed with a stream of data to investigate. Due to the fallibility of human beings, traces may be overlooked, and the backlog of work awaiting investigation frequently impairs the delivery of

[12] See Chapter 5

criminal investigation results in time for judicial proceedings (Montasari and Hill, 2019). A prolonged judicial process may end in an absolute acquittal, particularly if the defense attorney can demonstrate that the delay is the result of the prosecution's lack of significant evidence. This is especially true in jurisdictions where criminal procedure is accusatorial (as it is in the United Kingdom and former British Empire countries), requiring the prosecution report to be made available to the defense in order to ascertain agreement or otherwise prior to the actual trial (Sommer, 2018). In general, this pre-trial phase of civil litigation is referred to as "discovery," and it allows both parties in the lawsuit to request documents and other evidence.

Although there are forensic technologies capable of handling several terabytes of data, what these systems typically lack is the capacity to organize this material succinctly into relevant investigative clues. As (Garfinkel, 2010) noted, the majority of tools were built to aid experts in identifying a particular piece of evidence, not to aid with investigation. Garfinkel suggests the combination of process automation and forensic investigation, in his words, such automation "*should be able to detect and present outliers and other data elements that seem out-of-place. These systems will be able to construct detailed baselines that are more than simply a list of registry entries and hash codes for resident files.*" Thus, overcoming the challenges associated with identifying probative values in huge data would require the development of tools with the necessary engineering and visualization capabilities that can report possible digital clues to examiners in a standardized, unified manner for further investigation (Caviglione, Wendzel and Mazurczyk, 2017).

A strong probative analysis of a crime should be able to place objects, activities, and time in a single dimensional space, allowing for the reconstruction of events that may be suggestive of actors' prior activities. Historically, reconstruction tasks have been performed manually by examining a variety of disparate artifacts in order to establish relationships between objects, time, and crime, which explains why forensic investigations have taken an abnormally long time to complete. The term "event reconstruction" refers to the method of converting the state of a digital object to that of its causal event (Carrier and Spafford, 2004a). This simply entails establishing the occurrence of an event and its time of occurrence.

Several approaches have been proposed over the last decades in an attempt to address some of the numerous challenging problems involved with digital evidence reconstruction – the bulk of which are based on automated systems. The authors in (Khan, Chatwin and Young, 2007) monitored file system manipulations, captured the snapshots at specified intervals, and then

trained a neural network on the acquired data to build a post-event timeline on a hard disk to prove the execution of different software applications. In a similar task, but using a comparison technique, (Khan, 2012) evaluated the performance of a Bayesian network-based model vs a neural network-based model for reconstructing post-event timelines. The Bayesian network model outperformed the other models in terms of recognizing and detecting patterns in incomplete datasets. According to (Gladyshev and Patel, 2004), an incident state can be examined using a Finite State Machine to find all possible scenarios (FSM). Their FSM computes all plausible explanations based on a series of witness testimony. On a computer system, social network data, such as internet browsing cache, can be used to locate social media (Facebook, LinkedIn, e-mail sent and received, and so on) data. Turnbull and Randhawa (2015) demonstrated how low-level digital artifacts can be homogeneously fused to assist investigators in translating the status of file systems and reducing them to sequences of events, which can aid in making fact-based conclusions. In (Studiawan, Sohel and Payne, 2020), a DL technique in conjunction with a context and content attention model to identify and highlight terms with negative sentiments as events of interest using log (message) file data. "*The experimental result produced an F1 score of 98.43 percent and an accuracy of 99.64 percent, respectively.*"

Given the promising results of these proposed methods (and countless others) for reconstructing sequences of events, it may be sufficient to argue that intelligent automated systems have demonstrated sufficient effectiveness to stake a strong claim for inclusion in the domain of digital forensics – at least at the foundational level of DF investigation. However, the use of AI in forensics has been met with widespread criticism within the DF community (James and Gladyshev, 2013), with concerns that the quality of the results could be endangered, and expert expertise significantly diminished. Another area of concern that necessitates the adoption of AI in DF analysis is deep fake (Westerlund, 2019) — a deep/machine learning-based false synthetic images or videos created by substituting a human with a fictitious (non-existent) human being. This trend poses a significant threat to digital forensics and the possibility of identifying the perpetrator of a crime. Traditional forensic techniques will have a difficult time identifying false patterns in a deep fake image. In contrast to stakeholders' reservations about the deployment of closed-box models, we argue that for deep fakes, only an AI-based model that generated the image can learn and identify anomalous patterns in such images via a reverse engineering technique. It is safe to expect, however, that we will continue

to see such deceptively complex technologies in the future, for which AI can be instrumental in identifying.

## 1.5 Thesis Question

Considering the problems described in Section 1.4, this thesis aims to answer three (3) questions as follows:

**RQ1** – *How can we intelligently learn (detect) hidden patterns from the complex unstructured textual artifact?*

**RQ2** – *To what extent can Artificial Intelligence (AI) models aid in the analysis of a digital forensic investigation process?*

**RQ3** – *How can the (mis)trust in AI-based digital forensics analysis be mitigated through an effective evaluation of its results and an understandable presentation of its outcomes?*

The study discussed in this thesis aims to provide answers to the questions raised above.

### 1.5.1 Research Goal

The overall research goal of this thesis is to advance support for the adoption of AI methods in digital forensic investigation. Our approach is to begin by experimenting with a complex case scenario in which suspects corresponded by e-mail and suspiciously deleted some communication(s), ostensibly to conceal evidence. We observed that manually analyzing this case may be difficult and time consuming, and most available e-mail forensic tools lack the capability to detect latent behaviours. The objective is to demonstrate the usefulness of DNN in learning and recognizing communication patterns over time, and then predicting the possibility of missing communication(s). To accomplish this, we devised a novel approach and included several existing models. The correctness of our results is evaluated, and the performance of the model is measured when deployed on previously unseen data. Second, we intend to formalize the application of AI in digital forensics. The idea is to emphasize on the instruments that helps to achieve best evidential results in the process. Lastly, we seek to advance our notion in support of the application of AI in digital forensics by offering methodologies and mechanisms for bridging the trust gaps using interpretable approach in forensic analysis and presentation.

### 1.5.2 Research Approach

To guide the scope of this work, the broad definition of intelligent systems used in this thesis excludes the use of physical machines (such as robots) in the analysis or investigation of digital crime. To begin, examining the requirements for resolving the problems raised by the research questions (particularly **RQ1**) necessitates the need to modularize the entire method. Modularization[13] aids in determining which aspects of the problems require (or do not require) AI-based techniques. For example, the technique for processing and extracting text patterns from e-mails was developed using theoretical computing science[14] and formal language theory[15], not an AI model. This approach allows for the decomposition of the problem into sub-modules that can be tackled individually. A significant advantage of this modularization approach is that each sub-module can be independently evaluated, tested, and optimized. Combining the individual results yields the optimal solution to the problem.

Our approach focuses on identifying the most appropriate dataset and the most effective technique for pre-processing the data. Recognizing the need for relevance of extracted text, and the value of noise removal, while ensuring that the final dataset is not inappropriately skewed or biased. Our techniques enabled us to capture sequential snapshots of communications occurring at different points in time. We construct a Dynamic (communication) Graph (Siljak, 2008; Trivedi et al., 2018; Kazemi et al., 2020) in such a way that it allows for the smooth reconstruction of communication patterns over a pre-defined time period. While constructing a communication graph is a common technique in neural networks, training a Graph Neural Network (GNN) (Scarselli et al. 2008) model with multiple edge features has not been fully exploited in NN architectures. Our methodology is detailed in Chapter 4 of this thesis.

In the other part of this thesis, with reference to the practicality of our approach to solving the **RQ1** and a review of available methods, we discuss the evaluation techniques adopted in the majority of the methodologies. The purpose is to evaluate how effectively the methods used to investigate align with the objective. We discuss different standards and optimization techniques, and which ones are appropriate for a given task. Chapter 5 goes into greater detail

[13] Modular programming is the division of a computer program into distinct sub-programs.
[14] Theoretical computer science focuses on mathematical aspects of computer science such as computation theory, calculus, and type theory.
[15] Formal language is composed of words whose letters are drawn from an alphabet and which adhere to a specific set of rules.

on standards and optimization techniques, including an effective mode of verbally representing probabilistic outcomes in terms of strength of evidence.

Finally, we discuss the current issues of explainability and interpretability in the field of Digital Forensics AI, highlighting the reasons for the public's scepticism (and suspicion) of its acceptance (particularly among legal practitioners and courts in general), as well as the general efforts to make AI interpretable. The chapter 6 of this thesis details the techniques and the recommendations offered to mitigate the distrusts.

## 1.6 Contribution to Knowledge

The significance of this work is the approach utilized to address all of the research questions' components. The practical experimental approach, in particular, aims to demonstrate a use case that may be difficult to analyse with conventional forensic tools but is possible using AI. We can contextualize this work's contribution to knowledge in terms of the issues addressed in each part of the thesis. The following are the key contributions:

1. Pattern Recognition and Reconstruction of e-mail artifacts: In this scenario, we use a dynamic graph to describe the temporal evolution of communications, followed by a Variational Graph Autoencoder (VGAE) (Kipf and Welling, 2016b) to uncover probable e-mail deletions in the communication network. Our model represents node and edge attributes using multiple types of features, some of which are derived from the messages' metadata and others from the contents using natural language processing (NLP) (Manning and Schutze, 1999; Liddy, 2001; Chowdhury, 2003) and text mining (Tan, 1999) techniques. We use the autoencoder to identify missing edges; which we interpret as likely deletions, and to reconstruct their attributes, through which we infer hypotheses about the deleted messages' contents. The contributions made in this part, first , showcases the efficacy of edge reconstruction as a prediction technique in digital evidence mining. Second, our technique demonstrates the reconstruction process using a VGAE in conjunction with a graph convolutional recurrent network (GCRN) (Seo et al., 2018) and multidimensional edge features. Finally, we demonstrate the effectiveness of using topic models as feature vectors for edges in a dynamic temporal graph design. This part of the thesis is adapted from our work in (Solanke, Chen, and Ramírez-Cruz, 2021).

**2.** Digital Forensics AI standardization: As a discipline, digital forensics has developed certain standards that must be strictly adhered to in order to conduct a forensically sound investigation. However, approaches based on AI are being used that lack domain-specific standards. We present a "preliminary" conceptualization of Digital Forensics AI (DFAI) in this part of this thesis, with the goal of using it as a springboard for a more generic formalization. Additionally, as an addition to our proposal, we examine several evaluation methods, standardization procedures, and optimization techniques that are suitable to AI models employed for DF. The contributions here are twofold: first, contextualizing and presenting a preliminary conceptualization of Digital Forensics AI on which more refined formalization can be built; and second, the proposal of an adaptive confidence scale (C-Scale) for evaluating the strength of evidence generated by an AI algorithm. Finally, as a minor contribution to this part, we discuss and compare numerous AI algorithm optimization methods; highlighting their strength and drawbacks; their applicability to DF; including their time complexities, which may be crucial in determining the methodology to utilize in digital evidence mining.

3. Explainability and Interpretability of Digital Evidence Mining Techniques for Distrust Mitigation: The widespread belief that AI systems are closed-boxes and that it is difficult to analyse their inferential model has cast doubt on the legitimacy of decisions made by machines. Trust and confidence are much more crucial in a high-stakes domain like DF. Numerous research efforts have been made to improve the understandability of AI systems, and as a consequence, different methods have been proposed under the concept of explainable AI (XAI). However, as crucial as digital evidence is in proving or disproving an entity's guilt or innocence in a legal proceeding, efforts to make AI-based forensic analysis methods interpretable are still lacking (in practice or research). This chapter provides an overview of explainable and interpretable AI, as well as some approaches for making it more intelligible. Connecting this to DFAI, the contribution of this part of the thesis is majorly in the recommendations made to improve the interpretability of DFAI techniques and to mitigate mistrust. Additionally, a preliminary definition of explainable DFAI (xDFAI) is provided based on review of literatures and an understanding of the subject matter. Further along this line, a case for interpretable DFAI as a preferable approach is presented, as well as several use cases for different explanation approaches described.

## 1.7 Thesis Outline

We addressed digital forensics in detail in Chapter 2; the branches and process models, along with digital evidence and its admissibility in legal proceedings, as well as its challenges. Chapter 3 examines Artificial Intelligence (AI) and its relationship to the mining of digital evidence. We discuss symbolic and non-symbolic AI and their subcomponents, as well as the different ways in which AI might be used to extract digital evidence. Chapter 4 introduces our experiment with pattern recognition and reconstruction, in which we present an AI-based solution for detecting suspicious deletions in e-mail communications. In Chapter 5, we provide the preliminary formalization of the "Digital Forensics AI" concept, which aims to advance the application of AI in digital forensics by providing techniques for evaluating, standardizing, and optimizing digital evidence extraction. Chapter 6 discusses the importance of explainability and interpretability in AI-based digital forensics processes with recommendation for a trustworthy procedure. Chapter 7 concludes.

## 1.8 Chapter Summary

This chapter offers the groundwork for understanding the extent of our study in its entirety. We emphasized the problem statement that serves as the motivation for our work. We discussed our research goals and our proposed approach for addressing the issues. We summarized our contribution to knowledge for each part of the thesis. Meanwhile, the algorithmic description of our work's experimental part is available on: `https://github.com/spyderweb-abdul/Deletion-Detection-in-Unstructured-Data`

# Chapter 2

# Digital Forensics and Digital Evidence

As stated in Chapter 1, DF is the science that deals with the relationship and application of computers to legal issues. A DF investigation is usually instituted when there is a commission of crime — or civil case, through electronic means, with the purpose of finding probative values (referred to as evidence) to support or refute a claim. This chapter provides an overview of what DF is, the stages and procedures required, the challenges, and the many sorts of artifacts.

## 2.1 What is Digital Forensics?

The most widely acceptable definition of Digital Forensics to date was put forward by the Digital Forensics Research Workshop Technical Committee (DFRWS, 2001), which defines computer forensics as:

> *"The use of scientifically derived and proven methods towards the preservation, collection, validation, identification, analysis, interpretation, documentation, and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstructions of events found to be criminal or helping to anticipate unauthorized actions shown to be disruptive to planned operation."*

This definition is the most generic not only because it is "appropriate," but also because the sequential operations involved in DF; the medium; the resultant value; and the eventual consequence are all explicitly fused into a sentence. However, recent revisions to this definition (and to the broader notion of DF) now incorporate the legal aspects of these procedures, as well as rules of evidence. There are additional definitions of digital forensics available in a variety of literatures. In (Welch, 1997) computer forensics is defined as "*...the study of computer technology in relation to the law*." A slight modification to this definition is added in (Kuchta, 2000), defining computer forensics as the science that studies the interaction between computers (including its application) and legal issues. As one of DF's objectives is to identify and analyse digital evidence, Carrier in (Carrier, 2003) describes digital forensics as the process of identifying digital evidence using valid scientific methods in order to facilitate the reconstruction of events during an investigation. According to Caloyannides, "*computer*

*forensics is a collection of techniques and tools for locating evidence on a computer that can be used against one in a court of law*" (Caloyannides, 2004). The United Kingdom forensics regulator, as narrated by (White, 2010), defines digital forensics as *"…any scientific and technical knowledge that is applied to the investigation of a crime and the evaluation of evidence to assist courts in resolving questions of fact in court*." Brighi and Ferrazzano (2021) defines DF as the application of scientific and analytical methods to data stored in digital device or in transmission through a digital medium, to identify, process, and preserve the data in ways that makes it accessible as evidence at trial. DF also defines best practices for handling digital evidence, as well as the rigorous methodological requirements by the legal process.

Over the last decades, the DF discipline has gradually evolved into a forensic science discipline by adopting the major forensic science guidelines and standards. Although forensics refers to the application of scientific methodologies in justice systems, Casey (2004) defined it as "*the application of science to law and it is ultimately defined by use in court*." The scientific methods employed to collect and uncover evidence must be reliable, trustworthy, impartial, and the veracity of the resulting evidence must be provable. Essentially, any discipline purporting to be related to forensics must adhere to the standards of evidence collection, preservation, and analysis for presentation in a court. Thus, despite the fact that DF investigation procedures are arguably distinct from other scientific forensic investigations, they have consistently lent themselves to the scrutiny of hypothesis, verifiable approaches, reproducible outcomes, and standardized methodologies applicable in the practical scientific domain.

As stated previously, the purpose of digital forensic investigation is to uncover probative trails that could be used or presented in court as digital evidence. The majority of the discussion in this chapter about DF is devoted to the extraction of this probative value and the required procedures for identifying, analyzing, and presenting it in furtherance of, or conclusion of, a criminal or civil litigation. Any object discovered to be forensically valuable during a digital investigation is an artifact. Artifacts can take several forms, including log files, shell bags, database files, user dictionaries, registry keys, event logs, and timestamps. Other types include textual data, multimedia files, and GPS logs. Artifacts are incoherent data (sometimes, unintelligible) that, when analysed, can reveal the intent or state of mind of suspects or litigants, corroborate a certain content, or point to factual evidence. In this thesis, significant emphasis is placed on textual artifacts as a source of digital evidence.

To begin with, there are several words and concepts in digital forensics that needs explanation, and we define some of these words and concepts below:

## A. Digital Data

Digital data is data that is represented numerically. Often, they take the form of binary systems (traditionally, 0's and 1's that accepts a single value from a finite set) — which are represented as a string of discrete symbols and can be interpreted by a variety of machines. In a more formal sense, digital data is any data that is stored on or in a digital format. However, binary encoding is not a requirement; for example, network and keyboard cables carry electric signals that are converted to digital representation when instructions are passed (Carrier and Spafford, 2004b). Due to the unintelligibility of raw data, a sequence of processes must be performed to transform it to distinct (rather readable) outputs (Brighi and Ferrazzano, 2021), for example, output on paper or screen.

## B. Digital Object

*"An object composed of a set of bit sequence"* (CCSDS, 2012). Any digital entity or data structure composed of elements expressed in digital formats that is interoperable with other information systems is referred to as a data object (Gary, 2014). If a data object can be represented in digital format, and as a bitstream, it signifies that the data object is a collection of discrete symbols. Digital objects include hard disk sectors, network packets, and memory pages. A data object's characteristics are unique and can be identified differently; for example, a hard disk sector will store the content of an ASCII text document differently than it will store the content of a JPEG image (Carrier and Spafford, 2004a)

When the content of an object changes, the object assumes a new *'state'* as well. Therefore, when data is written to a computer memory, it modifies the state of the currently operating computer process.

## C. Digital Event

A digital event is an occurrence that modifies digital object's state (Carrier and Spafford, 2004a). The effect of an event causes the state of an object to change. When a digital object triggers an event, it is referred to as a cause. If a digital object modifies an object's event state, the object serves as evidence of the event.

A digital incident (or crime) is an occurrence or sequence of events that violates some established policy or law. To prove that an event occurred, hypothesis concerning the '*what, when, and where*' of such occurrence must be developed and tested — this process is referred to as investigation.

Forensic investigation is a phrase that refers to the process of examining digital objects using scientific methods in order to develop testable hypothesis about the occurrence of events that can be offered in furtherance of a court case.

The *digital evidence of an incident'* refers to any digital data that contains verifiable information supporting or refuting an incident hypothesis (Carrier and Spafford, 2004a).

Putting it all together, according to the authors in (Carrier and Spafford, 2004a) *"an object is evidence of an incident if its state was used to cause an event related to the incident or if its state was changed by an event that was related to the incident."*

## 2.2 Branches of Digital Forensics

Digital forensics is always evolving to keep up with the rapid advancement of digital technology. The miniaturization of digital devices has reduced the cost of technology, making it much easier to commit crimes or violate the law. To keep up with the fast pace of technological advancement, DF investigation has branched into a variety of discipline, each with its own set of unique specifics, methods, and guidelines. A few of these are described below.

### 2.2.1 Computer Forensics

This discipline encompasses computers, embedded digital systems, and storage media, with the objective of identifying, preserving, acquiring, analyzing, and presenting evidence discovered on these computer systems in court. Computer forensics is concerned with the reconstruction of events found to be incriminating (or potentially incriminating); this includes everything from internet activity logs to the primary files on a computer drive.

### 2.2.2 Disk Forensics

Disk forensics is a sub-field of DF concerned with the science of digital evidence extraction from storage media such as floppy disks, hard and USB drives/devices, and CDs. A disk drive's

forensic investigation often entails searching for active, modified, or deleted files on the drive's occupied and unallocated space. Additionally, disk forensics includes file recovery and carving from physically and logically damaged drives.

### 2.2.3 Network Forensics

This is a subfield of DF concerned with monitoring, capturing, storing, and analyzing computer network traffic (both local and internet) in order to gather information, collect evidence, and identify the source of an attack or intrusion. In comparison to other fields of DF, network data is generally volatile; thus, analysis always involves intercepting the network packet and filtering it in real time or storing it for later study. Because network traffic is frequently transferred and lost, it is prudent to investigate its pattern and capture data in transit between computing devices. The evidence gleaned through network analysis can be used in conjunction with other traces left on hard drives during a breach or intrusion attack. This also include wireless network forensics.

### 2.2.4 Mobile Forensics

This sub-branch is also known as 'mobile device forensics,' and it is concerned with the recovery of electronic evidence from mobile devices, e.g., cell phones, tablets, PDAs, GPS devices, and gaming consoles, among others. Mobile devices are critical in criminal investigations since they store many sorts of personal information such as contacts, messages, multimedia/image files, e-mails, chats, location information, and web browsing history. Typically, mobile forensics investigations focus on communication data, such as voice calls, e-mails, and short/instant messages, rather than on in-depth data recovery of deleted data (Casey, 2004). Although forensic analysis of mobile devices is becoming more difficult due to the usage of inbuilt end-to-end encryption to secure data, important information such as location — obtained via inbuilt GPS/location tracking — can be used as corroboration in a criminal case.

### 2.2.5 Malware Forensics

Malware is a type of computer program (code, script, or software) that is designed to disrupt or deny operational services, gain unauthorized access to systems, and exploit the system to exfiltrate data — frequently resulting in the loss of valuable resources, intellectual property, or privacy. As a result, discovering the source, functionality, and other properties of malware in

order to identify the perpetrators and motivations for attacks is referred to as 'Malware Forensics'. Malware analysis also includes determining the malicious program's entry point, mechanism of propagation, impact on the system, and the network port it attempts to use. There are numerous types of malwares, including rootkits, trojans, worms, viruses, backdoors, and keyloggers, and they are frequently classified based on various parameters such as the method of propagation; the program's intent; the mode of attack; and whether the program obtains user consent prior to execution.

### 2.2.6 Digital Image Forensics

Image forensics is a significant subfield in DF; it is concerned with resolving the origin and veracity of an image. The purpose of image analysis is to detect the presence (or absence) of anomalous traces inherent in a digital image as captured by the acquisition device and other operations involved in its creation (Piva, 2013). Typically, digital image forensics is prompted by two distinct events: 1) when a suspect argues their identity in an image; and 2) when a suspect repudiates the image's authenticity. However, the duty of investigators in these situations will be to refute these claims by presenting evidence to the contrary. A wide variety of evidence can be collected from image artifacts (Burns, 2020), including authenticity evidence (e.g., Exif data, metadata, pixel data), and content evidence (e.g., topography, sign language, and landmarks).

Differentiating between legitimate and illegitimate image processing is fairly challenging; for example, modifying, altering the compression ratio of an image, or editing an image to lessen its noise level are not considered illegal (Arshad, Aman and Abiodun, 2018). Consequently, it is necessary to specify a threshold point for quantifying the degree of legitimacy or alleged deception (Wong, et al., 2014). The admissibility of digital images as evidence in court has been complicated further by the recent emergence of deepfake[16], which calls into question the widely held belief that photographs are a true representation of reality — *"...there is more to a picture than meets the eye*" (Sencar and Memon (Eds.), 2012).

### 2.2.7 Multimedia Forensics

As with digital image forensics, multimedia forensics entails the analysis of multimedia signals such as audio, video, and games using techniques capable of extracting potentially probative

[16] While image manipulation is fairly common, deepfakes present a unique challenge by utilizing advanced AI algorithms to manipulate or synthesis visual and audio data with a significant potential for deception.

information that can aid in the authentication and estimation of the trustworthiness of digital multimedia content. The fundamental premise is that digital images have noise-like properties, which constitute an intrinsic trace of the acquisition device. Therefore, the investigation task will be to determine the source device that created the data, to verify the integrity of the content, and to extract key information from the multimedia signals. Statistical methods can be used to distinguish between computer-generated images, images produced with a camera, and images scanned. Similarly, by assuming that the acquisition device's traces can be manipulated, tamper detection algorithms can distinguish between patterns that match the image and those that do not. Forensics techniques can utilize a variety of signal enhancement techniques to increase the intelligibility of the data[17]. This forensic exploit allows the analysis of object's color, photogrammetric measurement of objects inside the content, and recognition of sound/visual patterns, among other things.

### 2.2.8 Memory Forensics

This is also referred to as 'live acquisition,' because it entails collecting evidence in raw form from system memory, such as RAM, cache, and system registers, which can subsequently be carved from the raw dump. Memory dumps may contain critical forensic data regarding the condition of the system before the occurrence of an incident, such as a security breach (Lord, 2020). Memory forensics has become vital as advances in cybercrime have enabled criminals to perpetrate crimes without leaving a trace on the hard drive. Typically, malicious programs must be loaded into memory in order to execute, which is why memory forensic analysis has proven critical in the investigation of sophisticated computer intrusions in which detectable trails on hard drives have been difficult to trace. Because the data in computer memory is volatile, care must be taken to protect the memory dump's integrity, much more so if the evidence collected will be utilized in a court case.

### 2.2.9 E-mail Forensics

E-mail forensics is a subfield of DF that focuses on the systematic analysis of e-mails in order to gather data that can be used as evidence in criminal investigations. E-mail has become one of the most commonly used modes of communication for sending messages, delivering documents (secret, legal, business, etc.), and transactions. However, e-mail has become a

[17] Multimedia Forensics – Laboratorio Elaborazione Segnali e Comunicazioni (LESC), Universitá Degli Studi Firenze.

primary conduit for a range of criminal activities, including the sharing of illegal files, the dissemination of hate messages, cyberbullying, and virus propagation, all of which have the potential to be lethal. In e-mail investigation, the major evidence is the e-mail header, which contains a wealth of metadata about the e-mail (Chirath, 2019). By analyzing the email header, it is possible to identify crimes such as internal data leakage, spam, spoofing, and phishing. Figure 2.1 illustrates an e-mail header graphically.


**Figure 4.1**: Graphical representation of a typical e-mail header
Image source: (Chirath, 2019).

Typically, when conducting an inquiry, the e-mail header is analyzed from bottom to top, as the sender's information is at the bottom and the receiver's information is at the top. Numerous techniques and tools exist for conducting forensic investigations into e-mail (Banday, 2011; Lazic and Bogdanoski, 2018; Chirath, 2019), the majority of which propose methodologies for analyzing e-mail headers. In Chapter 4 of this thesis, we discuss our method for detecting suspicious e-mail deletions between suspects.

### 2.2.10 IoT Forensics

This appears to be a new subfield of digital forensics that poses significant privacy concerns. It entails the forensic examination of IoT devices such as smart devices (lights, doorbells, speakers, cameras, and home appliances, etc.) and connected systems (e.g., in smart city). IoT forensics evidence can be used to corroborate claims made in court. For example, data retrieved from smart doors can be utilized to confirm the presence of a suspect in a particular location at a certain time.

## 2.3 Digital Forensic Process Model

The procedures involved in a typical forensic investigation are in phases which are expected to be diligently followed. These procedures, sometimes referred to as process model, consist of sequence of actions necessary to conduct forensic investigation. Process model usually begins with the notification of the occurrence of an incident through to presentation of findings (Casey, 2004). Generally, the basic standard DF process phases are identification, preservation, analysis, documentation, and presentation. However, several process models have been proposed – mostly in the form of frameworks — with additional phases, within which there could be sub-phases. The reason for these numerous (rather sophisticated) frameworks is because of the prevalent complex architectures of emerging technologies, which makes it almost impractical to adapt the general DF process model to a particular incident. For instance, (Du, Le-Khac and Scalon, 2017) analysed existing process model frameworks for cloud computing forensics, with focus on the benefit of Digital Forensics as a Service (DFaaS) (van Baar et al., 2014; van Beek et al., 2020) by leveraging the vast computing resources and storage capacities of cloud infrastructures. (Zia, Liu and Han, 2017) argued that, and introduced, the concept of application-specific DF process model as an important means to a sound forensics practice, especially in the context of IoT systems. They exemplified this importance in application scenarios such as Smart Home, Wearables, and Smart City. (Al Mutawa et al., 2018) proposed Behavioural DF Model which incorporates psychological approach to forensic investigation. The model is based on 'Behavioural Evidence Analysis (BEA)' (Turvey and Profiling, 2012) that involves the analysis of digital crime with respect to human interactions between the committer and the victim. The practical application of this model helped to identify logical paths to further relevant evidence and facilitated an in-depth understanding of the motivations of the offenders and other suspected collaborators. The capability of blockchain technology to comprehensively keep track of the entire events of transactions from inception is the basis of the work of (Lone and Mir, 2019). They proposed Forensic-Chain – a blockchain based process model to handle DF 'Chain of Custody' (Giova, 2011) in a way that ensures integrity of the process and maintain the evidence's authenticity. In one of several proposed frameworks to reduce instances of wrong convictions or exculpation, (Overill and Collie, 2020) proposed Digital Evidence Enhanced Process (DEEP), which is aimed at guaranteeing the reliability of presented evidence and ensure the standard evaluation of the competence of examiners.

Traditionally, DF (and mostly all forensic science) investigation begins with the investigation of the crime scene. The crime scene is any location connected to the incident (or where evidence may be found) which is of interest to the investigation. Investigators are often advised to be cautious about the management of a crime scene – its mishandling could totally nullify the entire forensic process (White, 2010). The prevalence of digital devices in incident/crime scenes particularly requires that standard digital forensic guidelines be applied, as electronic devices present in crime scene can also provide insight about an incident, or aid in the reconstruction of events that took place before, during, and after the crime. The modus operandi for the assessment of a crime scene is mostly dependent on the guidelines available to examiners according to predefined procedures established by their laboratory. Investigators should then be fully aware of these guidelines concerning pre-scene preparations and managements (ENFSI, 2015a). Described in section (2.3.1) are some important digital forensic investigation processes required to guarantee the integrity of an evidence for it to be admissible in court. It is important to also mention that the processes highlighted here are largely generic, and though the phases could be inexhaustive owing to numerous proposed models that tends to handle case scenarios differently, we support the notion that there is no one-size-fits-all process model.

### 2.3.1 Process Model Phases

#### A. Seizure

Before investigation can proceed, digital media involved in the incident(s) will be seized. Seizure means 'a dispossession of something against the will of the possessor[18].'Although, in some literatures, the first phase is identification, i.e., digital devices that can be used as exhibits are identified, before seizure can take place. However, if the case in contention is civil in nature, or involves a company's internal incident, the assumption is that the devices involved can be already identified. In a criminal case, sometimes, this phase can involve a search warrant[19], which is the needed legal authorization to search and seize exhibits. Several European nations, including Germany, Sweden, and Spain, require a court order before accessing or seizing stored data, whereas in Luxembourg, the significance of the information sought and its impact on the

[18] 'Seizure.' See Duhaime's Law Dictionary. Available online at: http://www.duhaime.org/LegalDictionary/S/Seizure.aspx

[19] A search warrant is a legal document issued by a magistrate or judge permitting law enforcement officers to search objects (i.e., people, devices, and locations) and seize any evidence associated with an incident. The regulations governing search warrants vary by jurisdiction — some are more considerate of the rule of law and the right to privacy, while others allow for search and seizure without obtaining authorization.

fundamental rights granted are considered. According to (ENISA, 2015), it is recommended that examiners have a flow chart detailing how to proceed in this phase, depending on the case scenario. In case the exhibit involved is a digital device, the general rule of thumb is to mirror the system and analyse in forensic lab — not at the incident location. However, before the mirroring can happen, the examiner needs to be aware of the general state of the system, and indeed, the environment. The photograph of the entire environment can be taken prior to any activities — the believe is that the prior state of the environment can be crucial to the general hypothesis derivable from the investigation. In most cases, it is recommended that the system be turned off, and isolated if on a network. However, turning off the system could impact the potential evidence on the device. Due to the volatile nature of computer memory, all data will be erased when the machine is shut off. To adequately preserve data (or potential evidence) in circumstances in which turning off might be undesirable, ACPO principle[20] (ACPO, 2012) encouraged that specialist's advice be sought, and audit trail of all processes be recorded and preserved. The (United States Department of Homeland Security, 2015) provided a first responders' pocket guide on steps to follow during an electronic evidence seizure. This guide offers a comprehensive guideline, depending on the device involved, and the state of the object.

To reflect what is largely available in digital forensics literatures, we can literally categorize the activities in the seizure phase into two main sub-phases, namely:

1. **Identification** (of exhibits): the activity here is to identify potential source of relevant evidence, the location, and the possessor of the data. The exhibits include devices and network configurations.
2. **Preservation:** this step includes the prevention of any activities that can challenge the integrity of the data or evidence collected, e.g., making sure all ongoing or scheduled jobs on the system, which might interfere with evidence collection, is stopped. It also consists of preserving the incident scene (usually by taking visual images of the scene) and all relevant Electronically Stored Information (ESI)[21].

[20] 'Good Practice Guide for Computer-Based Electronic Evidence' (Version 5). Available online at: https://library.college.police.uk/docs/acpo/digital-evidence-2012.pdf

[21] According to the Federal Rules of Civil Procedure (FRCP) in the United States, ESI is information that is created, manipulated, conveyed, or stored in a digital format using computer hardware and/or software.

## B. Acquisition/Collection

This process is also referred to as imaging, which is the creation of the exact sector level duplicate of the original digital media. The 'forensic duplication', which is a "*bit-for-bit copy of the data contained in the original device without any additions or deletions, even for the portions of the device that do not contain data* (Novak, Grier and Gonzalez, 2018)", is usually done with a write blocker[22], or software imaging tools such as Encase[23] and FTK imager[24]. Imaging is required because one of the most important rules in DF is to never perform direct forensic analysis on the original evidence. Analysis on original evidence will eventually change the properties (such as, date and time of modified, accessed and created) of the file, which will render the evidence inadmissible in a legal proceeding. Preserving the integrity of potential evidence is mostly essential during this process, so also the extensive documentation of the: entire event scenarios; software and hardware specifications; systems involved in the investigation procedure, and system being investigated. The acquired image is constantly verified with a method called hashing[25] — which uses hash functions, such as SHA-1[26] or MD5 (Rivest and Dusse, 1992), to tamper-protect the extracted data. Whatever methods used in the acquisition must not be intrusive — as this may result in a change in the physical features and destruction of evidence.

## C. Analysis/Examination

This is the phase during which an exhaustive systematic search for traces relating to the incident under investigation (Reith, Carr and Gunsch, 2002) is carried out, to elicit evidence that either supports or refutes specific hypotheses, intentions, or indicators of (deliberate — or not) data concealment. In this stage, investigators examine the content from the archived image files – using variety of methods and tools (EnCase, FTK, ILOOKIX[27], etc.) to recover deleted materials, search for relevant matches of some specific keywords or file types from massive

[22] "A write blocker is any tool that permits read-only access to data storage devices without compromising the integrity of the data. A write blocker, when used properly, can guarantee the protection of the data chain of custody." – Write Blocker, CRU. Available online at: https://www.cru-inc.com/data-protection-topics/write-blockers/

[23] https://security.opentext.com/encase-forensic

[24] https://www.exterro.com/ftk-imager

[25] Hashing is the process of applying a mathematical algorithm on an object (e.g., a string of text, a file, or a storage device) in order to obtain a unique alphanumeric value (hash value).

[26] 'Secure Hash Standard' – Federal Information Processing Standards Publication 180-2 (2002). Available online at: https://csrc.nist.gov/csrc/media/publications/fips/180/2/archive/2002-08-01/documents/fips180-2.pdf

[27] https://www.xtremeforensics.com/ilookix

archives of data, classify files based on timestamps, and identify suspicious files, such as encrypted or deliberately hidden files. The type of data recoverable from allocated and unallocated (deleted) disk space, systems-and-user-generated files include images, documents, internet browser history, e-mails, message logs, log files etc., however it varies depending on the investigation. With regards to classification based on timeline, it is extremely important to reduce data to list of points in time when a certain activity occurred. Typically, a list including date/time, sources, identities, and a description of the findings is used. Having this list can greatly aid investigators in tracking how files change over a period, and to have option for timeline searches (ENISA, 2015). Next after recovery of relevant information is the reconstruction of fragmented traces, events, and actions to draw conclusions. Usually at this stage, investigators work in close collaboration with LEAs, lawyers, and other stakeholders to understand every nuance of the case, map out the extent of the investigative actions, and agree upon the type of information that can serve as evidence. The investigative process must ensure that the conclusion is based on factual data and the expert's knowledge. Expert opinion is permissible in many jurisdictions, for example, according to the U.S. Federal Rules of Evidence (hereinafter referred to as FRE), with specific reference in this case to Rule 702[28], which states that "*expert (by knowledge, skill, experience, training, or education) may testify 'in the form of an opinion or otherwise', so long as: (1) the testimony is based on sufficient facts or data, (2) the testimony is the product of reliable principles and methods, and (3) the witness has applied the principles and methods reliably to the facts of the case*."

## D. Reporting

When investigation procedure is concluded, it is required that the information about the entire process be reported – usually in suitable form that is understandable by non-technical audience. (Casey, 2004) describes reporting as "*To provide a transparent view of the investigative process, final reports should contain important details from each step, including reference to protocols followed and methods used, to seize, document, collect, preserve, recover, reconstruct, organize and search key evidence.*" The reporting stage can be strategically sub-divided into two major sub-phases to reflect opinion of most literatures, namely:

[28] 'Rule 702 Testimony by experts' – Federal Rules of Evidence. Available online at: https://web.archive.org/web/20100819114909/http://federalevidence.com/rules-of-evidence#Rule702

1. **Documentation**: which includes the contemporaneous documentation of all activities relating to the investigation — such as the methods and tools used for testing, recovering, duplicating, and archiving of data, so also the software and hardware specifications of the system investigated, and the systems used in acquisition, examination, and assessment of evidence. The documentation of procedures is especially necessary to demonstrate: (1) the integrity of the preserved data, authenticity of the findings, reliability of the scientific methods involved, and admissibility, (2) that proper policies, rules, guidelines, and procedures have been adhered to by all parties, (3) that other competent forensic examiners can replicate the procedures and reproduce the same results.
2. **Presentation**: is the passing of documented procedures to those in whose capacity the investigation was commissioned, such as law enforcements (in a criminal case), or the employing company (in civil proceedings). The validity of the report can then help the commissioner in deciding whether, or not, to use the evidence in court. The best practice before presenting an evidence is to run a second reliable forensic tool or manually examining and comparing the original location of evidence with the original result (ENISA, 2015). The presentation should also clearly state (in details) the formulation of hypotheses and the inferences that led to the evidence and the expert's conclusion (Casey, 2004).

## 2.4 Uses of Digital Forensics Analysis

The analysis phase is very essential and key to the conclusions drawable from the investigative process. Often, digital evidence is intentionally obfuscated to make traces difficult or impossible, however, an intuitive DF analysis should be able to logically gather all fragmented or disjointed piece of evidence together within the same space of objects (people, devices, networks, data, etc.), time (access, create, modify, logs, etc.), activities, and location (GPS, maps, distributed systems, etc.), to infer or establish (retrospective or prospective) commonality or association. Digital forensic analysis can help in providing clues and trails which can be useful in leading investigators to the culprit. Casey and Rose (Casey, 2010) itemized some of the context in which DF analysis can be essentially useful.

### 2.4.1 Attribution

"*Attribution is more of an art than science.*[29]" This is particularly true because some aspects of forensic analysis involve, not only the use of scientific techniques, but also the knowledge of previous events – the tactics, and the tools and methods deployed – as well as assumptions, historical data, open-source intelligence, etc., to establish the confidence level of an investigative conclusion. Attribution in digital forensics literally refers to the tracking, identifying (individual, or device), and the assignment of some specific actions or responsibilities to a suspect or perpetrator. "*There is no simple technical process or automated solution to determine the responsibility of a cyber operation.*[30]" Rather, there are often times when psychological behavioural analysis (Ikuesan and Venter, 2019), language stylistics (Rosenblum, Zhu and Miller, 2011; McMenamin, 2020), content patterns found in e-mails (Himal, 2010), domain names/IP addresses, method of delivery of attacks, and other metadata uncovered during investigation have helped to conclude on certain assumptions or assertions of attribution. A typical scenario of attribution in digital evidence was the case of 'Maury Troy Travis[31]', an American serial killer who was tracked and arrested by the FBI simply by leveraging on the information Internet companies keep of visitors to their website. However, attributing a specific computer crime or activity to a certain individual might be challenging - since it may be difficult to prove that the owner of an internet account committed a crime that was perhaps done by someone else who gained unauthorized access to that account (Casey, 2010). Throughout attribution process, case catalogue of who, why, what, where and how is assembled with the aim to put together fragmented patterns across multiple investigations. Correlation of matching patterns is therefore established, and assumption of attribution can then be made. "*While attribution isn't an exact science, we can come close to attribution beyond a reasonable doubt – and we should continue trying.*[32]"

---

[29] Digital Forensics, Incident Response & Attribution – Cyber Forensic Intelligence, Technology Resilience (2017). Available online at: https://www.cybersecurityintelligence.com/blog/digital-forensics-incident-response-and-attribution-2022.html

[30] 'A Guide to Cyber Attribution.' – Office of the Director of National Intelligence (US) (2018). Available online at: https://www.dni.gov/files/CTIIC/documents/ODNI_A_Guide_to_Cyber_Attribution.pdf

[31] Peter Shinkle (2002). "Serial Killer Caught by his own Internet Footprint." St. Louis Post-Dispatch. Available online at: https://murderpedia.org/male.T/t/travis-maury.htm

[32] Justin Harvey (2017) "The shadowy – and vital – role attribution plays in cybersecurity." Security Blog, Accenture. Available online at: https://www.accenture.com/us-en/blogs/blogs-shadowy-vital-role-attribution-cybersecurity

### 2.4.2 Confirmation of Alibi

According to Merriam-Webster online dictionary, alibi[33] is "*the plea of having been at the time of commission of an act elsewhere than at the place of commission*". In a legal proceeding, alibi (or statement of alibi) is a defence to a criminal allegation that asserts the defendant was not present at the scene of the crime when it occurred. The most essential components in the confirmation of alibi, especially as regards digital evidence, are time and location. It is also noteworthy to recognize that the time/location evidence relates to the device(s) involved in the commission of the act — not the user. However, using these devices, and some other supporting evidence, can help to associate a case with an individual (Casey, 2011). The proliferation of digital devices, enhanced by the digitization and interconnections of everything, has increased the possibility of leaving, and the traceability of, digital footprints. Usually, suspects try to mislead investigators deliberately, or unwittingly, into believing they were somewhere (or engaged in something) different when the incident occur. Nevertheless, such information given by the suspect can be cross-referenced with the suspect's digital activities, to support or refute an alibi or statement. The challenge herein, however, is when the time/date configuration on a digital device is changed to manipulate traces, or the process of sending an e-mail is automated. The falsifiability of digital alibi is possible, and it has been demonstrated in (Castiglione et al., 2012), wherein a methodology showed how a typical individual actions (such as mouse clicks, writing of texts, pressure of key, pattern of daily online activities, etc.) on a computer can be automatically simulated in a way that is indistinguishable from those of original human activities. What is equally problematic is the use of a Virtual Private Network (VPN)[34] to mask the correct IP address of a device — allowing individual to purport to be connected from a different location. Such scenario complicates the investigation, because even the third-party remote internet providers may not be able to invalidate the location of the IP address used. In cases where an obscured piece of equipment or technique is suspected to be involved, it might be necessary to extend the approach of investigation by, for instance, contacting the manufacturer of a device with specific questions relating to the configuration of the device, interviewing other individuals who might be familiar with some components of the device or network used, or reconstructing the events surrounding the alibi and compare it with the original evidence (Casey, 2011). Most importantly, the

[33] "Alibi" – Merriam-Webster. Available online at: https://www.merriam-webster.com/dictionary/alibi

[34] A virtual private network (VPN) enables users to send and receive data over shared or public networks in the same way that they would if their computing devices were connected directly to the private network.

presence of evidence, or the lack thereof, to support or refute an alibi, is not sufficient to assert that a suspect's claim is false, it is consequently critical to substantiate all assertions with concrete evidence using other associated cybertrails, rather than simply concluding on the absence of evidence (Casey, 2011). A popular axiom in forensics science says, '*absence of evidence is not evidence of absence.*'

### 2.4.3 Determination of Intent

An exploratory forensic analysis can uncover some infrequent/unusual behaviour of a suspect which can be key for determining intent. For example, suspect's computer use can reveal plan or premeditation to commit crime at a particular moment in time. Analysis of internet browser searches have been very instrumental for this purpose. A popular case of William B. Guthrie[35], a presbyterian minister, who was convicted in 2000 for killing his wife, is a very good example of how internet searches might be suggestive of a criminal intent. William's wife's unconscious body was found drowned in a bathtub with autopsy showing the "presence of subtherapeutic amounts of two antianxiety agents, Diazepam and Lorazepam, and a sedative, Oxazepam". William's internet searches of words such as: "household accidents," "bathtub accidents," and "Temazepam" were presented as evidence, and he lost several appeals to exclude his internet searches as evidence in the court case. Consistent searching of the word "Child Pornography" might also be indicative of involvement (or an attempt to be) in such crimes.

Also, malicious activities such as backdating the digital clock on a computer system could be useful in the assumption of suspicious actions, the same way the possession of disk cleaning or encryption program can be used to demonstrate the plan to wipe or obfuscate incriminating evidence. However, exhibition of caution is recommended because some of these supposed activities may have unharmful explanations, therefore, conclusions in this case should only be based on strong assertions, rather than mere assumption of malicious intent.

### 2.4.4 Evaluation of Source

The embedded metadata of data object can provide a useful insight in evaluating the origin of a piece of evidence. According to (Casey, 2010), '*a piece of evidence may have been: 1) produced by the source; 2) a segment of the source; 3) altered by the source; 4) a point in space.*' Every data object has a traceable embedded characteristics which can be used to

[35] STATE of South Dakota, Plaintiff and Appellee, v. William Boyd GUTHRIE, defendant, and Appellant. Available online at: https://caselaw.findlaw.com/sd-supreme-court/1085831.html

identify the computer with which the data object was created. For instance, documents contain metadata such as, name of author, creation/modification date-time stamps, directory, printer names, etc., useful for tracing their sources. Likewise, the source of an incriminating image on the computer of a suspect can be traced to the digital camera (Kurosawa, Kuroki and Akiba, 2009; Alles, Geradts and Veenman, 2009) or scanner found in the crime scene. It is not unusual however, for such incriminating images to be downloaded or copied from another computer, that is why thorough investigation is needed. Information embedded by the image file such as, model, manufacturer, and date/time the photograph was taken, can be associated with the digital camera in possession of the suspect, or a flaw on the image file could be traced to the scratch on the screen of a flatbed scanner (Casey, 2010).

### 2.4.5 Digital Document Authentication

The chronological arrangement of the content of log files makes it possible to detect falsification by checking the inconsistences in timestamps of document's creation and modification. Digital stratigraphy (i.e., the arrangement of data on storage media) can provide analyst with the supporting evidence to demonstrate that a document has been altered to cover up some useful leads into a criminal investigation. For example, if a suspicious document purportedly created at an earlier date is found on top of a deleted document created later, suspicion of staging can be raised, as newer files should not be overwritten by older ones under stratigraphy (Casey, 2010). However, anti-forensics techniques such as disk optimization program or disk defragmentation, can reposition data on the storage media, thereby inhibiting document authentication process through stratigraphy. Like other intuitive evidence corroborative techniques, analyst should be able to demonstrate the correctness of their assertions for evidence to be admissible in court.

### 2.4.6 Recovering Deleted Data

One of the most critical aspects of forensic analysis is the recovery of data from storage media and the conversion of unreadable data to readable data. Often, investigators are faced with the task of retrieving purposefully deleted files, e-mails, images etc., which were destroyed to hide evidence. Unfortunately, when a file is deleted, the storage location is marked as "free" showing that the space is available for use, but the content of the storage location is only overwritten when another file is assigned to this location. This means that, files are hardly

entirely removed from a storage media (especially on windows system), its storage location only become part of an unallocated space.

According to Alexander[36], recovery of deleted files (on windows/NTFS) is achieved by looking up a file table for files that have not be overwritten. File location can be recovered if entries are still in place, likewise, the original file can be completely recovered, if the location have not been reused by a new file. Also, partial recovery of file may be possible if some, but not all the storage locations have been reused. However, data recovery will be impossible if all locations have been reused. In addition to looking up the file table, a method called "carving" is used for recovering deleted files by searching the unallocated space on storage media for header and footer values associated with different files.

Another significant indicator of data hiding during forensic investigation is when the total size of all visible partitions on a disk is smaller than the drive's capacity. This may indicate the existence of another hidden partition that has to be discovered. Similarly, as one of the techniques during forensic analysis is to filter extracted data and classify it based on relevance to the incident being investigated, or by logical categorization based on file types, a large unclassified or unknown file type may, therefore, be suggestive of the use of data obfuscation or encryption (Casey, 2010).

## 2.5 Evidence and its Admissibility in Legal Proceedings

In law discipline, evidence, which is crucial in both civil and criminal proceedings, is predominantly characterized by weight, relevance, admissibility, burden of proof, and sufficiency of any material that should be admitted into the record of a case proceedings. If evidence is presented to establish a fact, it is deemed material. Literally, the weight of an evidence is dependent on the degree of conviction of presented evidence on the triers of fact[37] — to either accept or reject a statement of fact. Evidence with strong weight can change the probability of the fact in issue. Evidence is said to have less weight if it is vague or indefinite

[36] See "Understanding Deleted Files and What They Mean" – Expert Witness Article, Trace Digital Forensics, LLC. Available online at: https://www.hgexperts.com/expert-witness-articles/understanding-deleted-files-and-what-they-mean-44950

[37] See "TRIER OF FACT." Legal Information Institute, Cornell Law School, cited on Jul. 27, 2021. Available online at: https://www.law.cornell.edu/wex/trier_of_fact

(Hewling, 2013). The "competence" (mostly considered as issues relating to weight) of evidence is measured based on its compliance with certain notion of reliability.

On the other hand, according to the FRE; an evidence is said to be relevant if it has the "*tendency to make the existence of any fact that is of consequence to the determination of the action more probable or less probable than it would be without the evidence*"[38], i.e., an evidence is relevant if it contributes to the determination of the fact in issue and is capable of assisting in the furtherance of an investigation to make the existence (or not) of a fact more probable. The relevance of evidence can be a necessary condition but, in some cases, may be insufficient condition for evidence to be admissible. For instance, Rule 403[39] of the FRE allows for the exclusion of relevant evidence "*if its probative value is substantially outweighed by the danger of*" any of: unjust prejudice; perplexing or misleading the jury; or unnecessarily prolonging the trial duration. Thus, this provision vests the trial court with considerable discretion over the determination of relevance. In Europe, however, where Member State have differing rules to determine the relevance of an evidence, with strict observance of Art. 6[40] of the European Convention on Human Rights (ECHR) and Art. 47[41] of the EU Charter of Fundamental Rights, requires that states must examine evidence that could call into question the overall fairness of the judicial proceedings. Certain legal jurisdictions in the EU grant the court the authority to (or not) disregard evidence based on a variety of considerations, including the gravity of the crime, the intention to commit crime (or not), and fairness, among others. While numerous others make relevance decisions non-discretionary, the adjudication of inadmissibility will be an automatic consequence of a violation of procedural rules (Garamvolgyi et al., 2021). Specifically in Italy, evidence is physically excluded from the court file to guarantee that the decision authority is not swayed by information that should have been obtained differently (Garamvolgyi et al., 2021).

[38] See "Rule 401 – Test for Relevant Evidence." U.S. Federal Rules of Evidence, cited on Jul. 27, 2021. Available online at: https://www.rulesofevidence.org/article-iv/rule-401/
[39] See "Rule 403 – Excluding Relevant Evidence for Prejudice, Confusion, Waste of Time, or Other Reasons." U.S Federal Rules of Evidence, LII, Cornell Law School, cited on Jul. 27, 2021. Available online at: https://www.law.cornell.edu/rules/fre/rule_403
[40] See ECHR Case Law, Council of Europe, Guide on Article 6 of the European Convention on Human Rights, updated on 30th April 2021. Available online at: https://www.echr.coe.int/documents/guide_art_6_criminal_eng.pdf
[41] See Article 47 – Right to an effective remedy and to a fair trial, EU Charter of Fundamental Rights. Available at: https://fra.europa.eu/en/eu-charter/article/47-right-effective-remedy-and-fair-trial

### 2.5.1 Admissibility of Evidence

Under certain common and statute law, before an evidence can be accepted in advancement of a court trial, it must pass the test of admissibility. The specificity of admissibility is dictated by law, and just like the notion of relevance, evidence is deemed admissible if the trier(s) of fact finds it useful to the resolution of the dispute to which it is a part, and relevant to the fact to be proven. In some jurisdictions, like the U.S., the Rule 402[42] (of the FRE), provides that all "relevant evidence is admissible" except with certain exceptions — some of which are related to constitutional exigency, or informally, on the broad basis upon which the concept of legal admission or exclusion is predicated (e.g., as provided in relevance of evidence). The same rule also provides that, all "irrelevant evidence is not admissible." However, the statement that all "relevant evidence is admissible" is rationally debatable – and this has been highlighted and argued by legal experts — as this would logically contradict the "exclusion of relevant evidence" provided in 403. With the rule of exclusion in mind, and several other rules such as the Rules of Civil[43] and Criminal[44] Procedure, Bankruptcy Rule[45], etc., it is particularly reasonable to intuitively assert that; 'not all relevant evidence is admissible.' This has since been held true and recognized by the U.S. congress committee on the Judiciary, which led to the amendment of the rule of evidence (and other rules where the reference appears) in 2011[46], to accommodate all other rules prescribed by the Supreme Court pursuant to statutory authority.

The European approach can be observed in two different folds, namely the "controlled systems" — with legal jurisdictions that strictly filters materials to be admitted at trial, and the "free proof systems" — where the judges are left to decide whether it is appropriate to dismiss or accept evidence that is obtained illegally. Likewise, the possibility to challenge the admissibility of a piece of evidence before a competent court, including the provision for rules of "nullity" (or validity) of evidence, varies amongst Member States (Spencer, 2010). The nullity or exclusion of evidence is mostly filed in cases where potential infringement on fundamental right is alleged. It is therefore not uncommon to see such request during legal

[42] See "Rule 402 – General Admissibility of Relevant Evidence." U.S Federal Rules of Evidence, LII, Cornell Law School, cited on Jul. 27, 2021. Available online at: https://www.law.cornell.edu/rules/fre/rule_403
[43] See Rule 30(b) and 32(a)(3) of the "Federal Rules of Civil Procedure" (2020). Available at: https://www.uscourts.gov/sites/default/files/federal_rules_of_civil_procedure_-_december_2020_0.pdf
[44] See Rule 15 of the "Federal Rules of Criminal Procedure" (2016). Available at: https://www.uscourts.gov/sites/default/files/rules-of-criminal-procedure.pdf
[45] See "Federal Rules of Bankruptcy Procedure" (2020). Available at: https://www.uscourts.gov/sites/default/files/federal_rules_of_bankruptcy_procedure_-_december_2020_0.pdf
[46] Available at: https://www.govinfo.gov/content/pkg/CPRT-112HPRT70817/html/CPRT-112HPRT70817.htm

proceedings in countries like Italy, Spain, and France where laws that protects the rights of suspects exist. Germany and the UK, however, use systemic integrity model where exclusion of evidence is granted only if important rights are violated, or when this exclusion would not significantly undermine the appropriate conviction for a serious crime (Garamvolgyi et al., 2021). There are also considerably noticeable differences among States on the applicability of the "fruit of the poisonous tree[47, 48]" doctrine (not recognized in the UK), which does not only exclude from trial evidence obtained illegally, but also any additional evidence derived via those illegal means (Garamvolgyi et al., 2021).

#### 2.5.1.1 Reliability

Reliability in evidence is used to denote something trustworthy — a material of value that can be relied on as accurate or truthful. In some situations, reliability is associated with the testimony of expert in a legal proceeding, which is expected to guarantee the accuracy and correctness of the scientific principles and methods used to arrive at a certain conclusion. Additionally, reliability is seen in the context of repeatability or reproducibility of the hypothetical assertions. Repeatability tends to answer the question like; will the same result be obtained using the same instrumentation or approach, if the same material is provided the second time? A method, testimony, or approach will be regarded as reliable if the same outcome is obtained when the same procedure is repeated multiple times. An expert witness can be unconsciously cross-examined multiple times just to ascertain the degree of confidence in the expert's hypothesis. Reproducibility on the other hand, is the measurement of accuracy of the instrumentation or approach used in drawing a certain conclusion, when introduced in a different situation — i.e., the same results should be achieved by different methods than those used initially (Brighi and Ferrazzano, 2021). This is commonly used to ascertain scientific hypothesis' correctness by testing with several different or uncorrelated situations.

#### 2.5.1.2 Authenticity

Authenticity of an evidence is the determination of its worth by confirming that: 1) the contents have remained unaltered or unchanged, and 2) that the content originated from the purported

[47] "Fruit of the Poisonous Tree." LII, Cornell Law School. Available online at: https://www.law.cornell.edu/wex/fruit_of_the_poisonous_tree

[48] See GAFGEN V. GERMANY. European Court of Human Rights no. 22978/05, § 25, ECHR. Available online at: https://hudoc.echr.coe.int/eng#{%22itemid%22:[%22001-99015%22]}

source. According to Rule 901[49] (of the FRE), authentication requires that evidence must be sufficient to support that the claim of the proponent is indeed what it purports to be. Authenticity is determined at two stages of a proceeding. The first being to determine the genuineness of the probative value so that its admission could assist the jury, and the second would be for the jury to carefully examine and determine the veracity of the evidence. Besides that, authenticity can be established through an expert witness attesting to the fact that a matter is what it claims to be, or through a non-expert opinion, for example, attesting to the genuineness of a material based on familiarity with it that was not acquired during the current litigation. Additional determinants of authenticity include evidence demonstrating the accuracy of a result by detailing the process or system that generated it, as well as conformance to any method permitted by statute or common law.

The satisfaction of authenticity requirement does not necessarily guarantee admissibility. Other rules of evidence such as those related to hearsay can exclude authenticated evidence from being admissible.

#### 2.5.1.3 Privilege

A privilege is a legal rule, under the law of evidence, that refers to a regulation that allows the right to non-disclosure information or evidence about a certain subject or to exclude such evidence from disclosure or use in a trial. Privilege is a right that can be exercised by an individual, businesses, spouse, government, etc. Notably however, is the solicitor-client privilege – also referred to as the attorney-client privilege[50] (in the U.S), or legal professional privilege[51] (in Australia and the EU) — that protects the confidentiality of communications between a client and their legal adviser to facilitate proper functioning of the justice system. Some legal jurisdictions favours public interest privilege which seeks to prevent the disclosure of information (usually of secrecy) that is of interest to government or against public interest. The EU court of justice also recognize the confidentiality of communication between lawyers and their clients, by ensuring that clients are free to consult their attorneys without fear that any confidences may be subsequently disclosed, and that commissions do not improperly use the

[49] See Rule 901 – Authenticating or Identifying Evidence. U.S Federal Rules of Evidence, LII, Cornell Law School, cited on Jul. 27, 2021. Available online at: https://www.law.cornell.edu/rules/fre/rule_901
[50] See Attorney-Client Privilege. Privileges, LII, Cornell Law School. Cited on Aug. 3, 2021. Available online at: https://www.law.cornell.edu/wex/attorney-client_privilege
[51] See Legal professional privilege. Information and Privacy Commission. Cited on Aug. 3, 2021. Available online at: https://www.ipc.nsw.gov.au/fact-sheet-legal-professional-privilege

content of a confidential document in investigation or trial. In many jurisdictions, privilege can also be seen in light of policies on trade secrets[52, 53] protection, which upholds the right to non-disclosure or secrecy of any element of intellectual property. This right has been exercised in several civil and criminal proceedings.

#### 2.5.1.4 Best Evidence Rule

The foundation of the Best Evidence Rule is rooted in the originality of a material issued as evidence in a trial. The type of this evidentiary materials include written documents, voice messages, photograph, recordings, etc. The rule holds an original document as superior evidence, and that secondary evidence (e.g., a copy of the original), will only be admissible if the original document is not obtainable or does not exist. In any case, the party presenting such document as evidence must provide a genuine excuse for its absence and prove that the content of the secondary evidence is the direct copy of the original. Over the years, the abolition of the 'best evidence' rules have been witnessed through common and statute declarations[54, 55, 56]. The move to abolish might not be unconnected to the need to review laws to accommodate provisions of digital evidence. There have been several arguments as to whether computer printouts — which is a copy of the original document – qualifies under the best rule of evidence. These arguments have been however settled by rule 1001(4)[57] (of the FRE) — which states that "*for an electronically stored information, 'original' means any printout — or other output readable by sight – if it accurately reflects the information.*" — and the provisions of the European Union guidelines on electronic evidence[58] which relates to reliability.

[52] See Trade Secrets. European Union. Cited on Aug. 3, 2021. Available online at: https://europa.eu/youreurope/business/running-business/intellectual-property/trade-secrets/index_en.htm
[53] See Trade Secret Policy. IP Policy, United States Patent and Trademark Office. Cited at Aug 3, 2021. Available online at: https://www.uspto.gov/ip-policy/trade-secret-policy
[54] See Section 51, Evidence Act 1995, Federal Register of Legislation (Australia). Cited on July 30, 2021. Available online at: https://www.legislation.gov.au/Details/C2018C00015/Html/Text#_Toc503517635
[55] See "Best evidence rule laid to rest" CMS Law-Now. Cited on July 30, 2021. Available online at: https://www.cms-lawnow.com/ealerts/2001/04/best-evidence-rule-laid-to-rest?cc_lang=en
[56] See also Masquerade Music Ltd. V Mr. Bruce Springsteen [2001] ECWA CIV 513. Cited on July 21, 2021. Available online at: https://www.casemine.com/judgement/uk/5a8ff71060d03e7f57ea7048
[57] See Rule 1001(D) – Definitions that Apply to this Article, Federal Rules of Evidence (2021 Ed.). Cited on June 25, 2021. Available online at: https://www.rulesofevidence.org/article-x/rule-1001/
[58] See "Guidelines of the Committee of Minister of the Council of Europe on Electronic Evidence in Civil and Administrative Proceedings." Cited on March 3, 2021. Available online at: https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=0900001680902e0c

#### 2.5.1.5 Hearsay

Generally, evidence rules favour testimony given under oath by witness(es) in the courtroom, where the witness' appearance and behavior may be evaluated, and remarks can be cross-examined. However, any "*testimony given by witnesses based on conversations held outside the courtroom are considered hearsay"* (Goodison, Robert and Brian, 2015). It is any 'statement'[59] made – out of court – by the 'declarant'[60], other than the testimony given at a trial to prove the truth of the matter asserted. According to rule 801[61] (of the FRE), hearsay means a statement that:

1) *the declarant does not make while testifying at the current trial or hearing; and*
2) *a party offers in evidence to prove the truth of the matter asserted in the statement.*

Generally, "hearsay is not admissible"[62] as evidence unless it is specifically allowed by exceptions provided in statutes, evidence rules, or other precedence rules. Some of the "exception rules"[63] upon which hearsay can be admitted include but not limited to declarant's unavailability; declarant's availability is irrelevant; record recollection; records of a regularly conducted activity; statements of facts contained in certificates (e.g., marriage certificate); reputations (as in personal or family history), etc.

## 2.6 Types of Evidence

Evidence is often classified according to the type of facts it tends to establish, its form, the role it plays in the case, and the applicable laws (Hewling, 2013). Basically, evidence is divided into two categories which can either be physical; which involves any tangible object, real or material evidence relevant to the case, or testimonial; that relates to statement made under oath by a competent witness. Notwithstanding, several types of evidence can be derived from these two categories. The types and classes depend on legal jurisdictions and the forms of evidence that the local laws permit. We describe some types below:

---

[59] "Statement means a person's oral assertion, written assertion, or nonverbal conduct, if the person intended it as an assertion." – Rule 801 (Federal Rules of Evidence).
[60] "Declarant means the person who made the statement" – Rule 801 (Federal Rules of Evidence)
[61] See Rule 801 – Definitions that Apply to this Article; Exclusions from Hearsay. LII, Cornell Law School. Cited on May 23, 2021. Available online at: https://www.law.cornell.edu/rules/fre/rule_801
[62] See Rule 802 – The Rule Against Hearsay. LII, Cornell Law School. Cited on Mar 23, 2021. Available online at: https://www.law.cornell.edu/rules/fre/rule_802
[63] See Rule 803 – Exceptions to the Rule Against Hearsay. LII, Cornell Law School. Cited on Mar 23, 2021. Available online at: https://www.law.cornell.edu/rules/fre/rule_802

*A. Direct Evidence*

This is regarded as the most powerful evidence. It requires no inference – not assumed, and it is alone a proof. Direct evidence is the evidence of a witness testifying to the truth of an assertion (of guilt or innocence) directly. The most perfect example of direct evidence is an eyewitness testifying to seeing a criminal offence take place.

*B. Circumstantial evidence*

Contrary to direct evidence, the circumstantial evidence is not drawn from direct observation of fact, rather, it is deduced from other events or circumstances from which the occurrence of the matter can be reasonably inferred. It involves making probabilistic and statistical deductions based on suggestions rather than personal knowledge or observation. "The more circumstantial evidence there is, the greater weight it carries.[64]"

*C. Inculpatory and exculpatory evidence*

The inculpatory and exculpatory evidence differs only in the way they favour the defendant or the prosecution. Inculpatory evidence is favourable to the prosecution because it establishes (or seeks to establish) the defendant's involvement in an act. On the other hand, exculpatory evidence absolves (or tries to acquit) the defendant of guilt or culpability. It is used to establish innocence.

*D. Hearsay evidence*

The hearsay is an oral or written statement made by someone out of court (or a particular trial) which is tendered as evidence for the assertion of truth. Always not taken under oath. Hearsay is a complex area of law of evidence because, generally, it is not admissible, but the principle has been subject to numerous exceptions. According to some common law jurisdiction (e.g., the U.K.'s Criminal Justice Act 2003[65]), if all parties to the proceedings agree on the admissibility of hearsay evidence or the court decides that it is in the interests of justice to be

[64] "Types of Evidence." Available online at: https://www.casdschools.org/site/handlers/filedownload.ashx?moduleinstanceid=7201&dataid=6177&FileName=02-TypesOfEvidence.pdf

[65] "Admissibility of hearsay evidence." Section 114 of the Criminal Justice Act 2003. Available online at: https://www.legislation.gov.uk/ukpga/2003/44/contents

admissible. The hearsay is usually about what one was told by someone who witnessed an act, so, it is a second-hand evidence.

*E. Documentary evidence*

Documentary evidence is introduced through documents — mostly considered to be written forms of proof, such as diary entry, contracts, letters, or wills etc., and it is offered to support a fact. Documentary evidence can also include digital media, such as video or audio recordings, and images. For documentary evidence to be admissible, it is essential to therefore establish that the document is authentic and from reliable source.

*F. Expert evidence*

This type of testimony involves an expert witness testifying on a matter based on formal expertise and/or experience in a particular field. It is commonly used in reference to the scientific analogy of a subject that is beyond the competence of the trier of facts. An expert testifier is assumed to possess the necessary qualifications (and, in certain cases, license) and expertise in the field in which they are to testify.

*G. Prima facie evidence*

This is the evidence presented as "a first appearance" in a court proceeding to prove a fact and it is held as sufficient until it is successfully rebutted or disproved. It is also called "presumptive evidence."

## 2.7 Legality of Digital Evidence

Digital evidence, also referred to as electronic evidence, is data or information that exist in electronic format, useful to prove or reveal the truth about a crime in a court of law. The major difference between digital evidence and other scientific forms of evidence is that digital evidence data exists in digital format of zeros and ones. This means that digital evidence data may be unintelligible at the initial point of collection and would require specialized tools and protocols to make it readable. As a result of this significant distinction, the collection, analysis, interpretation, and presentation of digital evidence in civil or criminal proceedings is unique and challenging.

It is worth noting that digital evidence can be relevant and useful in the prosecution of all types of crimes — including non-electronic crimes. Digital evidence is information that is leveraged by condensing objects, events, and time into a unified dimensional space to establish causation for criminal incidents (Novak, Grier and Gonzalez, 2018). For instance, files on suspect's electronic device could reveal critical evidence of intent, relationship, location, and timing of crime. This is particularly true in the story of a Wichita police officer in 2005, who used a floppy disk drive to uncover the BTK serial killer. Before the discovery, the serial killer had claimed ten (10) lives and had escaped capture since 1974 (Wenzl, 2014).

Formally, here are some of the definitions that have been put forward to describe digital evidence. (Casey, 2004) defines "*digital evidence or electronic evidence as any probative information stored or transmitted in digital form that a party to a court case may use at trial*." In (Novak, Grier and Gonzalez, 2018), "*digital evidence is information stored or transmitted in binary form that may be relied on in court.*" Griffin described digital evidence as data or information stored in a digital format that is sufficiently reliable to be used in a court trial to establish or reveal the truth about a crime (Griffin, 2018). According to the National Institute of Justice (US), Digital evidence is "*information and data of value to an investigation that is stored on, received, or transmitted by an electronic device*" (NIJ, 2008).

There are few unique points that are peculiar in these definitions that can be elaborated further. Firstly, the probative information or data could be e-mails, databases, transaction logs, digital multimedia files, internet browsers, printouts, Global Positioning System (GPS) track logs, instant messages, system log files, etc. (Casey, 2010). The list is inexhaustive, however, what can be regarded as a digital data are categorized into to two groups, namely: 1) data stored on computers or other electronic devices; and 2) data transmitted through electronic means over communication networks. This also corroborates the use of binary (as computer system stores data and perform calculations with 0s and 1s) and digital formats in the definitions. More data types will fall within this category as we continue to witness technological developments.

Secondly, the transmission or reception of information highlights the potential involvement of two or more parties – this can be the sender and/or the receiver. For example, in cases involving child pornography (Taylor and Quayle, 2003; Adler, 2001; Webb, Craissati and Keen, 2007) or cyber-terrorism (Gordon and Ford, 2002), multiple syndicates may be involved, resulting in extended bottlenecks during the investigation, particularly if other parties are located outside the territory where the crime was committed or investigated (Garamvolgyi et al., 2021). In

other circumstances, it may be the act of a single individual, as with identity theft (Hoar, 2001), e-fraud (Graycar and Russell, 2002), cyber-stalking (Ogilvie, 2000), and malware transmission, among others. Also, the unlawful interception of data in transmission/communications are often regarded (except if done in protection of national security, even though that could be invasive of privacy (Ryan and Shpantzer, 2010)), as an electronic crime. Interception or wiretapping (Westin, 1952) has been legally controlled by several national statutes, e.g., wiretap Act[66], Electronic Communication Privacy Act (ECPA)[67], the Pen/Tap statutes[68], and numerous court cases[69] either authorizing or restraining the act. The essential premise is that there is always someone or some group of individuals committing (potential) crime against another group of people or an entity (through transmission or interception of something).

Furthermore, whatever form this probative information takes — whether as raw data or analytically processed — it must be reliable (Kenneally, 2001), admissible (Goode, 2001; McKemmich, 2008; Ryan and Shpantzer, 2010), authentic (Lynch, 2000; Grimm, Capra and Joseph, 2017), and relevant (Grimm, Capra and Joseph, 2017) to the court case. The admissibility and relevance of digital evidence are not sharply different from the general principles of evidence which have been described in section 2.5.1, however, some courts have treated digital evidence differently in aspects that relates to authenticity, hearsay, privilege, and the best evidence rule. Most legal jurisdictions have specifically promulgated laws to guide the administration of digital evidence in civil and criminal proceedings,[70, 71, 21] some, e.g., the U.S., have modified rules of evidence[72] to accommodate the preservation and disclosure requirements for electronically stored evidence, while others have applied extant traditional rules of evidence in cases where laws are silent or non-existent – as is the case in developing countries with insufficient resources and capacity to properly investigate electronic crimes.

---

[66] The Wire Tap Act, also known as Title III, is 18 U. S. C. §§ 251022

[67] The Pen Registers and Trap and Trace Devices statute is 18 U. S. C. §§ 312127

[68] Electronic Communications Privacy Act is 100 STAT. 1848, PUBLIC LAW 99508, which inter alia amended the Wiretap Act and added 18 U. S. C. §§ 270110 dealing with stored communications.

[69] See Katz v. United States, 389 U.S. 347; 88 S. Ct. 507; 19 L. Ed. 2d 576; 1967 U.S. LEXIS 2 (1967).

[70] "Electronic Evidence in Civil and Administrative Proceedings." Guidelines adopted by the committee of Ministers of the Council of Europe (2019). Available online at: https://rm.coe.int/guidelines-on-electronic-evidence-and-explanatory-memorandum/1680968ab5

[71] "Electronic evidence in Criminal Matters." European Parliament Think Tank. Available online at: https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/690522/EPRS_BRI(2021)690522_EN.pdf

[72] "Federal Rules of Evidence" Available online at: https://www.uscourts.gov/sites/default/files/federal_rules_of_evidence_-_dec_1_2019_0.pdf

Reliability in the context of digital evidence extensively include the collector of the evidence — who must be recognised by the court – as well as the processes and procedures adopted in the collection of the evidence. According to the FRE rule 702[73], an expert witness may testify if he or she possesses reliable scientific, technical, or other specialized knowledge and is capable of forming an opinion regarding the principles and scientific methods used to interpret the evidence or determine the fact in issue. What is admissible may be dependent on what the court admits as relevant to the case. However, before the relevance or materiality of a digital evidence can be determined, it must survive the threshold test posed by Frye standard[74], or later, the Daubert standard[75] (Ryan and Shpantzer, 2010). The acceptance of scientific evidence was initially governed by a heuristic known as the "general acceptance" or Frye standard, which was based on a well-known 1923 Supreme Court of the District of Columbia decision (James Alphonso Frye vs United States)[76], to determine validity. The rule states that:

> *"While courts will go a long way in admitting expert testimony deduced from a well-recognized scientific principle or discovery, the thing from which the deduction is made must be sufficiently established to have gained general acceptance in the particular field in which it belongs."*

The Frye case established a literal precedent for the admissibility of evidence only if the scientific approach upon which it is based is widely accepted by the scientific community (Goodison, Robert and Brian, 2015). It also brought to the fore the power of the courts to decide what should (or not) be accepted in a legal proceeding.

More recently, precisely since 1993, the Daubert standard — which came into being as a result of a court ruling in Daubert v. Merrell Dow Pharmaceutical, inc.[77] – has been the standard adopted by most Federal and some state courts in the U.S., and it has also been the general benchmark of 'Good Scientific' process requirements around the world. The factors necessary

---

[73] See Rule 702. Testimony by expert witnesses, Federal Rules of Evidence. LII, Cornell Law School. Cited on March 5, 2021. Available online at: https://www.law.cornell.edu/rules/fre/rule_702

[74] See "Frye Standard." LII, Cornell Law School. Cited on March 5, 2021. Available online at: https://www.law.cornell.edu/wex/frye_standard

[75] See "Daubert Standard." LII, Cornell Law School. Cited on March 5, 2021. Available online at: https://www.law.cornell.edu/wex/daubert_standard

[76] Frye V. United States 293 F.1013 (D.C. Cir. 1923) Available online at: https://www.mass.gov/doc/frye-v-united-states-293-f-1013-dc-cir-1923

[77] Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993). Available at: https://supreme.justia.com/cases/federal/us/509/579/

for the consideration of the validity of scientific methods under the Daubert standard are whether:

(1) the theory or technique has been empirically validated (or tested);
(2) it has undergone peer review and publication;
(3) it possesses any known or hypothetical error rate;
(4) they are subject to set standards governing their applications; and
(5) the methodology has been widely accepted by a relevant scientific community

However, evidence and testimony that does not follow these criteria may still be accepted — as the Daubert requirements are not exhaustively or entirely conclusive (Arshad, Aman and Abiodun, 2018). In an attempt to systematically codify and organize the elements stated in the Daubert standard, the Rule 702 (of the FRE) has been amended twice — in 2000, and then in 2011 — thereby extending the rules on the testimony of expert witnesses. The rule 702 as amended now reads:

> *"a witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if:*
>
> *(a) the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue;*
> *(b) the testimony is based on sufficient facts or data;*
> *(c) the testimony is the product of reliable principles and methods; and*
> *(d) the expert has reliably applied the principles and methods to the fact of the case."*

With this rule, a process known as the "Daubert Hearing", or "preliminary question" may occur before the main trial begins. The Daubert hearing — usually done out of the jury's presence – is mostly required to help the judge evaluate the validity of an expert's testimony or evidence, and to decide whether or not it is admissible. During the hearing, the parties in the trial are allowed to present the scientific methodology behind their hypothesis or evidence, and the admission of such evidence must be based on the satisfaction of all the questions raised in the rule above.

The Rule 104[78] (of the FRE), even though deceptively straightforward about preliminary questions, is essential to understanding the legality of the preliminary procedure with respect to admissibility of digital evidence. The rule 104(a) states: *"…the court must decide any preliminary question about whether a witness is qualified, a privilege exists, or evidence is admissible. In so deciding, the court is not bound by evidence rules, except those on privilege*. " This rule then posits that, the decisions about relevance, qualification of expert witness and the possible allowance (or otherwise) of the testimony the expert gives, application of best evidence rule, and the general admissibility of evidence, are solely made by the judge. This rule is applicable to physical evidence, however, in the case of digital evidence, there is a greater likelihood that the judge may sought the assistance of experts and the jury regarding admissibility (Grimm, Capra and Joseph, 2017). From the European perspective, and relating it to seeking expert's assistance, the preliminary procedure varies amongst Member States' legal systems — depending on whether the judicial system of a particular State favours the accusatorial or inquisitorial tradition. In accusatorial tradition (also known as adversarial tradition) — usually oral — the requirement to advance the investigative procedure is provided by the parties in the litigation, in which both adduce evidence in support of their positions (Champod and Vuille, 2011). The versions of facts from the prosecution and the defence are presented to the jury who determines the accuracy of both versions. Also, during this stage, the cross-examination of experts and witnesses by the parties occur, and in compliance with set rules, the decision on which evidence to admit and those to exclude from the proceeding is made (Champod and Vuille, 2011). The judge's role (mostly passive) in this legal system is to uphold the principles of fairness and equality until the final verdict is issued. This tradition is mostly related to common law jurisdiction (such as England, Wales, and U.S.) where previous judgement made by higher courts serve as precedence (and therefore binding) for lower courts. In contrast, the inquisitorial tradition, commonly found in civil law countries such as Italy and France, is aimed at getting the fact in issue through extensive investigation and examination of all evidence. The entire trial in this case is conducted by the court and the role (mainly active) of the judge here — in a quest to establish the truth — is to seek incriminating and exonerating evidence, collect substantive evidence, cross-examine witnesses (in whatever order in which they are to be heard), and appoint experts if necessary (Champod and Vuille, 2011). Unlike the accusatorial legal system, all evidence are admissible *a priori*, and the judge is free to decide

[78] See Rule 104. Preliminary Questions, Federal Rules of Evidence. LII, Cornell Law School. Cited on June 5, 2021. Available online at: https://www.law.cornell.edu/rules/fre/rule_104

all principles of admissibility by applying statues, without taking precedence from previous judgements.

There are, of course, observed disadvantages on both legal systems, most notably, on the accusatorial system is that the tradition of interviewing and cross-examining expert witnesses inhibits their ability to express their results or knowledge freely. Hence, the process is distorted in one direction or another (Champod and Vuille, 2011) to favour the expert witness with the best communication skills or the best ability to convince the court (Spencer, 1992) — and not especially for their scientific expertise. On the inquisitorial tradition, it is possible that the best qualified experts work with the prosecuting authorities, that in itself can raise critical questions about neutrality, and seriously deprive the defence of valuable legal resources (Champod and Vuille, 2011).

Regardless of the culture of legal systems adopted by different States in the European Union, the Article 6, and of course all the provisions, of The European Convention on Human Rights[79] is binding on all members. The question therefore is whether these two systems and the associated procedures through which they take place, guarantees a fair trial to all parties involved. Nevertheless, the principle of "Equality of arms" provided in article 6 of the convention seeks to ensure that all parties to a litigation have balanced opportunities to present their cases, as well as the rights to equal access to information and resources. Strict compliance to the fundamentals of human rights have somewhat made admissibility of digital evidence in European courts a bit challenging, and State's judicial systems have had to constantly grapple with how to balance the intricacies of digital evidence, the EU rights provisions, and extant laws on evidence.

In summary, the principle of reliability of electronic evidence requires that nothing about the collection and handling of the evidence should make its authenticity or veracity doubtful[80].

As earlier mentioned, that hearsay is mostly not admissible as evidence, except with some tightly controlled conditions. Some of these conditions relate to digital evidence where the hearsay rules permit the admission of evidence if the source of the digital records is reliable and acceptable. For instance, statements made by defendants preserved in e-mails, text

[79] "European Convention on Human Rights." Cited on Aug. 3rd, 2021. Available online at: https://www.echr.coe.int/documents/convention_eng.pdf

[80] See "Electronic Evidence Guide: A Basic Guide for Police Officers, Prosecutors and Judges." Ver. 2.0. Council of Europe. Cited on June 21, 2021. Available online at: https://au.int/sites/default/files/newsevents/workingdocuments/34122-wd-annex_4_-_electronic_evidence_guide_2.0_final-complete.pdf

messages, or other digital media (Goodison, Robert and Brian, 2015). Digital records such as e-mails are permissible as evidence so long as it can be proved as authentic, and its integrity can be asserted. The rules 803(6) and (7) which gives conditions for the '*Records of a Regularly Conducted Activity*'[81] and '*Absence of a Record of a Regularly Conducted Activity*'[82] provides for explorable exceptions to the rule of hearsay, especially to corroborate assertions, that indeed – based on records – some actions were a frequent conduct (or characteristics) of the defendant, and therefore, the fact in issue (an act, event, opinion, diagnosis, knowledge, etc.) is attributable. For this corroboration to happen, something more than just a record of events/actions must be adduced. For example, in the case of an e-mail, the computer involved must be something only the defendant use; in his home or personal office, password protected, and only him know the password.

Authentication of digital evidence requires the proponent to provide sufficient facts to support the adduced evidence. To establish authenticity, the presented evidence must be indisputable and representative of the original form. Presumably, most digital evidence stands authenticated, so long as the proponent can (and able to) sufficiently pull in all the necessary resources (Bellin and Ferguson, 2014). Physical evidence is no different from digital evidence, therefore, the rules associated with former is also applicable to latter. However, cautions are required in handling and administering digital evidence, because of the ease with which it can be inadvertently or deliberately altered. Nowadays, courts are largely concerned about reliability of digital evidence rather than authenticity (Ryan and Shpantzer, 2010), mostly because, over the years, set of guidelines (also constantly updated) have been established on chain of custody. The "*ACPO principles*" is one of such guidelines for the authentication and integrity of evidence. Modern commercial forensic software solutions are also designed to preserve the original form of evidence to guarantee its authenticity and integrity. According to the Council of Europe's guidelines on 'electronic evidence in civil and administrative proceedings', and to support the presumption that 'most digital evidence stands authenticated', the section on 'reliability' provides that *"...electronic data should be accepted as evidence unless the authenticity of such data is challenged by one of the parties.*"[83] The section also provides that

[81] See Rule 803(6) – Records of a Regularly Conducted Activity, Federal Rules of Evidence. LII, Cornell Law School. Cited on June 5, 2021. Available online at: https://www.law.cornell.edu/rules/fre/rule_803
[82] See Rule 803(7) – Absence of a Record of a Regularly Conducted Activity, Federal Rules of Evidence. LII, Cornell Law School. Cited on June 5, 2021. Available online at: https://www.law.cornell.edu/rules/fre/rule_803
[83] See Reliability (21), Guidelines of the Committee of Ministers of the Council of Europe on Electronic Evidence in Civil and Administrative Proceedings. Cited on June 22, 2021. Available online at: https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=0900001680902e0c

*"…the reliability of electronic data should be presumed, provided that the identity of the signatory can be validated and the integrity of the data secured, unless and until there are reasonable doubts to the contrary*."[84] "*We can therefore deduce from the foregoing that, in most cases that involves digital evidence, the burden to counter the presumption of authenticity and reliability of the presented evidence is on the defendant."*

Other key important criteria to consider in the evaluation of digital evidence for admission into a trial are:

**Completeness**: That the whole story about the arrival at a certain hypothetical fact or opinion based on the evidence should be told without the intention to favour a particular perspective.

**Believability**: That the evidence must be sufficiently representative of the proof of facts and the trier of fact, and the court in general, should find it, clear, understandable, trustworthy, and reliable.

**Proportionality**: That the evidence collection methodology must not be intrusive – rather must be fair and not prejudicial. "*Not only should we collect evidence that can prove a suspect's malicious actions, but also evidence that could prove their innocence (Exculpatory evidence).*" (Krishnan and Shashidhar, 2021)

The admissibility of digital evidence is fairly precise and structured in United States law, however, the European legal systems seems to be ambiguous on the subject, as most issues about admissibility (or reliability) is linked to the manner in which the evidence is accessed – i.e., a challenge to the scientific reliability of an evidence will most likely diminish its probative value or totally nullify its admissibility.

Basically, for digital evidence to be admissible in a trial, it must conform with the series of laws and rules that ensure its acceptability in court. The preservation of digital evidence that will be admissible is inconclusively hinged on the following basic principles:

1. Investigators actions should not in any way alter or modify the original evidence.
2. Access to original digital evidence should be done by only competent persons.

[84] See Reliability (22), Guidelines of the Committee of Ministers of the Council of Europe on Electronic Evidence in Civil and Administrative Proceedings. Cited on June 22, 2021. Available online at: https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=0900001680902e0c

3. All investigation procedures, from identification to presentation, should be thoroughly documented, with processes and methodology repeatable by an independent third-party.
4. The individual or organization in charge of evidence custody is ultimately responsible for ensuring that all applicable rules are followed.

## 2.8 Challenges with Digital Evidence

Identifying and extracting digital evidence is not an easy task, especially while having to ensure that the evidence (or the hypothetical facts to support a certain claim) is admissible in a legal proceeding. Generally, the challenges digital forensics is faced with are the same with its resultant probative value. However, digital evidence have some peculiar challenges which can be observed from different perspectives. We describe below some of the most notable challenges identified.

### 2.8.1 Distributed Complexity

Identification and forensic analysis of evidence on a single digital source could, sometimes, be a complex task, however, these complexities have been exacerbated by the advent of distributed systems, in which data and resources are scattered among different physical or virtual hosts (Caviglione et al., 2017). The distribution of data across several platforms will introduce additional layer of complexity to evidence data analysis. Automated data analytics techniques have been suggested as potential solution – to handle huge data – however, that has been met with serious criticism from the digital investigation community, sighting probable deterioration of evidence quality (Caviglione et al., 2017).

### 2.8.2 Issues with Privacy

One of the most challenging aspects of digital evidence investigation is privacy. This has been exacerbated by the enforcement of GDPR in Europe and how it affects its citizens around the world. Many jurisdictions are taking the rights to privacy of its people seriously, and in many cases, it has been the stumbling block in the proper investigation, or the successful conclusion of criminal trials. Furthermore, it is a usual practice to reconstruct events in a criminal case such as cyberterrorism, and more often than not, this process may involve analysing social connections of suspects and other potential individuals or groups. Users' privacy can be

potentially violated during this process — which may totally jeopardize the investigation and render the evidence inadmissible.

### 2.8.3 Anti-forensics Techniques

Cybercriminals always try to hide footprints of their activities, and in so doing, they obfuscate, cloak, or encrypt data, to either hide incriminating evidence or make traces difficult. The intension is to mitigate the effectiveness of forensic investigation (Liu, 2016). According to Rogers (2005), anti-forensics is the "*attempts to negatively affect the existence, amount and/or quality of evidence from a crime scene or make the analysis or examination of evidence difficult or impossible to conduct*." Anti-forensics in form of legitimate solution to security/protection of privacy such as encryption is one of the most challenging problems in digital evidence investigation. Additionally, tools and techniques which makes anti-forensic possible are becoming easily accessible and available to malicious criminals.

### 2.8.4 Generalized Standard

There are several dynamics to digital evidence investigations challenges — from cross-border evidence information exchange; to formal knowledge-based representation; to uniform resource database of case scenarios and investigative solutions; and general methodological standard format. The lack of all these afore-mentioned, and many more, have been identified within the digital forensic community as a critical challenge to the development of a robust systems for digital evidence investigation. However, there have been several steps taken to solve issues of standardization and enhance information exchange across diverse legal jurisdiction, but productive success is yet to be achieved.

### 2.8.5 Verification of Error Rates

A recent digital forensic lawsuit study had identified 10 out of a random 100 cases as having issues that relates to errors in data collection and evidence analysis (Cole et. al, 2015). The fault being incorrect output and inaccurate timestamp in the tool used for analysis (Arshad, Aman and Abiodun, 2018). The lack of proper tools to measure the frequency of error(s), and the accuracy and reliability of the approach used in arriving at a certain conclusion during criminal investigation is a significant challenge. Although, the Daubert principle requires that the error rates of the tools, methods/techniques used in the analysis of digital evidence must be well-established before such evidence is admissible in a trial, however, there are instances

where results from these tools or methods have been erroneous and have had devastating impact on the outcome of the trial.

## 2.9 Chapter Summary

In this chapter, we explained the meaning Digital Forensics with detailed description of its branches, process models, as well as the significant use of forensics analysis. Furthermore, understanding that digital forensic investigation, and the consequent analysis of artifacts is necessitated by the need to mine evidence; which can be used for inculpatory or exculpatory defence in the court of law, we discussed extensively, the legality of traditional evidence and its digital components. We juxtaposed different jurisdictional provisions relating to digital evidence and the numerous challenges to balance legitimate investigation and the complexities of rights violation. Lastly, we highlighted the types of evidence and various sources of digital evidence.

Chapter 3

# Artificial Intelligence and Digital Evidence Mining

In chapter 1, we briefly discussed the concept of AI and its corresponding subfields. However, there is a notable ambiguity around AI and its several nuances which tends to create conversational confusion. In this chapter, we give a detailed description aimed at demystifying the conceptual misunderstanding around AI, ML, DL, and NN. Also, in extension to the in-depth explanation of digital evidence given in Chapter 2, particularly on the aspect of its extraction, analysis, and presentation. Here, we bring into perspective, the idea of the divergence of cognitive computing into digital evidence analysis. This chapter (and thereafter, subsequent ones) is motivated by the current advancement in research and development of AI-powered methods used in big data mining, which seek to find meaningful and explorable patterns in data. Digital artifacts are digital data, mostly voluminous, complex and heterogenous. It is therefore intuitive to presume that the same cognitive approach used in data mining will succeed if adapted to the analysis of digital artifact. We introduced the necessary background of AI — its components and relevance — that aligns with the processes involved in digital forensics. Also in this chapter, we reviewed several literatures that have extended the concept of AI techniques in the DF analysis, including the methods or frameworks, proposed to demonstrate the practicality and benefit of these approaches. The descriptions in this chapter will lay the necessary foundational background to understanding the concepts and components of our experiments, as well as the results obtained in subsequent chapters. It is also worth noting at this point that the thesis' discussion of AI/AI-powered systems in DF excludes techniques for robotics, which, as far as we know, have no direct impact on DF.

## 3.1 Artificial Intelligence

Finding the right definition for AI is not a simple task as there is no clear definition. Nonetheless, there are several explanations formulated to put the notion of AI in a general perspective. Literally, AI is the simulation of intelligence in machines to learn and mimic human behaviour. Some definitions have also tried to extend this idea of simulation to include the ability to mimic human “thought process”. This has however been rejected by researchers as too high-level, and should instead, be considered as human “rational behaviour” (Russell

and Norvig, 2009). The rationality herein means logical reasoning that involves acting (mostly influenced by the environment) in a certain way to attain optimality within a set of predefined goal(s). For machines (or computers thereof) to process this logical reasoning, it must be formally represented. This representation in AI is what is referred to as "Knowledge Representation" (KR) (Davis, Shrobe and Szolovits, 1993). KR contributes to the pragmatic efficiency of the computational process of reasoning by organizing information in such a way that makes inferencing easy. A formal way to structure the representation of this reasoning process is what is currently known as "Ontology" (Peter, 1987; Guarino, Oberle and Staab, 2009; Barry, 2012). Ontology is domain specific, in that, it formalizes the properties (classes, attributes, and relationships) of a particular subject area and how they relate. Literally, the process of designing an ontology involves modelling the properties pertaining to a domain as facts; how these facts are processed into rules and techniques that defines how the system behaves in a certain situation; or how the processed knowledge is applied (meta-knowledge) (Faye, 2010). From the individual set of facts, the ontology data model creates a knowledge graph[85] — which is a collection of entities, expressed as nodes and edges, where the entities represents the nodes, and the edges form the type of relationship between the entities. Fig. 3.1 shows a conceptual model of a simple animal kingdom ontological design.

Ontology design and modelling has, over the years, become widespread and advanced to a high-level state, and now, they are able to reason across multiple subject areas. Using RDF/XML[86], it is possible to create ontologies for multiple domains that are sharable amongst application and systems. Also, ontologies can be expressed with the Web Ontology Language (OWL). "*OWL is a semantic web computational logic-based language, designed to represent rich and complex knowledge about things and the relationship between them*.[87]"

In another interesting proposal, (Kaplan and Haenlein, 2019) defines AI as the "*system's ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation.*" This definition provides the right basis to discuss the categories of AI techniques, viz. the symbolic and sub-symbolic AI.

---

[85] "The knowledge graph represents the collection of interlinked description of entities – object, events or concepts." Cited on 25th July, 2021. Available online at: https://www.ontotext.com/knowledgehub/fundamentals/what-is-a-knowledge-graph/

[86] https://www.w3.org/TR/rdf-syntax-grammar/

[87] https://www.ontotext.com/knowledgehub/fundamentals/what-are-ontologies/

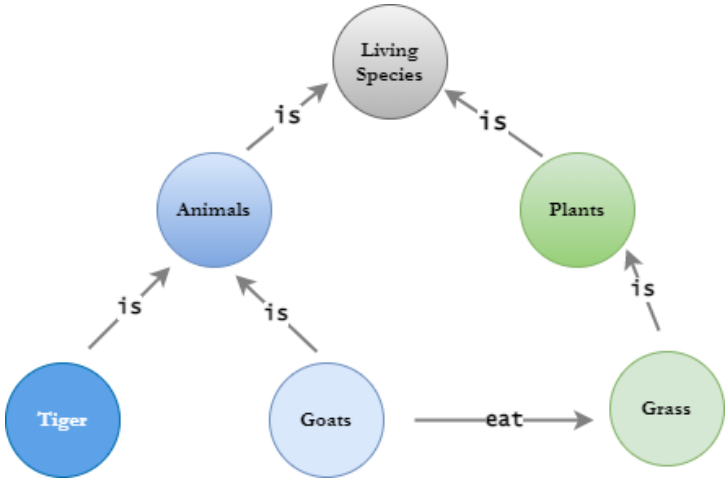


**FIGURE 3.1**: Conceptual model of a simple ontological design.

### 3.1.1 Symbolic AI

Symbolic AI, according to John Haugeland, is the "Good Old-fashioned Artificial Intelligence" (GOFAI) (Haugeland, 1989) that reasons based on first-order (predicate) mathematical logic, rules, and semantics. Symbolic reasoning are deductive, in that, conclusions are established based on set of logical inference rule, that is premised on certain consequence. The automation of the reasoning process involves using some sets of procedural axioms, defined declaratively to produce theorems. Symbolic systems are built on knowledge bases that consist of discrete entities through which logical reasoning can be inferred (Faye, 2010). A formal and popular example of a symbolic systems are the rule-based engines such as expert systems, or knowledge graphs which was described earlier.

#### 3.1.1.1 Expert Systems

Expert systems are predominantly stack of nested if-then statements used in drawing conclusions about entities and their relationships (this is an oversimplification of the meaning, however). They are designed to solve complex problems by reasoning through a knowledge base in a manner that mimics human experts' decision-making capacity (Peter, 1998). Expert systems have two (2) major subsystems, namely: 1) inference engine, and 2) knowledge base. The knowledge base contains facts and rules, and the inference engine applies these rules to existing facts to derive new facts. Consequently, at any point in time with expert system, explanations can be sought as to the reasoning behind a certain conclusion. This provides for an efficient debugging ability.

The drawbacks with expert systems, and generally all symbolic reasoning system is that, since all rules have to be explicitly stated, any non-existing rules will not be considered in the formulation of new facts. This is a serious disadvantage to the idea of autonomous logical reasoning. Expert systems is also monotonic; that is, it is one directional. This means the more rules added, the more knowledge is encoded. However, additional rules can not erase old knowledge. This could result in conflicting or error-prone decision-making system. Also, expert systems do not function well when fed with large quantities of data (Faye, 2010). This totally disqualifies them as a candidate of choice in areas of big data mining.

#### 3.1.1.2 Case-Based Reasoning

Case-Based Reasoning (CBR) was proposed in an attempt to solve the problems associated with the rule-based systems, such as expert systems. CBR's idea is deeply rooted in the concept of solving problems by adapting successful previous solutions to a similar problem. Indeed, it is a pervasive human behavioural reasoning to attempt to solve a problem based on past cases personally experienced. Basically, experts maintains a vast collection of case histories, drawn from several problem-solving techniques, inferences, and solutions. When a target problem is to be solved, experts use a metric to measure how close a similar problem in the case base matches the new problem. If a perfect match is found, then the solution to the previous case is implemented on the new one. However, if a perfect match is not found, the system tries to adapt to any solution in the case base that is closest (in measure) to the new situation. The CBR process has been formalized and can be summarized in a four-step scheme. Fig. 3.2 shows the diagrammatic representation of CBR design cycle.

According to (Aamodt and Plaza, 1994) the working cycle of a CBR comprises of:

1. **Retrieve**: for a given problem, retrieve all cases similar to it. A case is composed of a problem, its related solution, and annotations regarding the derivation of the solution.
2. **Reuse**: connect the solution(s) from the previous problem to the target situation — and adapt as necessary to fit the target situation.
3. **Revise**: after mapping the previous solution to the target problem, test the new solution and revise if necessary.
4. **Retain**: once the solution has been adapted to the target problem, store the derivation procedure in memory as a new case.

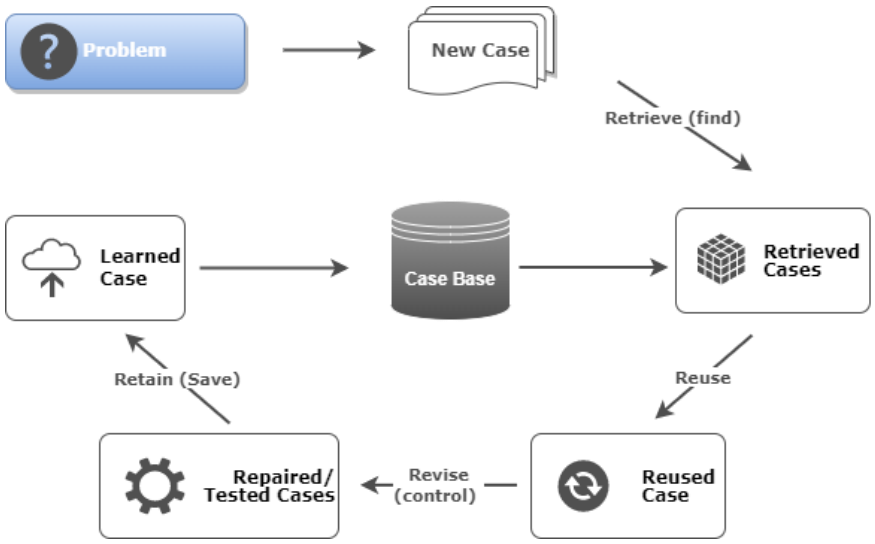


**FIGURE 3.2:** A typical Case-Base Reasoning Cycle
Image source: (Ludmila et al., 2021)

CBR systems offer the advantage of solving problems in ways that experts are acquainted to, are capable of dealing with previously unknown scenarios, and can also handle massive amounts of data. More significantly, they possess a slight ability to explain their reasoning process. The most observable limitations of CBR systems, however, is that the processes involved in the refinement of a solution requires the presentation (to user) of a lot of questions (to which the user is expected to provide answers) before an action can be triggered. Also, the encoding to machine-readable form is complex (Sally and Terrence, 1999).

### 3.1.2 Sub-Symbolic Reasoning

The sub-symbolic reasoning methods represent the connectionism movement in cognitive science that is trying to mimic the human brain and its interconnected neurons with the hope to explain the intellectual abilities using Artificial Neural Network (which we describe in section 3.1.2.5). Sub-symbolic methods establish highly complex relationships or correlations, often formalized by functions that maps input data to the output data or the target variables. Unlike the symbolic AI, the sub-symbolic AI are susceptible to noise, have high computing performance, can identify, and try to connect missing data. Furthermore, they are well-tuned for large scale (structure and unstructured) dataset and knowledge graphs. Connectionist methods often does not require any pre-knowledge of the subject domain, rather, they rely on their ability to independently learn meaningful deductions from data. This makes them a better candidate for perceptual problems.

Nevertheless, sub-symbolic methods have significant disadvantages. The fact that these methods rely on data, with a lot of parameters tuning, means they require a lot of computational powers and huge amount of quality data from which they can learn. Most often, this quality data is difficult to find. Furthermore, because of the complexity of their internal working architecture, it is mostly impossible to interpret or explain their results. This poses applicability bottlenecks particularly in sectors, such as legal, medical, and defence, where explanation and interpretation of results is significant to decision-making. Another difficulty is reproducibility — well-trained data may not be generalizable when extrapolated to previously unseen data that do not follow the training data's distribution. This may happen if the training data is not well labelled; can lead to a biased conclusion (Ntoutsi, et al., 2020).

Sub-symbolic AI include statistical learning methods, such as Bayesian learning, deep learning, genetic algorithms, and backpropagation learning methods. Other applications of sub-symbolic methods include Natural Language Processing (NLP) (Manning and Schutze, 1999; Liddy, 2001; Chowdhury, 2003), prediction, pattern recognition, classification of object and text, speech and text recognition, and clustering (Ilkou and Koutraki, 2020). Table 3.1 summarizes the major characteristics between the symbolic and Sub-symbolic AI.

Having described the methods underpinning the concept of AI, we discuss further other applications of sub-symbolic AI that will help us to create the right understanding and connection between AI and digital evidence extraction.

| **Symbolic** | **Sub-symbolic** |
|---|---|
| Symbols | Numbers |
| Logical | Associative |
| Serial | Parallel |
| Reasoning | Learning |
| Von Neuman Machine | Dynamic Systems |
| Localised | Distributed |
| Rigid and Static | Flexible and Adaptive |
| Concept composition and expansion | Concept creation, and generalization |
| Model abstraction | Fitting to data |
| Human intervention | Learning from data |
| Small Data | Big Data |
| Literal/Precise input | Noisy/incomplete input |

**Table 3.1:** Symbolic vs Sub-symbolic methods characteristics (Ilkou and Koutraki, 2020)

#### 3.1.2.1 Pattern Recognition

To begin with, we need to understand what pattern means. The term 'pattern' could have different meanings; depending on the domain or use case However, the definition that is mostly suitable, although high-level, is the one presented by Satosi Watanabe in 1985. According to (Watanabe, 1985), pattern is defined as "*the opposite of chaos; it is an entity, vaguely defined, that could be given name*." In another description of the term 'pattern', (Frawley, Paitetsky-Shapiro and Matheus, 1992) stated that: *"given a set of facts (data) F, a language L, and some measure of certainty C, we define a pattern as a statement S in L that describes relationships among a subset $F_S$ of F with a certainty C, such that S is simpler (in some sense) than the enumeration of all facts in $F_S$."* Summarizing this definition, it means pattern is an entity of interest (could be anything, in any context) which one needs to recognize and/or identify (Kpalma and Ronsin, 2007). Thus, given an input pattern, its recognition and/or classification entails either classifying it as a member of a predefined set of classes (supervised; descriptive) or assigning it to an undefined class and allowing a self-learning process based on pattern similarity (unsupervised; explorative). Given that it is a broad field that is constantly evolving, it is entirely logical that multiple definitions exist. Literally, pattern recognition implies automatic recognition of patterns in data. It is concerned with the identification of regularities in data with the use of computer algorithms and the application of these regularities to perform tasks such as classification into different categories. (Bishop, 2006). In broad terms, it refers to the study of how machines can analyze the world, learn to distinguish varied patterns of interest from their background, and draw logical inferences about the categories of the patterns (Boesch, 2021). While it has its roots in statistics and engineering, modern approaches integrate ML as a result of the increased availability of massive data and processing capacity. However, it is more of a loosely connected collection of knowledge or approaches than a singular technique.

Modern use cases of pattern recognition are based on AI technologies; with applications in domain such as: image recognition for security and healthcare, text pattern recognition, speech recognition, facial and movement recognition, deep video analysis, etc. Consequently, it is common to encounter any of these four methodologies in pattern recognition task, they include: 1) statistical method, 2) syntactic method, 3) template matching, 4) neural networks.

- **Statistical Method:** this method is the most extensively used because of its simplicity. In statistical pattern recognition, pattern features are converted to numerical vectors and

then grouped according to this number of features. The number of features determines how the pattern is represented on a multidimensional vector space. To compare or evaluate patterns in this method, distance measured between points in the vector space is calculated.

- **Syntactic Method:** this approach involves the representation of patterns in hierarchical perspective that considers the complex relationship between features (Venguerov and Cunningham, 1998). The syntactic approach relies on set of primitive sub-patterns (such as alphabets).
- **Template Matching**: is a popular pattern recognition technique that is commonly used in image processing to recognize and localize specific shapes within an image. By optimizing spatial cross-correlation or minimizing distance, template matching model attempts to identify the similarities between two entities by comparing the template function of the inputs. For each possibility, the matching rate is calculated, and the highest one that exceeds a predefined threshold is chosen. Typical real-world application of template matching can be found in face recognition. A significant downside of this method is its inefficiency in recognizing distorted patterns (Waweru, 2021).
- **Neural Networks:** this is currently the most popular pattern recognition method. Neural Networks (NNs) are based on massive interconnection of parallel neurons (or synapses) that simulates how the biological human brain works. It works by repeatedly supplying a set of inputs (samples), and the interconnected processing elements in the model are slowly adjusted until a desirable output, that matches the input, is achieved. Here, we only present a brief description of neural networks as a pattern recognition methodology under sub-symbolic reasoning. Section 3.1.2.5 discusses in detail, the idea of Artificial Neural Network.
- **Hybrid Method:** this is not a distinctive method because it involves the combination of different pattern recognition techniques. The powerful neural network approach is computationally intensive, while the other mathematical methods; though time consuming with heavy human resources involvement, are equally efficient. The hybridization of these models lead to optimized and efficient pattern detection results.
- **Fuzzy-based Method**: this is another approach that often does not get discussed. The fuzzy-based method (Pathak, Vidyarthi and Summer, 2005; Bayu and Miura, 2013; Orujov et al., 2020) applies the concept of fuzzy logic by utilizing the truth values between 0 and 1. Fuzzy models can be used as classifiers (Kuncheva, 2008) that assigns

a class label to an object, based on the object's description (in form of a vector containing attributes of the object). The model produces good results in uncertain domain, because in cases where dataset is not available, it can be designed based on prior knowledge and expertise.

#### 3.1.2.2 Genetic Algorithms

Genetic Algorithms (GA) is a metaheuristic[88] which belong to the larger class of evolutionary algorithm[89]. The GA technique is influenced by biological evolution and natural genetics mechanisms such as mutation, crossover, and selection. GA is stochastic; it is advantageous for solving optimization problems. The stochastic aspect of the search process is intended to guide it in such a way that the state of solutions explored are flexible and not solely controlled by the properties of the problem. To solve a problem using GA, the solutions are encoded as genes (which could be strings of characters from certain alphabets), which acts as an initial population of candidate solutions (chromosomes). The current population is then allowed to mutate by mating two solutions to create a new one. The breeding, mutation (or modification) process is iterated for a finite number of times (the population in each iteration is referred to as a generation), and the fitness of each candidate in the population is evaluated until the optimal solution is obtained, with the worst candidates being discarded (Shapiro, 2001). The evaluation of fitness is served by a fitness function, which is an objective function used to summarize, as a single figure of merit[90], how close a given design solution is to achieving the desired result.

In ML, GAs are critical for three reasons (Shapiro, 2001) — which are: 1) They operate on discrete spaces in which gradient-based approaches are inapplicable, also to search rule sets, neural networks architectures, etc. 2) unlike backpropagation methods, GA are essentially reinforcement learning, determined by a single fittest candidate. Thus, they are important in situations where performance is the only measurement for correctness. 3) sometimes, the desired solution to a problem can be group of solutions, instead of a single entity.

[88] To locate, produce, or select a heuristic (particular search algorithm) that can provide a good solution to an optimization problem, a metaheuristic is used.
[89] An evolutionary algorithm uses mechanisms inspired by biological evolution to address optimization problems.
[90] A 'figure of merit' is a metric used to describe how well a device, system, or approach performs in comparison to its alternatives.

#### 3.1.2.3 Knowledge Discovery in Databases (KDD)

In section 3.1.2.1, a high-level description of the term 'pattern' was given that includes the concept of language, and the measurement of certainty and interestingness (objective or subjective). The objective measure of interestingness is based on the structure of the discovered pattern (Hilderman and Hamilton, 1999), while the subjective measure deals with the measurement of *unexpectedness (surprising to the user)* and *actionability (if users can take actions to their advantage)* (Silberschartz and Tuzhilin, 1995). Extending this notion further in terms of knowledge discovery, we may deduce that a pattern that is both interesting and firmly certain (both according to the user-defined criteria) is referred to as *knowledge* (Frawley et al., 1992). Consequently, the *knowledge discovered* is the result from a monitored collection of facts in a database, and the patterns identified therein (Frawley et al., 1992). In 1989, at the inaugural KDD workshop, the term *"knowledge discovery"* was coined (Piatetsky-Shapiro, 1991) to underline that the end result of a data-driven discovery is knowledge. By definition, KDD is the systematic process of identifying genuine, valuable, and understandable, as well as previously unknown patterns within large and complex datasets (Maimon and Rokach, 2005). In (Frawley, Piatetsky-Shapiro and Matheus, 1992), the authors defined KDD as the "*non-trivial retrieval of implicit, previously unknown, and potentially useful information from data."* KDD involves the automatic exploration and modelling of large data repositories. The term 'automatic' means the process to sift through data and detect meaningful patterns requires minimal human input. This highlights the interconnection between KDD and AI. KDD's fundamental goal is to convert low-level data (voluminous, unintelligible) into more compact (a brief report), abstract (an approximated description), or valuable forms (a predictive value useful for future cases) (Fayyad, 1996).

At the centre of the KDD process is the data mining (Agrawal and Psaila, 1995) methods for pattern identification and extraction. Data mining is a subfield of computer science and statistics that is concerned with extracting information (through intelligent methods) from a collection of raw data and transforming it into a comprehensible structure for further use (Jiawei, Micheline and Jian, 2011). It encompasses data storage and access, scaling algorithms to large data set efficiently, visualization and interpretation of results, as well as the overall modelling and support for machine-interactions. Figure 3.4 shows the schematic representation of the steps KDD is composed of, and we describe the processes, according to (Brachman and Anand, 1994), below:

1. **Application domain understanding and task goals**. This step necessitates an in-depth analysis and specification of the end user's objective, as well as the context in which knowledge discovery will occur.
2. **Selection or creation of dataset**. It entails determining what data is available, acquiring any necessary additional data, including attributes, and merging all of the data into a single data collection.
3. **Pre-processing and cleansing**. This is a crucial element in any machine learning task since it improves data reliability. It include cleaning data, removing noise and outliers, and dealing with missing data or attributes.
4. **Data reduction and projection**. It entails data transformation (discretization of numerical attributes) and dimensionality reduction (feature extraction and selection). The effective number of variables can be reduced while ascertaining that data is well represented to fit the goal of the task.
5. **Matching goals with appropriate data mining task**. In this stage, the appropriate data mining methods depending on the goal of the knowledge discovery is determined. These methods include classification, clustering, or regression. However, the knowledge discovery goal can fall under two major data mining branches: predictive and descriptive.
6. **Models and Hypothesis Selection**. This is the stage at which the data mining algorithm and methodology for discovering data patterns are chosen. Additionally, this is goal-specific. Depending on the task's objective, it may make a trade-off between precision and understandability.
7. **Employing the data mining algorithm**. Finally, the mining algorithm is deployed on the prepared dataset. At this point, the algorithm may need to run multiple times before a good result is obtained.
8. **Evaluation**. This step evaluates and interprets the findings, in this case, the mined patterns (rules, dependability, etc) in relation to the pre-defined objectives. At this point, we can also evaluate the model's major components, such as the previous stages and their separate contributions to the final output.
9. **Acting on the discovered knowledge**. Put the knowledge to immediate use, incorporate it into another system for future use, or just document and communicate it to the appropriate parties are all options. Additionally, this approach entails identifying and resolving potential conflicts with prior (or extracted) knowledge.

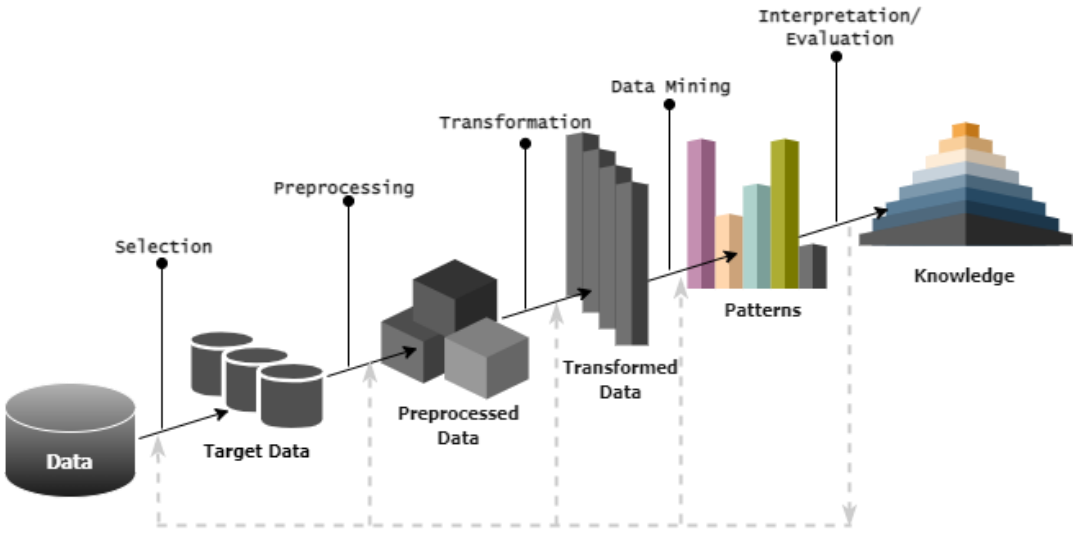


**Figure 3.3:** An Overview of the KDD Process Steps

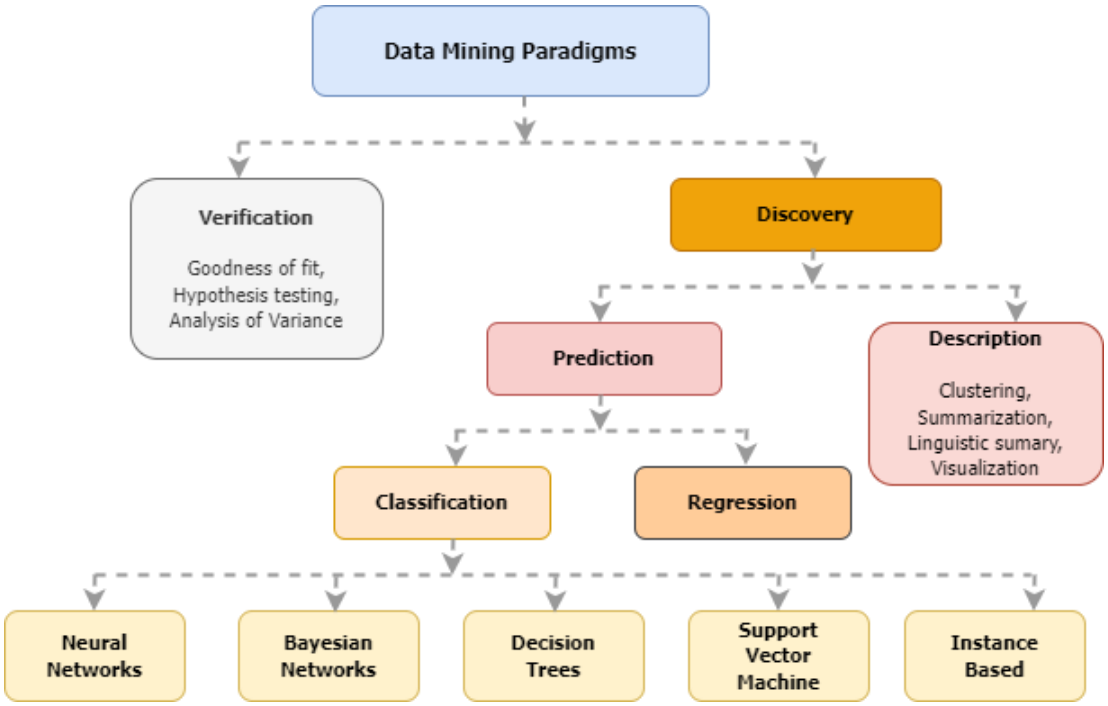


**Figure 3.4:** Data Mining Taxonomy

As previously stated, the discovery process is divided into two primary components: prediction and description (Maimon and Rokach, 2005). The prediction-oriented method develops a behavioural method for obtaining new and previously unseen samples and predict one or more sample-related variables. Additionally, the approach can aid in structuring the discovered knowledge pattern in an intelligible manner. Figure 3.4 diagrammatically describe the taxonomy of data mining. The descriptive method deals with interpretation, which focuses on how the underlying data relates to its components. Finally, KDD and data mining continue to

be the most effective artificial intelligence methods for dealing with vast amounts of data. However, because the KDD's reasoning process does not make use of previous knowledge or more complicated AI reasoning methodologies, it is probable that they will overlook more significant information in the process.

#### 3.1.2.4 Machine Learning

Machine Learning (ML) is a field of study that spans computer science, statistics (Harrington, 2012), and a variety of other disciplines concerned with continuous improvement, as well as inferences and decision-making in uncertain conditions. Several other fields are strongly related to the foundations of ML, including psychology, neuroscience, the study of human biological evolution, adaptive control theory, and educational methods (Jordan and Mitchell, 2015). In general, any field that requires the interpretation and processing of data can benefit from ML approaches (Harrington, 2012). ML started reorganizing in the 1990s, when it shifted focus from traditional AI, towards solving real-world problems. It shifted the emphasis away from the traditional logical, knowledge-based (symbolic) or AI-based approach and toward statistical and probabilistic models and techniques (Langley, 2011). There has been an increase in arguments concerning ML and its relationship to conventional AI. While some sources claim that ML is a subdivision of AI (Breiman, 2001; James et al., 2013; Mehryar, Afshin and Ameet, 2018), others believe that AI should refer to only a subset of ML that is intelligent (Bishop, 2006; Alpaydin, 2010). Regardless, ML learns and predicts through passive (or more recently, active) observation, however, AI is an intelligent agent that learns about its environment and undertakes activities to enhance its chances of success (Alpaydin, 2010). Additionally, ML, as stated in (Jordan and Mitchell, 2015), attempts to answer two interrelated questions: 1) How is it possible to construct a computer system that automatically learns from experience? 2) What are the computational and theoretical concepts underlying the learning processes of computers, humans, and organisations? These two questions appear to be interconnected across disciplines, including computational/mathematical science, statistics, psychology, philosophy, neuroscience, and economics.

Several definitions of ML exist, we state some of them here. According to the authors in (Mehryar, Afshin and Ameet, 2018), ML is a computational method that makes use of the learner's prior knowledge (in the form of electronic data) to improve performance or produce correct predictions. The digitized data may take the form of human-labelled training examples, or other information obtained via environmental interactions. However, the learner's success is

contingent upon the quality and volume of the training examples. IBM, a global leader in modern AI, defines ML as a branch of AI and computer science that focuses on using data and algorithms to replicate how humans learn, while continually improving the accuracy of the process[91]. In (Samuel, 2000), ML is a key aspect of AI that aims to equip computers with the ability to learn without explicit programming. Another definition in (Parth, 2017) states that ML is a collection of techniques that enables computers to automate the process of creating and programming data-driven models by discovering statistically significant patterns in available data. Daniel Faggella in his article (Faggella, 2020), sees ML as *"the science of getting computers to learn and act like humans do, and improve their learning over time in autonomous fashion, by feeding them data and information in the form of observations and real-world interactions."* By combining these definitions, it is intuitive to deduce that "*ML is a broad term that refers to a variety of algorithms that are powered by computational and statistical sciences, as well as biological human reasoning principles, and are designed to optimise the automatic discovery of latent behaviour in data that can aid in accurate predictions on unseen data."* A primary goal of ML is for it to perform accurately on new, unknown examples following exposure to a learning data set, i.e., to generalize from its experience (Bishop, 2006).

One key element crucial to the success of ML problem is the algorithm, or typically referred to as model, which is determined by the nature of the given problem, the features of the data (with primary consideration on the number of unique data point), and the type of the desired outcome (James et al., 2019). A large data set may necessitate a complicated ML method, but a smaller data set may be acceptable for a classical technique such as Decision Trees (Quinlan, 1986), or linear regression. While it is true that ML aims to replicate human reasoning abilities (which rely on common sense to filter out meaningless conclusions), transferring similar tasks to computers must be done with extreme caution. A set of well-defined principles must be provided to prevent the algorithm from reaching illogical or worthless conclusions (Shai and Shai, 2014). While significant attention is frequently placed on the learning algorithms, researchers have realized that some of the most intriguing problems originate from the training data, this also happen when working with ML in new domains (Faggella, 2020).

[91] What is Machine Learning? IBM Cloud Education. Cited on 23rd June 2020. Available online at https://www.ibm.com/cloud/learn/machine-learning

Consequently, the field of ML consist of various subdivisions, each of which deals with a distinct sort of learning task. Traditionally, ML approaches have been categorised into three (3) main groups depending on the signal (methods used to get training data) and feedback (test data used to evaluate) available to the learning system. The following sections address the three most often used techniques, as well as a few others that are rarely discussed in texts.

## A. Supervised Learning

Essentially, supervised learning is a ML approach that is distinguished by the usage of labelled datasets. It builds a mathematical model of set of input data (known as training data; with set of training examples) and the desired output (Russell and Norvig, 2009). The training examples are usually represented as an array of feature vectors, and the training data as a matrix. The datasets used in supervised learning aims to train or supervise algorithms to accurately classify data or predict outcomes. In general, supervised learning makes their predictions using a mapping function $f(x)$, which produces an output (or a probability distribution) $y$ for each input $x$. The training data consists of pairs of $(x, y)$ values, and the objective is to generate $y'$ in response to query $x'$. $x'$ can be a simple vector or a more complex objects (e.g., images, protein sequence, documents, etc.). Similarly, numerous types of output $y$ can be derived depending on the problem. For example, in classification problems, $y$ can be: binary, where the output take one of two values (e.g., 'spam' or 'not spam'); multiclass (where $y$ can take one of $k$ values); and multilabel (that takes several simultaneous $k$ values). In addition, the derived output $y$ could take the form of a partial order to solve ranking problem – a sets of constraints for generic prediction, or a set of real values or mixed with discrete values (Jordan and Mitchell, 2015). Predicting the output associated with a new (unseen) input involves iterative optimization of the objective function following the training phase. An objective function (or loss/cost function) is a function that intuitively maps an event or the values of one or more variables to a real number expressing the event's cost. The purpose of optimization is to minimize this loss function while maximizing the objective function. As a result, an optimal function of the ML model should be capable of effectively determining the output for inputs not included in the training sample.

We mathematically define the general idea of a learning model as follows: a supervised model is essentially a function $f(x, \theta)$, $g$iven $N$ items of input data $x_i$ and associated output $y_i$, where $i = \{1, \ldots, N\}$. $x$ is the input data and $\theta$ denotes a set (or sets) of parameters. The purpose is to iterate the loss function towards parameter $\theta$, producing output $\hat{y}_i = f(x_i, \theta)$that is as close to

$y_i$ as possible. After then, the model $y = f(x, \theta)$ can be used to make predictions about new and unseen data $x$. Further, we detail the types of supervised learning algorithms below, categorizing them according to the sort of problem they typically solve. They include the following:

**Classification** algorithms are used when the outputs are constrained to a finite set of values. It is a problem of assigning a category to each input item. For instance, classifying a document involves assigning categories such as business, politics, sport, or health to each input document, while classifying images consist of assigning categories to input images such as cat, car, shoes, or to predict cancerous conditions on a patient's skin. The algorithm predicts the category of a new, unseen input based on passive experience, which it has gained from the labelled training data. Classifier role is played by a dataset where each data point $x_i$ (e.g., vectors, objects) has the corresponding output $y_i$, describing which of $k$ possible classification $x_i$ belong to (James et al., 2019). In the case of a binary classification, $k = 2$ (representing 0 or 1). The output of a classifier is represented as a vector $\hat{y}_i = (\hat{y}_{i,1}, \dots, \hat{y}_{i,k})$ which is the probability that $x_i$ belongs to any of the class 1 to $k$, and the values sum to 1. In most cases if the value of $k >= 0.5$, then it is classified as 'positive', otherwise, if $k < 0.5$, then it is 'negative'. Nevertheless, the number of categories a classification problem is often less than few hundreds, however, it can be larger in some complex tasks and even unbounded as in OCR, text classification, or speech recognition tasks (Mehryar, Afshin and Ameet, 2018). *k*-Nearest Neighbours (KNN), Linear Classifiers (LC), Random Forests (RF) (Ho, 1995; Ho, 1998; Breiman, 2001), SVM (Cortes and Vapnik, 1995; Joachims, 1999; Noble, 2006), and Decision Trees are all examples of classification algorithms. Figure 3.5 shows a typical classification problem graph in a D-dimensional space.

We briefly describe below some classification algorithms:

> **Support Vector Machine (SVM)** are supervised machine learning models with an accompanying learning algorithm for classification and regression analysis. They are, however, mostly utilized to solve categorization problems. Given a set of training examples, each of which is labelled as belonging to one of two categories, an SVM model assigns subsequent samples to one of the two categories, thereby transforming it into a deterministic binary linear classifier. However, a technique known as "plat

scaling[92]" exists for using SVM in a probabilistic setting. SVMs are nothing more than the coordinates of individual observations; they do classification by mapping training examples to points in space and determining the hyperplane that separates the classes the most. Apart from linear classification, SVM may also be used for non-linear classification by applying the kernel trick[93] method to map inputs to a high-dimensional feature space. While SVM is mostly used to classify labelled data, the Support Vector Clustering (SVC) (Ben-Hur et al., 2001) algorithm can also be used to classify unlabelled data.

**Decision Trees** is a type of prediction model that is used in statistics, ML, and data mining (Lior and Oded, 2008). It is a subset of supervised learning wherein data is gradually split according to a specific parameter. Decision trees are a combination of computational and mathematical methods that are used in the description, categorization, and generalization of a given datasets. Three entities can be used to describe the tree: the root node, the leave node, and the decision node (or branch). The tree illustrates the evolution of observations (defined as branches) about an entity to conclusions about the entity's desired value (represented as leaves). The leaves represent class labels, whereas the branches describe the features that combine to form the class labels. A decision tree model can be represented as a classifier with target variables having discrete values (e.g., "true" or "false"). Alternatively, the decision tree model might be a regression tree with continuous (real number) values for the target variables. The process of creating a tree entail segmenting the source set (the root node) into subsets (the successor children) using a splitting rule based on classification features (Shai and Shai, 1994). This process is performed for each derived subset using recursive partitioning. The recursive process is complete when the continuous splitting no longer adds value to the prediction or when the node's subset value and the target variable are the same. Algebraically, data comes in the form: $(x, Y) = (x_1, x_2, x_3, \dots, x_k, Y)$. The dependent variable, $Y$, is the target variable for which classification is sought. The vector $x$, is composed of the feature $x_1, x_2, x_3$, etc., that are used for the predictive task. Algorithms for decision trees usually works *top-down*

[92] In ML, Platt scaling is a way of transforming the outputs of a classification model into a probability distribution over classes.

[93] Kernel tricks are a form of pattern analysis algorithm whose purpose is to discover and explore general types of relationships (through clustering and correlation) in datasets.

(*induction of decision trees*) (TDIDT) (Quinlan, 1986) choosing a variable at each step that offers the best split for the set of items. To measure the "optimal" split, several algorithms often assess the target variable's homogeneity within the subsets. The resulting values are then summed (and averaged) to produce a measure of the splits' quality. Some popular metrics used in decision tree algorithms are measurement of goodness, information gain (and the concept of entropy), variance reduction (for regression tree), Gini impurity, etc.

**Random Forests (RF)** during training, performs ensemble learning (a technique that combines numerous classifiers to solve a large number of complex problems) by assembling a number of decision trees that may be used for classification, regression, and other tasks. In a classification task, the output of RF algorithm is the class selected by the majority of trees, but in a regression task, the output is the average mean of the individual tree's prediction. While RF are frequently used in supervised learning, they can also be used to define a dissimilarity measure between unlabelled data by building an RF predictor capable of discriminating between real and synthetically generated data (drawn from a reference distribution) (Breiman, 2001) — which is similar to what unsupervised learning does. Unlike in a traditional Decision Tree, RF selects randomly, a subset of features at the node's splitting point. RF generates the required prediction by utilizing the "Bagging"[94] methodology — a bootstrap aggregation technique that utilizes several data samples. Depending on the RF algorithm deployed, the trees produce different output. The outputs are then ranked, and the highest is selected as the final output. RF often outperforms decision trees because it reduces overfitting and boosts precision (Hastie, Tibshirani and Friedman, 2008). However, the characteristics of the training data can have an effect on their performance (Piryonesi and El-Diraby, 2021). Explicitly, overfitting in decision trees is induced by trees that have grown to a depth and tend to learn irregular patterns. As a result, they exhibit little bias but a great deal of variance. To address this issue, RF averages numerous deep decision trees in an attempt to limit the variance (Hastie, Tibshirani and Friedman, 2008). The decrease in variance is offset by a little increase in bias and a loss of interpretability. RF are commonly characterized as "closed-box" model due to their ability to make significant

[94] Bagging is a ML ensemble method developed to increase the stability and accuracy of classification and regression tasks.

predictions across a wide variety of data with minimal configuration or without hyper-parameter tuning.

**Regression** algorithms are applicable when an ML task's target outputs are constrained to a finite set of values. It predicts the actual value for each input item. Regression algorithms seek to understand the relationships between dependent and independent variables. In regression, the penalty for making an inaccurate prediction is proportional to the magnitude of the difference between the predicted and actual values. For a regressor, we assume a linear function $f(x, \theta) = \beta x + m$, where $\theta = (\beta, m)$ is the parameter set; which contains the slope $\beta$ and line's intercept $m$. In the context of ML, training a regression model entails determining slope and intercept. A simple closed-form calculation is used to solve the $\beta$ and $\theta$, e.g., minimizing the sum of squares of the difference between actual and predicted values $(y_i - \hat{y}_i)^2$ (James et al., 2019) to find the degree of closeness. Regression algorithms are useful for the prediction of continuous values such as sales revenue projection in a given business. Some popular regression algorithms are logistic regression, polynomial regression, and linear regression. Mathematically, a linear regression relationship is given by:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

where:

$y$ is the response

$\beta$ are the model's coefficient; learned during model training/fitting step

$\beta_0$ is the intercept

$\beta_1$ is the coefficient of the $x_1$ (the first feature)

$\beta_n$ is the coefficient of $x_n$ (the $nth$ feature)

Unlike classification models, which are evaluated using "accuracy" metrics, regression models are evaluated by comparing continuous values (i.e., the predicted and actual values). The detailed description of regression model evaluation metrics are in chapter 5. Figure 3.6 shows a typical regression problem plotted as a graph.

**Similarity Learning** is a subfield of supervised learning that is comparable to regression and classification — albeit with distinct learning functions — with the objective to learn a (similarity) function that measures the similarity or relationship between two objects. Similarity learning is used in recommender and ranking systems, as well as visual tracking, face, and speech verification.

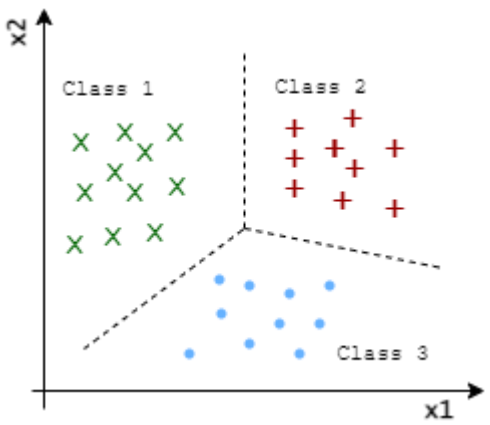


**Figure 3.5**: A Multi-class Classification problem illustration

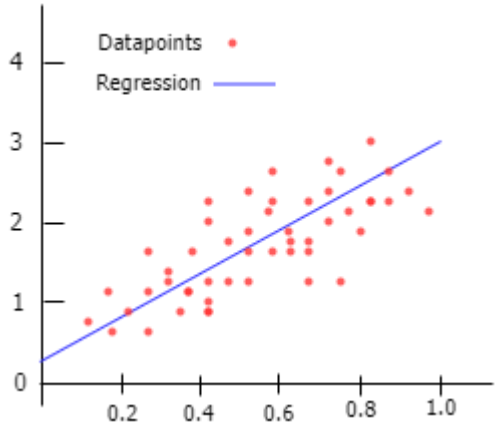


**Figure 3.6**: A regression problem plotted on graph.

## B. Unsupervised Learning

Unsupervised learning algorithms takes as input, a set of unlabelled data and, based on specific assumptions about the data's structural properties (e.g., algebraic, combinatorial, or probabilistic) (Jordan and Mitchell, 2015), learn to classify or categorize it by grouping or clustering the data points. The term "unsupervised learning" refers to a method of learning that requires no human supervision; yet human intervention is required to validate output variables. The learning algorithm studies how unlabelled data might be used to infer a function that describes the hidden structure. This is accomplished by identifying commonalities in data and reacting to their presence or absence in each new piece of data. Typically, the model does not produce the precise output, but it explores the data and can draw inferences from it to describe the data's hidden characteristics. While the tasks may include cluster detection or various types of pattern recognition (James et al., 2019), supervised learning is primarily concerned with summarizing and interpreting data features. In comparison to supervised learning, unsupervised learning is capable of extracting insights from large amounts of data; the goal is to determine what makes the dataset interesting and/or different. It is an excellent candidate for use in anomaly detection (Varun, Arindam and Vipin, 2009) and recommender systems. For example, the model can determine that online-buyers frequently purchase a specific group of products concurrently — such as a babysitting father purchasing diapers and beers because he prefers to stay at home (tending to the baby) rather than go to the pub with friends. Additionally, unsupervised learning is computationally complex — it requires a big training set to achieve a desired result — and might yield crazily erroneous results unless human intervenes to adjust some functional parameters (Delua, 2021). Unsupervised Learning are used for three (3) main

tasks: clustering, association, and dimensionality reduction. We define each learning model below:

## I. Clustering

Clustering is the process of partitioning a set of data into (homogeneous) subsets (occasionally referred to as clusters) in such a way that observations within identical clusters are comparable in terms of one or more predefined criteria, whilst the observations derived from other clusters are disparate. They are used to process raw, unclassified data into groups represented by structures or patterns in the information. Different clustering approaches make varying assumptions about the data's structure, which is typically expressed by some measure of similarity and evaluated in various ways, for example, by internal compactness or similarity between cluster members. A clustering task can have a varying input and output type. Mathematically, common setup of a clustering model is described below:

> **Input:** a collection of elements, $X$, and a distance function applied to them. That is, a function denoted by $d: X \times X \rightarrow \mathbb{R}^{+}$ that is symmetric, satisfies $d(x,x) = 0$ for all $x \in X$, and frequently also satisfies the triangle inequality. Alternatively, the function might be a symmetric similarity function $s: X \times X \rightarrow [0,1]$ that satisfies $s(x,x) = 1$ for all $x \in X$. Additionally, certain clustering algorithms require an input parameter $k$ (which specifies the number of clusters required).
>
> **Output:** a subdivision of the domain set $X$ into subsets. That is, $C = (C_1, \ldots, C_k)$ where $\bigcup_{i=1}^{k} C_i = X$ and for all $i \neq j, C_i \cap C_j = \varphi$. In some cases, clustering is "soft," i.e., the partitioning of $X$ into distinct clusters is probabilistic, with the output being a function that assigns a vector $(P_1(x), \ldots, P_k(x))$ to each domain point, $x \in X$, where $P_i(x) = P[x \in C_i]$ is the probability that $x$ is a member of cluster $C_i$.

Clustering algorithms can be categorized into specific type, viz. exclusive, overlapping, hierarchical, and probabilistic. We briefly describe these components below.

a) Exclusive Clustering

Exclusive clustering groups data points exclusively in one cluster, i.e., a data point can only belong to one group — not more than. Also referred to as the "hard" clustering. A classic example of an exclusive clustering algorithm is the "K-Means".

***K-Means Clustering*** (Hartigan and Wong, 1979): in which data points are assigned to $K$ groups depending on their distance from the centroid of each group. The data points that are closest to a specific centroid will be grouped together. A higher $K$ value indicates smaller groupings with increased granularity, whereas a lower $K$ value indicates bigger groupings with less granularity. It begins by defining a cost function over a parameterized set of possible clusters, and the algorithm's objective is to discover the cluster with the minimal cost (Shai and Shai, 2014). The clustering task is transformed into an optimization problem under this paradigm. $k$-means objective function is a pair of input, $(X, d)$, and a proposed clustering solution $C = (C_1, \ldots, C_k)$, to positive real numbers. Given such an objective function, denoted by $G$, the purpose of a clustering method is to identify a clustering $C$ such that $G((X, d), C)$ is minimized for a given input $(X, d)$. $k$-means clustering is frequently used in document clustering, image compression and clustering, and market segmentation.

b) Overlapping Clustering

The difference between this clustering technique and the exclusive clustering technique is that it allows for cluster overlap, which means that data points can be a member of different clusters with distinct membership degrees. The "soft" or fuzzy $k$-means (Dan, 2004) clustering algorithm is an outstanding example of an overlapping technique.

c) Hierarchical Clustering

This technique is also known as Hierarchical Cluster Analysis (HCA). It can be classified as agglomerative or divisive. Agglomerative clustering, on the one hand, is a bottom-up approach wherein data points are initially divided into various groups and then iteratively merged together based on similarity until a single cluster is obtained. The metric evaluation in this technique is performed by measuring similarities — usually the distance between two points within each cluster, often using the Euclidean distance[95], although the Manhattan distance[96] has been reported in certain publications.

d) Probabilistic Clustering

The probabilistic clustering model assists in resolving the problem of soft clustering or density estimation. Data points are clustered in this model based on their likelihood to belong to a given

[95] See https://www.sciencedirect.com/topics/mathematics/euclidean-distance
[96] See https://xlinux.nist.gov/dads/HTML/manhattanDistance.html

distribution. One of the most commonly used probabilistic clustering is the Gaussian Mixture Model (GMM).

### II. Association Rules

Association Rules (Agrawal, Tomasz and Swami, 1993; Agrawal and Srikant, 1994) are a rule-based approach for determining the relationships between variables in a given dataset. It is meant to identify strong rules in databases through the use of an interestingness metrics (Frawley, Piatetsky-Shapiro and Matheus, 1992). These techniques are constantly employed in market basket analysis to assist businesses in better understanding the relationships between various products, as well as consumer consumption habits, which aids in the development of cross-selling strategies and recommender systems (for instance, the strategies used by Amazon and Netflix). In (Agrawal, Tomasz and Swami, 1993), the authors presented association rules for recognising product regularities in massive amounts of transaction data generated by supermarket point-of-sale (POS) systems. For instance, a particular rule discovered in sales data indicates {potatoes, onions} → {burger} – which means that if a buyer purchases potatoes and onions, they are also likely to purchase hamburger meat. This method aids businesses in making promotional and product placement decisions.

The "*Apriori Algorithm*" is a frequently used approach based on the concept of association rules, and as a result, it has gained popularity as a result of its use in market basket analysis and recommender systems on online music platforms and e-commerce sites.

### III. Dimensionality Reduction

Dimensionality Reduction (DR) is the process of transforming data from a high-dimensional space to a low-dimensional space while retaining the significant features of the original representation; usually close to the intrinsic dimension. While it is true that more data produces more accurate results in machine learning, failing to transform to a lower dimensional space can have a detrimental effect on the learning model's performance (e.g., overfitting), make it computationally intractable, result in poor generalization, and make it difficult to visualize. DR is strongly related to the information theory idea of (lossy) compression. Additionally, dimensionality reduction can improve data interpretability — the ease with which significant structures can be discovered in the data (Shai and Shai, 2014). DR is frequently used in domains that deals with a large number of observations and/or variables, such as signal processing, bioinformatics, etc. (Laurens, Eric and Jaap, 2009). Additionally, it is effective for

use in noise reduction, data visualisation, and cluster analysis. The general idea of reduction involves the application of linear transformation to the original data. That is, if the data is in $\mathbb{R}^d$, such that we want to embed into $\mathbb{R}^n$ $(n < d)$. Consequently, we find a matrix $W \in \mathbb{R}^{n,d}$ that induces the mapping $x \rightarrow Wx$. The choice of $W$ will allow the reasonable recovery of the original $x$. However, the exact recovery of $x$ from $Wx$ is not possible (Shai and Shai, 2014).

Common methods in DR are divided into linear and non-linear approaches (Laurens, Eric and Jaap, 2009), also feature selection and feature extraction approaches (Pavel and Jana, 1998). Reduction is typically utilized during the pre-processing step of a learning model, and we will explore some of the most frequently used approaches below:

**Principal Component Analysis (PCA)**: is the major technique for dimensionality reduction. It performs a linear transformation on a set of data in order to create a new low-dimensional representation that maximizes the data's variance. In practice, the covariance (or correlation) matrix of the data is constructed, and its eigenvectors are computed. The corresponding eigenvectors to the biggest eigenvalues (principal components) are then utilized to reconstruct a considerable percentage of the variance in the original data. PCA reduces redundancies and compresses data through feature extraction. Mathematically, we describe the principal component reduction problem as follows:

Let $x_1, \ldots, x_m$ denote $m$ vectors in $\mathbb{R}^d$ .We would like to use a linear transformation to lower the dimensions of these vector. A matrix $W \in \mathbb{R}^{n,d}$ where $n < d$, induces a mapping $x \rightarrow Wx$, where $Wx \in \mathbb{R}^n$ is the lower dimensional representation of $x$. Then, using a second matrix, $U \in \mathbb{R}^{n,d}$, it is possible to recover (approximately) each original vector $x$ from its compressed version. That is, for a compressed vector $y = Wx$, where $y$ is in the low dimensional space $\mathbb{R}^n$, we can construct $\tilde{x} = Uy$, where $\tilde{x}$ is the recovered version of $x$ in the original high dimensional space $\mathbb{R}^d$.

**Singular Value Decomposition (SVD)**: is another dimensionality reduction techniques that factorizes a real or complex matrix, $M$, into three low-rank matrices, $UDV^*$. Particularly, the singular value decomposition of an $m \times n$ complex matrix, $M$, represent a factorization method of the form $UDV^*$, where $U$ is an $m \times m$ unitary matrix, $D$ denotes an $m \times n$ diagonal matrix consisting of non-negative real numbers on the diagonal, and $V$ is an $n \times n$ complex matrix. If $M$ is real, $U$ and $V$ are also guaranteed to be real orthogonal matrix. In such a case, the SVD is often denoted as $UDV^T$. The

diagonal entries $\sigma_i = D_{ii}$ of $D$ are regarded as singular values of $M$. The matrix $M$ rank is equal to the number of non-zero singular values. Summarily, unit vectors $V \in \mathbb{R}^{n}$ and $U \in \mathbb{R}^{m}$ are the right and left singular vectors of $M$ with corresponding singular value $\sigma > 0$ if:

$$MV = \sigma U \; and \; M^T U = \sigma V$$

Similar to PCA, SVD is also commonly used for noise reduction and compression of files, such as image files.

**Autoencoders (AE)**: literally utilize neural networks to compress data and then recreate its original representation. AE are used to learn (unsupervised) representations or features of unlabelled data (Kramer, 1991), mostly for dimensionality reduction. The data encoding is validated and refined by trying to regenerate the input from the representations, typically by training the network to disregard inconsequential (noisy) data. AE and its variants, such as Variational Auto-Encoders (VAE) (Welling and Kingma, 2019), Denoising Autoencoders (Vincent and Larochelle, 2010), and so on, are also referred to as generative models (Welling and Kingma, 2019), as they are capable of randomly generating new data that is identical to the input data, which is useful for learning representations for classification and prediction tasks. AE are applied to many machine learning problems such as feature detection (Géron, 2019), facial recognition (Hinton, Krizhevsky and Sida, 2011), word embedding for meaning acquisition (Liou, et al., 2014), anomaly detection (Varun, Arindam and Vipin, 2009; Farid and Rahman, 2010), etc. AE is composed of two major components: an encoder (input layer) that transforms the input to a low-dimensional representation (embedding) in the latent space, and a decoder (output layer) that converts the embedding into a reconstruction of the input. The diagram in Figure 3.7 illustrates a common autoencoder architecture. As illustrated in the diagram, the hidden layer serves as a bottleneck, compressing the input layer prior to reconstruction in the output layer. The input layer and output layer each have the same number of nodes (neurons). The term “encoding” refers to the stage that occurs between the input and hidden layers, while “decoding” is the stage that occurs between the hidden and output layers.

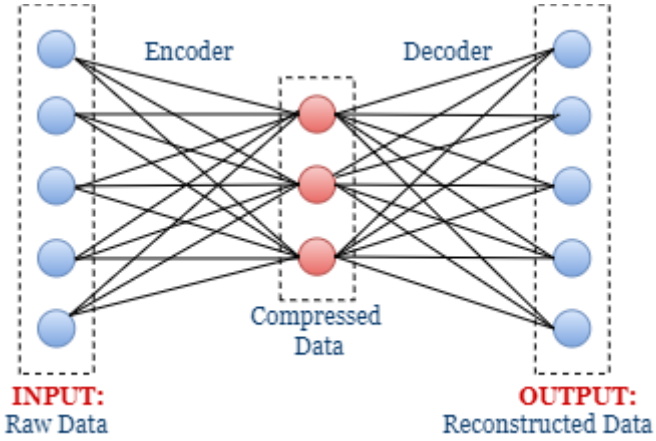


**Figure 3.7:** A simple Autoencoder Architecture

In its simplest form, AE is a feedforward (Zell, 1994), non-recurrent neural network identical to a single layer perceptron. However, it can also be used as a component of a multilayer perceptron (MLP)[97] with one or more hidden layers. As is the case with machine learning prediction problems, whereby target value $Y$ is predicted given input $X$, AE's goal is to learn (unsupervised) to reconstruct its input by minimizing the difference between the input and the output. Mathematically, the encoder and decoder are defined as a transition $\emptyset$ and $\psi$, respectively, such that:

$$\phi : X \rightarrow F$$

$$\psi : F \rightarrow X$$

$$\emptyset, \psi = argmin \, || X - (\psi \circ \emptyset) X ||^2$$

In a simple case of a single-layered autoencoder, the encoder takes the input

$x \in \mathbb{R}^d = X$ and maps it to $h \in \mathbb{R}^p = F$:

$$h = \sigma(Wx + b)$$

$h$ is knowns to as the latent (or hidden) variables/representation, $\sigma$ denote an element-wise activation function such as the sigmoid function[98] or Rectified Linear Unit (ReLU)[99], and $W$ and $b$ represent the weight matrix and bias vector, respectively (both

[97] A multilayer perceptron (MLP) is a form of feedforward ANN that creates outputs based on a set of inputs. Between the input and output layers, multiple layers of input nodes are connected via a directed graph.

[98] See https://deepai.org/machine-learning-glossary-and-terms/sigmoid-function

[99] See 'Rectified Linear Units (ReLU) in Deep Learning' By DanB. Available online at https://www.kaggle.com/code/dansbecker/rectified-linear-units-relu-in-deep-learning/notebook

are typically randomly initialized and are iteratively updated via backpropagation[100] during training (Rumelhart, Hinton and Williams, 1986; LeCun, 1988)).

The decoder stage maps $h$ to the reconstruction of $x'$ of the same shape as $x$:

$$x' = \sigma'(W'h + b')$$

It is worth noting that $\sigma'$, $W'$ and $b'$ are unrelated to the corresponding $\sigma$, $W$ and $b$ in the encoder part. AE is modelled to minimize the reconstruction (error) loss; therefore, the loss function is given by:

$$L(x, x') = ||x - x'||^2 = ||x - \sigma'\big(W'\big(\sigma(Wx + b)\big) + b'\big)||^2$$

where $x$ is typically calculated as an average over the training set.

Lastly, some of the most common real-world application of unsupervised learning include computer vision, anomaly detection, medical imaging, News section categorization, recommender system, and customer's persona.

### C. Semi-supervised Learning

Semi-supervised learning is a subset of ML whose functionality lies between supervised and unsupervised learning, and it combines a small amount of labelled training data with a large amount of unlabelled training data to create predictions for previously unknown data points. The goal is to significantly enhance learning accuracy by leveraging unsupervised learning's ability to intelligently detect hidden structures in massive amounts of unlabelled data and the accurate predicting abilities of the (labelled) supervised technique. It is prevalent in settings or sectors where unlabelled data is easily accessible but labelled data is expensive to collect due to a scarcity of human experts to perform the labelling. Numerous machine learning problems encountered in applications, such as classification, regression, or ranking, can be framed as semi-supervised learning problems in the hope that the learner's access to unlabelled data will result in better performance than supervised learning. However, the theoretical underpinnings and practical implementation of this approach continue to be a source of ongoing research interest (Mehryar, Afshin and Ameet, 2018). Furthermore, semi-supervised learning could be regarded as either transductive or inductive learning. The former relates to reasoning

[100] In contrast to a basic direct computation of the gradient with respect to individual weight, during ANN training, backpropagation effectively computes the gradient of the loss function with regard to the network's weights.

from observed, specific (training) cases and making predictions solely for unlabelled specific test cases, whereas the latter refers to reasoning from observed training cases to generic rules that are subsequently applied to the test case.

### D. Reinforcement Learning (RL)

Reinforcement learning is a subfield of ML concerned with the study of how intelligent agents should behave in a given environment in order to maximise the notion of cumulative reward. RL is a fascinating learning model because it is capable of not only learning how to map one input to an output (as native ML approaches do), but also mapping series of inputs to outputs dependencies, typically in the form of Markov Decision Processes (MDP). To collect information literally in RL, the learner actively interacts with, and in certain situations, impacts, the environment, and receives an immediate reward for each action (Mehryar, Afshin and Ameet, 2018). Typically, RL problems take place in a general control-theoretic framework, with the learning task being to develop a control policy for an agent acting in an unfamiliar dynamic environment, while also training the agent to choose actions (or responses) for any given state with the goal of maximising its expected reward over time (Jordan and Mitchell, 2015).

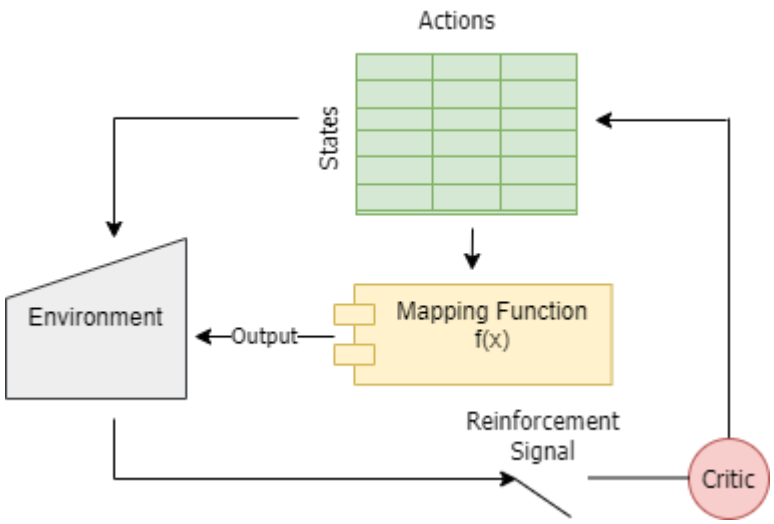


**Figure 3.8:** Reinforcement Learning model

However, because there is no long-term reward feedback from the environment, the learner is confronted with the challenge of “exploring” uncharted territories in order to gather additional information and "exploiting" the information already gathered. RL is distinguishable from supervised learning because it does not require the presentation of labelled input/output pairs or the explicit correction of suboptimal behaviours. During the learning phase, the algorithm

randomly explores state–action pairings in a specific environment (to create a table of state–action pairs), and then uses the state–action pair rewards to identify the optimal action for a certain state that leads to the desired result (Tim, 2017). Figure 3.8 illustrates the reinforcement learning model.

### E. Self-learning

Self-learning as a ML paradigm, along with the crossbar adaptive array (CAA) (Bozinovski, 1982), is an ANN capable of self-learning. It is a method of learning without external rewards or guidance from an instructor. The CAA self-learning algorithm computes both decisions about behaviours (such as emotions) and actions concerning consequence scenarios in a crossbar fashion. The system is propelled forward through the interaction of intellect and emotion (Bosinovski, 2014). Self-learning algorithm update a memory matrix $W = ||w(a,s)||$ such that, in each iteration, it runs the steps as seen in figure 3.9.

The figure algorithm illustrates a system with a single input, situation $s$, and a single output, action (or behaviour) $a$ - without receiving distinctive reinforcement from the environment. The emotion experienced in response to the consequence situation is the backpropagated value (secondary reinforcement). The CAA exists in two environments: one in which it is supposed to behave (behavioural environment), and another in which it acquires initial emotions about the situations it will encounter in the behavioural world for the first and only time (genetic environment).

```
>> In situation s, execute action a;
>> Receive consequence situation s';
>> Compute emotion in consequence situation v(s');
>> Update crossbar memory w'(a, s) = w(a, s) + v(s')
```

### F. Feature Learning

The term "feature (or representation) learning" refers to a set of techniques that enable a system to automatically learn the representations required for feature recognition or classification from data. This reduces the need for manual feature engineering and enables a system to learn and employ features to accomplish a certain task. As a pre-processing step prior to solving a classification or prediction problem, feature learning typically transforms data to useful

representations while retaining the information contained in the original input data. They can be either supervised or unsupervised. Examples of learning models that leverages feature learning under supervised architecture include Artificial Neural Networks (ANN) and Multilayer Perceptron (MLP). The unsupervised features are learned with unlabelled data, and example of such learning models includes Autoencoders and various forms of clustering algorithms.

## G. Deep Learning (DL)

Deep learning is a subset of ML that is designed to learn from massive amounts of data in ways that loosely mimic how the human brain works. DL is a subfield of ML, which is itself a subfield of AI. In other words, deep learning is a subset of a larger family of ML techniques that utilize Artificial Neural Networks (Discussed in 3.1.2.5). The recent advancements in emerging technologies such as generating captions for YouTube videos, speech/voice recognition on phones and smart speakers, machine translations, facial/entity recognition in images, as well as self-driving cars, are largely driven by deep learning. Because the majority of deep learning methods employ neural network design, DL models are frequently referred to as "deep neural network". Other common types of DL architecture include feedforward neural networks (FFNN) (Sanger, 1989; Bebis and Georgiopoulos, 1994), deep belief networks (DBNs) (Hua et al., 2015; Chen et al., 2015), convolutional neural networks (CNN) (O'shea and Nash, 2015; Albawi, Mohammed and Al-Zawi, 2017), recurrent neural networks (RNN) (Mikolov et al., 2010), and many more. While a neural network might have one or two hidden layers, the term "deep" in deep learning alludes to the deployment of multiple (dozens – or even hundreds) interconnected layers of nodes in the network. Each layer has an infinite number but a finite size, allowing for practical application and optimization while preserving theoretical applicability under optimal circumstances. The layers learn to transform by progressively extracting higher-level features and creating a composite representation of the raw input data. For efficiency, trainability, and understandability, DL allows layers to be heterogeneous and vary from biologically informed connectionist models. Although, depending on the nature of the task, increasing the number layers and nodes to the network may improve accuracy. Additional layers, on the other hand, incur a penalty in terms of parameterization and computational resources.

One significant distinction between DL and ML is that DL models take data from various data sources, analyze it, and learn patterns from it in real time without human intervention. It is

capable of ingesting unstructured data (text, images) and automatically determine the set of features that distinguishes various categories of the data. It can also use supervised (with labelled datasets) learning to inform its algorithm, but this is not necessarily required. Additionally, DL algorithms must be trained on huge datasets; the accuracy of their predictions is proportional to the amount of information available to them. For instance, the model will need to be ingested with thousands of cats' pictures with varying properties before it is able to classify new pictures of cats. Furthermore, weights — which are parameters that represent the strength of the connection between inputs — are used in the hidden layers to process the raw data supplied to a DL model. To improve predictions, the weights are adjusted during training (depending on the input). DL spends a lot of time training large amounts of data, which requires high processing power.

Uses of deep learning have been demonstrated in finance (Huang, Chai and Cho, 2020; Lee and Yoo, 2020, Heaton, Polson and Witte, 2017), health (da Silva et al., 2021; Rezaeianjouybari and Shang, 2020; Sahoo, Pradhan and Das, 2020) cybersecurity and digital forensics (Salih et al., 2021; Qadir and Noor, 2021), and social media (Pathak, Pandey and Rautaray, 2021; Singh and Sharma, 2021). Their applications have also been used in aerospace and defence engineering, industrial automation, electronics, and so on.

#### 3.1.2.5 Artificial Neural Networks (ANN)

To begin, this section presents an extended description of Artificial Neural Network (ANN) in order to supplement the prior discussions about neural networks. Additionally, in this thesis, the terms "Artificial Neural Network (ANN)" and "Deep Neural Network (DNN)" may be used interchangeably to refer to the same concept. However, when the term "DNN" is stated specifically, it refers to networks with several interconnected layers (usually two or more).

In 1943, Warren McCulloch, a neuropsychologist, and Walter Pitts, a mathematician, published the first article on how neurons could work (McCulloch and Walter, 1943). To characterize the neurons, they used electrical circuits to model a simple neural network. Donald Hebb's 1949 work "*The organizing behavior*" (Hebb, 1949) demonstrated that neural pathways are strengthened each time they are utilized — a principle that is crucial to how humans learn. His reasoning was predicated on the hypothesis (*mechanism of neural plasticity*) that when two nerves fire simultaneously, their connection is reinforced. As research in this field progressed, a psychologist named Frank Rosenblatt built the perceptron (Rosenblatt, 1957; 1958; Haykin, 2008), the first artificial neural network. A simple perceptron is an artificial neuron that uses a

unit step function (with 0 indicating a negative argument and 1 representing a positive argument) as an activation function. It is an algorithm used to learn a binary classifier (or a threshold function) that maps it input $\boldsymbol{x}$ (a real-valued vector) to an output value of the function $\boldsymbol{f}(\boldsymbol{x})$:

$$f(x) = \begin{cases} 1 & if\ w.x\ +\ b\ >\ 0, \\ 0 & otherwise\ 0 \end{cases}$$

Where $w$ is a real-valued weights vector, $\boldsymbol{w}.\boldsymbol{x}$ is the dot product:

$$\sum_{\boldsymbol{i}=1}^{\boldsymbol{m}} \boldsymbol{w_i}.\boldsymbol{x_i}$$

$\boldsymbol{m}$ is the number of perceptron inputs and $\boldsymbol{b}$ denotes the bias. The bias causes a shift of the decision boundary away from the origin that is independent of the input data. In a typical binary classification, the value of $\boldsymbol{f}(\boldsymbol{x})$**,** which is either 0 or 1, indicates the class of input $\boldsymbol{x}$ as either a positive or a negative example. A perceptron may be single-layered (using a simple feedforward neural network) or multi-layered (which is often misunderstood as a complex neural network). Rosenblatt's work sparked interest in the field until Minsky and Papert in (Minsky and Papert, 1969) demonstrated that a basic perceptron could handle a relatively limited class of linearly separable problems and that machines then, lacked sufficient powers to process useful neural networks, resulting in a period of stagnation in the field (Anders, 2008). However, distinct research continued in a relative direction until 1986, when Rumelhart, Hinton and Williams demonstrated that back-propagation (Rumelhart, Hinton and Williams, 1986) can be used to train rather complex networks of simple neurons to learn from examples by using word's internal representations as feature vectors to predict next word in a sequence. Unsupervised pre-training, combined with advances in computer processing power (GPU) — courtesy of computer-game industry — and distributed computing, paved the way for the deployment of larger networks capable of learning to recognize higher-level concepts such as objects from unlabelled images (Quoc V. Le et al., 2012) and in visual recognition, a process dubbed "deep learning" (Goodfellow, Bengio and Courville, 2016). For example, in object recognition, the systems can be fed thousands of labelled images of automobiles, plants, houses, and animals, and the model will look for visual patterns consistent with the labels in the images. Figure 3.9 illustrates the building block of a deep neural networks.

In general, the NN architectural underpinning is based on the concept of an input vector getting mapped to an output value and being optimized using real-valued weight vectors, and the

constant bias value (that helps in a way that best fit the given data). A Neural Network is a computational model inspired by the structure and connectivity of neurons in the human brain (Shai and Shai, 2014). They consist of networks of nodes termed artificial neurons (synapses) that are connected (by edges) in order to transmit signals. The signals are in the form of real values, and the neuron's output is computed using some non-linear functions of the sum of its inputs. The weights are modified as learning progresses to increase or reduce the signal strength. Typically, neurons are aggregated into layers, with each layer performing a unique transformation on their inputs. Similar to a linear regression, the algebraic formula of a simple neural network is represented as:

$$\sum_{i=1}^{m} w_i x_i + bias = w_1 x_1 + w_2 x_2 + w_3 x_3 + \cdots + w_n x_n + bias$$

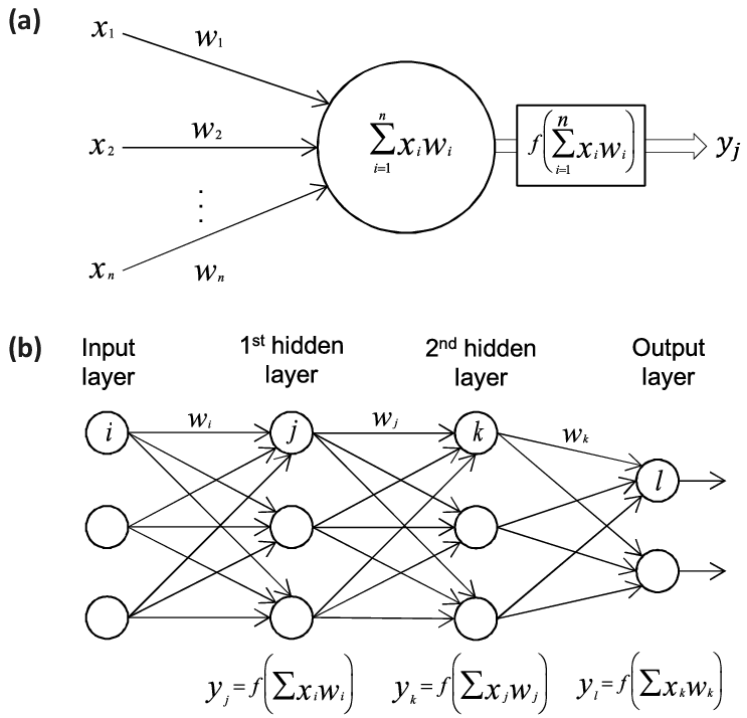


**Figure 3.9**: (a) Illustrate a typical deep neural network building block. (b) A feedforward multilayer neural network is depicted (also known as multilayer perceptron)
Image source: (Vieira et al., 2017).

For a classification problem, a neural network objective function is given by:

$$output = f(x) = \begin{cases} 1 & if \sum w_i x_i + b > 0 \\ 0 & if \sum w_i x_i + b < 0 \end{cases}$$

Three interconnected layers constitute an ANN: an input layer, an inner layer, and an output layer. The input layer receives external data, the output layer generates the desired result, and

between them are zero or more hidden layers. The hidden layer may be one-layered, unlayered, or multi-layered. The connections between layers can be *fully connected* (i.e., each neuron in one layer is connected to every neuron in the next layer), or *pooling* (in which group of neurons in one layer connects to a single neuron in the next layer) (Ciresan et al., 2011). Neurons connected in this manner form an acyclic graph, often known as a *feedforward network* (Zell, 1994). Additionally, *recurrent networks* are those that permit connections between neurons in the same or preceding layer (Miljanovic, 2012). Each layer aids the ANN in understanding an object's complex structure.

Depending on the learning rule, an ANN may employ the *backpropagation* process to adjust and compensate for any errors found during training. Backpropagation makes sure that any output with incorrect-labels is routed back through the layers and that the weights are updated/recalibrated, via stochastic gradient descent (Bottou and Bousquet, 2007) (or other techniques), in proportion to their contribution to the errors. Technically, backpropagation computes the gradient of the loss function with respect to the weights. As training progresses, the NN learns how to minimize the likelihood of errors while accounting for the difference between the desired and actual results.

Below, we briefly discuss other components of an ANN

a) **Hyperparameter (HP):** is a constant parameter that is initialized before neural network training begins. In some cases, the values of these parameters are determined by training. Hyperparameters are important because they have a crucial impact on the performance and the behaviour of the training model. Common HPs include, batch size, learning rate, and number of hidden layers (Lau, 2017). Certain HPs may be reliant on the values of others; for example, the size of some layers could be dependent on the overall number of layers.
b) **Learning:** entails modifying the network's weights in order to improve the result's accuracy. Typically, this is accomplished by minimizing observed error, and when observed errors cease to decrease, learning is complete. While errors in the learning process do not have to be zero to be optimal, if they are, the network should be redesigned. Learning involves defining a loss (cost or objective) function that is evaluated on a periodic basis throughout the learning process and continues as long as the error rate keeps decreasing.
c) **Learning rate (LR):** is a tuning parameter that sets the size of the corrective steps made by the model to compensate for observed error at each iteration while trying to

minimise the loss function (Murphy, 2012). It represents the rate at which a model learns; a high LR reduces training time but results in lower accuracy, whereas a low Lr increases training time but results in potential higher accuracy. However, there is a trade-off between a high and a low LR. A too high LR causes the learning to hop over minima, while a low LR either takes too long to converge or causes the learning to get trapped in an unwanted local minimum (Buduma and Locascio, 2017). To eliminate alternating connection weights and to improve convergence, it is currently common to adopt an adaptive LR that increases or decreases appropriately throughout training (Li, 2009).

d) **Cost Function:** or a loss function, is the function that must be minimized in a leaning model to obtain the optimal solution. It is usually used to measure the performance of ML models. A cost function, depending on the desirable properties of the problem (e.g., regression, binary classification, multi-classification, etc.), takes both predicted and actual output and calculates how far off target the model was in its prediction. A high-valued cost function may indicate how far the prediction is from the actual value. As the model hyperparameters are tuned during training, the cost function represents the degree to which the model has improved.

e) **Modes:** the learning mode can be stochastic or batch. The former produces a weight adjustment for each input, but the latter creates weight adjustments across the batch, accumulating errors. Given that adjustment is conducted in the direction of the batch's average error, batch mode is more stable moving toward the local minimum. A popular compromise is to utilize simple batches with samples stochastically selected from the dataset collection for each batch. This is referred to as "mini-batch."

## 3.2 Digital Evidence Mining (with AI)

After discussing the detailed background of AI and its corresponding components, this section will focus on the application of AI to digital forensics — specifically, on the extraction and analysis of evidence. We refer to "digital evidence mining" in this thesis *as the process of automatically identifying, detecting, extracting, and analyzing digital evidence with AI-driven techniques*. The phrase "*mining*" is borrowed from the notion of data mining, which embodies procedures and components that can be applied to the analysis of digital evidence. Certain significant challenges in digital forensics necessitate the adoption of AI techniques. AI methods have proved promising on a variety of complex tasks; particularly with ML and

NN algorithms which are capable of discovering hidden patterns in massive heterogeneous data. These challenges include the following: 1) increasing availability of distributed systems; 2) exponential growth in the size and capacity of storage devices; 3) the pervasiveness of electronic devices; and 4) the inability of existing forensic tools to keep up with the diversity and complexity of cybercrimes.

In a conventional manual procedure, investigators will often familiarize themselves with the domain before generating reports that summarize their insights about the acquired artifacts (Mohammad, 2019). Consequently, investigators mostly rely on their experience to interpret findings. For example, manually comparing samples to determine if they share enough attributes to indicate a common source, or whether a particular set of data exhibits a persuasive pattern to establish probative facts, might be subjective. Subjective assessments imply that various examiners analyzing the same evidence will reach different conclusions or make varying assessments at different points in time. It is, however, worth highlighting that experience should never be used in place of experimentation, particularly when the ground truth is known (Carriquiry et al., 2019).

The collection, categorization, and revision of evidence is the first step in evidence analysis, after which the probability of a crime (and prospective perpetrators) is hypothesized. The possible proofs that support these hypotheses are elicited, and the proof is then presented in a court-acceptable format. Examining fragmented knowledge and establishing complex scenarios that often incorporate time, uncertainty, causality, and possibilities is what evidence analysis entails (Constantini, Giovanni and Olivieri, 2019). Most DF analysis methods are incapable of extracting evidence in a holistic manner that considers all of the significant components in the artifacts, including the causal events. As a result, the analysis may be incomplete, and vital evidence may be overlooked. Another issue with existing investigative methods and practices is that they take an inordinate amount of time and require unnecessary human involvement to accomplish.

As stated earlier in section 1.4, that a strong investigative analysis should place objects, activities, and time in a space that allows for complete representation of the data so that meaningful reasoning can be inferred. AI methodologies can apparently provide a reasonable approach to tackle most of the DF challenges highlighted above. Additionally, AI's ability to discover potential regularities (or irregularities thereof) in a vast amount of complex, widely disparate, data in a reasonable length of time (Faye, 2010), with little or no human involvement,

makes them a choice candidate in mining evidence. Arguably one can subtly agree to the notion that the stages involved in conducting a DF investigation are similar to the steps needed in creating ML models (Mohammad, 2019). Various learning algorithms exist that can interact with digital artifacts in ways that enable reproducible experimentation and the identification of critical clues from data in a meaningful and visualizable manner. In what follows, we examine several prospective applications of AI algorithms to digital evidence mining, as well as existing approaches that have been proposed or implemented for the same goal. In this context, the term AI algorithms refers to all forms of AI approaches, including symbolic and sub-symbolic.

### 3.2.1 Network Data Analysis in Evidence Mining

Until recently, cyber security experts dealt with threats (or responded to incidents) such infiltration/intrusion in network, cloud, IoT, and mobile devices, or fraudulent activities by using anti-malware or antivirus software, as well as a firewall with specific rules (Brighi, Ferrazzano and Summa, 2020). However, today's threats are so advanced that they may be able to circumvent standard security measures. This security failure could be ascribed to a lack of expertise, the accuracy of the malware detection systems, or the time required to detect or investigate daily attack threats, among other things. AI algorithms, with their advanced pattern recognition abilities, can analyse millions of log files in a reasonable amount of time and find in any data cluster, an "atypical" behavior in files (Brighi, Ferrazzano and Summa, 2020). In fact, in a well-configured environment, a machine learning classifier trained on massive historical, labelled data can detect a malicious or anomalous entry in real-time. There exist a number of literatures proposing different AI techniques to solving several intrusion attacks. Anuradha (Anuradha and Padmavathi, 2019) detected botnet intrusion using DNS query data modelled with ML. Recognizing that signature-based or white/blacklist methods are ineffective in combating botnets on digital devices and social platforms, Singh, Singh and Kaur demonstrated the ability to detect "Bot infections" as opposed to bot detection, in a network by identifying anomalous patterns in DNS traffic using DNS fingerprint for each host, modelled with Random Forest classifier in (Singh, Singh and Kaur, 2019). In a similar vein, Alauthman et al. (2020) claimed that bot detection in a real-time, high-speed, voluminous network may be unachievable in the absence of a method to reduce the dimensionality of network traffic data. As a result, they introduced a network traffic reduction technique based on neural networks that reduces training time while also increasing the learning rate of newly extracted features in an online system. They also modelled a decision tree classifier to

accurately detect new bots. Pour et al. have introduced a novel large IoT data dimensionality reduction technique based on $\ell^1$-norm PCA (Kwak, 2008) which is applied on passive measurement data to infer, describe, and report maliciously exploited and coordinated probing activities on IoT devices (Pour et al., 2019). In network packets or logs, indicators of compromise (IOC) are also evidence artifacts indicating that a system is compromised (Xiaoyu et al., 2020). The authors in (Chiadighikaobi and Abdullah, 2017) developed an approach to identify the source of an attack utilizing a behavior malware analysis framework and a *k*-means clustering algorithm to generate IOC rules to detect malware. Murtaz et al. (2018) used multiple ML classifiers such as Random Forest, *k*-Nearest Neighbour, Decision Tree, and Regression to detect malware in Android network traffic.

Depending on the detection system design, ML/ANN models are capable of detecting unusual behaviour, and consequently either flag instant report, prohibit malicious entry, or log such reports so that human experts (or the system itself) can analyse its performance and perhaps provide reward in the form of reinforcements. As far as we understand, there is no detection tool with the aforementioned capacity that does not rely on AI learning techniques. This underscores the importance of implementing AI in such a domain.

### 3.2.2 Timeline/Event Reconstruction in Digital Evidence Mining

The conversion of the state of digital objects to their causal events is referred to as event reconstruction (Carrier and Spafford, 2004). Typically, a data object will reveal functional, relational, and temporal relationships with various events (Mohammad, 2019). To reconstruct events during a DF investigation, various factors must be addressed, including possible (causal) correlations, the context of the suspicious activities, the suspects' "*modus operandi*," and, most significantly, the time of the event (Constantini, Giovanni and Olivieri, 2019). Most digital items (or artifacts) have recoverable timestamps, though can be volatile because they change often in reaction to activities/events on the digital object (influenced by users or systems). However, a more extensive examination of sources such as the Windows Registry, event logs, database logs, etc., can reveal usable activity history, based on timelines, that can aid in event reconstruction — that is, what happened when, and sometimes, by who. Event reconstruction methodologies on huge volumes of artifacts, according to Chabot et al. (Chabot et al., 2015), must meet the following requirements:

- Automatic reconstruction and analysis of multiple event timelines. This will, however, necessitate the encoding of data in a machine-readable format, as well as a comprehensive approach to dealing with multiple heterogeneities. Heterogeneity, in this case, refers to identifying relationships by examining data from numerous sources such as file systems, Windows event logs, file metadata, server logs, web browser history, Memory dumps, and so on.
- Making timeline data easier to understand by using tools that can interpret, analyze, and detect correlations between events, as well as draw conclusive inferences from the artifacts.
- The availability of tools that allow for the easy and intuitive search and visualization of data.

Event reconstruction could be complex, especially given the temporal component of digital artifacts. For example, if timestamps are not expressed as a vector, they may have a wide range of variations. Scaling the data within the range [0...1] or [-1...1] is a common technique (Mohammad, 2019). Due to unsynchronized clocks, time zone variances, and differing system file time formats, examining artifacts from different sources may cause timing issues (Chabot et al., 2014). Nevertheless, several methods have been proposed (Chen et al., 2003; Chabot et al., 2014; Chabot et al., 2015) to address such issues, some of which are also based on symbolic or rule-based representations of digital artifacts as presented in (Turnbull and Randhawa, 2015) to detect timeline inconsistences. In this area, there are numerous prospects for AI applications. In fact, AI algorithms can meet all of the requirements outlined by Chabot et al. above. We briefly explore various related literatures on the application of AI to event reconstruction during an investigation. Khan et al. presented a method for characterizing the use of various application programs by monitoring and capturing file system changes at discrete timelines. The recorded data was then used to train a feedforward and RNN model to distinguish instances of application execution. The RNN model showed improved accuracy because the network could frequently correlate different inputs, which is critical for recognizing time series relationships (Khan, Chatwin and Young, 2007). Similarly, Khan presented the comparative effectiveness of Bayesian probabilistic networks and NNs to identify file system manipulation during a specific time period, in an experiment to reconstruct post-event timelines of unauthorized system access. The Bayesian network appeared to be more suited for such task, because of its ability to stochastically represent data to learn from prior knowledge and detect hidden patterns from an incomplete dataset (Khan, 2012). Studiawan et al. (2020) employed a

sentiment analysis technique modelled with DL (word embedding with context and content attention layer) to detect aspect phrases and the associating sentiments in a forensic timeline to identify "Events of Interest (EOI)" from message logs (Studiawan, Sohel and Payne, 2020). The objective is to establish a class of positive and negative messages, with negative sentiments signalling the presence of EOI and are being highlighted in the timeline. This can provide investigators with insights to further investigate the activities within the surrounding timeline.

There are numerous areas where AI could be applied to event timeline reconstruction, particularly in the detection of anomalous or deviant behavior in log files, databases, system files, etc. Nonetheless, anomalous behavior in a digital system is mostly dependent on a heuristic definition of what is legal or illegal – the distinction between which can only be determined via additional analysis of surrounding activities.

### 3.2.3 Pattern Recognition in Digital Evidence Mining

Given that the notion of pattern recognition was fully covered in subsection 3.1.2.1, we will focus on its divergence into DF and its application to evidence extraction in this part.

Pattern recognition is a fundamental tenet of AI; not necessarily a distinct stage in DF. Nonetheless, it is a valuable approach for examining digital artifacts. Crimes are characterized by a series of acts that, at times, follow a self-consistent, traceable pattern, while at other times, particularly when experienced criminals are involved, the patterns can be complex and difficult to trace. Additionally, there may be some distinct dynamics underlying the events surrounding the commission of a given crime, which may appear disparate or disjointed in the real sense (particularly when examined manually) but are connected in unimaginable ways. Digital crime analysis relates to identifying and correlating fragmented pieces of digital artifacts that can aid in establishing facts. Consequently, when digital artifacts are as large as they are today, it is nearly impossible to manually search for evidence clues. It thus necessitates the need for a robust mechanism capable of sifting through the data and identifying significant hidden relationships that can lead investigators to factual information. Pattern recognition is a suitable method for such complex task. Literally, it is a scientific discipline that deals with the automatic detection of regularities in data and the classification of data into various categories (Bishop, 2006). Pattern recognition has proved successful in examinations involving texts, images, audio, deep video, among others. For example, it can be used to determine whether a pattern in a disk image indicates that it is a component of a sound file (Faye, 2010) or to identify text patterns that appear frequently in phishing (Morovati and

Kadam, 2019) or SPAM (Santos et al., 2012) e-mails. Other areas of application include the detection of textual stylistics (linguistic analysis) in e-mails for the purpose of authorship attribution (Farkhund et al., 2008; Bogawar and Bhoyar, 2016; Emad et al., 2019), as well as the detection of objects, incriminating content, or abnormal behavior in video surveillance (Jianyu, Shancang and Qinglian, 2019). A full survey of digital video forensics can be found in (Javed, et al., 2021).

Historically, pattern recognition systems were classifiers (which are largely supervised); that is, they determine if a piece of data is a member of an "object of interest," X. Then they attempt to match all possible (or as close as computation allows) pieces of similar data to X, with enough generality to match all positive examples but enough specificity to exclude all negative examples (Faye, 2010). However, the breakthrough with DNN means that the same operations may now be accomplished with greater accuracy even without prior knowledge of the characteristics of the digital artefacts. It is simply a matter of feeding a neural network algorithm multiple disparate, but properly structured, digital artifacts and let the model deduce meaningful patterns from them. The model's accuracy and the interpretability of the regularities detected would therefore be contingent on the quality of the fed examples and, to a considerable extent, the interpreter's (investigator's) experience. The following are some of the techniques that may be used singly or in combination to recognize patterns in digital artifacts:

- ***Entity extraction*** is though heavily reliant on a massive amount of input data, it gives basic information for crime analysis. It can be used to extract patterns from textual data, image files, audio files, among other sources. In literatures, a neural network-based entity extractor was utilized in conjunction with entity extraction techniques to extract personal information from police reports, such as a person's attributes, addresses, drug history, and so on (Chau, Xu and Chen, 2002). As described in (Spafford and Weeber, 1993), viruses typically leave a trail of codes in infected systems; hence, code that remains after an attack may include source code, object code, executable scripts, and so on. As a result, MacDonell et al. stated that programmers, to a degree, have different coding styles that are readily identifiable by code analysts, given sufficient coding samples (MacDonell, et al., 2002). The authors also demonstrated that software metrics[101] combined with psychological and linguistic analysis of codes, and ML

[101] Measurement made on software program to assess user satisfaction, degree to which comments is source code matches comments, ratio of statement lines to blank lines, etc.

algorithms (to recognize patterns) are sufficient for identifying, characterizing, and discriminating malicious code writers.

- ***Clustering techniques*** assists in grouping data items with comparable characteristics in ways that minimize or maximize similarity overlap. Crime analysis may require that suspects (or perpetrators of crime) be identified using the same *"modus operandi",* or that groups belonging to distinct networks be distinguished. Due to the fact that the majority of clustering approaches are unsupervised, they do not require predefined class labels. Rather, the statistics-based algorithm identifies associations and links between items such as organizations, individuals, crime patterns, among others (Hauck et al., 2002; Gani, Hacid and Skraba, 2012; Shao et al., 2019). Clustering is also used in link analysis, which is crucial for investigating financial crimes and money laundering (Senator et al., 1995; Rouhollahi, 2021, Dasaklis and Arakelian, 2021), as well as drug cartels and extremist networks (Florea et al., 2019). Section 3.1.2.4 (B) has a detailed description of the clustering technique (and its variants).
- ***Association rule mining*** can identify patterns in sets of regularly occurring objects in digital artifacts and report them as rules. Its application in network intrusion detection has proved important in identifying profiles and detecting potential network attacks by deriving association patterns from users' interaction history (Lee, Stolfo and Mok, 1999; Mabu et al., 2010; Hyeok, Cholyong and Ryang, 2016; Safara, Souri and Serrizadeh, 2020). Similarly, "*sequential pattern mining"* falls under this category, as it identifies a sequence of frequently occurring items over a series of events that occurred at various times. This approach could be extremely valuable for detecting and reconstructing intrusion patterns with respect to timestamps.
- ***Classification*** as a pattern recognition technique is a widely used "digital forensic AI" methodology for identifying common characteristics among various crime entities and organize them into predefined classes. For example, classification has been used to determine the origin of e-mail spam based on the senders' linguistic styles (De Vel, 2001; Santos et al., 2012); differentiate genuine and fake multimedia files in order to detect the presence of deepfake content (Ferreira, Antunes and Correia, 2021a); and to determine whether a computer file system has been manipulated by a specific software program (Mohammad, 2019). Additionally, predictive crime analysis can make use of classification approaches. However, the technique requires both predefined class labels and high-quality training and testing data (Chau, Xu and Chen, 2002).

- ***Social network analysis*** enables visualization of criminal networks, by detailing the roles and interconnections of nodes in a conceptual network. Numerous criminal activities occur in cyberspace, including identity theft, public defamation, cyber stalking, personal data theft, fraud, and gangsterism (Baca, Cosic and Cosic, 2013). Some of these crimes (or the perpetrators' interactions) take place on social media platforms and determining the entities responsible for these acts would involve an investigation of multiple links, roles, the movement of tangible and intangible objects, as well as the association between these entities. A thorough forensic examination of the suspects' interactive behaviours can identify crucial roles, subgroups, and penetration loopholes in the network (Chau, Xu and Chen, 2002).
- ***Deviation detection*** is also known as "outlier detection" — is a technique that employs certain techniques to analyze data that deviate significantly from the rest of the data. It is commonly used in fraud and intrusion detection, as well as to spot missing patterns in data. However, depending on the learning algorithms and the data structure, a deviating behavior may appear normal when visualizing results, making it difficult to discover anomalies.
- ***String comparators:*** is mostly used to analyse textual data — it compares pairs of textual records and calculates their similarity. The approach is capable of detecting misleading information in criminal records, such as names, Social Security numbers, and dates of birth (Wang, Chen and Atabakhsh, 2004).

### 3.2.4 Knowledge Discovery in Digital Evidence Mining

Knowledge discovery, also known as Knowledge Discovery in Databases (KDD), is another branch of artificial intelligence that benefits digital evidence mining. It is also commonly used in electronic discovery — which is the process of organizing forensically acquired data into information that is understandable, replicable, and available to all parties involved in a court proceeding (Krishnan and Shashidhar, 2021). KDD is a term that refers to the process of extracting valuable data from a big collection of data (or in this case, digital artifacts). It encapsulates the integration of artificial intelligence, statistics, and probabilistic methods with the purpose of discovering meaningful representations of data that can aid in the detection of valid, novel, valuable, and meaningful patterns in huge and complex datasets (Maimon and Rokach, 2005). Unlike the majority of knowledge representation methods, KDD requires little or no background knowledge of the domain for which digital evidence is sought (Faye, 2010).

Additionally, the heavy lifting involved in a standard AI technique's complex representation of knowledge may be ineffective in KDD (Faye, 2010). Moreover, KDD is exploratory in nature; it frequently requires human (in the middle) intervention to issue query-like instructions such as "identify item with *X* properties that is connected to event *Y*." If the output is well structured, the human expert can perceptually detect patterns, albeit this can occasionally result in the identification of non-existent patterns. In reality, KDD continues to be one of the most helpful AI approaches in the legal domain (DF inclusive) and for large-scale data analysis. They should, however, be utilized with extreme caution, as they may overlook significant evidence due to the lack of background knowledge or advanced reasoning abilities.

It is typical in knowledge discovery (and, of course, DF) to examine more of the object's metadata because it is commonly not visible to the users and, unless tampered with, provides a rich artifact for DF investigation. Metadata, which literally translates as "data beyond data," is structurally embedded within digital files. It is frequently created and updated automatically by application programs (unless altered explicitly by a human), operating systems, or Malware. Metadata contains a variety of properties that vary according to its source — which could be file systems, images, documents, or the Internet (browsers and web pages). Explicitly, a document's metadata (which varies depending on the application program) includes information such as time stamps, last changed timestamps, the time period during which an edit occurred, the file hash, the author, and the computer that created the document. Similarly, an image file will have embedded metadata such as product/manufacturer details, lens information, time/date of creation, geographical coordinates, pixel information, among others (Larry and Lars, 2012). Meta-tags, programming language, page rendering intent, static or dynamic content production are all examples of web page metadata that may be relevant for digital investigations. However, the metadata associated with web pages is not as relevant to investigators as that of internet browsers (Krishnan and Shashidhar, 2021). For instance, browsing history, cached passwords, and search records are all examples of browser metadata that can be extremely relevant for investigations. The primary goal of the KDD technique in DF is to analyse all of these components in order to extract meaningful knowledge that may be used to derive inculpatory or exculpatory conclusions.

### 3.2.5 Fingerprinting in Digital Evidence Mining

For provenance analysis and object authentication, device or machine fingerprinting is frequently utilized. Device fingerprinting is the process of obtaining information about an

electronic device that uniquely identifies it. Typically, device fingerprinting is used for legitimate purposes such as preventing fraud or unwanted access to systems. When systems with fingerprinting mechanism are accessed, information such as the browser version, IP address, OS version, system fonts, screen resolution, HTTP cookies, and GPS location are recorded. Subsequent access information is compared with the previously logged data to validate identity or authenticate transactions. There are instances where malware bots are programmed to repeatedly access systems, and when this occurs, equivalent countermeasures to identify and block these malwares should be implemented. Gandotra et al. (2014) provided an overview of numerous malware classification techniques based on the behavior of static or dynamic malware (Gandotra, Bansal and Sofat, 2014), as well as malware author attribution (Alrabaee et al., 2017). In multimedia provenance, images can be analyzed to determine whether or not it was created by a particular camera. A minor, unnoticed defect in the sensor of a digital camera might leave imprints on the images produced (Xiaoyu et al., 2020). As a result, when analyzing photographs, it is possible to connect their content with a specific camera sensor (Lukas, Fridrich and Goljan, 2006; Freire-Obregon, 2017). Tsai, Lai and Liu demonstrated how to train and categorize image characteristics using SVM on similar image scenes captured with a conventional and a mobile phone camera (Tsai, Lai and Liu, 2007). Their model correctly identified the image's originating camera. CNN have also demonstrated encouraging results on a variety of image, sound, and video recognition/analysis tasks (Karpathy et al., 2014; Hijazi, Kumar and Rowen, 2015).

## 3.3 Chapter Summary

We discussed the history of artificial intelligence in this chapter, from its inception to the concept of symbolic and sub-symbolic reasoning. The former, which is predicated on logical rules premised on certain consequence, exemplifies expert systems and case-based reasoning. While the latter refers to a broader concept of complex relationships or correlations between data points. We discussed in depth various types of sub-symbolic reasoning AI, such as ML, DL, DNN, and ANN, as well as introduced learning algorithms using supervised and unsupervised techniques. Furthermore, we discussed the potential divergence of machine learning algorithms for DF analysis and evidence extraction. Additionally, we cited cases from the literature in which learning algorithms and neural networks aided in the mining of evidence or provided indications for further analysis.

# Part II

# Pattern Recognition and Reconstruction for Evidence Mining

Chapter 4

# Pattern Recognition and Reconstruction: Evidence Mining from Unstructured Data

Textual communication data continues to be a valuable source of evidence; it comes in a variety of forms, including text messages, tweets, documents, and e-mails. In this chapter, we describe our neural network-based approach to pattern recognition in unstructured textual communication data, as well as the reconstruction of time series event to detect potential evidence concealment. Our method is based on the Variational Graph Autoencoder (VGAE) (Kipf and Welling, 2016), constructed with a multi-featured dynamic graph to represent the temporal evolution of e-mail exchanges between multiple user pairs. Our approach makes use of e-mail metadata and contents extracted from the body with NLP and text mining (Tan, 1999) techniques. The primary goal of our work is to aid the partial automation of the detection of suspicious e-mail deletions during an investigation. Thus, the constructed model is aimed at detecting missing graph edges, which we interpret as probable deletions, and reconstructing the edge attributes from which we infer the deleted messages' topics.

In subsequent sections, we briefly introduce the context of the problem including the related works. We then discuss our approach which can be divided into three key parts as follows:

- Methodology for building graphical representation of e-mail collections
- VGAE-based method for detecting e-mail deletion and topic reconstruction
- Experiment, evaluations, and results.

## 4.1 Mining Evidence from E-mails: the Context

E-mail is a fundamental mode of communication, as it enables the transmission of text messages, documents (confidential, legal, commercial, etc), and the conduct of transactions. However, it is a major source of a number of criminal activities, including illegal file sharing and malware transmission.

Discovering probative evidence in a vast volume of data is a difficult process that requires systems capable of identifying and visualizing digital cues (Caviglione, Wendzel and Mazurczyk, 2017). As a result, mining evidence from a large pool of semi-structured and

unstructured data, such as e-mails, will require analytical methods capable of fusing objects (e.g., e-mail users), activities, and time into a multidimensional space that enables the reconstruction of events that may be suggestive of prior activities (Solanke, Chen and Ramírez-Cruz, 2021). The majority of proposed efforts on e-mail forensics have focused on the analysis of e-mail headers (Miyamoto, Hazeyama and Kadobayashi, 2008; Guo, Jin and Qian, 2013), which has proved beneficial in detecting crimes such as spam, phishing, and spoofing (Aparna and Dija, 2015; Shukla, Misra and Varshney, 2020). Meanwhile, certain textual analysis approaches have made use of NLP to discover and prove the attribution of crime or criminal intent (Studiawan, Sohel and Payne, 2020). In general, we can analyze a few literatures covering e-mail analysis and the proposed automated approaches. In (Miyamoto, Hazeyama and Kadobayashi, 2008), a method for detecting virus-infected e-mails using e-mail headers was proposed. The authors in (Morovati and Kadam, 2019) discussed techniques for detecting and classifying "phished" and benign e-mails using a variety of machine learning classifiers (SVM, k-nearest neighbours, etc.). The classifiers exploited features such as common text patterns in phishing e-mails and potentially purposeful spelling errors. Additionally, (Santos et al., 2012) presented a spam filtering algorithm based on word frequency in the subject and body. However, the technique omitted two crucial parts of the metadata: the sender and the time stamp. ‘Holmes’ (Peilun, Fan and Hui, 2021) is a recently proposed semantic-based e-mail anomaly detection method that intends to discover e-mail threats that usually evade enterprise anti-spam systems’ detection. Their strategy is based primarily on the information contained in e-mail headers. In (Mrityunjay, Chauhan and Gupta, 2017), an autonomous e-mail analyzer architecture was proposed that can monitor the content of e-mail headers in real-time and raise a suspicion flag if a certain collection of strings is identified. Uma and Nikkath developed a machine learning-based prototype dubbed “Enhanced Forensics Fuzzy C-Means Clustering (EFFCM)” (Uma and Nikkath, 2021) for the purpose of detecting potentially illicit materials in e-mails. Similarly, several authors (Farkhund et al., 2008; Bogawar and Bhoyar, 2016; Emad et al., 2019) have presented techniques for determining authorship or intent in e-mails through stylistics and discourse analysis.

Criminal intent or activities may occasionally involve accomplices attempting to conceal incriminating messages; thus, having a tool that can detect e-mail deletions may be useful in determining criminal intent. From a novel standpoint, our work benefits from the uniqueness of the manner in which we integrated the semi-structured information from the headers with the semantic features extracted from the (unstructured) textual content of the bodies. This

supports the richness of our representation of the e-mail collections. Additionally, and in contrast to other works utilizing e-mail artifacts, we seek to expose dishonest persons who conspire to delete evidence of discussion about a certain subject. Our approach is also distinct in the formalism with which the problem is addressed: VGAE – a sort of ANN utilized for creating genuinely synthetic data, reconstructing partially degraded images, and so on.

Hypothetically, we assume a case scenario where investigators are given access to specific e-mail artifacts, presumably after authorizations have been obtained. They consequently aim to ascertain: i) whether an e-mail exchange between two suspects was deleted; and ii) whether the deletion was made with the intent of concealing evidence of communication about a particular subject.

Below, we detail our thorough approach to resolving the issue.

## 4.2 Building Graphical Representation of E-Mail Collections

Our method combines the semi-structured information in the e-mail header and information extracted from the unstructured text in e-mail bodies to construct an attributed dynamic graph that represents the collection of e-mail exchanges. The metadata extracted from the e-mail headers include sender and receiver addresses, timestamps (at which e-mail was sent and received), labels (e.g., Cc, BCc), the subject (if any), etc., Explained below are the steps taken to construct the graphical representation of the collection of e-mails.

### 4.2.1 E-mail Collection Pre-processing

The e-mails are processed incrementally, while a log of triples of the form (sender, receiver, timestamp) is updated at each iteration. As a result, before extracting the content of a new e-mail, a check is made to ensure it is not a duplicate. This is important since a copy of the message sent will be stored in the sender's outbox (or sent items, as the case may be) and the receiver's inbox. Having both will amplify the noise.

In the message parsing process, we use Regular Expressions (RE) (Thompson, 1968) to match string patterns. RE are specially encoded text strings or sequences of characters that specify/match a string pattern; they are extremely valuable for extracting entities such as e-mail addresses, social security numbers, phone numbers, and web collection entities (Yunyao et al., 2008). Using RE to extract entities may be relatively simple; this is especially true for extracting popular strings such as those found in e-mail headers, as multiple samples exist to

accomplish this. Due to the complexity of matching patterns in the body of an e-mail, the string searching algorithm must be robust and adaptive.

To ensure that duplicates are removed correctly, we use RE to search the list of triples for patterns with the same (sender, receiver) pair. If a matched pair is found in the log, the new message is ignored if the absolute difference between the messages time stamps is less than two seconds. Mathematically, given two messages, $m_1$ and $m_2$ with corresponding timestamps $m_{t_1}$and $m_{t_2}$, respectively. The time difference $\mathcal{T} = m_{t_1} - m_{t_2}$; therefore, if $|\mathcal{T}| \leq 2\ seconds$, then the message is a duplicate. This method has been extensively validated and is highly effective in identifying duplicates in e-mails exchanged between pairs. Additionally, with RE, we eliminate messages that match the patterns of typical automated e-mails, as well as the ASCII representation of attachments included in the e-mail body, and textual contents attached to the end of forwarded or replied-to messages. These text fragments constitute duplicates of the original messages. Furthermore, we label each message to indicate if the receiver was the primary recipient or was added in Cc or Bcc, as well as whether the e-mail initiates a conversation, or it is a reply or forwarded message. Finally, we pass the de-noised contents of the non-removed e-mails to the text processing pipeline.

### 4.2.2 Text Processing Pipeline

We employ domain-specific heuristics and specialized tools from the NLTK (Natural Language Toolkit) library[102] to process the unstructured texts contained in the e-mail bodies. The steps illustrating the specific details of our text processing is represented in Figure 4.1 and the pipelines are described below.

1. **Punctuation removal.** Considering our use case, we carefully selected the punctuations that needed to be discarded. We utilize a specialized Python module ("*strings*") that searches for list of punctuations (or special characters) in text and automatically identify and remove them. The list of removed punctuations include:
2. **Text tokenization**. This is the process of segmenting text into smaller components known as tokens. It entails the breakdown of a complete sentence (or the entire content of a document) into simple or complex lexical terms/words, acronyms, abbreviations,

[102] https://www.nltk.org/

and alphanumeric expressions. Tokenization can be accomplished using specific tools such as sentence or word tokenizers[103] available in the NLTK libraries.

3. **Part-of-Speech (POS) tagging**. This is the process of associating a word in a textual corpus with its associated part of speech, as defined and contextualized by the word. POS tagging can be rule-based (Brill, 1992) or stochastic, and it is based on computational linguistics, which correlates discrete terms and hidden parts of speech with a collection of descriptive tags. We accomplish this task by utilizing the Stanford POS tagger[104] (Manning et al., 2014). In our use case, we retain all nouns (NNP, NNS, NN), verbs (VB), adjectives (JJ), and adverbs (RB), and then eliminate the other POS tags.
4. **Lemmatization.** Typically, this phase could be replaced by (or added to) stemming — a well-known technique for reducing words to their morphological roots. Lemmatization ensures that words are not severely stemmed to the point of meaning loss, and we chose it over stemming to avoid many terms with common stems collapsing into a single term. This is accomplished through the usage of the NLTK's WordNet-based lemmatizer[105].
5. **Stopword removal.** As with the terms that were already removed from the e-mail during the pre-processing phase, stopwords are further collection of frequently used words that do not contribute to the semantic understanding of the sentence or provide any value to its analysis. We employ NLTK's standard stopword list[106], supplemented with a few terms that are ubiquitous but uninformative in the context of e-mail processing, such as 'dear', 'attached', 'regards', and so on.

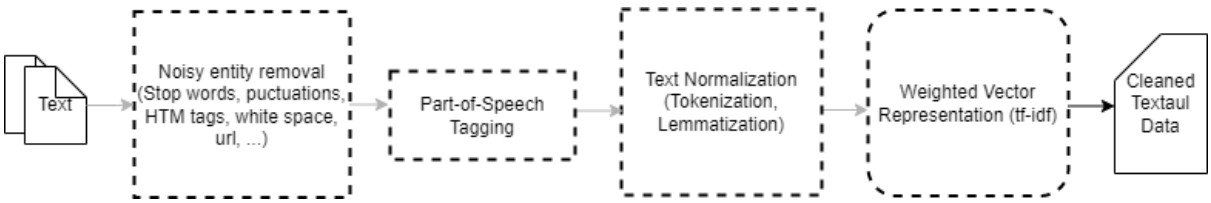


**Figure 4.1**: Text Pre-processing Pipeline

Following the text processing pipeline, we construct a weighted vector representation of each e-mail. To do this, we employ the widely used Term Frequency – Inverse Document Frequency (TF-IDF) technique (Juan, 2003; Robertson, 2004). TF-IDF was developed

[103] https://www.nltk.org/api/nltk.tokenize.html
[104] https://nlp.stanford.edu/software/tagger.shtml
[105] https://www.nltk.org/_modules/nltk/stem/wordnet.html
[106] https://www.nltk.org/nltk_data

primarily for the purpose of document search, keyword extraction, and information retrieval. It is defined as follows:

- **Term Frequency (TF)** is given by the number of times a term occur in a document – the more a term is represented in a document, the important it is. In our case, we compute it as the relative frequency of the term in each e-mail.
- **Inverse Document Frequency (IDF)** is conceptualized in terms of the frequency with which a word occurs in an entire collection of documents. It is based on the principle that the relevance of a term is inversely proportional to the number of documents in the corpus as a whole. The closer a term's $idf$ is to 0, the more common or important it is. Our inverse document frequency is computed as $idf(t) = \log\left({}^{N}/_{D_t}\right)$, where $D_t$ is the number of e-mails containing term $t$ and $N$ is the total number of e-mails.

Generally, TF-IDF is mathematically given by:

$$W_{t,d} = f_{t,d} * \log\left({}^{N}/_{D_t}\right)$$

Where $W_{t,d}$ is the weight of term $t$ in in document (e-mail) $d$; $f_{t,d}$ is the frequency of term $t$ in document (e-mail) $d$.

### 4.2.3 Dynamic Attributed Graph-Based Representation of E-mail Collections

We represent e-mail collections as a dynamic graph $\mathcal{G} = \{G^{(1)}, G^{(2)}, \dots, G^{(T)}\}$, where every snapshot $G^{(t)} = (\mathcal{V}^{(t)}, \mathcal{E}^{(t)}),\ t \geq 2$, represents all e-mails exchanged from time-step $t-1$ (not inclusive) and time-step $t$ (inclusive). That is, $G^{(1)}$ represents all communications until time-step 1 (inclusive). Each node $v \in \mathcal{V}^{(t)}$ represents an e-mail address, while with each edge $(v, \omega) \in \mathcal{E}^{(t)}$, we represent the e-mail(s) exchange between the addresses $v\ and\ \omega$ during the snapshot $G^{(t)}$. Our dynamic graph is undirected and organized in such a way that new nodes and edges (with their corresponding attributes) can appear and disappear at various time-steps. The graph is undirected because our model aggregates the entire body of e-mails exchanged during each time step, rather than individual messages, for each user pair.

All nodes at snapshot $t$ have a pre-defined number of attribute vectors represented as $F_t$, and we denote the node attribute by $\mathcal{X} = \{X^{(1)}, X^{(2)}, \dots, X^{(T)}\}$, where $X^{(t)}$is an $N_t \times F_t$ matrix such that $X_{ik}^{(t)}$ contains the value of the *k*-th entry of the F-dimensional feature vector of the *i*-th node

at time-step $t$. Similarly, the edge attributes are denoted by $\mathcal{E} = E^{(1)}, E^{(2)}, \dots, E^{(T)}$, where the $P$-dimensional vector in $E_{ijp}^{(t)}$ represent the probability distribution of a specific number $n$ of latent topics according to a topic-based language model (which we describe later in this chapter) fitted to the collection of e-mails exchanged between $v_i$ *and* $v_j$ in $G^{(t)}$. For every $(v_i, v_j) \notin E^{(t)}$, we set $E_{ij.}^{(t)} = [0]^n$.

### 4.2.4 Obtaining Semantic Edge Features with Probabilistic Language Models

Topic model (or Probabilistic Language Model (Bengio, 2003)) is a statistical method used for discovering 'latent' topics or hidden semantic structures present in a collection of documents. It is a non-rule-based, unsupervised techniques used for finding pattern of co-occurring terms in a large corpus of text. Fundamentally, topic models are generative models which are based on the rationale that the generation of a document is through the sampling of a collection of topics, from which words are sampled with topic-specific distributions.

We employ topic models to generate low-dimensional abstract feature vectors that describe the contents of each pair of users' collections of exchanged e-mails. Originally, we explored the option of encoding e-mail subjects to enrich the feature vectors but observed that it made an insignificant contribution and was ultimately unsuitable due to the large number of e-mails with blank, uninformative, or deceptive subjects. The abstract vectors contain information from the weighted terms (*tf-idf*) and the bag-of-words, both of which represent the entire e-mail collections exchanged between user pairs over a predefined time period.

Numerous techniques for topic modelling have been proposed, including *Latent Semantic Indexing (LSI)* (Dumais, 1994), *Probabilistic Latent Semantic Indexing (pLSI)* (Hofmann, 1999), *Latent Dirichlet Allocation (LDA)* (Blei, Ng and Jordan, 2003), and *Non-negative Matrix Factorization (NMF)* (Xu, Liu and Gong, 2003). *A priori*, any topic model should fit well with our model; nonetheless, based on empirical evaluation (which we describe later in this chapter), we experimented with LDA and NMF. Literally, LDA is modelled using Dirichlet distributions; it is based on a topic per document and words per topic model. It represents any collection of documents in a lower-dimensional vector space as a document-term matrix. The matrices are denoted as document-topic matrix $(N, K)$ and topic-term matrix $(K, M)$; where $N$ denotes the number of documents, $K$ is the number of subjects, and $M$ is the vocabulary size. Thus, for a collection of documents, LDA generates a probability distribution between 0 and

1, indicating the percentage of correctness for each of the topic in the selected number of topics. Figure 4.2 is a pictorial representation of the LDA model.

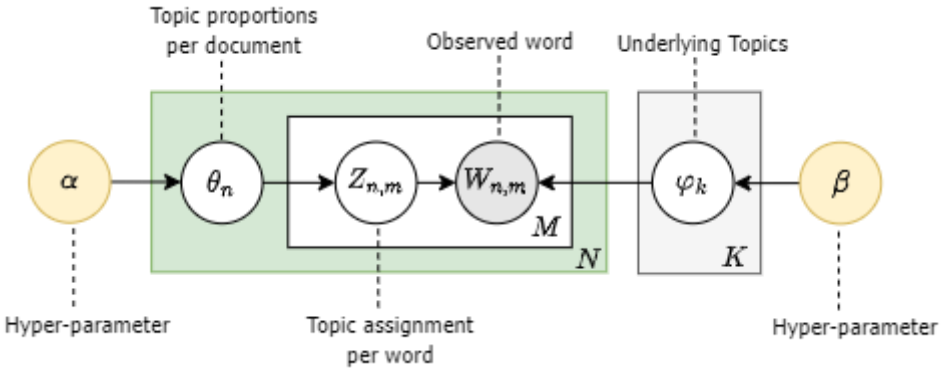


**Figure 4.2**: Graphical representation of an LDA Model

On the other hand, NMF is a statistical method that employs factor analysis to reduce the dimension of inputs by assigning a lower weight to words with a minimal degree of coherence. It is an unsupervised ML technique for identifying latent or hidden structure in data. One of the key features of NMF is its ability to automatically extract interpretable factors from sparse data. NMF operates by decomposing a $words\ X\ documents$ matrix $A$ into the product of a $words\ X\ topics$ matrix $W$ and a $topics\ X\ documents$ matrix $H$. Figure 4.3 graphically describes the NMF rationale. Mathematically, NMF is given by:

$$A = W * H$$

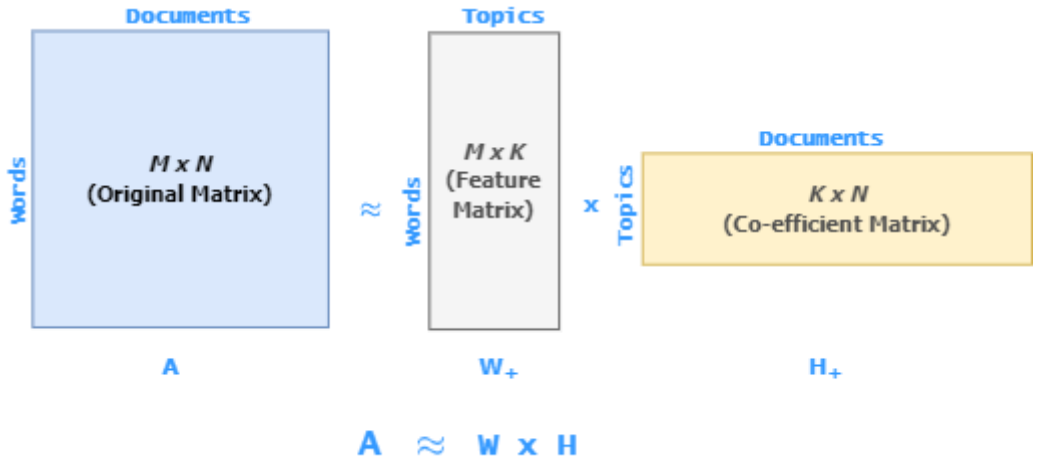


**Figure 4.3**: Graphical representation of the NMF model

### 4.2.5 VGAE-Based Method for Detecting E-mail Deletions

The theory is that by combining a VGAE with recurrent graph neural networks, it is possible to detect infrequent activities in e-mails (such as deletions) and partially reconstruct the edge vectors encoding the subjects of deleted e-mails. Our model, which takes e-mail exchange as input, is constructed using standard social network analysis and anomaly detection techniques, as well as autoencoders' reconstruction capabilities. Precisely, our approach uses conventional missing link prediction as a basis for deletion detection, and the interpretation of topic reconstruction is based on edge attribute reconstruction.

The VGAE model is an extension of the VAE (Kingma and Welling, 2013) model, with the addition of the requirement to reason about and act on graph-related structures. VAEs are a subset of AE. AEs are NN models widely used for synthesising realistic data, de-noising images, etc. The encoder in AE takes a data point $\boldsymbol{X}$ as input, converts it to a lower-dimensional representation $\boldsymbol{Z}$, *or embedding (or latent variables)*. Then the decoder takes the embedding $\boldsymbol{Z}$ and return the reconstructed representation of the original input $\widehat{\boldsymbol{X}}$. The distinction in architecture for VAEs is that the encoder learns a multivariate Gaussian distribution $\boldsymbol{q\varphi(z|x)}$ of the input data. The decoder samples the embedding from the latent space and reconstructs it to generate the output $\widehat{\boldsymbol{X}}$, which is a variational approximation $\boldsymbol{q\theta(x|z)}$. VGAE takes as input the adjacency matrix[107] and feature vectors representing the nodes (and in our model, also the edges) of a graph. Our initial intuition is that when a damaged communication graph is given into a VGAE-based model, the model learns the pattern of communications as it evolves over time during training. The model's output shall then reflect the reconstructed communication graph that will serve as the foundation for our deletion detection tasks. Additionally, while a standard VGAE can be represented with multi-dimensional node features, our model can include multi-dimensional edge features as an addition. The input to our model are:

*i.* the dynamic adjacency matrix $\mathcal{A}$ of the (dynamic) graph representing the e-mail exchange, is a sequence of matrices $\{A^{(1)}, A^{(2)}, \dots, A^{(T)}\}$, where $A^{(t)}$ is an $N_t \times N_t$ matrix, and $A_{ij}^{(t)} = 1$ if an edge exists between nodes $v_i$ and $v_j$ at the $t$-th snapshot, otherwise $A_{ij}^{(t)} = 0$;

*ii.* the node attributes $\mathcal{X}$, described in section 4.2.3;

[107] The adjacency matrix, also known as the connection matrix, of a simple labelled graph is a matrix with rows and columns labelled by graph vertices and a value of 1 or 0 in position ($v_i$, $v_j$) depending on whether $v_i$ and $v_j$.

*iii.* the edge attributes $\mathcal{E}$, also briefly described in section 4.2.3, is a sequence of matrices, where $E^{(t)}$ is an $N_t \times N_t \times P_t$ tensor, and we set the $p$-th entry of the edge attribute $E_{ijp}^{(t)} = [0]^n$, if $(v_i, v_j)$ does not exist, otherwise, the probability vectors (in this case, 10 values) of the latent topics (derived using probabilistic language model) between $v_i$ and $v_j$ are used.

To reduce the dimensionality of the graph and extract meaningful features, we use a variant of the Graph Convolutional Networks (GCN) (Kipf and Welling, 2016a) as proposed in (Chen, 2020). Below, we describe the convolutional architecture of the multi-dimensional weighted edge as proposed by Chen.

*i. Normalization of the edge weights*

The node degree matrix $D^{(t)}$ at time-step $t$ is denoted by an $N_t \times N_t$ diagonal matrix[108], where $D_{ii}^{(t)}$ gives the degree of the $i$-th node. Therefore, the edge features are normalized in a symmetric manner using the node degrees. However, contrary to the normalization technique in GCN, the adjacency matrix is replaced with $N_t \times N_t \times P_t$ edge weight tensor. Mathematically, the normalization method at time-step $t$ is given by:

$$\hat{E}_{ijp}^{(t)} = D_{ii}^{(t)\,-\,0.5} \,.\, E_{ijp}^{(t)} \,.\, D_{ii}^{(t)\,-\,0.5} \qquad (4.1)$$

Representing equation (4.1) in matrix form, we have:

$$\hat{E}_{..p}^{(t)} = D^{(t)\,-\,0.5} \,.\, E_{..p}^{(t)} \,.\, D^{(t)\,-\,0.5} \qquad (4.2)$$

*ii. Edge weights as convolution co-efficient*

$H^{(l)(t)}$ denotes the $N_t \times d_t^{(l)}$ matrix of hidden node states at the $l$-th layer at time-step $t$; with $H_{i.}^{(l)(t)}$ of the $i$-th row representing the ${d_t}^{(l)}$ dimensional hidden state vector of the $i$-th node in the $l$-th layer at time-step $t$. At the initial $l$-th layer $l = 0$, hidden layer $H^{(0)(t)} = X^{(t)}$ (the node attribute matrix at time $t$). At other $(l+1)$-th layer, for each of the $P$ channels of the edge weights, the hidden node state is obtained iteratively by performing a weighted convolution operation on $H^{(l)(t)}$ with $\hat{E}_{ijp}^{(t)}$ as convolution co-efficient, along with the usual weight matrix,

[108] diagonal matrix is a matrix in which the entries outside the main diagonal are all zero

$W^{(l,p)}$, of dimension $d^{(l)} \times \hat{d}^{(l)}$. This implies that, for each $p \in \{1, \dots, P\}$, an $N\ x\ \hat{d}^{(l)}$ matrix, $\widehat{H}^{(l,p)}$ is computed at time-step $t$ as:

$$\widehat{H}^{(l,p)(t)} = \sigma\left(\hat{E}_{..p}^{(t)} .\ H^{(l)(t)} .\ W^{(l,p)}\right)$$

*(4.3 )*

where $\sigma\ is\ non-linear\ activation\ function$; in our case, we use the Rectified Linear Unit (ReLU) (Nair and Hinton, 2010). Precisely, when given a negative input, the ReLU function returns 0, but when given a positive input, it returns the same number. It can be written as: $f(x) = \max(0, x)$.

*iii. Edge weight channel aggregation*

At each snapshot, we propagate information across the different edge channels and update the nodes by combining $\widehat{H}^{(l,p)(t)}$ for $p \in 1, \dots, P$ into the hidden node states of the next layer, $\widehat{H}^{(l+1)}$. In our task, we use the sum aggregation method which is given as:

$$H^{(l+1)(t)} = \sigma\left(\sum_{p=1}^{P} \widehat{H}^{(l,p)(t)} . W_{sum}^{(l)}\right)$$

*(4.4 )*

where $W_{sum}^{(l)}$ is a learnable weight matrix of size $d^{(l)} \times \hat{d}^{(l+1)}$.

Therefore, the entire update rule for a sum-aggregated multi-dimensionally weighted edge convolution method (with a bias) is:

$$H^{(l+1)(t)} = \sigma\left(\sum_{p=1}^{P} \sigma\left(\hat{E}_{..p}^{(t)} .\ H^{(l)(t)} .\ W^{(l,p)}\right).\ W_{sum}^{(l)} + b\right)$$

*(4.5 )*

The bias $b$ parameter enables the activation function to be shifted (to the right or left) by adding a constant to the input, analogous to the role of a constant in a linear function.

Finally, in the last layer $l = L$, the last hidden state at time-step $t, H^{(L)(t)}$ is then pass through a linear layer to obtain a global graph-level output.

We use Graph Convolutional Recurring Networks (GCRN) (Seo et al., 2018) to simulate the time series evolution of the communication network. GCRN combines GCN with Recurrent Neural Network (RNN) (Mikolov et al., 2010) to capture spatial temporal patterns in data.

Precisely, GCRN takes as input an adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$, and a sequence of node attributes $\mathcal{X} = \{X^{(1)}, X^{(2)}, \ldots, X^{(T)}\}$. GCRN then takes the $F$-dimensional node attributes $X^{(t)} \in \mathbb{R}^{N\times F}$, at each time step $t$, and updates its hidden state $h_t \in \mathbb{R}^p$; that is:

$$\boldsymbol{h}_t = f(\boldsymbol{A}, \boldsymbol{X}^{(t)}, \boldsymbol{h}_t - 1)$$

*(4.6 )*

$f$ represents a deterministic deep neural network in equation (6). In this experiment, we employ a recursive neural network known as the Gated Recurrent Unit (GRU), which was introduced in (Cho et al., 2014), to control the flow (remember and forget component) of temporal information across the hidden units of the network. Other recursive neural networks, such as the Long Short-Term Memory (LSTM) (Gers, Schmidhuber and Cummins, 1999), can be used. However, our choice of GRU was based on empirical evaluation (Chung et al., 2014; Gruber and Jokisch, 2020) that GRU performs better on certain smaller and less frequent datasets. A fully gated unit is given as:

$$z_t = \sigma_g(W_z x_t + U_z h_{t-1} + b_z)$$
$$r_t = \sigma_g(W_r x_t + U_r h_{t-1} + b_r$$
$$\hat{h}_t = \emptyset_h(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h$$
$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \hat{h}_t$$

*(4.7 )*

The variables in the equation(s) above denotes:

$x_t : input\ vector$

$h_t : output\ vector$

$\hat{h}_t : candidate\ activation\ vector$

$z_t : update\ gate\ vector$

$r_t : reset\ gate\ vector$

$W, U\ and\ b : parameter\ (weight) matrices\ and\ vector$

$\sigma_g : Sigmoid\ function; given\ as: \dfrac{1}{1+e^{-x}}$

$\emptyset_h : Hyperbolic\ tangent\ (tanh); given\ as: \dfrac{1-e^{-2x}}{1+e^{-2x}}$

$\odot : Hadamard\ product\ (i.e., an\ element - wise\ product\ of\ two\ matrices)$

Additionally, we combine the models discussed previously with a variation of VGAE proposed in (Hajiramezanali et al., 2020). The model, dubbed "Variational Graph Recurrent Neural Network (VGRNN)," combines GCN and RNN to produce GRNN (Hajiramezanali et al., 2020); which is a dynamic graph autoencoder model, as well as VGAE. GRNN can get different adjacency matrices at different time snapshots and reconstruct the graph and edge attributes at time $t$, by utilizing an inner-product and an edge attributes reconstruction decoder on the hidden state $\boldsymbol{h}_t$. The edge attribute reconstruction decoder is generated by passing the concatenated embedding of the edges $(v_i, v_j)$ through a fully-connected linear layer. In our implementation, $\boldsymbol{h}_t$ is designed as node and edge embedding of the dynamic graph at time $t$. The VGAE is integrated to further improve not just the time dependence of graphs, but also to represent nodes and their associated edge attributes in latent space using a stochastic distribution.

We extend the proposed model – which previously considered just one-dimensional edge features, to handle multi-dimensional edge features. Nonetheless, our strategies are sequentially based on the methods described in this work, but with a modified implementation that addresses our objectives explicitly. Noteworthy is that the following equations extend the rationale described in (Hajiramezanali et al., 2020).

A standard VGAE consists of three models: generative, inference, and learning, and we present our formulations, as well as the idea behind deletion detection and topic inference, as follows:

*A. Generation*

The generative model is conditioned on the recurrent hidden state variable $\boldsymbol{h}_{t-1}$, which accounts for the dynamic nature of the graph's topology and the time dependency of its node and edge features. The recurrence equation of the GRNN is then given by:

$$h_t = f\left(A^{(t)},\ \delta^x\left(X^{(t)}\right),\ \delta^z\left(Z^{(t)}\right),\ \mathcal{E}^{(t)},\ h_{t-1}\right) \tag{4.8}$$

where the function of $f$ is as described in equation (6). Likewise, $\delta^x$ and $\delta^z$ DNN used for independent feature extraction from $X$ and $Z$, at time $t$, respectively. Therefore, based on the recurrence of hidden state $\boldsymbol{h}_t$, the prior and the generating distributions can be factorized as:

$$p(A^{(\leq t)}, E^{(\leq t)}, Z^{(\leq t)} \mid X^{(\leq t)}) = \prod_{t=1}^{N_t} p\left(Z^{(t)} \middle| X^{(\leq t)}, E^{(\leq t)}, Z^{(\leq t)}\right)\ p(A^{(\leq t)}, E^{(\leq t)} \mid Z^{(t)})$$

*(4.9 )*

The prior distribution of the generative model at the first snapshot is $p\left(Z_i^0 \mid -\right) \sim \mathcal{N}(0,1)$ for $i \in \{1, \ldots, N_0\}$ and the hidden state at $t = 0$ is $\boldsymbol{h}_0 = 0$.

*B. Inference*

Mathematically, the values of the latent variables are inferred based on observed data as follows:

$$q\left(Z^{(t)} \middle| X^{(t)}, A^{(t)}, E^{(t)}, h_{t-1}\right) = \prod_{i=1}^{N_t} q\left(Z^{(t)} \middle| X^{(t)}, A^{(t)}, E^{(t)}, h_{t-1}\right) = \prod_{i=1}^{N_t} \mathcal{N}\left(\mu_{i,enc}^{(t)}, diag\left(\left(\sigma_{i,enc}^{(t)}\right)^2\right)\right);$$

$$\mu_{enc}^{(t)} = E\_GCN\_CONV_{\mu}\left(A^{(t)}, E^{(t)}, \odot\left(\delta^x\left(X^{(t)}\right), h_{t-1}\right)\right)$$

$$\sigma_{enc}^{(t)} = E\_GCN\_CONV_{\sigma}\left(A^{(t)}, E^{(t)}, \odot\left(\delta^x\left(X^{(t)}\right), h_{t-1}\right)\right)$$

*(4.10 )*

where $\mu_{i,enc}^{(t)}$ and $\sigma_{i,enc}^{(t)}$ are the i-th rows of $\mu_{enc}^{(t)}$ and $\sigma_{enc}^{(t)}$, respectively; $\mu_{enc}^{(t)}$ and $\sigma_{enc}^{(t)}$ denote the parameters of the approximated posteriors modelled as a standard multivariate Gaussian distribution $\sim \mathcal{N}(0,1)$; $\delta^x$ is a fully-connected neural network; and $E\_GCN\_CONV_{\mu}$ and $E_GCN_CONV_{\sigma}$are the encoder functions. Although they can be any pair of feature extraction models, we employ the variation of GCN specified in equation (5) to strategically model the multi-dimensional edge features in our implementation. If the hidden state variables $\boldsymbol{h}_{t-1}$ are not utilized, then the prior becomes independent across snapshots, resulting in a conventional VGAE model. Lastly, $\odot$ represent the concatenation of the node attributes at time $t$ with the recurrent hidden state variable.

*C. Learning*

Our objective function encodes the parameters of the generative and inference models used for edge reconstruction by summing all the joint maximizations of the expected likelihood of the input data with respect to its parameters at each time step. Due to the fact that the edge features are continuous, we define the objective function in terms of regression model; thus, we employ

the $\ell^2 - norm$ as a regularization term in the reconstruction loss. Learning model is described mathematically as:

$$\mathcal{L} = \sum_{i=1}^{N_t} \Big\{ \mathbb{E}_{Z^{(t)} \sim q\left(Z^{(t)} | A^{(\leq t)}, X^{(\leq t)}, E^{(\leq t)}, Z^{(\leq t)}\right)} \ell^2\left(E^{(t)} \middle| Z^{(t)}\right) - KL\left(q\left(Z^{(t)} \middle| A^{(\leq t)}, E^{(\leq t)}, Z^{(\leq t)}\right) || \, p\left(Z^{(t)} \middle| A^{(\leq t)}, X^{(\leq t)}, E^{(\leq t)}, Z^{(\leq t)}\right)\right) \Big\}$$

*(4.11 )*

Similarly, the objective function for the reconstruction of the adjacency matrix is the binary cross entropy between the target $\boldsymbol{A}$ and the output $\widehat{\boldsymbol{A}}$ – which is the log-likelihood of the original adjacency matrix, and it is given as:

$$\mathcal{L} = \sum_{i=1}^{N_t} \Big\{ \mathbb{E}_{Z^{(t)} \sim q\left(Z^{(t)} | A^{(\leq t)}, X^{(\leq t)}, E^{(\leq t)}, Z^{(\leq t)}\right)} \log p\left(E^{(t)} \middle| Z^{(t)}\right) - KL\left(q\left(Z^{(t)} \middle| A^{(\leq t)}, E^{(\leq t)}, Z^{(\leq t)}\right) || \, p\left(Z^{(t)} \middle| A^{(\leq t)}, X^{(\leq t)}, E^{(\leq t)}, Z^{(\leq t)}\right)\right) \Big\}$$

*(4.12 )*

Consequently, we adopt the inner-product decoder in our model, due to its wide use in the prediction of missing links (edges). Thus,

$$p\left(A^{(t)} \middle| Z^{(t)}\right) \prod_{i=1}^{N_t} \prod_{j=1}^{N_t} p\left(A_{i,j}^{(t)} \middle| \, Z_i^{(t)}, Z_j^{(t)}\right)$$

*(4.13 )*

with

$$p\left(A_{i,j}^{(t)} = 1 \middle| \, Z_i^{(t)}, Z_j^{(t)}\right) = \frac{1}{1 + e^{-x}} \left( Z_i^{(t)} \left( Z_j^{(t)} \right)^T \right)$$

Similarly, edge attributes prediction in our model is given by:

$$p\left(E^{(t)} \middle| Z^{(t)}\right) = \prod_{i=1}^{N_t} \prod_{j=1}^{N_t} p\left(E_{ijp}^{(t)} \middle| \, Z_i^{(t)}, Z_j^{(t)}\right)$$

with

$$p\left(E_{ijp}^{(t)} \middle| \, Z_i^{(t)}, Z_j^{(t)}\right) = \eta(\odot[\, Z_i^{(t)}, Z_j^{(t)}])$$

*(4 14 )*

where $\eta$ is a fully-connected neural network; $Z_i^{(t)}, Z_j^{(t)}$ are the corresponding embedding of nodes $v_i$ and $v_j$ at time $t$, respectively; and $\odot$ represents the concatenation of the embeddings.

*D. Detecting deletions and inferring probable topics with the model*

The outputs of the decoding functions in (13) and (14) forms our hypothesis to determine whether e-mails were deleted, and reconstruction of the topic vectors affected by the deletions. The e-mail deletion detection is modelled as a missing link prediction task, based on the edge occurrence probabilities given in equation (13). To predict links on the $(t + 1) - th$ snapshot, our model takes as input the sequence $\{G^{(1)}, G^{(2)}, \dots, G^{(T)}\}$ of previously observed snapshot. If the model predicts the existence of a link between two nodes $v$ *and* $\omega$ in the $(t + 1) - th$ snapshot but the actual graph does not have such a link, we report this as potential evidence that the users represented by $v$ and $\omega$ may have deleted e-mails. Similarly, our model also use the same sequence $\{G^{(1)}, G^{(2)}, \dots, G^{(T)}\}$ of previously observed snapshots as input for reconstructing topic vectors in the $(t + 1) - th$ snapshot. We use the probability vectors defined in equation (14) to make inferences about the possible deleted e-mails' topics. As a result, we deduce that the edge attributes associated with the predicted missing links are likely to be the subjects discussed. Figure 4.4 is a representation of our model's architecture.

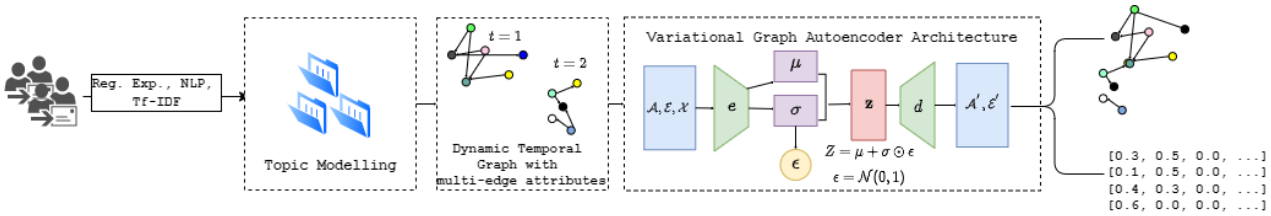


Figure 4.4: Image representation of the E-mail deletion detection and topic inference model architecture

### 4.2.6 Experiments

Our experiments will evaluate the effectiveness of our deletion detection method, as well as our model's capacity to reconstruct edge attributes, on a widely used real-life e-mail corpus. We detail the steps and approach used in our experiments below.

*A. Dataset*

We used the publicly available Enron e-mail dataset[109], which accurately portrays the environment in which our approach is intended to be applied, as it contains e-mails made available in pursuant to a court order. Another advantageous feature of this dataset is that we can process just e-mails sent by Enron employees, which eliminates worries regarding the

[109] https://www.cs.cmu.edu/~enron/

privacy of third-party individuals. The collection is organized into 149 unique folders, one for each employee. Additionally, each user's folder has been appropriately classified into sub-folders such as ‘inbox’, ‘sent-items’, and ‘deleted messages’.

*B. Selecting the most appropriate topic model for the Enron dataset.*

We conducted an empirical evaluation of two widely used topic models, LDA and NMF. We chose these two candidates based on previously reported strengths and weaknesses in the literature (Suri and Roy, 2017; Yong et al., 2019; Rania, Tet and Morad, 2020): LDA is quite successful on large corpora, although it performs best when each document is substantial in and of itself, whereas NMF has been shown to perform well on collections of short documents. Given that the efficacy of both models is highly dependent on the number of topics chosen correctly, our experiment aims to decide both the model to use and the number of topics. The coherence measure (Roder, Both and Hinneburg, 2015) of the topic distributions for each pair is derived from the form (model, number of topics). The figures (4.5) and (4.6) illustrate the coherence values of LDA and NMF, respectively, when a number of topics ranging from 2 to 50 is estimated. As shown in the figures, NMF with ten (10) topics is the most appropriate topic model for our experiment.

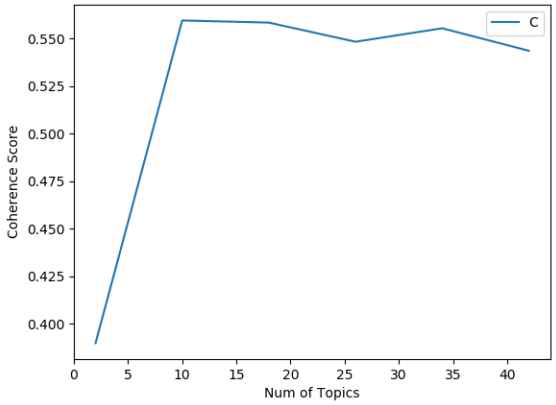


**Figure 4.5**: Coherence values of LDA for different numbers of topics.

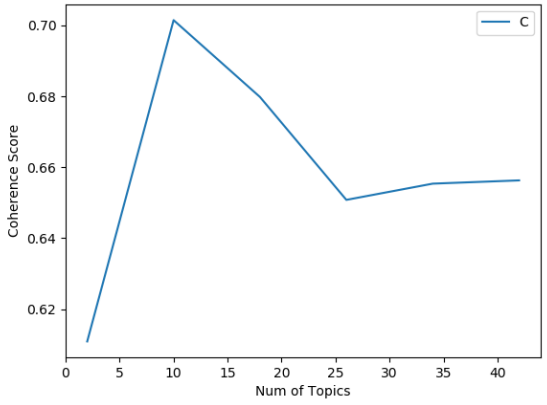


**Figure 4.6:** Coherence values of NMF for different numbers of topics.

*C. Experimental Setup*

We used the networkX[110] library to create the dynamic graph that represents the e-mail exchanges in the dataset; each snapshot represents one-month period of communications. This

[110] https://pypi.org/project/networkx/

creates a dynamic graph with 27 snapshots. We added an edge for each pair of users who exchanged at least one e-mail, and as indicated previously, we build the edge feature vectors using NMF with ten topics. Additionally, for each node, we manually set the following six attributes:

1) the number of messages sent by the user during the time period represented by the snapshot;

2) the number of e-mails received by the user as the primary addressee (the user is listed in the 'To' field);

3) the number of e-mails received as 'Cc';

4) the number of e-mails received as 'Bcc';

5) the degree of the node; and

6) the node's betweenness centrality.

The recurrent network in the model uses a 32-unit hidden layer. Moreover $\mu_{enc}^{(t)}$ and $\sigma_{enc}^{(t)}$ are modelled with 32-dimensional hidden layers and a latent space with 16 variables. All functions that make use of deep or fully connected neural networks have hidden layers with a depth of 32. The VGAE has a latent variable dimension of 16. We train the model with a learning rate of 0.01 using the Adaptive Moment Estimation (Adam) optimizer (Kingma and Ba, 2014). We run the models for 1000 epochs for both missing edge prediction and edge reconstruction tasks. The model is run on a Google cloud server (Colab) equipped with an NVIDIA K80/T4 GPU and 16GB of RAM.

### *D. Evaluation*

We trained our model on the 26 first snapshots of the dynamic graph and evaluated it on the 27-th snapshot for link prediction and edge reconstruction. We evaluated the model's ability to accurately predict the existence of all edges in this snapshot of the real graph. These are referred to as *positive examples*. Additionally, we randomly selected a set of non-edges in the real graph to examine the model's ability to not predict their existence, i.e., to assign extremely small probability values. The elements of this set are referred to as *negative examples*. We selected an equal number of negative and positive examples. Additionally, we evaluated edge reconstruction by comparing the reconstructed feature vectors to those in the real graph.

The message deletion detection is based on the prediction of possible missing links, as such, the evaluation is based on assessing if the predictor ranks a positive example higher than a negative example. The Area Under (ROC) Curve (AUC) (Hanley and McNeil, 1982; Bradley,

1997) and Average Precision (AP) (Zhang and Zhang, 2009) are used in this context. A detail description of AUC and AP is given in section 5.1.1.1.

The edge reconstruction quality measures are based on the closeness of the estimated feature values to the actual values. As a result, we evaluate our model's performance on this task by measuring the difference between the predicted and actual values. Our regressor performance metrics in this instance includes the Mean Squared Error (MSE) ) (Toro-Vizcarrondo and Wallace, 1968; Alen, 1971; Sammut and Webb, 2010), Root Mean Squared Error (RMSE) (Nevitt and Hancock, 2010), Mean Absolute Error (MAE) (Sammut and Webb, 2010), Mean Absolute Percentage Error (MAPE) (De Myttenaere et al., 2016), and the Error loss. The error loss is a model-based error evaluation. It is a quantitative measurement of the difference between the predicted and actual output. It estimates the network's error in predicting the output. The meaning and mathematical representations of the other metrics are discussed in Chapter 5.

Given these error metrics, the closer the values are to 0, the more accurate the approach can be claimed to be. Due to the intrinsic randomness of VGAEs, the findings reported in this thesis are obtained by repeating the experiment multiple times (the results do not vary significantly in all situations) and averaging the values obtained for each measure.

*E. Results and discussion*

Table 4.1 summarizes the results of link prediction in terms of AUC and AP. The AUC value of 0.7477 is encouraging, as it indicates that our model is capable of accurately predicting links in the snapshot under study (which coupled with the inspection of the real graph, in our application scenario, is available to the investigators, enables them to accurately predict a significant number of edge deletions. Moreover, the fact that AP is 0.7275 implies that the model's capacity to detect deletions correctly does not come at the expense of issuing an excessive number of false positives. Furthermore, the edge reconstruction results, which are provided in Table 4.2 in terms of all four error measures, demonstrate that the reconstruction capability of our model is rather good, as all error values are considerably close to 0.

| **Metrics** | **Scores** |
|---|---|
| **AUC Mean** | 0.7477 |
| **AP Mean** | 0.7275 |

Table 4.1: Link prediction results on the 27th snapshot of the dynamic graph

| Metrics | Scores |
|---|---|
| **Error Loss** | 4.115 |
| **MSE** | 0.0155 |
| **RMSE** | 0.1236 |
| **MAE** | 0.0913 |
| **MAPE** | 9.1365 |

Table 4.2: Reconstruction results on the 27th snapshot of the dynamic graph

To supplement the previous findings, we ran an additional experiment in which we deleted a number of edges from the real graph and evaluated our model's ability to reliably reconstruct the corresponding feature vectors and detect deletions. At each time step, we randomly chose 80%, 90%, and 95% of node pairs from our communication graph and then deleted 25% of the communications between these node pairs. The results of this experiment, shown in Table 4.3 and 4.4, demonstrate that the model is still capable of detecting deletions and accurately reconstructing feature vectors in this context.

| Percent of affected users | AUC Mean | AP Mean |
|---|---|---|
| **95%** | 0.6457 | 0.6743 |
| **90%** | 0.7932 | 0.8205 |
| **80%** | 0.7235 | 0.7402 |

Table 4.3: Results of the random removal experiment on link prediction

| Percent of affected users | Error Loss | MSE | RMSE | MAE | MAPE |
|---|---|---|---|---|---|
| **95%** | 9.308 | 0.0143 | 0.1194 | 0.0616 | 6.1671 |
| **90%** | 10.8207 | 0.0206 | 0.1435 | 0.0621 | 6.2117 |
| **80%** | 6.6501 | 0.0145 | 0.1203 | 0.0601 | 6.0122 |

Table 4.4: Results of the random removal experiment on edge reconstruction

## 4.3 Chapter Summary

In this chapter, we introduced a new tool to partially automate the task of forensic investigators examining e-mail collections. Particularly, we presented techniques for detecting possible

malicious deletions of e-mails exchanged between suspects. Our solution is based on Variational Graph Autoencoders and takes advantage of the model's ability to reconstruct partially missing patterns. The autoencoders (coupled with other feature extraction and graph recurrent methods) manage a rich dynamic attributed graph-based representation of the e-mail collection, which incorporates metadata from the e-mail headers and semantic information derived from the e-mail contents, all while accounting for the communication's temporal properties. Our technique demonstrated the efficiency of combining e-mail header metadata with natural language body elements. Additionally, our graph neural network model performs well in dynamic temporal contexts with multiple edge properties. Finally, we presented our findings, which are both encouraging and promising in terms of stimulating additional research into this technique.

# Part III

# Digital Forensics AI: The Concept

# Chapter 5

# Digital Forensics AI: Evaluation, Standardization, and Optimization of Digital Evidence Mining Techniques

The development of research methodologies for big data mining which seeks to discover meaningful and explorable patterns in data, has motivated its application in DF investigations. Despite concerns regarding closed-box AI models' ability to produce reliable and verifiable digital evidence (Pasquale, 2015), the idea that cognitive approaches employed in big data analysis will work when applied to DF analysis has fueled a decade-long research surge into the application of AI in DF.

To begin, a misunderstanding exists regarding the colloquial use of the terms "*Forensics AI*" and "*AI Forensics*" within the forensics community (and beyond), with some using the phrases interchangeably as referring to the application of AI in DF. While both phrases are self-explicit, in this thesis, we propose the conceptualization of "Digital Forensics AI" and to draw a preliminary distinction between the common misconceptions and distinguish the two concepts. On the one hand, according to (Doowon, 2020), a word preceding 'forensics' in DF domain denotes the target (too or device) to be analyzed (e.g., memory forensics, network forensics, cloud forensics, etc.). Hence, the author refers to "AI Forensics" as forensic analysis of AI tools or methods, rather than forensic investigation applying AI techniques. In the same vein, as proposed in the paper of Baggili and Behzadan (2020), refers to "AI Forensics" as the "*scientific and legal tools, techniques, and protocols for the extraction, collection, analysis, and reporting of digital evidence pertaining to failures in AI-enabled systems.*" To summarise, AI Forensics is the examination of the sequence of events and circumstances leading up to the failure of an intelligent system, including the determination of whether the failure was caused by malicious activity and the determination of the responsible entity(ies) in such situations. On the other hand, a thorough search of academic databases such as Google Scholar, IEEE Explore, and Scopus for the phrases "Forensics AI" or "Digital Forensics AI" reveals that the vast majority of resources are based on DF analysis methodologies augmented with AI techniques. However, in this thesis, we refer to Digital Forensics AI (hereafter referred to as DFAI) "*as a generic or broader concepts of automated systems that encompasses the scientific and legal tools, models, methods; including evaluation, standardization, optimization,*

*interpretability, and understandability of AI techniques (or AI-enabled tools) deployed in digital forensics domain*."

As this, to our understanding, is the maiden attempt to conceptualize this idea (in terms of definition) — it is merely intended as a preliminary proposal that could serve as a springboard for a more refined formalization of this concept, either as a sub-domain of DF or as an essential part of the existing framework.

Indeed, despite concerns about the "closed-box" AI models, their success in other domains has made its adoption in DF procedures unavoidable. The majority of DF tools are designed to report solely on data that exists in digital artifacts; not the non-existing ones (SWGDE, 2018). Unsupervised ML algorithms, for instance, can interact with artifacts, extract meaningful cues, and cluster or classify them appropriately. Consequently , it is necessary to chart a new course in order to standardize all sub-components of DFAI in ways that are adaptable and consistent with the scientific and legal intricacies of the DF and Law domain. Furthermore, the majority of digital artifacts are unstructured; they have an irregular and ambiguous data model, making them difficult to understand. This adds additional layer of complexity to DF examination process. However, with a well-structured AI algorithm and appropriate accuracy metrics, this complexity can be reduced to the point where insightful clues can be deduced and the degree of accuracy/correctness of this deduction can be measured.

Therefore, this chapter discusses (at a foundational level) several approaches for evaluating, standardizing, and optimizing AI methodologies used in DF investigation. As such, while all subsequent references to the aforementioned DFAI components are also relevant to conventional DF procedures (without AI algorithms), in this chapter (as well as the following ones), they are framed within the context of AI and the mining of unstructured evidentiary data, to which these components apply.

We will discuss some techniques and scientific principles that we consider relevant for evaluating, standardizing, and optimizing AI-driven approaches to digital forensics procedures in the sections that follow.

## 5.1 Methods for Evaluating DFAI Analysis

During an investigation, examiners define hypothesis as an initial proposition based on informed supposition from observed data that is evaluated against other competing assertions (Pollitt et al., 2018). The issue is that, as highlighted in (Sunde and Dror, 2019), in an attempt

to make sense of what is observed (sometimes by coercively ensuring that it fits the initial assumption), investigators subconsciously: 1) seek findings that support their assertions; 2) interpret relevant and vague data in relation to the hypothesis; and 3) disregard or assign less weight to data that contravene the working hypothesis. Numerous factors could account for these biases, including but not limited to: confidence (as a result of the presumption of guilt), emotional imbalance, concern about long term implications (e.g., loss of prestige), personality characteristics (e.g., dislike for uncertainty or a predilection to over-explore various scenarios) (Ask, 2005), and the expert's support for the party's position (*adversarial allegiance* – for prosecution or defence) (Murrie et al., 2009). Thus, before conclusion is reached in a forensic investigation, each component of the initial hypothesis must be independently and wholly tested (or evaluated) in order to ascertain the degree of confidence in the processes that led to the fact. Evaluation, therefore, is the process of determining the strength of evidence supporting competing claims, as well as their relative believability and probability (Lau and Biedermann, 2020). Expert examiners can evaluate the outcomes of a forensic analysis through a variety of methods, some based on predefined scientific parameters and others entirely on logical deductions supported by experience or subjective reasoning. We emphasized the pitfalls of subjectivity in Section 3.2, and while it may be necessary at times, it is not a recommended scientific practice. (Sunde and Dror, 2019) discussed further issues with subjectivity and the problems with human cognitive factors in forensic investigations. However, in the context of DFAI, forensic evaluation is viewed through the lens of the assessment of the AI techniques used to accomplish the DF analysis such as identification, classification, reconstruction, and presentation. Such deployment necessitates metrics and measurements that are compatible with the evaluation of AI models. The evaluation of DFAI models can be performed on the functional parameters of the algorithm (i.e., the evaluation of individual modules) or on their outputs. Unlike typical methods for evaluating ML/DNN models, which employ standard metrics related to the task or learning algorithm, establishing confidence in the outcome of a DFAI investigation may require extra human observation of the output. Numerous studies in DF have also revealed that forensic practitioners frequently issue inconsistent or biased results (Ask, 2005; Graham et al., 2006). In addition, the majority of AI-based approaches lack the necessary clarity and replicability to allow investigators to assess the accuracy of their output (Bollé, Casey and Jacquet, 2020). Thus, a forensically sound process[111], is one that integrates automated investigative analysis —

[111] Transparent digital forensics procedure that preserves the data's true context for use in a legal proceeding.

evaluated through scientific (accuracy and precision) metrics — with human assessments of the outcome. For example, a DF investigation into Child Sexual Exploitation Material (CSEM) (Anda et al., 2019) may seek to automatically detect and classify images of people found on a seized device as adult or underage (based on automatic estimated age). Given the complexity and bias in the training dataset, the learning algorithm may pick up on errors/biases, resulting in misclassification (i.e., false positive[112]), misinterpretation of features, and the possibility of missing critical features during the classification process that could have served as evidence (false negative[113]; e.g., an underage wearing adult facial makeup) (Anda et al., 2019). In this scenario, merely addressing bugs in codes may not be sufficient, as the classification errors may be subconsciously inherited and propagated through data. Similarly, in Chapter 4 of this thesis, we described a temporal analysis of e-mail exchange events to detect whether suspicious deletions of communication between suspects occurred and whether the deletions were intended to conceal evidence of discussion about certain incriminating subjects. One significant drawback of that analysis is the system's inability to thoroughly investigate if the suspicious message(s) were initiated or received by the user or were deliberately sent by an unauthorized hacker remotely accessing the user's account and sending such incriminating message. To reach a factual conclusion in this case, various other fragmented unstructured activity data (unrelated to e-mail, maybe) must be analyzed and reconstructed. Depending on the design, a robust AI-based system can uncover various heretofore unrecognized clues. If these new revelations (even though relevant) are not properly analyzed and evaluated, they may lead investigators to believing that the outputs dependably fulfil their needs (Bollé, Casey and Jacquet, 2020). As a result, an extensive review of the output of DFAI will be required (supposedly provided by human experts) to arrive at a factually correct conclusion. This has also been highlighted as a significant instrument for analyzing digital evidence in (ENFSI, 2015a; Pollitt et al., 2018).

As with the output of any other forensic tool capable of extracting and analyzing evidence from digital artifacts, which frequently requires additional review and interpretation that are compatible with the working hypothesis, the results of forensic examinations conducted using DFAI should be viewed as "recommendations" that must be interpreted within the context of the entire forensic investigation (Bollé, Casey and Jacquet, 2020). In a typical analysis of unstructured evidence data, DFAI models are arranged in sub-modules, with each module

[112] See section 5.1.1.1
[113] See section 5.1.1.1

addressing a particular aspect of the overall problem. Thus, at each decision point in the investigation process, or within each of the DFAI algorithm's sub-modules, an evaluation procedure may be carried out to determine the confidence in the decision taken at that level, with the aim to address the pre-defined proposition. In addition, the evaluation apparatus must be verifiable, appropriate for the task it seeks to solve, and compatible with the other contextual analysis of the investigative model. Taking this into consideration, DFAI evaluation can be viewed in terms of two significant instruments: performance and forensic evaluation. We discuss below, the significance and components of each of these instruments. In our opinion, for a sound forensic process based on DFAI, these two instruments are essentially required.

### 5.1.1 Methods for Evaluating the Performance of DFAI Analysis

In a machine-driven system, evaluation produces value as a measure of the model's performance in accomplishing the task for which it was commissioned, which may be used to influence decision-making (Pollitt et al., 2018). Depending on the problem the model attempt to solve, evaluation may be: a set of thresholds formulated as binary (i.e., 'yes' or 'no', or 0 or 1) or categorical (qualitative; one of a possible finite outcome) as the case maybe; discrete (enumeration of strength; e.g., range between 0 to 10); or continuous (e.g., probability distributions of real values between 0 and 1). Consequently, evaluating the performance of a DFAI model built to recognize specific faces in a CSEM is distinct from evaluating the performance of a model meant to classify faces as underage or adolescent. Similarly, distinct metrics are required for models that detect SPAM e-mails and those that attempt to infer intent from an e-mail content. The majority of DFAI tasks will fall into one of three categories: classification, regression, or clustering. There may be instances when regression problems or continuous data are transformed into classification tasks (through categorical data discretization and dichotomization[114]) and evaluated as such for the sake of clarity and conciseness. While we discussed both categories in Section 3.1.2.4, including the corresponding algorithms, the next sections provides further details on the evaluation process of a classification, regression, and clustering models.

[114] See 'Discretization in data mining.' Data Mining, JavaTpoint. Cited on January 20, 2022. Available online at: https://javatpoint.com/discretization-in-data-mining

#### 5.1.1.1 Evaluating Classification Algorithms in DFAI

Classification models[115] are predictive in nature, identifying the class to which a set of input samples belongs. A classification task is commonly modelled in ML as a binary representation that predicts a Bernoulli probability distribution for each sample. Bernoulli distributions (Dai, 2013) are a type of discrete probability distribution in which events have binary outcomes such as 0 or 1. The model's performance is measured by its ability to correctly predict (assign a high probability value to) the class of positive samples and to assign a very low probability value to non-existent samples.

Prior to deploying a DFAI model, it is important to examine the characteristics and complexities of the investigation to determine whether the model is suitable for that purpose. Apart from binary classification, there are varieties of other classification models that are tailored to tackling particular DF problems. Practitioners are expected to be aware of the unique characteristics of learning algorithms and to use them appropriately. For instance, in a forensic investigation involving facial materials, which requires facial classification. There are two main models that can be applicable: verification and identification. Verification entails comparing an unknown face to a known face directly (One-vs-One) (Gidudu, Hulley and Tshilidzi, 2007) and computing their similarity score. This can be adapted as a binary classification problem, in which the system predicts whether or not two faces share a high degree of similarity, based on a predetermined threshold. On the other hand, identification involves One-vs-Rest (Hong and Cho, 2008) comparison, in which an unknown face is compared to the faces in a database of known persons. The Identification task is a typical "***Multi-Class Classification***" (Wu, Lin and Weng, 2004) problem, in which samples are classified into one of a set of known classes. Other classification models are:

**Multi-Label Classification** (Tsoumakas and Katakis, 2007): This technique is employed in specialized classification tasks that require the prediction of one or more class labels (often two or more) for each example. In contrast to previous classification techniques, Multi-Label Classification can predict many outputs, each of which follows a Bernoulli probability distribution adapted as binary classification. This classification type is extremely beneficial for identifying objects in crime scene images, particularly when there are numerous materials to analyse. It has the capability of infer the presence of several objects within an image. For this

[115] See 3.1.2.4(A)

task, specialized multi-label variants of classification algorithms, e.g., Multi-label Decision Tree (Vens, 2008) and Multi-label Random Forest (Liu et al., 2015) can be utilized.

**Imbalanced Classification** (Tang et al., 2008; Zou et al., 2016)**:** refers to classification problems in which the number of examples per each class is unevenly distributed. They are basically binary classification wherein the majority of training examples fall into the positive (normal) class and the remaining minority fall into the negative (abnormal) class. As such, in a cost-sensitive setting, certain specialized techniques are used to under-sample the class with majority examples and oversample the minority class, ensuring that the minority class receives more attention during model fitting on the training dataset. This classification type can be quite effective in detecting fraud or anomalies when the majority of sample data appear to be legitimate (or behaving normally), with the exception of a few bad samples that should be addressed.

Metrics such as accuracy, precision, recall, and F-Measure are all relevant depending on the investigation's characteristics.

The measure of "***accuracy***" can be seen as the validity measure of a model. It is the ratio of the correctly classified samples to the total samples, and it is given as:

$$Accuracy = \frac{number\ of\ correctly\ classified\ examples}{total\ number\ of\ cases}$$

The accuracy metric can tell us whether a model was correctly trained and how well it will function in general. However, caution should be exercised when using this information alone to reach a general conclusion in forensic investigation, as it provides little information about its application to the problem and performs poorly in circumstances of severe class imbalance. That is, if the dataset is asymmetric, e.g., if the proportion of false positives is not (or nearly) equal to the proportion of false negatives. Accuracy is calculated in terms of a confusion matrix while performing a binary classification task, such as predicting whether an e-mail is "SPAM" or "NOT-SPAM." The confusion matrix is applied to a set of test data, for which the true values are known. The confusion matrix depicted in Table (5.1) is an illustration of one.

| | | Predicted Class | |
|---|---|---|---|
| | | SPAM | NOT-SPAM |
| Ground Truth Class | SPAM | True Positive (TP) | False Negative (FN) |
| | NOT-SPAM | False Positive (FP) | True Negative (TN) |

Table 5.1: Confusion Matrix of a Typical SPAM e-mail Classifier

From Table (5.1), "*True Positive*" and "*True Negative*" are the samples that are correctly predicted. What a classifier seek to minimize is the number of "*False Positives*" and "*False Negatives.*"

A *true positive (tp)* is one in which the model accurately predicts the positive samples, while a *true negative (tn)* indicates the result of correctly predicted negative samples. Similarly, a *false positive (fp)* outcome occurs when the model incorrectly predicts positive samples, whereas a *false negative (fn)* outcome occurs when the model inaccurately predicts negative samples. Therefore, in terms of confusion matrix, an accuracy measure is represented as:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

**Precision** – this metric is critical, particularly in the domain of DFAI — can be regarded as reliability measure of the model. It provides additional assurance by posing the question: "how frequently is the model correct when it predicts a positive sample?" With precision, we affirm the classifier's ability not to label a negative sample as positive. Given that the outcome of a forensic investigation may be critical to the outcome of an inculpatory or exculpatory proceeding, the cost of a high rate of false positives may be detrimental. For example, in the United Kingdom, a study by (Smit, Morgan and Lagnado, 2018) identified cases in which digital evidence had a role in around 32% of the 235 wrongful convictions. Thus, to avoid misleading errors, a precision score must be persuasively high (i.e., low false positive rate). Precision is calculated as:

$$Precision = \frac{tp}{tp + fp}$$

**Recall –** this is crucial in DFAI as well, especially when the cost of false negative could be catastrophic. For example, a facial recognition algorithm could be used to analyze criminal materials through training examples. While the system may be capable of identifying and classifying many positive samples, we may need to determine how many true positives were correctly identified from the predicted true positives. This is critical for evaluating working hypothesis and assisting in the answering of some potentially damning questions during court proceedings. Recall allows informed decisions concerning false negatives - such as relevant details that should not be overlooked. Recall is calculated as:

$$Recall = \frac{tp}{tp + fn}$$

**F-Measure** - this metric combines precision and recall for determining the model's overall accuracy. It accounts for both false positives and negatives, i.e., a low false positive and negative rate is indicative of a good F-measure, and it can aid in the reduction of false claims during forensic investigation. The F1 score is the weighted average of precision and recall; an ideal fit has a value of 1 (or approximately), while the worst-case scenario has a value of 0. F1 score is mathematically given as:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

Another relevant metric for measuring a classifier's capacity to distinguish between classes is the **Area Under the Curve (AUC)**, which serves as a summary of the Receiver Operating Characteristic (ROC) curve. The AUC and **Average Precision (AP)** are the quality measures used in link prediction. AUC reflects the likelihood that the model ranks a positive example higher than a negative example in terms of probability of existence. AUC values range between 0 and 1. A model that makes 100% inaccurate predictions has an AUC of 0, whereas a model that makes correct predictions 100% of the time has an AUC of 1.0. Mathematically, the AUC is given as:

$$AUC = \frac{\sum_{t_0 \in \mathcal{D}^0} \sum_{t_1 \in \mathcal{D}^1} 1[f(t_0) < f(t_1)]}{|\mathcal{D}^0| \cdot |\mathcal{D}^1|}$$

where $1[f(t_0) < f(t_1)]$ denotes an indicator function which returns 1 *iff* $f(t_0) < f(t_1)$, otherwise, returns 0. $\mathcal{D}^1$ and $\mathcal{D}^0$ are the set of positive and negative examples, respectively. A high AP value suggests that a model is capable of detecting a large number of positive cases effectively without wrongly classifying an excessive number of negative examples as positive. It is used to quantify the average difference between precision and recall at different decision points. AP is given by:

$$AP = \sum_{n} (R_n - R_{n-1}) P_n$$

where $P_n$ and $R_n$ are the precision and recall at the *n*-th threshold.

There are instances when accuracy is preferred over F-measure, particularly when the cost of false positives and negatives is similar, implying that the consequences are not negligible. If, on the other hand, the situation is reversed, it is fair to evaluate the F1 score. Additionally, caution should be exercised when evaluating performance on classified samples that involves the assignment of a threshold (as is the case in some logistic regression models). Increases or

decreases in the threshold have a significant effect on the precision and recall values[116]. In contrast to a model designed to optimize business decisions, it may not be prudent to include any threshold in DFAI — as it would be appropriate to have a realistic picture of the analysis' outcome — unless we are convinced that doing so will have no detrimental impact on the outcome. Nonetheless, accuracy is critical; so, if the trade-offs can be quantified and justified satisfactorily, the threshold can be considered.

#### 5.1.1.2 Evaluating Regression Algorithms in DFAI

In contrast to classification models, which predict the classification of input samples, regression models[117] predict an infinite number of possible (continuous; real-valued such as integer or floating point) outcomes. In DFAI, regression analysis can be utilized for two conceptually distinct purposes: forecasting and prediction; and inference of causal relationships between dependent (observed) and independent (predictors) variables. Before a regression analysis may be commissioned, the examiner must be convinced that the correlations present in the data possess the predictive power to infer a new context or that these correlations can induce a causal interpretation based on observational data (Cook and Weisberg, 1982; Freedman, 2009). This is particularly important for forensic investigations. To elucidate this point further, consider the e-mail experiment described in Chapter 4, the component of the model that predicts the likely topic of discussion between two suspects is based on regression analysis. The choice of regression model was influenced by the dataset's characteristics, which include temporal dynamics, a large number of e-mail exchanges across time, and the e-mail text from which abstract topics were derived. As a result, the model gained insight into how communications evolved over time, enhancing its prediction capabilities in an unknown scenario. Furthermore, by delving deeper into the conversations between various user pairs, we may be able to deduce the reason for their interactions or, as was the case in our experiment, the reason for deletion.

A critical element that might enhance a regression model's predictive potentialities is when the input variables are arranged chronologically (with event time), a concept referred to as time series forecasting. This can be advantageous for forensic purposes involving the detection of deviations (anomalies), crime forecasting, the prediction of potential connections between

[116] For examples, see: https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall
[117] See 3.1.2.4(A)

data, and event reconstruction. Importantly, examiners should be wary of interpolation and extrapolation when using regression. In many circumstances, the former is appropriate, as it entails the prediction of values within the range of data points in the dataset used to fit the model. However, the latter is frequently undesirable. It is based on regression assumptions and entails predicting values that are not within the range of observed data. An extrapolation over a range that is far away from the observed data involves danger and is an indication of possible model failure.

A regression model's performance is measured as an error in prediction, i.e., how close the predictions were to the ground truth. To do this, the following error measures are frequently used: MSE, RMSE, MAE, and MAPE. Although there are several other error metrics available; the choice of which is determined by the type of error being evaluated. We present a brief discussion about these metrics below.

**MSE**: or Mean Squared Deviation (MSD) measures the average squared difference between predicted and observed values. MSE can be used to evaluate the quality of a predictor or an estimator[118]. However, in DFAI, it better-off as a predictor since it can map arbitrary input to a sample of random variables. A MSE of zero indicates a perfectly accurate prediction, however this is rarely possible (Lehmann and Casella, 1998). MSE values that are close to zero and strictly positive (as values are squared) are considered ideal. As a loss function, MSE is optimized using least squares, i.e., to minimize the mean squared difference between prediction and expected values. Unfortunately, other measures have been sometimes preferred to MSE due to its disproportionate weighting of outliers. This occurs as a result of magnification of large errors than on small ones, due to each value being squared. Mathematically, MSE is given by:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

where $n$ denotes the number of data points. $Y_i$ and $\hat{Y}_i$ are the observed and the predicted values.

**RMSE**: is an extension of the MSE, except that the square root of the averaged squared error is calculated. Also, it is a measure of the differences between predicted and actual values. RMSE is always non-negative, therefore, like MSE, a value of zero (0) is almost unrealistic; and if it does occur, it is a hint that the model is trivial. The RMSE is sensitive to outliers, as

[118] a mathematical function that maps a sample of data to an estimate of a population parameter

larger errors are weighted more heavily. For a DFAI task, it may be prudent to create a baseline RMSE for the working dataset by predicting the mean target value for the training dataset using a naive predictive model. This can be accomplished by transforming or scaling (i.e., normalization) the dataset's feature vectors between 0 and 1. Most GNN models perform this transformation during feature extraction. If the evaluating RMSE achieves a better value than the baseline RMSE, it is said to be well-fit. Mathematically, RMSE is given as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \| y_i - \hat{y}_i \|^2}{N}}$$

where $N$ denotes the number of data points, $y_i$ and $\hat{y}_i$are the actual values and the corresponding predicted values.

**MAE:** measures the differences in error between paired observations expressing the same phenomenon, i.e., it is scale-dependent; it employs the same scale as the data being measured[119]. MAE is defined mathematically as the average of the absolute (vertical or horizontal distance) errors $|e_i|$ between predicted and actual values. The absolute, or *abs()* in mathematical notation, simply ensures that the results are not negative. In contrast to the previously stated error measures, which require squaring the differences, MAE changes are linear, intuitive, and interpretable; they simply represent the contribution of each error in proportion to the error's absolute value. MAE does not give greater or lesser weight to errors and hence provides a realistic view of the main prediction errors; thus, it is strongly recommended for DFAI. Additionally, it is a frequently used metric for forecasting error in time series analysis (Hyndman and Koehler, 2006). Mathematically, MAE is given as:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

where $n$ denotes the number of data points. $y_i$ is the predicted value and $\hat{x}_i$ is the actual value.

**MAPE:** is the mean or average of a regression forecast's absolute percentage errors (MAPE, 2000). Due to its relatively intuitive interpretation in terms of relative error, MAPE is frequently used for model evaluation and as a loss function for regression problems. MAPE has been asserted to be very well-suited for prediction, especially when sufficient data is available (De Myttenaere et al., 2016). Caution should be exercised, however, to avoid the occurrence of the 'one divided by zero' problem. Additionally, MAPE penalizes errors with

[119] See "Evaluating Forecast Accuracy." OTexts. Cited on Aug. 5, 2021. Available at https://otexts.com/fpp2/accuracy.html

negative values substantially more than positive values; hence, when used in a prediction task, it is biased towards methods with very low forecasts, rendering it inappropriate for evaluating problems where large errors are expected (Ren and Glasure, 2009; De Myttenaere et al., 2016). Arithmetically, MAPE is given as:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

where $n$ denotes the number of data points. $A_t$ denotes the real values, while $F_t$ are the predicted values.

There are other error measures for regressors such as Max Error (Garofalakis and Kumar, 2004); that calculates the maximum residual error and detect worst case errors (Bollé, Casey and Jacquet 2020), and $R^2$ (also known as R-Squared, Goodness of fit; Co-efficient of Determination ) (Wright, 1921; Barrett, 2000; Di Bucchianico, 2008), which is the measure of variance proportion in the regressor. Arithmetically,

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(X_i - Y_i)^2}{\sum_{i=1}^{m}(\bar{Y} - Y_i)^2}$$

The numerator denotes the squared sum error of regression line, while the denominator is the squared sum of error of mean line.

After describing each of these error measurements for regression problems, along with their associated limitations in some circumstances, determining which one is most appropriate for forensic investigation can be fairly daunting. However, according to literature, (Armstrong and Collopy, 1992) indicated that the RMSE is unreliable and unsuitable for measuring correctness in a time series analysis. In addition, RMSE was discovered to have "disturbing characteristics" in (Willmott and Matsuura, 2005; Willmott, Matsuura and Robeson, 2009), making it ineffective as an error measure. MSE and all other squared errors were equally stated as unfit for evaluation purpose. However, by offering arguments in support of RMSE, (Chai and Draxler, 2014) partially disputed these conclusions. Nonetheless, MAE was recommended in vast majority of instances; which is understandable. As previously stated, the MAE measure is a consistent and compatible evaluation technique with DFAI; it is a more natural measure of average error magnitude (Willmott and Matsuura, 2005) that accurately depicts the model's performance. The $R^2$ is another metric that deserves a role in DFAI. A recent comparison of regression analysis error measures is discussed in (Chicco, Warrens and Jurman, 2021). The R-squared value can be in the range $[-\infty, 1]$, with an upper bound of 1 or close reflecting a

good fit regardless of the scale of measurement, and despite the fact that it is not lower bounded, a value of 0 refers to a trivial fit (Chicco, Warrens and Jurman, 2021). $R^2$ exhibit desirable features, including interpretability in terms of the data's information content and sufficient generality that span a relatively broad class of models (Cameron and Windmeijer, 1997). Although a negative $R^2$ indicates a worse fit than the average line, this representation may be critical in determining how the learning model fits the dataset. Moreover, regardless of whether an examiner reports the $R^2$ score, it is a highly effective technique for evaluating the performance of a regression analysis and should be considered.

#### 5.1.1.3 Evaluating Clustering Algorithms in DFAI

Evaluating a clustering method, particularly in an unstructured data, can be challenging because it is mostly used in unsupervised learning, which means that no ground-truth labels are available. Clustering in a supervised setting, on the other hand, can be evaluated using supervised learning metrics. One significant difficulty with unsupervised learning is that applying clustering analysis to a dataset blindly would categorize the data into clusters (even if the data is random), as this is the algorithm's expected function. As a result, examiners must check the data's non-random structure before deciding on a clustering approach. Three critical factors that should be considered in clustering are: 1) Clustering tendency; 2) Number of clusters, $k$; and 3) Clustering quality.

***Clustering tendency:*** quantifies data's spatial randomness by calculating the probability that a given dataset is produced by a uniform distribution. Clustering techniques may be meaningless if the data is sparsely random. This preliminary assessment is key to DFAI, particularly because it can help to reduce the time spent analysing artefacts. A common method for assessing a dataset's cluster tendency is to utilize the Hopkins statistic (Hopkins and Skellam, 1954), which is a type of sparse sampling test. The Hopkins statistic is used to test the null hypothesis ($H_0$), which states that the data is generated by a Poisson point process[120] and are, therefore, uniformly distributed (Banerjee and Rajesh, 2004), and the alternative hypothesis ($H_a$), which states that the data is not generated by uniform distribution (i.e., contains meaningful clusters). If the Hopkins statistic is close to 1 or $H > 0.5$, we can reject the null hypothesis and infer that there are considerable clusters in the data. A value close to 0 indicates a uniform distribution,

[120] A Poisson point process is a type of random mathematical object composed of randomly located points in a mathematical space.

and thus the absence of clustering. However, the way values are assigned differs according to the tool used.

***Number of clusters:*** obtaining the ideal number, *k*, of clusters is critical in clustering analysis; while there is no definitive method for doing so, it can rely on the shape of the distribution, the size of the data set, and the examiner's preference. If *k* is set to a value that is too high, each data point has a chance of forming a cluster, whereas a value that is too low may result in inaccurate clusters. Additionally, the following approaches can help to determine the cluster number:

> *Prior domain knowledge* - prior domain knowledge can provide insight into the optimal number of clusters to choose.
>
> *Data driven approach* – employs mathematical methods to determine the correct value, such as *rule of thumb method* (using the formula $k \approx \sqrt{\frac{n}{2}}$, where *n* is the number of data point), *elbow method* (Ng, 2012; Kodinariya and Makwana, 2013), and *gap statistics* (Tibshirani, Walther and Hastie, 2001).

***Clustering quality***: is defined by the minimum intra-cluster distance and the maximum inter-cluster distance.

To evaluate the performance of a clustering task, two validation statistics are key, namely: internal cluster validation and external cluster validation.

**Internal cluster validation** evaluates the goodness of a clustering structure without referring to external data. It frequently reflects the clusters' compactness, connectivity, and separation. The silhouette coefficient (Rousseeuw, 1987; Aranganayagi and Thangavel, 2007) and Dunn index (Dunn, 1973) can be used to evaluate the algorithm's performance in relation to its internal clusters. There are additional indices (e.g., Davies-Bouldin index (Davies and Bouldin, 1979)); nevertheless, the silhouette coefficient and Dunn index show the most compatibility with DFAI in general, and specifically in terms of ease of interpretation.

By measuring the average distance between two observations, the **Silhouette Coefficient** determines how well they are clustered. In this approach, for each observation $i$, the average dissimilarity $a_i$ between $i$ and other points within its own cluster is calculated. Similarly, for all other clusters $C$, the average dissimilarity $d(i, C)$ (calculated as $b_i = \min_{C} d(i, C)$) between $i$ to all other points in $C$ is also calculated. Therefore, a silhouette width of point $i$ is defined by:

$$S_i = \frac{(b_i - a_i)}{\max(a_i - b_i)}$$

If $S_i$ is close to 1, then the data points are well clustered. A value close to 0 indicates that the data point is located between two clusters, whereas a negative value suggests that the data point is most likely placed in the wrong cluster.

If computational cost is not an issue, the **Dunn index** can be utilized. Simply compute the distance between each point in a cluster and the points in the other clusters, and then use the pairwise distance with the minimum value as the inter-cluster separation (*min. separation*). Additionally, to determine intra-cluster compactness, compute the distance between points within the same cluster and then use the maximum intra-cluster distance (*max. diameter*). Hence, the Dunn index is defined as:

$$Di = \frac{min.separation}{max.diameter}$$

**External cluster validation** compares and quantifies a cluster analysis' results against externally known benchmarks (e.g., externally provided gold standard labels). Such benchmarks are made up of a collection of predefined classes of items, which are often created by human experts. The evaluation approach quantifies the degree to which the clustering analysis result corresponds to predefined ground truth classes. To evaluate the performance of external cluster, the Rand index (Rand, 1971), the Purity index (Manning, Raghavan and Schutze, 2008), the F-measure (with precision and recall; as indicated in the classification task), and the Fowlkes-Mallows index (Fowlkes and Mallows, 1983) can be utilized. This thesis does not go into detail regarding the evaluation techniques used in this approach in order to keep the scope focused on components critical to DFAI.

As a matter of fact, it remains unclear how external cluster validation could improve DFAI. To elaborate on this fact, given the majority of digital artifacts from which evidence can be derived are sparse, unconventional, and previously unseen, having a ground truth label with which to compare may be impracticable. In addition, because the majority of DF analysis are crime-specific (or specific case related), the question is whether it is permissible to compare a crime-related data analysis to general task ground truth labels. However, if gold standard, case-based labels are available, such as those for videos and images in (Ferreira, Antunes and Correia, 2021b) or (though limited in scope and diversity) the "Computer Forensic Reference Dataset

Portal CFReDS)[121]" or "Datasets for Cyber Forensics[122]," then suitable comparisons can be established.

### 5.1.2 Forensic (Decision) Evaluation

After deriving facts from a forensic investigation, decision-making follows; which is the adoption of a hypothesis as a conclusion (Lau and Biedermann, 2020). Whilst evaluation of forensic decisions is frequently discussed in court contexts, it is applicable at all stages of forensic investigation (Casey, 2020). It begins with the evaluation of the individual hypothesis against all competing claims; the accuracy (including quantification of error rates) of the results obtained through automated tools used in the analysis; the extent to which experience, and domain knowledge were helpful; and to the simplicity with which the entire investigative process can be explained to a non-expert. Because automated systems are not self-contained and thus cannot take everything into account (Bollé, 2020), it is possible that multiple DFAI approaches were used to find solutions to all competing hypotheses. As a result, forensic evaluation in this case will entail weighing the differing claims against the overall investigative problem. One way of determining this is to assign an evidential weight (strength of evidence) or "Likelihood Ratios" (LRs) (Berger et al., 2011, Kerkhoff et al., 2013; ENFSI, 2015b) to all contending claims. Although LR was originally created as a framework for evaluating forensic evidence, the concept can be adopted to help make the DFAI model's evaluation outcome more intelligible. Contrary to the factually deterministic requirements of evidence in a criminal or civil case, the majority of AI-based algorithms and their outputs are mostly probabilistic. However, forensic examiners do not pronounce judgments or issue final decisions; they rather provide expert testimony (or an opinion) or report of their findings to fact finders (attorneys, judges, etc.). Succinctly reporting forensic investigation findings remains a challenge (Thompson, 2017), and while it may be comprehensible to state an opinion on a hypothesis and its alternatives as true (or false), this approach lacks the transparency and logical correctness necessary to reach a verdict in a legal proceeding. As a result, reporting DF findings in terms of weights or LRs enables the decision maker to assign the evidence an appropriate level of confidence (Bollé, 2020). Particularly with LRs, which are frequently used by forensic investigators in Europe (also in the U.S), represent their assessment of the relative probability of observed features under various hypotheses concerning a particular case. To mathematically

[121] https://cfreds.nist.gov/
[122] https://datasets.fbreitinger.de/datasets/

express the concept of LRs in a typical source identification analysis, let $E$ denote the observed features of two items to be compared; let $H_s$ represent the hypothesis that the items originate from the same source; and let $H_d$ denote the hypothesis that the items belong to a different source. The LR is thus defined as "the ratio of the probability of $E$ given $H_s$ to the probability of $E$ given $H_d$." That is: $\frac{p(E|H_s)}{p(E|H_d)}$. Using databases and statistical models, LRs have been utilized to reflect the strength of forensic evidence (in voice comparison) (Morrison and Thompson, 2017). Furthermore, the ENFSI recommends LR (simply in terms of numbers) even when examiners must make subjective decisions (ENFSI, 2015b), because it makes the examiner's belief and inferential process explicit and transparent, facilitating the evaluation of strengths and weaknesses for those who rely on it (Thompson, 2017). While expressing subjective decision in terms of LRs has grown widespread in Europe, doubts have been raised in support of empirical data instead (Berger et al., 2011). In other contexts, verbal expressions of LRs have been proposed; for example, consider an LR expression in the form: "*at least 1,000 times more likely*" and "*far more probable*." The former is likely to receive scepticism regarding the basis for that figure, whereas the latter has a stronger possibility of acceptance. (Berger et al., 2011).

Consequently, given the probabilistic (or stochastic) nature of the results of DFAI models, and the fact that these models have been empirically verified as accurate and well-suited for analytical purposes[123], as well as the inclusion of an "expert-in-the-middle[124]," it is still necessary to find the most suitable method to report the results in the clearest and most understandable manner possible, albeit as recommendations. A table of recommended LRs that is commensurate with the accuracy values of a typical AI-based evaluation of forensic investigation results could look like table (5.2)*. The table illustration reflects the Association of Forensic Science Providers (AFSP) in the United Kingdom's recommendation on the "standard for the formulation of evaluative forensic science expert opinion" (AFSP, 2009).

In Table 5.2, an equivalence of the common LR with corresponding verbal terminology that express the strength of support for a claim or evidence is presented with a simple AI-based accuracy score. Importantly, the false positive/negative rates are presented to demonstrate the significance of the false identification and exclusion rates in forensic outcome report.

---

[123] Via published studies, surveys, experiments, and peer review

[124] Either by way of having human expert verify the results manually, or with a rule-based expert system.

* This is just for insight purposes, does not reflect what is feasible in the real sense of likelihood ratio; which cannot be expressed in terms of probabilistic uncertainty.

| Likelihood Ratio | AI Accuracy Score (%) | False Positive Rate (%) | False Negative Rate (%) | Verbal Expression (Strength of Support) |
|---|---|---|---|---|
| 1 – 10 | 0 – 40 | 60 – 100 | 60 – 100 | Weak or limited support |
| 10 – 100 | 40 – 50 | 50 – 60 | 50 – 60 | Moderate support |
| 100 – 1,000 | 50 – 60 | 40 – 50 | 40 – 50 | Moderately strong support |
| 1,000 – 10,000 | 60 – 70 | 30 – 40 | 30 – 40 | Strong support |
| 10,000 – 1,000,000 | 70 – 90 | 10 – 30 | 10 – 30 | Very strong support |
| > 1,000,000 | 90 – 100 | 0 – 10 | 0 – 10 | Extremely strong support |

Table 5.2: A combined AI-Adaptive Likelihood Ratio with associated verbal support for reporting forensic outcomes.

The FP and FN rating scales in Table 5.2 can be adjusted according to investigative tasks, as there are instances when a 50% to 60% false positive/negative rate would indicate "weak support." In 2016, the US President's Council of Advisors on Science and Technology (PCAST, 2016) recommended that forensic examiners reveal the error rates observed in closed-box validation when reporting or testifying on forensic comparisons. Thus, error rates have become an intrinsic element of investigative outcome reporting, and with it, factfinders have a greater logical and empirical understanding of the probative value of the examiner's conclusion (Berger et al., 2016). However, it is evident that this method of evaluation is only appropriate when the investigation result is categorical (or discrete); the same may not be practicable when the outcome values are continuous; this is especially true for regression analysis. An alternative approach is as proposed in (Morrison, 2011) which based on the combination of prior probabilities and the likelihood ratio. It is not straightforward to express likelihood ratios in ways that are consistent with probabilistic distributions or error estimates (usually real values between 0 and 1). When the conditional components of a hypothesis are transposed, evaluating its probability might be logically fallacious (Casey, 2020). Probabilities are hardly acceptable in judicial decisions, as an 80% probability implies that one in every five cases would be decided incorrectly (Atkinson et al., 2020). Given that probability is relative to certainty (or otherwise), we can align our DFAI evaluation intuition with the "Certainty Scale", or "Confidence Scale" (C-Scale) proposed in (Casey, 2002; 2011; 2020), which is reasonably appropriate for assigning strength of evidence to continuous values with respect to the hypothesis. As noted by (Casey, 2020); *"...the strength of evidence does not exist in an abstract sense, and is not an inherent property of the evidence; it only exists when a forensic practitioner assigns value to the evidence in light of the hypothesis*." Therefore, in light of each working

hypothesis resolved via DFAI, table 5.3 represent a proposed C-Scale for expressing the strength of evidence in evaluation of a typical DFAI task.

| C-Value | AI Accuracy Score (%) | False Positive Rate (%) | False Negative Rate (%) | Verbal Expression (Strength of Support) |
|---|---|---|---|---|
| C0 | 0 – 20 | 55 – 100 | 55 – 100 | Erroneous/Incorrect |
| C1 | 20 – 30 | 50 – 55 | 50 – 55 | Extremely weak evidence |
| C2 | 30 – 40 | 40 – 50 | 40 – 50 | Very weak evidence |
| C3 | 40 – 55 | 30 – 40 | 30 – 40 | Weak evidence |
| C4 | 55 – 70 | 20 – 30 | 20 – 30 | Strong evidence |
| C5 | 70 – 90 | 10 – 20 | 10 – 20 | Very strong evidence |
| C6 | 90 – 100 | 0 – 10 | 0 – 10 | Extremely strong evidence |

Table 5.3: A proposed AI-adaptive C-Scale evaluation of strength of evidence for DFAI

This is by no means a standard evaluation, but rather a tentative proposition that will need to be refined as research in this field progresses. Additionally, unlike the LR recommendation and the C-Scale proposals, which are based on hypothesis (or strength of hypothesis) about source identification during a forensic investigation, the DFAI C-scale evaluation approach is fairly generic (for hypothesis and AI models) and applicable in a wide variety of situations, including strength of evidence.

As previously noted, human expert interpretation and evaluation are key components of DFAI in a partially automated setup because it is difficult to predetermine all of the reasonings required to do a forensic investigation work (Bollé, 2020). However, in a fully automated scenario, learning algorithms in conjunction with contextually structured expert systems can incorporate domain-specific knowledge-derived rules. The expert system can also be configured to evaluate every hypothesis at each modular level and make recommendations based on codified likelihood ratios.

## 5.2 Standardization in DFAI

The issue of standardization in digital forensics has persisted for several years; first because standard guidelines have been unable to keep up with the dynamic pace of technological sophistication, and second, because forensic stakeholders have been unable to agree on certain rules and standards, resulting in conflict of interest (Bennet, 2012). Additionally, the distinctiveness of investigation, the domain's diversity, and the existence of disparate legislative frameworks are all reasons cited as impediments to the standardization of the DF

field (Palmer, 2001; Reith, Carr and Gunsch, 2002). Nowadays, when it comes to standardization, the majority of what we encounter (with guidelines) are boxes to be ticked - since the belief is that the more details, the better the standard (Sommer, 2018). Nonetheless, the "Forensic Science Regulator" in a 2016 guidance draft highlighted the validation of forensic methods as a standard, rather than the software tool (FSR, 2016). This method validation entails a number of assessments, including the evaluation of data samples, which are relatively small in DF (Arshad, Aman and Abiodun, 2018). Standardization in DF (as well as DFAI) is a broad and intricate area of study, as every component of DF requires it. However, within the limits of this thesis and as part of the preliminary advancement of DFAI (for which further study is envisaged), we examine standardization in the context of forensic datasets and error rates.

### 5.2.1 Datasets Standardization in DFAI

Datasets (or data samples) are a critical component of AI, as they define the validity of an AI model to a great extent. A dataset is a set of connected, discrete items that, depending on their context, have varied meanings and are employed in some form of experiment or analysis (Grajeda, Breitinger and Baggili, 2017). To evaluate or test novel approaches or to replicate existing procedures, similar data sets are required; for example, investigations on facial recognition require human facial sample data. Similarly, an inquiry into message spamming necessitates the collection of e-mail samples. Datasets are often beneficial in the following ways, according to the National Institute of Standards and Technology (2019)[125]:

i. *For training purposes*: dataset is generated for training purposes, i.e., simulation of case scenarios to train a model to learn the specifics of that environment, and to facilitate practitioner's training on case handling so that their ability to identify, examine, and interpret information can be assessed.
ii. *Tool validation:* wherein dataset is utilized to determine the completeness and correctness of a tool when it is deployed in a given scenario.
iii. *Familiarity with tool behavior:* for instance, a dataset collected from users' software interaction traces. These datasets are critical for decoding how certain software behaves on a computer and helping in the understanding of digital traces left by usage (Horsman and Lyle, 2021).

[125] National Institute of Standards and Technology, 2019. The CFReDS Project. Available at https://www.cfreds.nist.gov/. (Accessed 20 June 2021).

The process of creating a dataset is critical, even more so in the domain of DF, where each component must be verifiable, fit for purpose, and compliant with some set of standards. Therefore, the created dataset must be realistic and reliable (Gobel et al., 2020). This also includes having a high-quality, correctly labelled dataset that is identical to the real-world use case for testing and evaluation purposes, sufficient enough in quantity for adequate learning, and available to ensure reproducibility (Grajeda, Breitinger and Baggili, 2017). In the context of DFAI, there are a few considerations that must be made in order to conduct a forensically sound operation with respect to datasets.

Due to limited availability of datasets in DF, practitioners frequently overuse a single data corpus in developing several tools and methodologies, resulting in solutions gradually adapting to a dataset over time. For example, in Chapter 4, we used the Enron dataset to develop a particular forensic solution on e-mails. The Enron corpus has developed into a research treasure for a variety of forensic solutions, including e-mail classification (Miyamoto et al., 2008; Guo, Jin and Qian, 2013; Morovati and Kadam, 2019), person of interest identification (Noever, 2022), and other forensic linguistics works (Farkhund et al., 2008; Bogawar and Bhoyar, 2016; Emad et al., 2019). However, proving that a solution based on a single corpus is sufficiently generalizable to establish a conclusion in a forensic investigation will be difficult. Nevertheless, this is a widely recognized issue among stakeholders, and while it may be excusable in peer reviews, it is a major issue in the standardization of DF that requires immediate resolution. Similarly, while a sufficiently workable DF dataset is being sought, caution should be exercised when using a (single) dataset as a benchmark for a tool or method's validity.

Datasets are created as a "mock-up" of a specific scenario, representing the activities/events that occur within an environment; supposedly within a specified time period. Each use case is time-dependent; as such, the continued relevance of a particular use case (from a previous period) in a future period may be debatable. This is particularly true in the domain of DF. For instance, given the advancements in computer network architecture, it may be illogical to use a dataset of network traffic from the 1990s to model an intrusion detection system today. This is also a point made in (McHugh, 2001). Similarly, it may seem counterintuitive to argue that a model trained on images retrieved from an older (e.g., 2000) CCTV footage or camera is helpful for identifying objects in a contemporary crime scene image – technology has improved. However, in an ideal circumstance and for a robust model, updating the dataset with

a collection of new features compatible with recent realities, rather than completely discarding the old dataset, should be viable.

Criminal cases are predominantly local in nature, and while they may have a global dimension, investigations should consider local nuances. For instance, in a typical forensic linguistics investigation (e.g., cyberbullying), a language corpus plays a vital role. However, native speakers' use of language (for example, English) may differ greatly from those of non-native speakers. Language, in usage and writing, varies across borders. An AI model trained to identify instances of bullying using a message corpus extracted from British databases may not be fully representative of the same use case in Anglophone Africa – there are some English terms that are offensive to native speakers but inconsequential to non-natives. As such, a training dataset for DFAI should accurately reflect the use case (in terms of location and dimension) to which it is meant to be applied.

Lastly, the demand for synthetically generated datasets is increasing in the DF domain, and rightly so. The issues of privacy, unavailability, and non-sharing policy continue to be a barrier to getting forensically viable datasets for the purpose of training, testing, and validating forensic tools. Synthetic data, first introduced in (Rubin, 1993; Little, 1993), is described as an artificially generated data that contains statistical features of the original data. While synthetic data can be extremely beneficial for research and education, the question is whether any novel technique can be tested on fictitious data (Baggili and Breitinger, 2015), and particularly for DF; whether a perfect simulation of a crime event can be achieved. Nonetheless, several research (not related to DF) have demonstrated the usefulness of synthetic data in comparison to actual data (Heyburn et al., 2018; Rankin et al., 2020), in which a model was trained on synthetic data and tested on real data. The results indicated that the accuracy of numerous ML methods were slightly decreased and varied as compared to when the real data set was used. Synthetic data can be used to enrich or expand an existing dataset or to correct for data imbalances caused by limited occurrence of an event. In DFAI, modelling with synthetic data can be advantageous in some circumstances, but not in all. Synthetic data generation involves a purpose-built dataset that may be too specific for general-purpose solutions; demonstrating the results' suitability for real-world crime data may be problematic. This point is highlighted in (Yannikos et al., 2014), while some other challenges are emphasized in (Horsman and Lyle, 2021). Furthermore, synthetic datasets are randomised, which means that the data do not follow a regular pattern. We foresee an extended challenge if the dataset is used to train an unsupervised neural network model - the model may learn non-interpretable patterns. While it

is natural to assume that random data is less biased, there is no means to verify this claim. Thus, while synthetic datasets may be advantageous for solving specific ML problems, their usage in DFAI should be carefully addressed.

### 5.2.2 Error Rates Standardization in DFAI

As critical as accuracy is in determining the correctness of an evidence mining process, so also is the error rate. The error rate not only indicates the probability that a particular result is correct, or the strength of a technique, but also its limitations. According to the Scientific Working Group on Digital Evidence (SWGDE), the term "error" does not allude to a mistake or blunder, but rather to the inevitable uncertainty inherent in scientific measurements. Numerous factors can influence these uncertainties, including algorithmic flaws, statistical probability, physical measurements, and human error (SWGDE, 2018). In Chapter 2, we discussed the Daubert standard, and one of the criteria for validating scientific methods under Daubert is error rate. Indeed, some of the other requirements are largely contextualized around error rate. For instance, the Daubert standard involves testing of a theory or technique; how can we test a hypothesis and its alternatives or a method without assessing the rate of uncertainty? Likewise, peer review publishing of the method is essential. How scientifically valid is a published work that does not acknowledge methodological uncertainties? This highlights how crucial error rates assessment is to forensic methods.

In alignment with the guidance offered in (SWGDE, 2018), the uncertainty associated with any DFAI technique can be assessed in two ways: random and systematic. Random uncertainties are related with the algorithmic component of the technique and are frequently associated with measurements, while systematic uncertainties are typically associated with implementation — they occur in tools. DF tools represents implementation of a technique, and their functionality varied according to the task they seek to resolve. It is not uncommon for software to possibly have intrinsic bugs (Walker, 2011) — triggered by logical flaws or incorrect instructions. For instance, an erroneous string search algorithm can cause a tool to report certain critical evidence incompletely. In this instance, the tool will extract some relevant strings but may underreport them. Because these flaws are not random, the tool will frequently generate the same result when given the same input, which may be inadvertently deceptive to an examiner. As a result, additional error mitigation strategies may be necessary to detect and fix it.

Due to the probabilistic nature of DFAI algorithms (the outcome of which may be random), the error rates are estimated in terms of false positive and false negative rates (which we

discussed earlier in this Chapter). Depending on the percentages of these errors, and as long as there is sufficient trust in the algorithm's optimality, the error rates will merely describe the technique's limitations; not its true efficiency. Reporting and publishing error rates in a technique should be encouraged in DF domain, and this should be particularly true for DFAI. This increases method's transparency and ensures that the expected outcome is known in the event of method replication. Additionally, disclosing error rates provides prospective researchers with a baseline understanding of the components that function efficiently, where improvements are anticipated, as well as prevent potential biases in interpretation. Mitigating this error may not be straightforward scientifically, as it is dependent on a variety of factors; however, algorithm optimization, sufficient datasets, accurate labelling (in supervised settings), and strong domain knowledge (for proper interpretations) are some of the ways to achieve a fairly reasonable success. Additional mitigating strategies for systematic errors include training, written procedures, documentation, peer review, and testing (SWGDE, 2018).

## 5.3 Optimization of DFAI Techniques

Optimizing an AI algorithm can be a difficult challenge, all the more so when the approximation function contains a large number of inputs, an unknown functional structure, non-differentiable elements, and noise. The key aim of ML is to solve some kind of optimization problem. Thus, constructing a ML model entails initializing and optimizing weight parameters using an optimization algorithm until the objective function tend towards minimum value or towards a maximum value in terms of accuracy (Sun et al., 2020). In addition to learning in predictive modelling, optimization is necessary at several stages of the process, and it includes selecting: 1) the model's hyper-parameters (HPs); 2) the transformation techniques to apply to the model prior to modelling; and 3) the modelling pipeline to apply. This section is not intended to discuss the depth of optimization in AI, but to briefly discuss hyper-parameters optimization (HPO) (Steinholtz, 2018) as a critical component in optimizing a DFAI model.

Two parameters are crucial in ML models: 1) the model parameters, which are initialised and updated throughout the learning process; and 2) the HPs, which define the model's structure, not directly estimable from data, and must be set prior to training (Kuhn and Kjell, 2013). The traditional method, which is still used in research but requires knowledge of the ML algorithm's HP value configurations, entails manually tuning the HP until the desired result is achieved (Abreu, 2019). This is ineffective in some cases, particularly for complex models

with non-linear HP interactions (Yang and Shami, 2020). However, HPO is an automatic technique that improves the effectiveness of applying ML to practical problems (Elshawi, Maher and Sakr, 2019). Numerous circumstances may necessitate the application of HPO techniques (Hutter, Kotthoff and Vanschoren, 2019); we zero-in on a few of them below, focusing on forensic investigation tasks in the context of DFAI.

i. Conducting a digital forensic investigation takes an inordinate amount of time. Over the years, reducing this time has been a key focus of research in this domain. Similarly, machine-driven techniques can be time intensive, depending on the size of the dataset or the number of HP, and applying AI techniques to already difficult forensic investigation will almost certainly increase the complexity. HPO can significantly reduce the amount of human effort required to tune these HP, hence reducing the overall analysis time.
ii. We already discussed the importance of performance in the overall scheme of DFAI operations. ML methods have a variety of HP settings necessary to achieve optimality for different dataset and problem. Numerous techniques exist in HPO that can aid in optimizing the performance of AI-based models by searching across multiple optimization spaces in pursuit of the global optimum for a particular problem.
iii. As previously stated, reproducibility is a fundamental need for a standard DF approach. HPO can assist in achieving this goal in a variety of ways. For instance, when comparing the efficacy of various AI algorithms on a specific analysis, using the same HP settings across all models creates a fair comparison mechanism. This can also aid in determining the most appropriate algorithm for a given problem. Reporting these HP configurations can be beneficial in the event of replication.

As with conventional AI models, developing a DFAI model with HPO in mind entails the following steps: defining an estimator (a classifier or regressor) and its objective function, defining a search (configuration) space, defining an optimization method for determining suitable HP combinations, and defining an evaluation function for comparing the performance of different HP configurations (Yang and Shami, 2020). A typical HP configuration can be discrete (for example, the number of clusters, *k*), continuous (as for multiple LR values), categorical (e.g., optimizer type), or binary (for early stopping or not) all of which can be combined to produce an optimized model. Because the majority of ML algorithms have well-defined open-source frameworks (such as *scikit learn*) that can assist in solving problems by tuning (changing values) some already pre-set HPs, we will focus on

HPOs related to DL models because they have more parameters to set. HP in DL are configured and tuned in accordance with the complexity of the dataset and task, and they are proportional to the number of neurons in each layer (Koutsoukas et al., 2017). The initial parameter setting for a DL model is to specify the loss function (binary cross-entropy, multi-class cross-entropy, or squared error loss) appropriate for the problem type. Then comes the type of activation function (e.g., ReLU, sigmoid, tanh, SoftMax, etc.) that describes how the weighted sum of the input is transformed into the output. Finally, the optimizer type is specified, which may be stochastic gradient descent (SGD) (Goodfellow, Bengio and Courville, 2016), Adam, or root mean square propagation (RMSprop) (Tieleman and Hinton, 2012). In what follows, we describe several optimization techniques that can be vital to the optimization of a DFAI model.

### 5.3.1 Optimization Methods in DFAI

#### 5.3.1.1 Manual Tuning

This method involves tuning parameters manually. It entails testing a large number of HP values based on past experience, guesswork, or analysis of prior results. The approach is to improve parameter guesses iteratively until a satisfying result is obtained. This approach may be impractical for a variety of issues, particularly those involving DF analysis, that could involve large number of HP or complex models (Yang and Shami, 2020). However, this technique can improve interpretability by allowing for the assessment of the model's various working parts as the parameters are tuned.

#### 5.3.1.2 Grid Search (GS)

This is a frequently used technique for exploring the HP configuration space (Injadat et al., 2020). It does an extensive, parallel (i.e., independent of time sequence and previous search) search of the configuration space, and it is suitable within a limited search space; otherwise, it may suffer from the "curse of dimensionality" (Bach, 2017). This strategy is nearly always preferred when the examiner has sufficient knowledge of the HP to specify a finite set of values (Hutter, Kotthoff and Vanschoren., 2019) for the search space (for example, recognizing that no more than three HPs should be tuned concurrently) (Yu and Zhu, 2020). Due to the computational complexity of GS (Lorenzo et al., 2017), its application in DFAI is mostly focused on comparing different algorithms (Rami and Mohammed, 2019) in order to determine

the one that performs the best on a particular forensic task. A botnet detection method using GS optimization technique is described in (Gonzalez-Cuautle et al., 2019).

#### 5.3.1.3 Random Search (RS)

RS was proposed in (Bergstra and Bengio, 2012) as a way to circumvent GS's limitations. Unlike GS, however, RS randomly selects a predefined number of candidate samples between a specified upper and lower bound and trains them until the target accuracy is achieved or specified budget is exhausted. RS explores, in parallel, a bigger space on a limited budget by allocating resources to best performing regions (Krivulin, Dennis and Charles, 2005) to discover optimal configurations set (referred to as the Monte Carlo technique (Harrison, 2010)).

Due to the simplicity with which RS parallelizes, it is an excellent candidate for DFAI tasks involving CNNs, such as multimedia forensics (e.g., audio and video), image forensics, and others, in which (low-dimensional) features are mapped from one layer to the next. This method is time and memory consuming. To optimise the process, a batching technique (Pavlo, 2014) is used that makes use of the batch size and learning rate to reduce training time without compromising on performance. RS may be advantageous in this scenario for determining the optimal range of values for these parameters (Ari and Heru, 2020) because just the search space must be specified. Furthermore, RS' application in optimising multimedia forensics analysis suggests that it may be crucial for RNN, however RS has the disadvantage of not factoring-in previous results during evaluation (Yang and Shami, 2020). Therefore, when used in recursive tasks such as event reconstruction in DFAI, RS may produce less-than-optimal outcomes.

#### 5.3.1.4 Gradient Descent (GD)

The gradient descent (Bengio, 2000) optimization computes the gradient of variables so as to determine the most promising path to the optimum. Gradient-based optimization techniques converge faster to the local minimum than the previously described techniques, but they are only applicable to continuous HP, such as the learning rate in NN (Maclaurin, Duvenaud and Adams, 2015), as other types of HP (e.g., categorical) lack gradient direction. The application of GD is almost ubiquitous in DFAI, as it is utilised in virtually all DL models. This is one of the simplest optimization architectures to comprehend and interpret. However, the findings in (Goodfellow, 2014b) established the existence of "Catastrophic Forgetting" when GD is

utilized, most notably in reproduction. That is, when trained on a new task using solely gradient descent, ML models may forget what they learned on a previous assignment. However, a combination with dropout (Dahl, Sainath and Hinton, 2013) is recommended.

#### 5.3.1.5 Bayesian Optimization (BO)

BO (Jones et al., 1998; Snoek et al., 2012) is an iterative algorithm that calculates future evaluation points based on the prior results. It is a typical model for all sorts of global optimization, with the goal of becoming less incorrect with more data (Koehrsen, 2018). BO attempts to locate the local optimum with the fewest possible trials which enables it to operate faster regardless of whether the objective function is stochastic, continuous, convex, or non-convex. BO, on the other hand, is a sequential approach that is difficult to parallelize (Yang and Shami, 2020). Gaussian process (GP) (Seeger, 2004), Sequential Model-based algorithm configuration (SMAC) (Hutter, Kotthoff and Vanschoren., 2011), and Tree Parzen Estimator (TPE) (Bergstra et al., 2011) are also examples of common BO algorithms.

BO is particularly advantageous with tools such as the Waikato Environment for Knowledge Analysis (WEKA) (Hall et al., 2009), an open-source set of ML and data processing algorithms. Numerous DF analysis approaches (Bhat et al., 2011; Nirkhi, Dharaskar and Thakare, 2012; Uma and Nikkath, 2021) have been proposed or implemented using WEKA – exploiting its extensive data mining capabilities and the ability to choose from or compare a diverse set of extensible, base learning algorithms for a specific forensic task. Selecting the best method and HPs for a WEKA-based DFAI analysis might be hard. In this case, BO's excellent properties can assist in selecting the appropriate ML approach and HP settings to minimize analytical errors. The works presented in (Thornton et al., 2013) and (Kunang et al., 2020) demonstrates how BO can be utilized as meta-learning (specifically, with SMAC and TPE) to guide the selection of ML algorithms and HPO settings that outperform conventional selections on a classification task.

#### 5.3.1.6 Multi-fidelity Optimization (MFO)

MFO techniques are frequently used to overcome the time constraints limitations imposed by other HPO due to huge configuration space and datasets. MFO evaluates practical applications by combining low- and high-fidelity measures (Zhang et al., 2016). In low-fidelity evaluation, a small subset is examined at a low cost, resulting in poor generalization performance; while in high-fidelity evaluation, a larger subset is examined at a higher cost, but with improved

generalization performance (Yang and Shami, 2020). MFO techniques include “Bandit-based” (Karnin, Koren and Somekh, 2013) methods that allocates computational resources to the “best arm” (most promising) HP configurations. The two most commonly used bandit-based techniques are successive halving (SHA) (Jamieson and Talwalkar, 2015) and hyperband (HB) (Jamieson and Talwalkar, 2015; Li et al., 2017). Transfer learning (TL) (Zhan et al., 2017); which is the process by which previously stored knowledge is used to solve unrelated but related problems, is a technique for applying MFO in DFAI. TL has been applied to a range of DFAI problems (Zhan et al. 2017; Al Banna et al., 2019), most notably image forensics and detection problems using labelled samples. Thus, depending on the size of the stored information (dataset), the investigative problem, and available computational resources, low or high fidelity optimization can be beneficial for determining the optimal solution. Prasse et al. (2019) illustrates how to detect (signature-based and unknown) malware-infected domain based on HTTPS traffic, using TL and HB optimization. Additionally, a state-of-the-art hyper-parameter optimization technique called Bayesian Optimization Hyperband (BOHB) (Falkner, Klein and Hutter, 2018), which combines BO and HB to maximize the benefits of both, is gaining attention, and it will be interesting to see how DF research employs this promising technique in the future.

#### 5.3.1.7 Metaheuristic Algorithms

Metaheuristic algorithms are a popular type of optimization technique that are primarily inspired by biological evolution and genetic mutations. Metaheuristic techniques are capable of resolving problems that are not continuous, non-convex, or non-smooth (Yang and Shami, 2020). Population-based optimization algorithms (POAs) (Eggensperger, 2013) are an excellent example of metaheuristic algorithms since they update and evaluate each generation within a population until the global optimum is found. The two most frequently utilized types of POA are genetic algorithms (GA) and particle swarm optimization (PSO) (Shi and Eberhart, 1998). PSO, specifically, is an evolutionary algorithm that functions by allowing a group of particles (swarm) to traverse the search space in a semi-random fashion (Steinholtz, 2018), while simultaneously discovering the optimal solution through information sharing across swarms.

PSO is well-suited for network forensics, as training such models can be time-consuming due to the requirement of identifying complex patterns from vast volumes of data. Iterative reverse engineering of parser and network traffic logs is required to discover network intrusion or

attack; this might be challenging for humans (Koroniotis, Moustafa and Stinikova, 2020). The work described in (Koroniotis, Moustafa and Stinikova, 2020) demonstrates the efficacy of PSO as a useful tool for minimizing/maximizing an objective function and determining the best HPs (such as epochs, LR, and batch size) that contribute to the AUC accuracy and false alarm rate reduction of the deep forensic model.

### 5.3.2 Discussion

It is worth emphasizing that the techniques discussed here are by no means exhaustive in terms of definition, components, and applicability. These few are chosen for their popularity and as a means of briefly discussing optimization techniques in the context of DFAI-models. As such, in depth discussions about HPOs are available in (Sun et al., 2020; Yu and Zhu, 2020; Yang and Shami, 2020). In general, depending on the size of the data, the complexity of the model (e.g., the number of hidden layers in a neural network or the number of neighbours in a KNN), and the available computational resources, an HP configuration may lengthen the time required to complete a forensic task. Further along this line, in most cases, only a few HP have a substantial effect on the model's performance in ML methods (Yang and Shami, 2020). Hence, having many HP configurations exponentially increases the complexity of the search space. However, with deep learning, HPO techniques will require significant resources, particularly when dealing with large datasets. Considering all of these complexities, especially in the context of DFAI, where timeliness, transparency, and interpretability are critical, a well-chosen HPO technique should aid in rapid convergence and avoid random results. Given that DF analysis are case-specific, often distinctive, with interpretability as a fundamental requirement, decomposing complexity should be a priority. Summarily, unless forensic investigators have sufficient computing resources and a working knowledge of the parameter settings for various HPO techniques, they may choose to consider the default HP settings in major open-source ML libraries, or make use of a simple linear model with reduced complexity, where necessary. In case of a self-defined DNN model, basic HP settings and early stopping techniques can be considered.

Finally, to summarize the various HPO algorithms mentioned thus far, Table 5.4 compares these HPO algorithms and their respective strengths and drawbacks, as adapted from (Yang and Shami, 2020).

| HPO Technique | Strengths | Drawbacks | Time Complexity | Use case in DFAI |
|---|---|---|---|---|
| **GS** | • Simple<br>• Independent (Parallelization)<br>• Exhaustive use of the search space | • Effective with categorical HP<br>• Time-consuming<br>• HP grows exponentially<br>• Possible overfitting | $O(n^k)$ | Comparison of DFAI algorithms, botnet detection, etc. |
| **RS** | • Effective parallelization<br>• Improvement over GS<br>• Better with low-dimensional data<br>• Reduce overfitting<br>• No HP tunning except for specifying search space | • Less-effective with conditional HP<br>• Ignores previous result during evaluation<br>• Potential for variance since it is random | $O(n)$ | Multimedia forensics, identifying HPs (e.g., batch size, learning rate, etc.) for optimal forensic analysis. |
| **GD** | • Fast convergence speed for continuous HP such as learning rates | • Support only continuous HP<br>• Detects only a local optimum | $O(n^k)$ | Event reconstruction, text analysis, multimedia forensics, etc. |
| **BO (BO-GP, SMAC, BO-TPE)** | • Fast convergence speed for continuous HP<br>• Effective with all types of HP (in SMAC and BO-TPE cases)<br>• Computes mean and variance | • Poor parallelization capacity<br>• Slow convergence with dimension > 1000<br>• Specification of prior is difficult | $O(n^3)(BO-GP)$, $O(nlogn)$ (SMAC, BO-TPE) | Useful for WEKA, identifying best algorithm for forensic task, Meta-learning, etc. |
| **HP** | • Better parallelization | • Less effective with conditional HP | $O(nlogn)$ | Malware detection, Network intrusion analysis, Transfer |

| | | | | |
|---|---|---|---|---|
| | | • Subset with small budget required | | learning for forensics, etc. |
| **BO-HB** | • Effective with all types of HP<br>• Better parallelization | • Subset with small budget required | $O(nlogn)$ | |
| **GA** | • No initialization<br>• Effective with all types of HP<br>• Produces multiple optimal solutions<br>• Possible global optimal solution<br>• Large solution space<br>• Supports multiple objective function | • Poor parallelization capacity<br>• Computational complexity | $O(n^2)$ | Best performing parameter selection for forensic task |
| **PSO** | • Better parallelization<br>• Effective with all types of HP<br>• Efficient global search algorithm<br>• Insensitive to scaling of design variables | • Initialization required<br>• Weak local optimum search space | $O(nlogn)$ | Network forensics, optimal HP configuration for DFAI analysis. |

Table 5.4: The comparison of HPO techniques (*n* denote the number of HP values and *k* is the number of HP)

## 5.4 Chapter Summary

In this chapter, we lay the groundwork for defining "Digital Forensics AI" (DFAI) and dispelling the common notion that it is synonymous with Forensic AI. As a result, we covered the primary evaluation techniques for AI-based methods such as classification, regression and clustering that are utilized in DF analysis, as well as the critical metrics that should not be overlooked when reporting forensic results. Additionally, we reviewed some fundamental DFAI standards. In this case, we focused on two fundamental factors: datasets and error rates,

both of which are critical to DFAI. We stressed the significance of exercising caution while working with synthetic datasets and of reporting error rates in DFAI. Finally, we compared the strengths and drawbacks of numerous hyperparameter optimization techniques.

# Chapter 6

# Mitigating Mistrust in Digital Forensics AI: on Explainability and Interpretability of Evidence Mining Techniques

## 6.1 The Concepts

Over the last few decades, objectivity in fact finding has switched from human to machine-generated proofs, and to an extent, with improved accuracy (Roth, 2015). Just as human experts' analysis of the same case can result in divergent opinions, machines have also expressed different viewpoints on the same scientific evidence, raising serious concerns about the challenges that machine-generated evidence poses for the legality of digital evidence. Further along this line, just as out-of-court testimony, such as hearsay (Goodison et al., 2015); poses a risk to the justice system in terms of ambiguity; dishonesty, misconceptions; and memory loss (Morgan, 1948), machine testimonies (sources) may indeed present closed-box challenges (Carr, 2014; Pasquale, 2015) that may lead fact finders to drawing incorrect/incomplete inferences (Roth, 2015). When the output of a machine-driven analysis is imprecise or ambiguous, or in a situation where an event is incorrectly interpreted, several factors may be responsible, including but not limited to design (e.g., erroneous algorithms/code), input (e.g., skewed, or disproportionate dataset), model (e.g., defective functional components of the system), and environmental forces (e.g., OS, distributed platforms, etc.). All of a machine's vital components (design, input, and operational modules) are designed and structured by humans, which is why some scholars argue that machine's credibility is highly dependent on human. Thus, human being is the true declarant[126] of any output conveyed by a machine (Wolfson, 2005). While the designer or operator of a machine may share or bear moral responsibility for the assertions made by the machine, she is not the only source for the assertions (Roth, 2015). She is only regurgitating the output of a machine-driven operation. As the expert's opinion is the product of "distributed cognition" between the expert and other experiential influences (Dror and Mnookin, 2010), so is the result of a

[126] "Declarant" is a term used in the context of hearsay as a label for the witness tendering evidence statement as truth of the matter asserted.

machine-driven forensic investigation; it is driven by the distributed cognition of humans and technology (Dror and Mnookin, 2010).

The foregoing demonstrates the connections between humans and machines, as well as the impact of the former on the entire range of machine-generated evidence, the closed-box, and the determination of responsibility. The subject of AI and its inscrutability (opaqueness) is still an area of ongoing research; and given the widespread misconceptions about whether AI systems should be explainable or interpretable, the road to a unifying consensus may be further away. AI/ML-powered systems are being deployed in a variety of sectors of our daily lives, with diverse consequences in each of these sectors. There is a rising concern over AI systems' inexplicability, particularly in fields where decisions have significant consequences for individuals affected, and where transparency, accountability, or legal compliance are required (such as health and law) (Coyle and Weller, 2020). Particularly, in DF, AI-based technologies have been instrumental for identifying or detecting interesting clues that are subsequently analyzed to support or rebut a certain claim. The outcome of a forensic investigation must be presented in an understandable and comprehensible manner, but when the processes that produced them are debatable scientifically, or are insufficiently transparent to establish a conclusion, then the additional level of complexity must be addressed. Conventional (i.e., non-AI-based) methods utilizing specialized forensic tools (that lawyers, jurors, and others are familiar with) have been helpful in the extraction, analysis, and reporting of digital evidence over the years; however, the sophistication of technology today, and the manner in which it is used to commit crime, necessitates the deployment of more robust, autonomous, and equally intelligent systems such as AI to identify potential evidence.

This chapter's primary objective is to examine, first, the differing views on explainability and interpretability in AI, with a particular emphasis on how they affect DF and evidence mined using AI algorithms. This is necessary to provide the proper foundation for these ambiguous concepts. To put things in the right perspective, guidance through literature will, presumably, assist to draw the right conclusions, particularly as they pertain to DFAI. Second, the concerns about closed-boxes are examined, as are the numerous approaches and attempts to discover a practical solution (even though that remains elusive). After discussing the numerous work-around proposed, domain-specific recommendations to mitigate mistrust in AI-powered digital forensics analysis are then offered. Furthermore, a formal pre-concept for explainable digital forensics AI is offered, along with a number of relevant methods for

delivering comprehensible interpretations for AI models and their applicability to AI-based DF analysis.

While the promise of AI was that it would enable better decision-making, as seen in some forms of medical diagnostics (De Fauw et al., 2018) or monitoring attempted financial frauds (Aziz and Dowling, 2019), questions have been raised concerning its usage in critical contexts such as the justice system and policing (Coyle and Weller, 2020). The underlying issue or demand is the need to explain to interested parties, the conclusions reached as a result of algorithmic decisions. Explainability of AI closed-box systems (Samek, Weigand and Muller, 2017; Pedreschi et al., 2018; Samek et al., 2019; Guidotti et al., 2019), also known as Explainable AI (XAI), is a field of research devoted to making AI systems and the data they use transparent (Gross-Brown, 2015) by "*glass-boxing*" the system's operating components. As a result of AI's pervasiveness across several disciplines, explanation connotes different things to different fields, and emphasis is assigned based on technical requirements and the consequences of the outcomes. For example, while the decision-making process of a recommender system will require little or no explanation, questions about the decision-making mechanism of a crime prediction or recidivism algorithm will continue to be raised. XAI carries a huge amount of weight in law and everything that surrounds the justice system, including policing, law enforcement, and DF, because the implications of a wrong machine-generated decision in these areas could be grave. As a result, arguments have developed about whether the results of a closed-box AI system should be explainable (Arrieta et al., 2020) or interpretable (Rudin, 2019); whereas some argue in favour of understandable or responsible (Benjamins, Barbado and Sierra, 2019) systems instead. However, these notions — particularly, interpretable and explainable AI — have been used interchangeably across literatures. To demonstrate these misconceptions over time and the gradual shift in reasoning toward interpretability in literatures, a simple search in the Scopus®[127] database was conducted for publications whose keywords, title and abstract contained the terms "Interpretable Artificial Intelligence", "Explainable Artificial Intelligence", or "XAI." The outcome is depicted in Figure 6.1. In a similar vein, figure 6.2 illustrates the application of these concepts across multiple domains of knowledge. This is to demonstrate the critical nature of transparent AI across these disciplines, as evidenced by the volume of literature devoted to it. Further discussions about the figures are presented later in this chapter. But first, to gain a better

[127] https://www.scopus.com/home.uri

understanding of these concepts, definitions and distinctions among terms may be necessary; and in what follows, the summary of the most commonly used nomenclatures are presented.

**Explainability**: is a term that refers to the concept of explanation as a link between humans and a machine's decision-making process that is both accurate and understandable to humans (Guidotti et al., 2019). It embodies the notion that an AI model's high abstraction level and its output can be rationally explained in a way that is acceptable and understandable to humans.

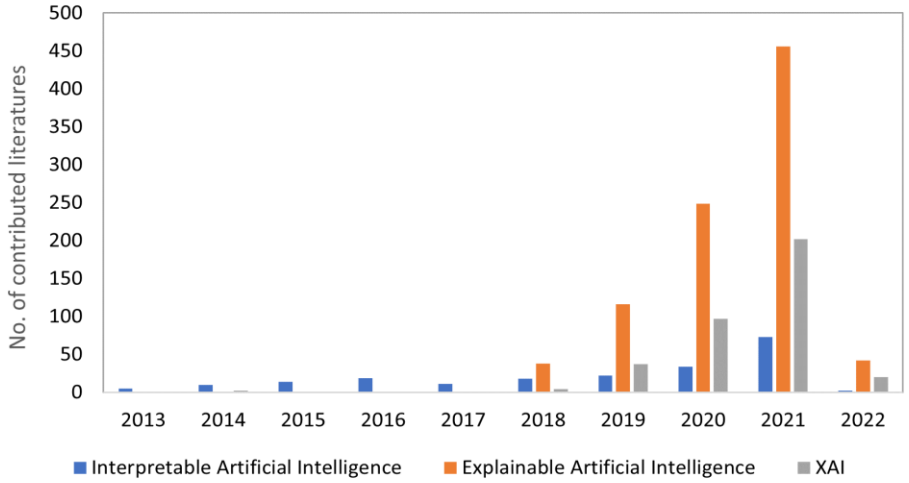


Figure 6.1: represent the evolution of publications with titles, abstracts and/or keywords that refers to explainable/interpretable AI over the last years. This chart depicts the numbers as of 10th January, 2022 as retrieved from the Scopus database. The legends indicate the exact search terms used in the query.

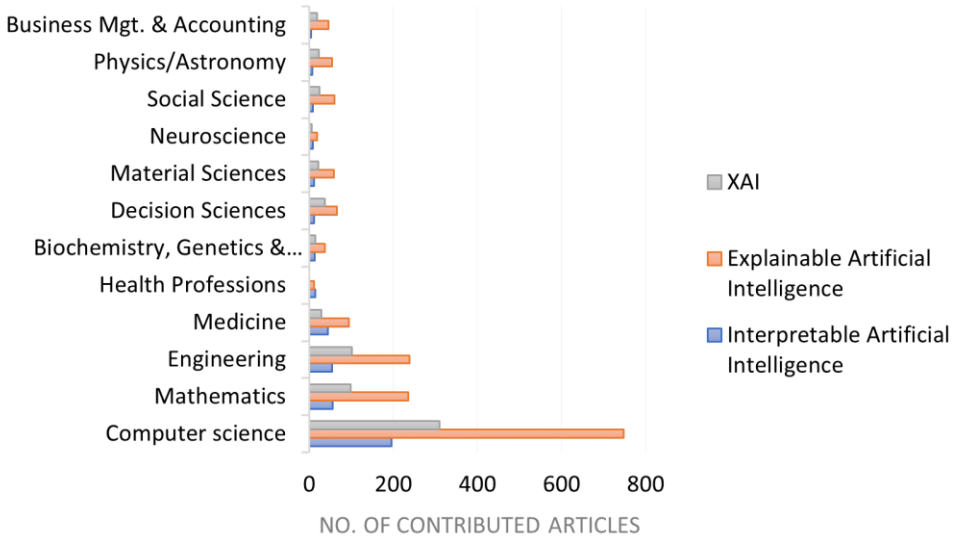


Figure 6.2: depicts the number of published works with title, abstract and/or keywords that refers to the terms in the legend. The figures are as derived from the Scopus database as of 10th of January, 2022.

Classical ML models tends to be readily explainable; albeit with less performance, while others such as DNN/DL performs better, but remains much harder to explain. Inexplicability may lead to misunderstanding and distrust in AI-enabled systems.

To contextualize explainability within the framework of AI, a truly explainable AI makes use of knowledge bases during analysis and provides a technique for deconstructing the output in a way that logically justifies the interpretation provided to the input data (Hall et al., 2021). According to Gunning in (Gunning, 2019), "*XAI will create a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generations of artificially intelligent partners*." The emphasis on XAI gained traction when it became apparent that in order for AI systems to be trusted, they must always prioritize the plight of end users (e.g., different stakeholders in a legal system). It is self-evident that when a machine-derived decision has an impact on humans' lives (as is the case in medicine, law, or defense), satisfactory explanation becomes imminent (Lipton, 2018).

**Interpretability**: refers to the capacity to convey an explanation or meaning in a human-comprehensible manner (Arrieta et al., 2020). Due to the fact that interpretability is domain-specific (Rupin, 2006; Huysman et al., 2011), a universal definition is unattainable. Nonetheless, interpretability in the context of machine-generated output should be viewed in terms of conformance to structural domain knowledge, causality, or physical constraints (Rudin, 2019), as well as sparsity (of data); the latter of which can be measured in terms of human cognitive capacity (all at once) (Miller, 1956; Cowan, 2010).

Not only does an interpretable system enable users to visualize the model, but also to study and comprehend the mathematical underpinnings of how the input is mapped to the output (Doran, Schulz and Besold, 2017). It implies transparency and understandability. Interpretable consideration for an AI model is an additional driver that, according to (Arrieta et al., 2020), can improve its implementation in three (3) ways: 1) ensure objectivity in decision-making, i.e., transparent management of bias in the training dataset; 2) resilience towards adversarial perturbations that may impede prediction; and 3) ensure that only correct variables are used in output's inferencing, i.e., assurance that model reasoning is premised on true causality. The above indicates that the practicality of an interpretable AI system is contingent upon the understandability of its predictions, the visualization of its discriminating rules, or the disclosure of factors that might perturb the model (Hall, 2018).

**Understandability**: or intelligibility, concerns model's characteristic that enable it to be self-explicit in terms of its operational functionality — without the need to explain its internal structure or the underlying algorithms used to process data (Montavon, 2018).

**Comprehensibility**: is typically measured in terms of the model's complexity (Guidotti et al., 2019), which includes the model's capacity to describe its learning process in an understandable manner (Craven, 1996; Gleicher, 2016). According to Michalski's theory (Michalski, 1983) model's comprehensibility and as quoted in (Arrieta et al., 2020), which states that: *"the result of computer induction should be symbolic descriptions of a given entities, semantically and structurally similar to those a human expert might produce observing the same entities. Components of these descriptions should be comprehensible as single 'chunks' of information, directly interpretable in natural language, and should relate quantitative and qualitative concepts in an integrated fashion."*

In contrast to a closed-box system, a comprehensible system adds additional remarks to its output, sometimes, in the form of confidence scores (e.g., the C-scale measure described in Chapter 5) — which fosters trust — or by offering contrastive insights into the model's operational functions (Chari et al., 2020).

**Transparency**: a transparent model is self-explanatory; it possesses characteristics such as simulatability (i.e., the simplicity with which the system may be replicated), decomposability (i.e., chunking, and easy analysis of the functional components), and algorithmic transparency (Lipton, 2018). Transparency in a system may be fairly uneasy to achieve; nonetheless, what could be attainable is a post-hoc explanation that tends to justify the rationale for a system's decision rather than the system's real operational structure (Preece et al., 2018).

Having defined all of these core tenets, which are critical for the advancement of the viewpoints this chapter seeks to develop, few points are up for discussion, particularly regarding DFAI. As a result, subsequent discussions are contextualized within the paradigm of explainable/interpretable AI's application to DF.

To begin, explainable AI (or XAI thereof) widely appears to be a generalized notion of explanation or a constant effort to minimize (or eliminate entirely) the opacity of AI systems through deconstruction of complex variables, while maintaining a good balance between transparency, performance, and correctness. Meanwhile, as supposedly observed, the idea underlying all of these concepts appear to be intertwined, with all highlighting the importance of AI models being understandable, precise, and objective in their decision-making process. It

is easy to misinterpret the fundamental meanings of these concepts, and of course, they are used interchangeably in this chapter — albeit with consciousness of the exact thought they intend to convey. Most significantly, this chapter places considerable emphasis on two concepts: explainability and interpretability, and while the other concepts are mentioned accordingly, the goal is to identify which is more fundamental to DFAI. As illustrated in figure (6.1), it is apparent (from the literatures) that "interpretable AI" became more prevalent over time until 2018, when it appears that explainability began to get formalized. The sheer volume of articles devoted to it attests to the gradual tremendous shift. However, as research becomes more critical in that direction, and reasoning in the domain tends northward, there is considerable evidence that interpretability is gaining traction and receiving the attention it deserves. Similarly, as illustrated in Figure 6.2, which depicts the use of interpretable AI (IAI)/XAI across multiple disciplines of study, the consistent trend toward the adoption of IAI is readily apparent. Although XAI receives more research attention and mention, this is presumably because it is the primary generic phenomenon whose critical reasoning gave birth to the expansion of others, i.e., the ambiguities and arguments over how explainable systems should work gave rise to other forms of white-boxing concepts. However, it should be noted that larger body of works may have more references to explainability/interpretability than the title, abstract, or keywords that we considered from a single database. What is clear from the illustration is that, apart from the main fields from which the idea of AI originates (e.g., computer science, mathematics, engineering, and so on), AI research in fields such as medicine and decision sciences (to which, believably, DF belongs) has begun to focus on XAI and IAI. These domains have seen a substantial deployment of AI in numerous areas of its operations over the years, and most crucially, they deal with humans, on whom the impact of the machine decision may be severe.

## 6.2 AI, Law, and the Right to Explanation: A Brief

It is self-evident that courts do not produce evidence; they are not witnesses and are not bound by evidentiary rules. Likewise, Law and case law are not evidential. Nonetheless, the court exists to enforce rules and interpret evidence (Marcinowski, 2021). This means that while the prosecuting and defence parties (most of whom are attorneys) present evidence in a legal proceeding, the duty of finding this evidence falls on law enforcement agencies or forensic practitioners (in this case, digital forensics experts). The commissioner is therefore required to prove (with convincing explanation) the validity of the methods and approaches employed to

establish the facts presented as evidence. When these techniques entail implicitly complex application (e.g., closed-box system), both the prosecution and defence have one critical right: the *right to explanation* (Doshi-Velez, 2017).

In a practical legal context, "justice must not only be done, but also seen to be done;" hence, the necessary transparency that establishes the veracity of a case's outcome may be missing without explanation (Atkinson, Bench-Capon and Bollegala, 2020). It is fairly arguable that, before mathematical science fields (where modern AI algorithms originate) recognized the necessity to explain in AI systems, the Law discipline did; and it has been the driving factor in that direction in recent times. Miller (2019) outlined four crucial characteristics of explanations (in AI) in his fascinating review of AI from a social science perspective (though naturally relates to Law too), which he claimed the majority of AI researchers are unaware of. According to the author, explanations should be:

- *Contrastive*: often in the form of a counterfactual hypothesis; for example, if a predictive analysis classifies certain image as containing CSEM, a balanced explanation for this classification will explain what influences such inference (and why not something else). The HYPO (Rissland and Ashley, 1987; Ashley, 1991) is an excellent example of a contrastive, case-based system in law, as it examines whether hypothetical alterations on cases would affect their conclusion. This approach, however, is already ubiquitous in forensic science. Indeed, as discussed in Chapter 5, establishing facts in DF includes the definition of working hypothesis, one or more of which include alternative counter-hypothesis to test the outcome's veracity in an unlikely scenario.
- *Selective*: frequently impacted by cognitive biases — implying that a perfectly detailed explanation of the cause of an event is rarely offered logically. Rather, on the assumption of common background knowledge among stakeholders — which is sometimes not the case — a few (salient; supposedly only persuasive) causes are selected for explanation from an infinite number of causal events.
- *Rarely probabilistic*: while truth and likelihood (in ratio terms) are crucial in forensic science, employing "*most likely*" as a semantic explanation for a causal event may be unsatisfying. Thus, explanations based on probabilities or statistical relationships as a generalization of the rationale for the occurrence of an event are ineffective unless a causal explanation for why that generalization is typical is provided. This is compatible with the discussion on standardization of probabilistic outcomes in DF in Chapter 5.

Because, even when Bayesian reasoning (or a probabilistic model) is used in DF, it could be more effective to explain degree of certainty/uncertainty in terms of scenarios (Vlek et al., 2016) or arguments (Timmer et al., 2015).

- *Social*: entails the conveyance (or transfer) of knowledge through conversation or interaction. Thus, the explanation is offered in light of the explainer's assertions regarding the audience's beliefs.

Explanation as a right can be expressed in the form of examples (Atkinson, Bench-Capon and Bollegala, 2020), i.e., in order to persuade juries or judges, it is a common law tradition to present distinguishing precedent cases in a contrastive manner (with positive and negative examples), which may favour one side over the other. Additionally, presenting hypothetical features of a prior case that provide an argument that the outcome of a case would have changed had the features been different is a sort of explanation by example (Rissland and Ashley, 1987). Likewise, explanations can be expressed as rules; this is especially true in European Civil Law culture. The practice entails the extraction of conceptual features and definitions from statutes (as well as cases and commentaries), such as those governing immigration, copyright, labour, benefits, etc., and formalizing them as logical rules (Sherman, 1987; Johnson and Mead, 1991) that law practitioners can easily use to bolster their arguments in court. Alternatively, knowledge gathered from domain experts can be codified as rules (in an inference engine) for a rule-based or expert system (Schlobohm and Waterman, 1987). In the case of rule-based systems, explanations are provided through the expert system in the form of queries — as in *how*, *why*, and *what-if* (Atkinson, Bench-Capon and Bollegala, 2020).

The above discussion demonstrates that Law is a significant domain for studying explainable AI because explanation is an inevitable requirement for all fielded legal applications to which DF belongs. DFAI, in particular, must consequently learn to deal with the intricacies of the legal domain.

In what follows, the concerns with closed-box systems and why they pose additional challenges for DF are discussed.

## 6.3 Key Concerns with Closed-box Models and Explanations

To guide the scope of this chapter, the reference to "closed-box" system is viewed in the light of DL/DNN models (not necessarily in ML models) employed in DF. While the emphasis is on neural networks, other ML models with considerably complex algorithmic structures, such

as SVMs or Random Forests, are included in the closed-box category as well. Further on this, the reference on closed-box herein excludes (even though sometimes regarded as same) proprietary systems whose internal working structures and codes are safeguarded to protect trade secrets or illicit copyrighting. One reason for making the majority of proprietary systems closed-box is to prevent them from being gamed (or exploited) or reverse engineered (Rudin, 2019). Nonetheless, closed-box refers to a system (or algorithmic function) that is incomprehensible to humans. Apparently, we employ machines because they possess superhuman abilities to detect patterns, discriminate, and draw conclusions. Our understanding of these processes, however, is contingent upon the model's output, which we cannot follow (Yampolski, 2020). DL fall into this category due to their high recursive nature (Rudin, 2019) and deeply nested structures. A closed-box system does not necessarily imply inefficiency; it frequently works for the purpose for which it was designed. The concern is that if the system claims to possess significant reasoning powers comparable to those of humans and the ability to make decisions almost as accurate as humans in various situations, it should be able to offer explanations about how it arrived at a certain conclusion. To audiences in a high-stakes domain such as law, a low-fidelity explanation of a system's decision-making process undermines trust in both the system and the explanation. The crucial point here is that explanation is just as important as the model itself, and this is an area that requires immediate attention in DFAI. As Rudin correctly points out, if an explanation for a model is true 90% of the time, a tenth of the time, it is still incorrect, lowering the level of confidence in the explanation and model (Rudin, 2019). The challenge with closed-box systems' unexplainability stems from their architecture — they are modelled on the natural neurological impulses of humans, and as such, humans can also be thought of as "closed-boxes" (Yampolski, 2020). For example, a series of split brain experiments in (Gazzaniga, 2015) demonstrate that humans instinctively invent explanations (or justification) for already decided actions. That is, humans have an unconscious tendency to rationalize their choices or the process by which they made them after the fact (Shank, 2006). Can we then deduce that machines are only exhibiting human's cognitive rationalization? The problem that forensic practitioners must be aware of is that by introducing an additional layer of distrust through unconscious irrational explanations, the full adoption of AI in DF can be impeded.

Another worrisome trend that may be misleading in the explanation for closed-box systems is the provision of explanation mainly for correctly classified labels. An excellent use case is the description of the saliency map (Li, 2002; Underwood et al., 2006; Alqaraawi et al., 2020) in

a typical object detection/recognition task. In computer vision, a saliency map highlights the region of an image that attracts the most attention. It aims to convey the significance of a given pixel to the human visual system. In a saliency map, multiple edges are highlighted or segmented, and in most cases, explanations given for each class are identical; this also occur even if the classes are incorrect. The explanation on the saliency map reflects the "likely" imprecision of closed-box predictions; as the reason behind them may be unknown (Rudin, 2019). Additionally, a recent study on medical imaging (Arun et al., 2021; Saporta et al., 2021) discovered that using saliency to interpret DNNs failed to meet several critical utility and robustness criteria. These works make apparent the critical issues that DF analysts should be aware of while providing explanations for evidence or testifying as an expert. It may be difficult to justify a decision made by saliency maps, and a lack of explanation will make troubleshooting the closed-box even more difficult (Rudin, 2019).

Despite their expressiveness, research has demonstrated that DNN models can, nevertheless, learn counter-intuitive solutions (Szegedy et al., 2013). Specifically, by introducing a small but deliberate undetectable perturbation to examples of a deep learning-based classifier showed erroneous predictions with "high confidence" when a minor but deliberate undetectable perturbation is introduced to the examples (Goodfellow, 2014a). The authors demonstrate that given a particular example, a correctly classified example with a confidence level of 57% can be disrupted with adversarial examples (such as noise), resulting in the model making a false prediction with a confidence level of 99%. Similarly, a slight change in the stop signal of an autonomous vehicle's computer vision system, invisible to human eye, caused vehicles to interpret it as a 45mph signal (Eykholt et al., 2017). Neural networks operate by multiplying and summing sample features with weight coefficients in a recursive fashion. Then, based on whether or not the weighted-sum exceeds a predefined threshold, a prediction is made (Atkinson, Bench-Capon and Bollegala, 2020). It is possible to discover considerably different adversarial examples in the latent space that have identical weighted sums, making discrimination challenging for the model. This could (and most likely will) have a seemingly daunting implication on legal decisions. Consider a counterfactual claim (such as the impact of adversarial examples) made by an opposing party showing that a forensic conclusion may be inaccurate, and that the decisions deduced using the same technique are unreliable. A reasonably informed audience about AI, much alone the less informed, can readily be convinced by such an example. Although this adversarial discovery has resulted in the development of more robust deep generative models such as the Generative Adversarial

Network (GAN) (Goodfellow et al., 2014b; Goodfellow et al., 2020) and VAE. A generative model is trained by introducing small amounts of noise to the input vector, which is subsequently represented as latent variables (embeddings). These latent representations can be manipulated and their contributions on the output examined. It is possible to derive insights and detect specific patterns about the predicted class from the output. The generative models have also continued to present unique challenges, one of which being GAN's game-theoretic foundation.

There is a growing concern that machines may augment their operating parameters in ways that result in analytical inaccuracies (Roth, 2017). This may occur when training sets contain fewer samples, are less reflective of current real-world use cases, or are insufficient to make inferences on future observations. Accounting for an excessive number of variables may potentially cause the machine to learn illogical representations. For example, in reinforcement learning, an AI model is trained to react to its environment in order to solve complex problems that are intractable using conventional ML techniques. While the technique requires large data for robustness, it may suffer from learning state overload, diminishing its results overtime. Consider, for example, a predictive crime-detecting algorithm[128] deployed in surveillance cameras that tracks criminal movements and alerts officers before or right as crime is being committed. The technology was developed by simulating specific patterns associated with crime. According to reports, the algorithm learned to distinguish three handshakes in succession as possible narcotic transactions by examining crime-related sample. While this decision could be reasonable based on training samples, it may overlook future real-world drug-related cases if the pattern does not occur (Roth, 2017). Exemplifying this as a justification for drawing inference in a court case will only serve to increase public distrust of machine-generated evidence.

Consider an NLP and ML-based (SVM) classifier developed to predict the outcome of EU Court of Human Rights cases, as illustrated in (Aletras et al., 2016). The predictions are classified according to the most likely topics related to the article. Specifically , the terms that represent topic 23 (on the most predictive topics) for Article 6 violations are:

[128] See SmartSensory - https://www.govtech.com/public-safety/smart-cameras-aim-to-stop-crimes-before-they-occur.html

*"court, applicant, article, judgement, case, law, proceeding, application, government, convention, time, article convention, January, human, lodged, domestic, February, September, relevant, represented"*

Apparently, looking at these terms, one would notice that they do not provide any explanation. Perhaps some phrases, such as court, law, judgment, and proceedings, show a common phrase associated with decisions or violations, but others, such as month names, appear to have been randomly picked from the dataset based on frequent occurrence. As previously mentioned, explanation is relative; while an expert in topic modelling may easily comprehend and deduce the phenomenon to which these phrases refer (perhaps, with low confidence), layperson(s) would undoubtedly fail to comprehend this. Topic modelling (covered in Chapter 4) is one of the most intuitive components of ML and NLP for identifying abstract topics within a collection of documents. It can model topics with great accuracy if a large amount of data is available. However, factfinders or forensic experts may be required to demonstrate how it operates when used to infer evidence.

## 6.4 Explainable DFAI: the Goal

The resulting value of a digital forensics investigation is the evidence; mined (extracted, uncovered) by a forensic expert and communicated to fact finders (e.g., legal practitioners, law enforcements, organizations, etc.). Evidence is mostly presented as facts, inferred from a series of correlations of causal relationships; which involves decoupling intricate interrelationships between multiple heterogeneous artifacts. The court or commissioning agency determines the weight, relevance, and substantiveness of the evidence. However, it is the forensic expert's responsibility to present an intelligible explanation of the methodologies and hypothetical approaches used in reaching the conclusion.

Explaining an AI-based DF analysis may entail balancing, comparing, or persuading the audience via logic-based formalization of (counter) arguments (Besnard and Hunter, 2008), or reducing the complexities to simplify the output. Whichever way, given the high-stakes audiences in an evidence-oriented context for whom presentation is crucial, an explainable DFAI *"is an AI-based digital forensics method(s) that provides explicit and intelligible (as well as assessable) rationale for its functions and the specifics of its inferential reasoning.*"

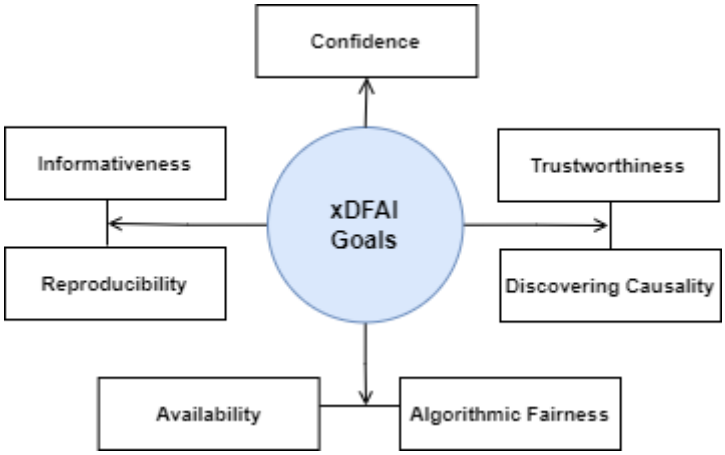


Figure 6.3 xDFAI Goals

This description can serve as a preliminary (tentative) conceptualization of explainable DFAI (xDFAI) idea, with a more refined and generic formalization envisaged as research in the domain advances. According to Clancey (1983) notion of explanation, which is adaptable to DFAI, xDFAI should aim to provide explanations for the following : *why was a specific fact used? Why was a certain fact ignored? How was a particular conclusion arrived at? How was a different conclusion not reached?* Notably, research outputs on the application of AI in DF have placed much premium on the performance and accuracy of the methodologies employed, with little concern for the interpretability of the process. This seem an unhealthy practice, especially that the process' outcome is the interest of law and society — the domains that are primarily driving AI system's transparency. The pursuit of an explicable DFAI can aid in further improving the practicality of the techniques. Bearing the foregoing in mind, we can elaborate a bit more on the goals of xDFAI by relating it to concepts that have been frequently associated with XAI in research. Specifically, these goals are adapted from the work of Arrieta et al. (2020), and while the authors contextualized them to fit a different narrative, they are expressed here in alignment with the requirements of DFAI. The following are the goals that an xDFAI should pursue during the examination process and while reporting/presenting the derived results:

- **Trustworthiness –** While this is not a guarantee that a model is explainable, it is an assessment of a model's ability to behave (at all times) as expected (or defined) in a certain context. Trust in a model grows over time as long as it continues to behave consistently in accordance with the stakeholder's mental model and provide consistent accurate and verifiable predictions (Bhatt et al., 2020). If an unanticipated failure occurs in a trusted system, stakeholders may overlook it, as it will not significantly erode their confidence.

However, in the case of DFAI, where a system is expected to work optimally at all times due to the grave consequences of its failure, then the popular Russian adage "*trust but verify*" may apply: i.e., even if the model behaves appropriately, it still requires a human gatekeeper (Desai and Kroll, 2017). As a result, if an unexpected error occurs, it should be adequately reflected in the output and sufficiently reported. The stakeholders can then assess the extent to which the failure is acceptable given the circumstances.

- **Discovering Causality –** This involves a significant amount of prior knowledge and is a key quality expected from an experienced investigator. Causality is the process of establishing (or inferring) causal relationships between observed data (Pearl, 2009). Explainable models may aid in inferring causal relationships between variables (Rani et al., 2006). While a ML model can be instrumental in identifying correlations between learned data, such correlations does not imply causation. Thus, a robust xDFAI should provide intuitive evidence of causal relationships within observed artifacts or aid in validating the output of a causality inference technique. Nevertheless, a human validator should not be dismissed.
- **Reproducibility –** The training and testing (as well as validation) phases in a learning process are used to confirm the model's applicability and its parameter reusability in various circumstances. Thus, explainability in this case entails elucidating the model's operational functionality in order to facilitate the understanding of its constraints (or boundaries) and the seamless transfer of knowledge for reproduction in another system (Arrieta et al., 2020). The absence of explanation, on the other hand, may influence incorrect assumptions about the model (Kim et al., 2017). Transferability also drive improvement on the performance of a system. This is particularly evident in the ML research domain, where the explanations provided in literature have inspired improvements of the state-of-the-arts. Consequently, confidence in DFAI models is likely to increase when the functional parameters are explicitly elucidated, and its method extensively reproduced.
- **Informativeness –** The output of a DFAI model is almost always numerical (probabilistic of some sort). It will take much work to connect these values to the real-world problem for which a solution is sought. To avoid misinterpretation, xDFAI should provide thorough details of how these values are represented and how they aid in inferring facts through investigative analysis. This is exemplified in Table (5.1), where we proposed a verbal expression for the evaluation of strength of probabilistic evidence. Explanation and information are complementary; neither can exist without the other. Once the model's

capacity to predict reliably in multiple scenarios is established, the extent to which it is credible will slightly depend on the quantity of information provided regarding its inferential processes and the accuracy of its output.

- **Confidence –** This is nearly synonymous with trustworthiness; it is a characteristic of a stable system. Confidence is relative; it is tangible in cases where reliability is demanded (Arrieta et al., 2020). It might be expressed by the one who presents the facts or by the person to whom the facts are presented. As with trustworthiness, confidence in DFAI might not easily lend itself to the notion of explainability because it is earned via operational and result consistency — not necessarily by explanation of its operational parameters. Nonetheless, an xDFAI can be critical in providing information on the confidence level of each modular component of the system. This way, each component of the decision-making process can be evaluated, and appropriate confidence scales assigned.
- **Algorithmic Fairness –** One of the aims of explainability in AI might be seen to be ensuring fairness in relation to the system's specified objectives. Fairness is considered in the legal domain in terms of adherence to ethical principles, the right to be informed, and the right to contest decisions (Goodman and Flaxman, 2018; Wachter, Mittelstadt and Floridi, 2017). This may be accomplished by clearly visualizing the relationship between hypothetical components affecting the decision. Algorithmic fairness (or unfairness) is largely connected to decision-making biases, and it is widely believed that enforcing XAI will help mitigate this. While biases are spontaneously learned from data, it may be vital to preserve the alleged biased features in order to maintain the quality of the original data. Several elements that may contribute to algorithmic decision-making being unfair or bias include skewed data, limited features in the data, disparities in sample sizes, an erroneous problem definition, and the presence of correlated variables that generate bias even when sensitive features are eliminated (Barocas and Selbst, 2016). One critical part of DF fairness that should not be ignored is the presentation of the results of AI-based analysis in the most justifiable manner possible. Occasionally, an investigator's subjective inclination toward a particular outcome may cause her to disregard evidence to the contrary. As a result, unjustified conclusions are reached. If this conclusion is obtained algorithmically, it has the potential to further undermine trust in machine-generated results; this should be avoided. Thus, xDFAI should be considered as a mechanism for avoiding unethical or unfair algorithmic conclusions (Arrieta, 2020).
- **Availability –** This is connected to transferability, and it entails considering explainability as a way of involving end users in the process of improving certain AI models (Miller,

Howe and Sonenberg, 2017). While open-sourcing the algorithm and publishing it with peer review will ideally help technical users better grasp the technique, xDFAI will almost likely lessen the difficulty that non-technical users will face when interacting with the algorithm. Thus, if a forensic expert is required to report (or testify) on an algorithm's decision in a legal proceeding, an already available open-sourced and/or peer-reviewed procedure is likely to be understood and accepted.

## 6.5 Explainable DFAI: the Methods

This section discusses numerous explainability techniques for AI models. The goal is to elaborate on XAI and to establish meaningful connections with xDFAI when necessary.

There has been debate over whether it is appropriate to oversimplify AI models in order to make them more interpretable at the expense of performance and accuracy. XAI approaches aim to balance interpretability and model performance. As a result, post-hoc explanation has grown in popularity. Conversely, the intrinsic approaches (not discussed in detail) that are based on simpler, self-explicit models (e.g., rule-based, linear models, Decision Trees, etc.) are possible. Figure 6.4 is an illustration of the xDFAI structural model.

### 6.5.1 Post-hoc Explanation Approaches

A post-hoc explanation sheds light on a model by elucidating its salient features (Ribeiro, Sigh and Guestrin, 2016; Lundberg and Lee, 2017; Davis et al., 2020), training points (Koh and Liang, 2017; Yeh et al., 2018), counterfactual reasoning (Wachter, Mittelstadt and Rusesell, 2018), or decision boundaries (Bhatt et al., 2020). Some post-hoc explanations exist to convey information about the model to stakeholders; however, few solutions allow for the model to be toggled and updated in response to stakeholders' perceptions (Bansal et al, 2019; Lee et al., 2020). To increase the practicality of AI models and to foster greater trust in their decisions, a variety of post-hoc explanation approaches have been proposed for deep learners, including explanation by: *model simplification, visualization, feature importance estimation, localization, text, and example*. Post-hoc explanation can be examined in two unique context: model-agnostic and model-specific.

The model-agnostic approach, one the one hand, incorporates interpretability within its internal mechanism, is independent of the model's internal structure and is applied after the model's training (Molnar, 2019). This generic technique is intended to extract information about a

model's prediction procedure (Arrieta et al., 2020). On the other hand, model-specific approach are restricted, and only applicable to specific algorithm types. In fact, most intrinsic explainability method are model-specific. This chapter discusses model-specific approaches from the perspective of their application to DNNs — the emphasis is mostly on methods applicable to deep-layered neural networks, but shallow models (e.g., SVM, RF, etc.) are mentioned in a few instances. It is worth emphasizing that the explainable models covered in this section are by no means exhaustive; they represent only a subset, and their selection is motivated by their possible applicability for DFAI. Tables 6.1 and 6.2 summarise model-agnostic and model-specific post-hoc approaches and their probable applicability for DFAI tasks. The next sections describe post-hoc explanation methods for improving the interpretability of closed-box models.

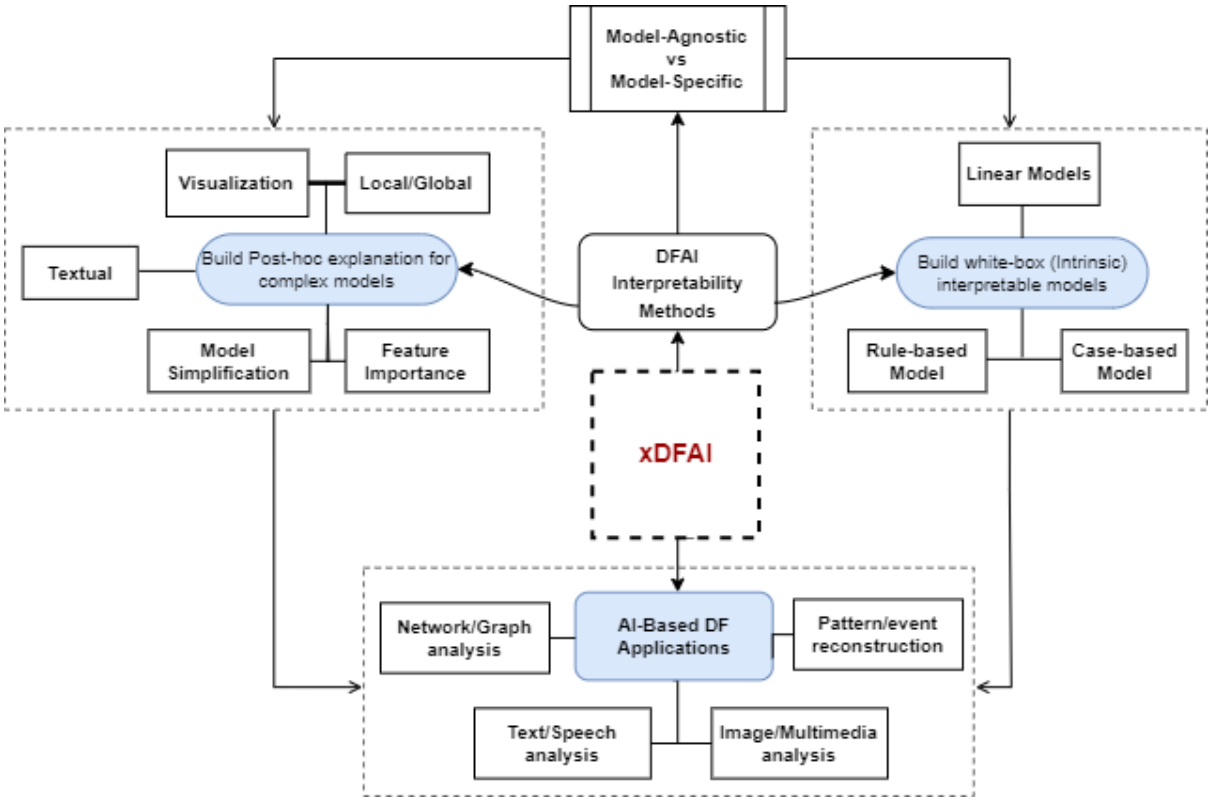


Figure 6.4: Mind map representing an illustration of the explainable digital forensics AI (xDFAI) Model

### *A. Explanation by model simplification*

Model simplification appears to be the broadest of the model-agnostic post-hoc explanations. They are largely based on rule extraction techniques, however, (Bastani, Kim and Bastani, 2018) proposed a model extraction process based on approximating a transparent model to a complex one. Popular techniques for extracting information in the form of rules to improve

interpretability includes the Genetic Rule Extraction (G-REX) (Johansson, Kong and Niklasson, 2004a; Johansson, Niklasson and Konig, 2004b; Konig, Johansson and Niklasson, 2008), which is based on genetic algorithms, and CNF (Conjunctive Normal Form) or DNF (Disjunctive Normal Form) (Su et al., 2016).

### *B. Explanation by feature importance*

This approach elucidates a closed-box model's operationality by quantifying and evaluating the influence, relevance, and significance of each training variable on the model's prediction. The SHAP (SHapley Additive exPlanation) SHAP (Lundberg and Lee, 2017) framework, and approach for explainable image analysis based on saliency detection method proposed in (Dabowski and Gal, 2017), offers a significant contribution to feature importance. Additionally, the Automatic STRuctured IDentification (ASTRID) (Henelius and Ukkonen, 2017; Henelius, Puolamaki and Ukkonen, 2014) is a useful tool for determining feature importance in a predictive model. However, several alternative approaches have been proposed that go beyond the relevance measure, e.g., in (Koh and Liang, 2017), an influence function is used to trace (back to training data) a model's prediction through its learning algorithm, therefore identifying the feature points most responsible for a given prediction. Basically, the approaches mentioned here provides highly valuable techniques for xDFAI, which can be explored further in future research.

### *C. Explanation by visualization*

Visual explanation is likewise a technique for achieving model-agnostic explanations, but it is highly effective, and most common in model-specific approaches; particularly with DNNs. In a typical model-agnostic settings, developing visualizations based just on the inputs and outputs of an opaque model may be a difficult task (Arrieta et al., 2020). A frequently used technique in this approach is to use feature importance techniques to provide explanations. Notable methods for visualization of shallow ML models (e.g., SVM, RF, etc.) are proposed in (Cortez and Embrechts, 2011; 2013), based on Sensitive Analysis (SA), and Individual Conditional Expectation (ICE) (Goldstein et al., 2013) for estimating any supervised learning techniques. While feature relevance is advantageous for xDFAI, visualization techniques offer an intuitive way to visualize the interaction of influential variables during training. Although the approach is relatively complicated, it demonstrates a promising research direction from which xDFAI can benefit.

*D. Local explanation*

Considering that DL models have a high degree of dimensionality and curvature, the concept of local explanation stems from the fact that insight-generating interpretable methods can be applied to a tiny region with detectable changes in individual or grouped features. Using the network's feature space to represent each case (data point) or its neighbours, local explanation provides a semantic explanation for specific cases (Leslie, 2019). However, a *global explanation* entails capturing the internal logic and function of each prediction or classification made by an opaque model as a whole (rather than a tiny region) (Leslie, 2019). The technique, known as LIME (Local Interpretable Model-Agnostic Explanations) (Ribeiro, Sigh and Guestrin, 2016) is an example of a model-agnostic approach designed to simplify explanations, which explains model predictions by learning interpretable models locally and modeling them as a sub-modular optimization problem. In the local context, an explanation for a single prediction is provided, which could enhance user's confidence in the result (Kelly et al., 2020).

*E. Text explanation*

Although it is not often discussed in the literature, this method entails enhancing closed-box models to provide explanations, most likely in natural language. With this approach, naive methods may include associating text with each model's decisional components. In some cases, text explanations are incorporated in a rule-based (or if...then) style, in which all decision-making components are semantically explained. This approach, when combined with other approaches (e.g., feature importance and visualization), can be quite beneficial for xDFAI.

### 6.5.2 Methods for Explaining Deep Learning Models

This section discusses the explainability of DNNs briefly. Three distinct neural network architectures are considered: multi-layered networks (MLNNs), CNNs, RNNs. The criteria for selection are their utility/applicability to DFAI. However, because the descriptions provided here are mostly limited in scope and depth, detailed reviews are available in Linardatos, Papastefanopoulos and Kotsiantis (2021) and Arrieta et al. (2020) for a comprehensive survey of explainable approaches.

MLNNs are a type of closed-box, yet adaptable AI model that excels at inferring intricate relationships between data variables but is frequently unable to justify their underlying assumptions. Three fundamental explainable methodologies are utilized to explain multi-layer

neural networks: model simplification through rule extraction from hidden layer of a neural network (DeepRED) (Zilke, Mencía and Janssen, 2016; Sato and Tsukimoto, 2001) feature importance of contributing elements with models such as Deep Taylor (Montavon et al., 2017) and DeepLift (Shrikumar, Greenside and Kundaje, 2017) and visualization for which TreeView (Thiagarajan et al., 2016) was proposed. DeepLift uses backpropagation to evaluate the contributions of input components similarly to deep Taylor, but compares each neuron's activation to its reference activation and assigns scores to contributing elements depending on the difference. Due to the fact that DeepLift and deep Taylor are exemplified with image classification, they may be excellent xDFAI options for forensic image analysis as well as pattern recognition-based investigations.

CNNs structure reflects DNN's extremely complex internal cores. They lay the groundwork for computer vision's unique underpinnings — from object identification and image classification to instance segmentation (Arrieta et al., 2020). Due to the visual nature of CNN's representations, they connect well with human reasoning, making them somewhat explicable. To explain CNN functionality, one can either map the output back to the input to determine which input data were discriminative of the output, or create interpretations depending on how the layers see the external world. A common feature importance and local explanation method

| **Explainability Technique** | **Post-hoc Explanation** | **Tools** | **Potential Applicability t DF** |
|---|---|---|---|
| **Model-Agnostic** | Model Simplification | G-REX, CNF or DNF | Pattern recognition, digital file forensic analysis, text analysis etc. |
| | Feature importance | SHAP, ASTRID, Influence function, Saliency detection (Koh and Liang, 2017; Dabowski and Gal, 2017) | Image forensics, object classification, predictive analysis, etc. |
| | Visualization | SA and Global SA, ICE | Pattern recognition, object identification/classification, document classification, etc. |
| | Local | LIME, Fairness (Dwork et al., 2012), L2X (Chen et al, 2018), AIX360 (Dhurandhar et al., 2018) | Object classification, predictive analysis, multimedia forensics, etc. |
| | Text | TextAttack (Gao et al, 2018), HotFlip (Ebrahimi et al., 2018) | Spam message detection, e-mail forensics, attribution, malware detection, etc. |

Table 6.1: An overview of some model-agnostic explainability methods, proposed tools, and their potential applications to digital forensics

| Explainability Technique | Post-hoc explanation | Tools | | Potential Applicability in DF |
|---|---|---|---|---|
| Model-specific | MLNN | Model simplification | DeepRED | Forensic image classification, object identification/detection, pattern recognition, CSEM analysis, etc. |
| | | Feature importance | Deep Taylor, DeepLift, Deconvnet | |
| | | Visualization | TreeView | |
| | CNN | Visualization | LRP, DGN, Grad-CAM, CNN+CRF_bi-LSTM (Ma and Hovy, 2016) | Forensic image/video reconstruction, forensic data visualization, object identification, source identification, deep fakes analysis, image recognition, etc. |
| | | Text | CNN+RNN (Xu et al., 2015) | |
| | RNN | Feature importance | RETAIN | Speech recognition, authorship attribution, determination of intent, forensic linguistics, timeline/event reconstruction, malware detection, email forensics, e-Discovery, IoT Forensics, Network intrusion detection, etc. |
| | | Visualization | Finite n-gram horizon+RNN | |
| | | Local | RNN+Hidden Markov Model (HMM) | |

Table 6.2: An overview of some model-specific explainability techniques based on DNNs, proposed/developed tools, and their potential application to digital forensics

is Deconvnet (Zeiler et al., 2010; 2011; Zeiler and Fergus, 2013) that repeatedly occludes sensitive region of an image during training to determine which portion produces desired impact. Another approach based on feature importance and localization is the Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017). Layer-wise Relevance Propagation (LRP) (Bach et al., 2015) proposes a method that visualizes relevant elements that contributes to prediction. Other methods (Dong et al., 2017; Xu et al., 2015) combines CNN models and RNN such as bi-directional LSTM encoder (Ma and Hovy, 2016) for the purpose of describing visual material via textual explanations. As presented in (Zhou et al., 2015), a simple and intuitive method identifies image regions that are related to a particular object class by interposing a global average pooling layer between the final convolution and the fully-connected layer that predicts the object class. Perhaps an excellent and easily interpretable approach is the deep generator network (DGN) (Nguyen et al., 2016), which not

only generates an incredibly realistic synthetic image, but also reveals the features learned by each neuron. Given that certain DF analysis will require object identification, the DGN approach appears to possess both quality and suitable characteristics for the development of xDFAI.

RNNs are one of the most important techniques for DFAI because they are capable of solving prediction problems using sequential data — which is critical for forensic event reconstruction (Solanke et al., 2021). RNNs take pride in their capacity to retain information about data's time-dependent relationships. There have been two approaches to explaining RNN models: 1) through feature importance techniques that seek to understand what the model has learned over time; and 2) by providing insights into (or explanations of) the model's decision-making process through modification of its architecture (local explanations) (Arrieta et al., 2020). Numerous proposals are offered in this respect, which may spark the interest of DFAI professionals. With RNN, some explanation approaches (Donadello, Serafini and Garcez, 2017; Donadello, 2018; Garcez et al., 2019) have demonstrated the possibility of merging probabilistic and logical reasoning (Manhaeve et al., 2021) (based on background knowledge) in a symbolic/sub-symbolic (Haugeland, 1989; Ilkou and Koutraki, 2020) fashion. Some other approaches include visualization approach based on finite horizon n-gram models (Karpathy, Johnson and Fei-Fei, 2016) to study predictions, combination of RNN with a simple and transparent hidden Markov Model (HMM) (Krakovna and Doshi-Velez, 2016) to interpret speech recognition representations, and the RETAIN (Reverse Time Attention) model introduced in (Choi et al., 2016) for detecting influential past visit patterns and significant variables within the patterns. This technique could be useful, for example, in performing forensic analysis on users' log history (e.g., internet browsing history) during a CSEM investigation.

In contrast to the preceding methods, which are either model-agnostic or model-specific, a novel technique dubbed Contextual Importance and Utility (CIU) is proposed (Framling, 2020; Anjomshoae, Framling and Amro, 2019; Framling, 2022). It is based on Contextual Importance/Influence (CI) and Contextual Utility (CU) theory. CIU appears promising as it is applicable to both linear and non-linear models and may be represented visually or in natural language. Additionally, feature representations can be read and validated directly from input-output graphs. Although the CIU approach is just developing, its features indicate that it has the potential to considerably aid in xDFAI.

## 6.6 Interpretability in DFAI: the Case

To be trustworthy, a system must go beyond accuracy evaluation — which, in many cases, does not correctly reflect the real-world use case. Trust is determined in part by how users perceive the system's decisions. On the other hand, users' perceptions of a system are contingent upon how interpretable, or easily comprehensible its features are. Incorporating interpretable components into a closed-box model might be challenging due to the model's domain-specific constraints. In general, constrained problems are more difficult to solve than unconstrained problems. Thus, when the complexities of DF investigations are considered, particularly when AI models are used, interpretability practically translates to a set of application-specific constraints. As such, domain expertise will be required for the model to incorporate interpretable features. Interpretability not only provides an answer to the question of what was predicted (which is only a partial solution to the problem), but also to the question of why such predictions were made (or what caused them). By incorporating interpretable features into DFAI, it is possible to harmonize and update gaps in domain knowledge, as by attempting to answer why a particular decision was made, new dimensions to the problem or solution can be uncovered, and methods for debugging or auditing can be established. Additionally, an interpretable model aids in determining the root cause of an error and may also suggest strategies to resolve it. For example, during an investigation of child sexual exploitation material, a classifier may incorrectly classify a person as an adult rather than an underage. We may discover through interpretable models that the misclassification was caused by an underage person wearing adult facial makeup. In an inquisitorial tradition, opposing parties may request access to the tool used to infer facts; in this situation, interpretable models will ensure simulatability (of the model's reasoning being provable and reproducible), decomposability (sub-component interpretability), and algorithmic transparency. It should be highlighted that constructing interpretable models can be time-and-resource-intensive; yet, for high-stakes decisions such as those involving digital evidence, it is less expensive than the expense of creating a flawed model (Rudin, 2019) that could result in eventual exculpation or inculpation of the wrong entity. Which indicates that, even as timeliness remains an issue in DF, which is one of the reasons automation systems are necessary in the first place, dedicating additional effort and cost to building a high-quality interpretable model would be worthwhile.

## 6.7 Interpretable DFAI Model: Recommendations for Mitigating Distrust

As traditional (non-AI) forensic investigation requires clarity, conciseness, and understandability of the techniques used to arrive at a conclusion, interpretability in DFAI approach is critical. However, because interpretations vary across disciplines, it is important to consider the interests, demands, and expectations of the stakeholders whose lives are impacted by the design, consumption, and subsequent consequences of the decision. AI literally refers to making computer do things that require intelligence when performed by a human. Keeping this in mind, explaining algorithmic decisions will involve a cognitive reasoning process. This explanation should be explicit about the factors that influenced the outcome and their contributions to the conclusion reached. The following paragraphs contains a set of recommendations that may be critical for achieving robust interpretability in DFAI. They are partly adapted from the guidelines offered in (Leslie, 2019).

First, prior to implementing AI models in DF, it is essential to contextualize the scenario, potential impact, and available AI tools for analysis while assessing the investigation's interpretability requirements. This implies that deployment should be preceded by an examination of the context in which the application will be used — for example, a civil, or criminal case. Apparently, there is a significant difference (in terms of techniques and interpretation requirements) between analyzing e-mails for suspicious deletions intended to conceal incriminating activities, and determining responsibility in an e-contract agreements between two or more parties, concluded via e-mails. This contextual understanding provides a clearer picture of the stakes involved and the scope of interpretability requirements. Furthermore, apart from having the technical capability to analyse artifacts using the appropriate tools, acquiring domain knowledge to gain insight into domain-specific explanation standards would be vital to the interpretability approach. Seeking domain knowledge might also provide pertinent insight into previous use cases. Another factor to consider before deployment is whether to use pre-existing AI algorithms or to develop a new one. In any case, employing existing algorithms may necessitate a thorough study or assessment of their functionality, expressiveness, complexity, performance, and interpretability. Alternatively, a custom algorithm addressing the aforementioned components as well as the investigative task could be considered.

Obviously, the DF domain and its constituents are sensitive — they are task-critical and require transparency and accountability. Thus, when DFAI is required, less-sophisticated, non-opaque,

interpretable evidence mining techniques (such as decision tree, linear/logistic regression, case-based reasoning, rule-based list, etc.) should be considered. Simple interpretable models are usually preferred when forensic data is well-structured, sufficient domain knowledge with meaningful representations is present, or if computational resources are constrained. This is also highlighted in (Rudin, 2019). The situation in which "when there is a hammer, everything else becomes a nail" should be avoided. Closed-box models should be a matter of choice influenced by the nature of task, not a necessity. Which implies that, unless inefficiencies with native ML models are observed, relying on closed-box models (such as deep learners) to improve performance and accuracy may not be appropriate.

Majority of digital forensic investigations may require the use of complex, opaque systems that typical linear models are incapable of handling. Cases involving image classification, speech recognition/audio analysis, or object identification in video footage, as well as anomaly detection in unstructured data, typifies the tasks in a DF investigation. Given that only non-linear DL models are viable for these purposes, investigators are urged to consider available options for interpretability (some of which are mentioned in Section 6.5), or to incorporate features into a custom-built model that: fits the specifics of the case; assesses the impact of the decision; and addresses the need of the audience, prior to deployment. To break this down, the foregoing suggests that:

- the potential impact of the decision and the risks associated with incorrect interpretations should be thoroughly considered in advance, ensuring that the design promotes fairness and accountability;
- the use of supplemental interpretability tools should ensure that: the semantic explanation it provides meet the needs of the stakeholders; the technical method of explanation-support satisfies both interpretability need and appropriate for the algorithmic approach of the use case. The tool should also provide reasonably sufficient level of mitigation against unethical/unfair outcome or interpretations.

Considering that interpretable methods will be assessed based on their ability to articulate the logical rationale behind their decisions and behaviours in a given scenario, as well as on their users' ability to give account of the generated output in a decent, coherent, and reasonable manner, a few critical questions should be asked prior to selecting a method. They are: 1) what is the affected audience's mental capacity for comprehending the outcome? 2) will the method assist decision-makers (e.g., judges, organizations, etc.) in making informed/justifiable

evidence-based judgments? 3) will the method generate counterfactual, misleading, or confusing explanation?

A critical point to emphasis is the modularization of design. Digital investigation, without a doubt, entails the study of digital data artifacts that may be heterogeneous and unstructured. Prior to imbuing data into a DL model, it must be pre-processed. Ordinarily, the pre-processing stage does not involve AI techniques, and even when it does, as in the case of NLP or probabilistic language models, the techniques are fairly interpretable. Additionally, in a communication-related investigation, it may be essential to generate a graph of subjects' relationships; this is not AI, and the construction can be simply understood. In case the entire process involves AI, submodules with independently interpretable methods can assist in rapidly identifying deviations. This implies that modularization enables the building of structured applications where AI application is responsible for a certain component of the investigative tasks rather than for the entire process (Asatiani et al., 2020). This can ensure proper control over the functionalities, reduce the investigator's explainability burden, and enhance the understanding and confidence of the audience.

To leverage on the benefits of cloud computing, Digital Forensics as a Service (DFaaS) is projected to impact the future of forensics (Van Baar, van Beek and Van Eijk, 2014; van Beek et al., 2015; Du, Le-Khac and Scanlon, 2017; van Beek et al., 2020). In such situation, DFAI as a service may also include online learning, in which a model learns to adapt to environmental changes and continuously updates its best predictor. While online learning can be advantageous for reconstructing events — especially with data that is generated as a function of time — it becomes more difficult to monitor and explain variable interactions in the feature space over time. Online learning issues may involve the inability to control the working parameters of the model, which may be problematic in high-stakes domains (Asatiani et al., 2020). The same might be said for transfer learning (TL) (especially when offered as a service), which entails applying previously learned knowledge to a different but related problem. They may benefit DF in terms of sample efficiency (Karimpanal and Bouffanais, 2018), investigation time reduction, and decreased false positives and negatives. However, they provide little information about how the models were trained or implemented, or how trustworthy are the platforms that host them (Aditya, Grzonkowski and Le-khac, 2018). Explainable TL methods are still limited, and their selection for DFAI should be done cautiously.

Legal practitioners are generally conversant with symbolic algorithms (e.g., expert systems, case-based reasoning, etc.) because they are used in legal rule mining and in the modelling of philosophical norms. It may not be difficult for laypeople to comprehend the logical foundation upon which they are built. As such, DFAI techniques that make use of symbolic algorithms should find it relatively straightforward to explain its findings. However, symbolic algorithms have a number of limitations, rendering them insufficient for most forensic investigations. Researchers have proposed a hybridization (Zeleznikow and Stranieri, 1995; Mao et al., 2019) of non-symbolic (such as NN models) and symbolic approaches that takes advantage of the former's robust unsupervised capacity to learn from complex data and the latter's ease of explanation to produce an explainable model. Neurosymbolic AI (Garcez and Lamb, 2020) is one of such methods. Although pioneers in DL, such as Yoshua Bengio, have argued strenuously against this method[129], stating that future DLs will be able to perform inferential reasoning in the same way that symbolic models can (Atkinson, Bench-Capon and Bollegala, 2020). While such systems are still in their infancy, hybrid techniques are likely to give the necessary level of interpretation for predictive decisions. Furthermore, an equally helpful method is to incorporate a "human-in-the-loop" or "man-machine" approach (Nguyen and Choo, 2021) with the hybrid technique. That way, automated decisions can be verified by the gatekeeper (Desai and Kroll, 2017) at different levels and appropriate validations performed prior to reaching a final conclusion.

Finally, the description of generative models was presented earlier in this chapter, and they do offer a potentially beneficial solution to interpretability problems. Generative models can be extremely advantageous for DFAI when it comes to solving specific tasks, given their robustness in terms of performance and accuracy. With an appropriate visualization mechanism, the latent features, which are the direct random low-dimensional representations of the input data, can be examined and tracked during training to determine which features contribute to a particular prediction. In this case, providing interpretations for such glass-box operations should be straightforward. Therefore, the use of generative models for complex DF analysis (such as pattern/speech recognition, object classification, event reconstruction, etc.) is highly recommended.

[129] See https://bdtechtalks.com/2019/12/23/yoshua-bengio-neurips-2019-deep-learning/

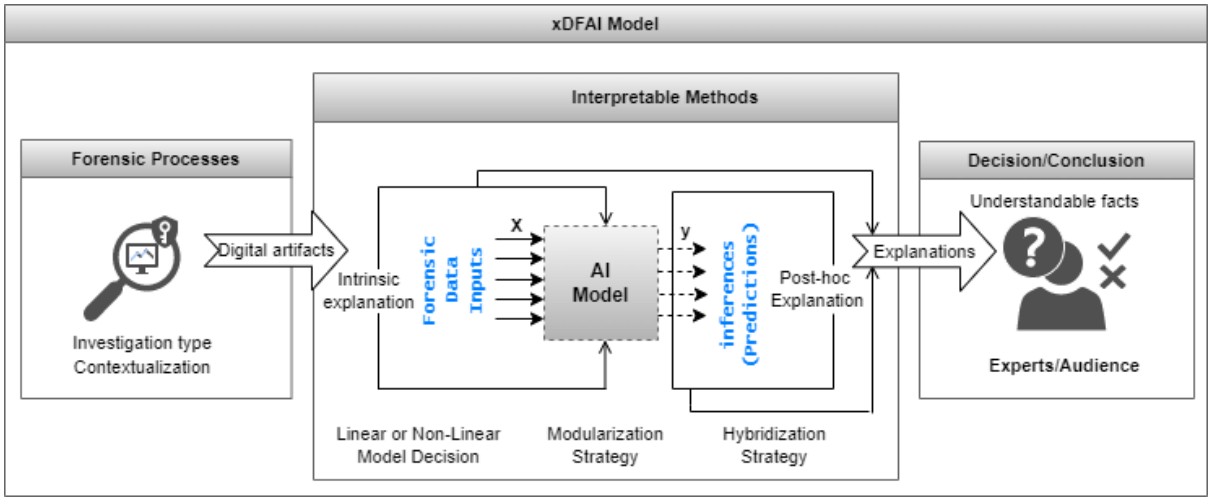


Figure 6.5: A typical structure of an interpretable DFAI model

## 6.8 Discussion

According to the surveyed literatures on XAI and interpretable AI, it is apparent that numerous attempts have been made to dissect, demystify, or improve the transparency of closed-box AI models. Thus, from a technical standpoint, it is arguably obvious that AI researchers now have a good grasp of the fundamental underpinnings of AI algorithms, which explains why there has been a surge of research output introducing novel approaches or improving on existing state-of-the-arts. However, the non-technical people, that makes up the majority of AI system users or those who are impacted by AI decisions, seems to struggle to comprehend the intricacies upon which AI systems are based. In a slightly trivial opinion, one can assume that, while algorithmic biases have been reported and confirmed in some AI-generated decisions — which are more related to training data than to the technicalities of data processing (and of course, deserves the attention it is getting) — the distrust is "partly, arguably" influenced and amplified by the discovery of a new research gold mine. While advocacy for transparent and explainable AI (led largely by the Social Science discipline) has aided its penetration and understanding across disciplines, it is hoped that, along-side calls for regulations or explainability from the business side of AI, we will continue to push for a more standardized and responsible approach to designing an AI-powered systems. One of these standards could be to make proprietary AI-based technologies that affect the public (DFAI falls into this category) more programmatically transparent (this, of course, has been well pushed in the EU), or to mandate that no closed-box should be used for certain high-stakes decisions when an interpretable model with the equal performance ability exists (Rudin, 2019). This, however, may be challenging, particularly in light of legislation protecting trade secrets and the recent innovations enabled by AI that were

hitherto considered practically unimaginable. Nonetheless, science advances rapidly, responding to (internal or external) reactions along the way. What could be concerning is an attempt to oversimplify science for the sake of comprehension. This is why explanations based on simplification should be used with caution. "*Some things in life are too complicated to explain...Not just to explain to others but to explain to yourself. Force yourself to try to explain it and you create lies.*[130]" While there is a significant difference between comprehending and nearly comprehending something, a correct explanation may result in decreased comprehensibility; conversely, a more comprehensible explanation may result in decreased accuracy (Yamploskiy, 2019). It may, therefore, seem illogical or counter-intuitive to presume that technical explanations provided post-hoc, or modelled with the internals of AI models will be understood by the intended audience even after simplification. Perhaps at that point, an evaluation of comprehensibility will be required. As a result, explanation of an AI-enabled outcome should justify not just the mathematical basis, technical underpinnings, and social context, but also the impact on people.

Lastly, it is worth emphasizing, however, that the discussion in this section is a trivially expressed opinion of the author; based entirely on personal social observations. They are merely offered to lessen the escalation of debate about whether AI (with its perceived opacity) should be applied to DF investigation.

According to a famous Albert Einstein quotation, which reads as follows:

> *"It would be possible to describe everything scientifically, but it would make no sense. It would be a description without meaning – as if you described a Beethoven symphony as a variation of wave pressure.*[131]*"*

## 6.9 Chapter Summary

The chapter explored the human-machine relationships that involve explaining machine-generated output, while demystifying the interchangeable usage of many words such as explainability, interpretability, and understandability. The relationship between artificial intelligence and law, as well as the right to explanation, is briefly discussed. Additionally, the goals and methods of explainable AI were elaborated on by deviating the concepts into Digital Forensics AI (DFAI). During this process, a working definition for explainable DFAI was

[130] Quote of Haruki Murakami - https://bukrate.com/quote/544024
[131] Quoted in Max Born, Physik im Wandel meiner Zeit, (Braunschweig: Vieweg, 1966)

proposed (xDFAI). The case for interpretability in DFAI was discussed, and some recommendations for an interpretable DFAI model were offered. Finally, a trivial discussion was added to express the author's personal opinion.

# Chapter 7

# Conclusion and Future Work

Digital forensics as a field has advanced significantly over the last few decades, charting its own path while simultaneously lending itself to the intricacies of forensic science. More criminal and civil cases are being resolved on the basis of evidence derived from digital data analysis, while digital forensic practitioners are providing expert testimony in high-stakes cases, including those involving national security. Hence, it is fair to assume that digital forensics is approaching the maturity level expected of a scientific field. However, there are obstacles along the way, and so many more are being unravelled on a regular basis. Notably, the sophistication of modern digital devices and the rapid growth of technological innovations have posed the most significant challenges to digital forensics as a domain to date. As criminals develop new techniques of crime, digital forensics as a domain has been grappling with how to respond. The advent of AI and its associated models (such as data mining, machine learning, and deep learning, among others) has proven to be a game changer when applied to a variety of complex tasks for which a solution had previously appeared elusive. While the advent of AI has created a new avenue for crime (with rapidly growing dynamics), it has also demonstrated that it may be beneficial in combating, identifying, and preventing crimes in many instances. Nevertheless, the forensic science community have expressed concerns about inferred facts using an AI-based methods; and rightly so, the broader social discussions have been centred around whether AI systems (sometimes referred to as closed-boxes) are sufficiently transparent or trustworthy to foster belief. The same issues have permeated the legal system, where evidence obtained using probabilistic algorithm procedures has been subjected to intense review.

This work takes a contrasting position and provides a practical use case to demonstrate that AI has shown promising and persuasive outcomes on complex tasks and is sufficiently robust to be considered as a suitable tool for “digital evidence mining.” Numerous studies have proposed AI-based approaches for analyzing digital evidence, the majority of which did not consider the intricacies of the law. We view the application of AI to digital forensics as a distinct body of knowledge that bridges three

domains: digital forensics, artificial intelligence, and law. As a result, we advocated formalizing “Digital Forensics AI” (DFAI) as a subfield or as an integral component of the existing framework in digital forensics. This enabled us to develop a holistic view of the proposed field’s sub-components.

In the first part of this thesis, we explored in depth the subject domain’s intersecting components: digital forensics, digital evidence and its legality, and AI. This is intended to provide the required context for the appropriate comprehension of the concept that this work aims to promote.

The second part demonstrates an automated approach for the examination of e-mail artefacts to demonstrate the efficacy of AI in seemingly complex forensic tasks that would take an unimaginable amount of time (with the possibility of errors) if performed manually. To be specific, we hypothesised that investigators are searching for evidence of a suspected fraud cover-up. We described a temporal analysis of e-mail exchange events in the experiment to determine whether suspicious deletions of communication between suspects happened and whether the deletions were intended to conceal evidence of discussion about certain incriminating subjects. In the use case, our model was able to identify with “*strong evidence*” that deletions did occur, while also accurately predicting the possible subjects they discussed. This method present a novel event reconstruction approach to address such investigation.

In the other parts, a preliminary conceptualization of DFAI was offered that could serve as a springboard for a more refined advancements in the future. Additionally, the conceptualization attempts to distinguish between “Digital Forensic AI” and “AI Forensics,” those of which have been used interchangeably or colloquially to mean the same thing. Further on this line, several techniques for evaluating AI models were discussed, with a major emphasis on the most important metrics for evaluating digital forensics tasks conducted using AI-based methods. In a similar vein, we examined the standardization of DFAI processes as they relate to the Daubert standards, with a particular focus on the processing and reporting of forensic datasets and error rates.

Numerous optimization techniques for AI models were examined, as well as their potential consequences, time complexity requirements, and their potential suitability for a typical DFAI task.

In this part, additionally, the significance of explaining the processes that resulted in the conclusions reached during a digital forensics investigation as a prerequisite for the admissibility of digital evidence is elaborated. The technique for machine-generated inferences must be transparent, trustworthy, fair, and justifiable. Explainable AI is a concept that encompasses all methods and proposals for making closed-box artificial intelligence models understandable. Our emphasis was on the domain-specific nature of explanation in digital forensics, and therefore the divergence of explainable AI into DFAI was established. The importance of interpretability was highlighted, and recommendations were offered on how to mitigate mistrust in AI-based digital forensic investigation through interpretable approaches.

In the future, it is envisaged that some of the use cases highlighted in this research will be refined through practical experiments in order to establish the validity of some of the methodologies (shallowly) described in a general sense. Equally as promising as the results of the e-mail analysis are, the method does not account for the possibility of a malicious use of a suspect's email account to falsely implicate them or conceal detection. Future works will attempt to build upon this reality. In light of the fact that the application of AI to digital forensics lacks a standardized approach and the domain-specific requirements have not been fully conceptualized, future research will seek to review developments in the law and AI domains in order to adapt practical, standardized, and interpretable mechanisms.

# Chapter 8

# Resources, tools, and sources

## 8.1 Datasets

| Dataset | Chapter | Source |
|---|---|---|
| Enron Dataset | 4 | https://www.cs.cmu.edu/~enron/ |

## 8.2 Software /other tools

| Software | Source |
|---|---|
| Experiment Codes | https://github.com/spyderweb-abdul/Deletion-Detection-in-Unstructured-Data |
| NLTK | https://www.nltk.org/ |
| Tokenizer | https://www.nltk.org/api/nltk.tokenize.html |
| Part-of-Speech Tagger | https://nlp.stanford.edu/software/tagger.shtml |
| NLTK Wordnet | https://www.nltk.org/_modules/nltk/stem/wordnet.html |
| NetworkX | https://pypi.org/project/networkx/ |
| LDA/NMF (Gensim) | https://github.com/RaRe-Technologies/gensim |

# Bibliography


Aamodt, A., & Plaza, E. (1994). CBR: Foundational Issues, Methodological Variations and System Approaches. *AI Communication, 7*(1), 39-59.

Abreu, S. (2019). Automated Architecture Design for Deep Neural Networks. *arXiv preprint arXiv:1908.10714*.

Aditya, K., Grzonkowski, S., & Lekhac, N.-A. (2018). Enabling Trust in Deep Learning Models: A Digital Forensics Case Study. *IEEE Intl. Conf. on Trust, Security and Privacy in Computing and Communications*, (pp. 1250-1255).

Adler, A. M. (2001). The Perverse Law of Child Pornography. *Columbia Law Review, 101*(2).

Agrawal, R., & Psaila, G. (1995). Active Data Mining. *Proceedings of the First Intl. Conf. on Knowledge Discovery and Data Mining (KDD '95)*, (pp. 3-8).

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. *Proceedings of the Intl. Conf. on very Large Databases (VLDB)*, *1215*, pp. 487-499.

Agrawal, R., Tomasz, I., & Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases. *Proceedings of the 1993 ACM SIGMOD Intl. Conference on Management of Data*, (p. 207).

Al Banna, H., Haider, A., Al Nahaian, J., Islam, M., Taher, K., & Kaiser, S. (2019). Camera Model Identification using Deep CNN and Transfer Learning Approach. *Intl. Conf. on Robotics, Electrical and Signal Processing Techniques (ICREST)*, (pp. 626-630).

Al Mutawa, N., Bryce, J., Franqueira, V., Marrington, Andrew, & Read, j. (2018). Behavioural Digital Forensics Model: Embedding Behavioural Evidence Analysis into the Investigagtion of Digital Crimes. *Digital Investigation, 28*, 70-82.

Alauthman, M., Aslam, N., Al-Khasassbeh, M., & Khan, S. (2020). An Efficient Reinforcement Learning-Based Botnet Detection Approach. *Journal of Network and Computer Application, 150*, 102479.

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017). Understanding of a Convolutional Neural Network. *Intl. COnf. on Engineering and Technology (ICET)* (pp. 1-6). IEEE.

Aletras, N., Tsarapatsanis, D., Preotiuc-Pietro, D., & lampos, V. (2016). Predicting Jusicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective. *PeerJ Computer Science, 2*, e93.

Allen, D. M. (1971). Mean square error of prediction as a criterion for selecting variables. *Technometrics, 13*(3), 469-475.

Alles, E. J., Geradts, Z., & Veenman, C. J. (2009). Source Camera Identification for Heavily JPEG Compressed Low Resolution Still Images. *Journal of Forensic Sciences, 54*(3), 628-638.

Alpaydin, E. (2010). *Introduction to Machine Learning* (2nd ed.). MIT Press.

Alqaraawi, A., Schuessler, M., WeiB, P., Constanza, E., & Berthouze, N. (2020). Evaluating saliency map explanations for convolutional neural networks: a user study. *Proceedings of the 25th Intl. Conf. on Intelligent User Interfaces*, (pp. 275-285).

Al-Qershi, O. M., & Khoo, B. E. (2013). Passive detection of copy-move forgery in digital images: state-of-the-art. *Forensic Science International, 231*(1-3), 284-295.

Alrabaee, S., Shirani, P., Debbabi, M., & Wang, L. (2017). On the Feasibility of Malware Authorship Attribution. *arXiv preprint arXiv:1701.02711*.

Anda, F., David, L., kanta, A., Becker, B., BBou-Harb, E., Le Khac, N.-A., & Scalon, M. (2019). Improving the Accuracy of Automated Facial Age Estimation to Aid CSEM Investigation. *Digital Investigation, 28*, S142.

Anda, F., Lillis, D., Le-Khac, N.-A., & Scalon, M. (2018). Evaluating Automated Facial Age Estimation Techniques for Digital Forensics. *IEEE Security and Privacy Workshop*, (pp. 129-139).

Anders, K. (2008). What are artificial neural networks? *Nature Biotechnology, 26*, 195-197.

Anjomshoae, S., Framling, K., & Amro, N. (2019). Explanation of Black-Box Model Predictions by Contextual Importance and Utility. *Intl. Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, (pp. 95-109).

Anuradha, D. B., & Padmavathi, B. (2019). BotHook: A Supervised Machine Learning Approach for Botnet Detection Using DNS Query Data. *Intl. Conf. on Communication and Cyber Physical Engineering (ICCCE). 570*, pp. 261-269. Springer.

Aparna, J., & Dija, S. (2015). Detection of Spoofed Mails. *IEEE Intl. Conf. on Computational Intelligence and Computing Research (ICCIC).* IEEE.

Aranganayagi, S., & Thangavel, K. (2007). Clustering categorical data using silhouette co-efficient as a relocating measure. *Intl. Conf. on Computational Intelligence and Multimedia Applications (ICCIMA 2007). 2*, pp. 13-17. IEEE.

Ari, N., & Heru, S. (2020). Hyper-parameter Tuning Based on Random Search for DenseNet Optimization. *Intl. Conf. on Information Technology, Computer, and Electrical Engineering.*

Armstrong, J., & Collopy, F. (2009). Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting, 8*(1), 69-80.

Arras, L., Montavon, G., Muller, K.-R., & Samek, W. (2017). Explaining Recurrent Neural Network Predictions in Sentiment Analysis. *Proceeding of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social media Analysis*, (pp. 159-168).

Arrieta, A. B., Diaz-Rodriguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., . . . Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Towards Responsible AI. *Information Fusion, 58*, 82-115.

Arshad, H., Jantan, A. B., & Abiodun, O. I. (2018). Digital Forensics: Review of Issues in Scientific Validation of Digital Evidence. *Journal of Information Processing Systems, 14*(2).

Arun, N., Nathan, G., Praveer, S., Ken, C., Mehak, A., Bryan, C., . . . Jayashree, K.-C. (2021). Assessing the (Un)trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, e200267.

Asatiani, A., Malo, P., Nagbol, P. R., Penttinen, E., & Rinta-Kahila, T. (2020). Challenges of Explaining the Behaviour of Black-Box AI Systems. *MIS Quaterly, 19*(4), Article 7.

Ashley, K. (1991). Reasoning with cases and hypotheticals in HYPO. *International Journal of Man-Machine Studies, 34*(6), 753-796.

Ashton, K. (2009). That 'Internet of Things' Thing. *RFID Journal, 22*(7), 97-114.

Ask, K., & Granhag, P. A. (2005). Motivational Sources of Confirmation Bias in Criminal Investiagtions: The Need for Cognitive Closure. *Journal of Investigative Psychology and Offender profiling, 2*, 43-63.

Association of Chief Police Officers of Englan, Wales & Northern Ireland (ACPO). (2012). *Good Practice Guide for Digital Evidence (Version 5).* Retrieved from https://library.college.police.uk/docs/acpo/digital-evidence-2012.pdf

Association of Forensic Science Providers (AFSP). (2009). Standards for the formulation of evaluative forensic science expert opinion. *Sci Justice, 49*(3), 161-164.

Atkinson, K., Bench-Capon, T., & Bollegala, D. (2020). Explanation in AI and Law: Past, Present and Future. *Artificial Intelligence, 289*, 103387.

Atzori, L., Lera, A., & Morabito, G. (2010). The Internet of Things: A Survey. *Computer Networks, 54*(15), 2787-2805.

Aziz, S., & Dowling, M. (2019). Machine Learning and AI for Risk Management. In T. G. Lynn, J. Mooney, P. Rosati, & M. Cumming (Eds.), *Disrupting Finance: Fintech and Strategy in the 21st Century* (pp. 33-50). Palgrave Pivot.

Baca, M., Cosic, J., & Cosic, Z. (2013). Forensic Analysis of Social Networks (Case Study). *5th Intl. Conf. on Information Technology Interfaces.* IEEE.

Bach, F. (2017). Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research, 18*(1), 629-681.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Muller, K.-R., & Samek, W. (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise relevance Propagation. *PLos One, 10*(7), e0130140.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., & Muller, K.-R. (2010). How to Explain Individual Classification Decisions. *Journal of machine Learning Research, 11*, 1803-1831.

Baggili, I., & Breitinger, F. (2015). Data sources for advancing cyber forensic: what the social world has to offer. *Proceedings of the 2015 AAAI Spring Symposium Series, Palo Alto, CA.*

Bagilli, I., & Bezadhan, V. (2020). Founding the Domain of AI Forensics. *Proceedings of the SafeAI@AAAI.* arxiv preprint: arXiv:1912.06497.

Banday, M. T. (2011). Techniques and Tools for Forensic Investigation of E-mail. *International Journal of Network Security & Its Application, 3*(6), 227.

Banerjee, A., & Rajesh, N. D. (2004). Validating clusters using the Hopkins statistic. *IEEE Intl. Conf. on Fuzzy Systems* (pp. 149-153). IEEE.

Bansal, G., Nushi, B., Kamar, E., Weld, D. S., Lasecki, W. S., & Horvitz, E. (2019). Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. *Proceedings of the AAAI conference on AI*, *33*, pp. 2429-2437.

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *104 California Law Review, 671*.

Barrett, G. B. (2000). The Coefficient of Determinantion: Understanding r squared and R squared. *The Mathematics Teacher, 93*(3), 230-234.

Barry, S. (2012). Ontology. *The Future of the World, 9*, 47-68.

Bastani, O., Kim, C., & Bastani, H. (2018). Interpretability via Model Extraction. *arXiv preprint arXiv:1706.09773.*

Bayu, B. S., & Miura, J. (2013). Fuzzy-based Illumination Normalization for Face Recognition. *IEEE Workshop on Advanced Robotics and its Social Impacts.*

Bebis, G., & Georgiopoulos, M. (1994). Feed-forward neural networks. *IEEE Potentials, 13*(4), 27-31.

Beebe, N. (2009). Digital Forensics Research: The Good, the Bad and the Unaddressed. *IFIP Intl. Conf. on Digital Forensics* (pp. 17-36). Springer, Berlin.

Beek, V., M, H., van den Bos, J., Boztas, A., van Ejik, E. J., Schramp, R., & Ugen, M. (2020). Digital Forensics as a Service: Stepping up the Game. *Forensic Science International: Digital Investigation, 35*.

Bellin, J., & Guthrie, A. F. (2014). Trail by Google: Judicial Notice in the Information Age. *NULR, 108*(4).

Bengio, Y. (2000). Gradien-based optimization of hyperparameters. *Neural Computation, 12*(8), 1889-1900.

Bengio, Y., Ducharme, R., Pascal, V., & Janvin, C. (2003). A Neural Probabilistic Language Model. *Journal of Machine Learning Research, 3*, 1137-1155.

Ben-Hur, A., Horn, D., Siegelmann, H., & Vapnik, V. N. (2001). Support vector clustering. *Journal of Machine Learning research, 2*, 125-137.

Benjamins, R., Berbado, A., & Sierra, D. (2019). Responsible AI by DESIGN in Practice. *arXiv preprint arXiv:1909.12838*.

Bennett, D. (2012). The challenges facing computer forensics investigators in obtaining information from mobile devices for use in criminal investigations. *Information Security Journal: A Global Perspective, 21*(3), 159-168.

Berger, C. E., Buckleton, J., Champod, C., Evett, I. W., & Jackson, G. (2011). Evidence Evaluation: a response to the court of appeal judgement in R v T. *Sci Justice, 51*(2), 43-49.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research, 13*(1), 281-305.

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for gyper-parameter optimization. *Proceedings of the Neural Information Processing Systems*, (pp. 2546-2554).

Besnard, P., & Hunter, A. (2008). *Elements of Argumentation.* MIT Press.

Bhat, V. H., Rao, P., Abhilash, R. V., Shenoy, D. P., Venugopal, K. R., & Patnaik, L. M. (2011). A Data Mining Approach for Data Generation and Analysis for Digital Forensics Application. *Intl. Journal of Web Engineering and Technology, 2*(3), 313-319.

Bhatt, U., Andrus, M., Weller, A., & Xiang, A. (2020). Mahine Learning Explainability for External Stakeholders. *arXiv preprint arXiv:2007.05408*.

Biasiotti, M. A., Bonnici, J. P., Cannataci, J., & Turchi, F. (2018). Handling and Exchanging Electronic Evidence Across Europe. *Law, Governance and Technology Series, 39*.

Biasiotti, M. A., Epifani, M., & Turchi, F. (2018). The Evidence Project: Bridging the Gap in the Exchange of Digital Evidence Across Europe. *Conference on Systematic Approaches to Digital Forensic Engineering.*

Bishop, C. M. (2006). Pattern Recognition. *Machine Learning, 128*(9).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research and Development in Information Retrieval, 51*.

Boesch, G. (2021). *What is Pattern recognition? A Gentle Introduction*. Retrieved from Visio.ai: https://viso.ai/deep-learning/pattern-recognition/

Bogawar, P. S., & Bhoyar, K. K. (2016). A Novel Approach for the Identification of Writing Traits on Email Database. *1st India Intl. Conf. on Information Processing (IICIP)*, (pp. 1-6).

Bollé, T., Casey, E., & Jacquet, M. (2020). The Role of Evaluations in Reaching Decisions Using Automated System Supporting Forensics Analysis. *Forensic Science International: Digital Investigation, 34*, 301016.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. *Proceedings of COMPSTAT*, (pp. 177-186).

Bottou, L., & Bousquet, O. (2007). The tradeoffs of large scale learining. Advances in neural information processing systems. In S. Sra, S. Nowozin, & S. J. Wright (Eds.), *Optimization for Machine Learning* (pp. 351-368). Cambridge: MIT Press.

Bozinovsk, S. (1982). A Self-learning system using secondary reinforcement. In R. Trappl (Ed.), *Cybernetics and Systems Research: Proceedings of the Sixth European Meeting on Cybernetics and Systems Research* (pp. 397-402). North Holland.

Bozinovski, S. (2001). Self-learning agents: A connectionist theory of emotion based on crossbar value judgement. *Cybernetics and Systems, 32*(6), 637-667.

Bozinovski, S. (2014). Modelling mechanisms of cognition-emotion interaction in artificial neural networks, since 1981. *Procedia Computer Science*, 255-263.

Brachman, R., & Anand, T. (1994). The Process of Knowledge Discovery in Databases: A Human-Centered Approach. *Advances in Knowledge Discovery and data Mining*, 37-58.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145-1159.

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32.

Breiman, L. (2001). Statistical Modelling: The Two Cultures. *Statistical Sciences, 16*(3), 191-231.

Brighi, R., & Ferrazzano, M. (2021). Digital Forensics: Best Practices and Perspective. *Digital Forensics Evidence: Towards Common European Standards in Antifraud Administrative and Criminal Investigations*, 25-60.

Brighi, R., Ferrazzano, M., & Leonardo, S. (2020). Legal Issues on AI Forensics. *I-Lex, 13*(9), 19-42.

Brill, E. (1992). A Simple Rule-based Part of Speech Tagger. *Proceedings of the 3rd Conference on Applied Natural Language Processing* (pp. 152-155). ACM Digital Library.

Brunelli, R., & Poggio, T. (1997). Template matching: Matched Spatial Filters and Beyond. *Pattern Recognition, 30*(5), 751-768.

Bryan, K. (2017). *Psychologist Michelle Theer, her internet affair with John Diamond, and the murder of air force Captain Marty Theer*. Retrieved from https://soapboxie.com/military/Michelle-Theer- John-Diamond

Buduma, N., & Locascio, N. (2017). *Fundamentls of Deep Learning: Designing Next-Generation Machine Intelligence Algorithms.* O'Reilly.

Burns, M. (2020). *A Quick Guide to Digital Image Forensics in 2020*. Retrieved from Camera Forensics Blog: https://www.cameraforensics.com/blog/2020/03/06/a-quick-guide-to-digital-image-forensics-in-2020/

Caloyannides, M. A. (2004). Privacy Protection and Computer Forensics. *Artech House*.

Cameron, C. A., & Windmeijer, F. A. (1997). A R-Squared Measure of Goodness of Fit for some Common Nonlinear Regression Models. *Journal of Econometrics, 77*, 329-342.

Carbonell, J. G., Michalski, R. S., & Mitchell, T. M. (1983). An Overview of Machine Learning. *Machine Learning, 1*, 3-23.

Carr, N. (2014). *The glass cage: automation and us.* WW Norton & Co.

Carrier, B. (2003). Defining Digital Forensic Examination and Analysis Tools Using Abstraction Layers. *International Journal of Digital Evidence, 1*(4), 1-12.

Carrier, B. D., & Spafford, E. H. (2004). An Event-Based Digital Forensic Investigation Framework. *Digital Forensic Research Workshop (DFRWS USA).*

Carrier, B. D., & Spafford, E. H. (2004). Defining Event Reconstruction of a Digital Crime Scence. *Journal of Forensic Science, 49*(6), JFS2004127-8.

Carriquiry, A., Heike, H. T., & Vanderplas, S. (2019). Machine Learning in Foresnics Applications. *Significance, 16*(2), 29-35.

Casey, E. (2002). Error, uncertainty and loss in digital evidence. *International Journal of Digital Evidence, 1*(2).

Casey, E. (2004). *Digital Evidence and Computer Crime* (2nd ed.). Elsevier.

Casey, E. (2010). *Handbook on Digital Forensics and Investigation.* Academic Press.

Casey, E. (2011). Digital Evidence and Computer Crime. In *Forensic Science, Computers, and the Internet* (3rd ed.). Academic Press.

Casey, E. (2019). The Chequered Past and Risky Future of Digital Forensics. *Australian Journal of Forensic Sciences, 51*(6), 649-664.

Casey, E. (2020). Standardization of Forming and Expressing Preliminary Evaluative Opinion on Digital Evidence. *Forensic Science International: Digital Investigation, 32*, 200888.

Casey, E., Barnum, S., Griffith, R., Snyder, J., Harm, v. B., & Nelson, A. (2017b). Advancing coordinated cyber-investigagtions and tools interoperability using a community developed specification language. *Digital Investigation, 22*, 14-45.

Casey, E., Biasiotti, M., & Turchi, F. (2017a). Using Standardization and Ontology to Enhance Data Protection and Intelligent Analysis of Electronic Evidence. *Discovery of Electronically Stored Information Workshop.*

Castiglione, A., Cattaneo, G., De Maio, G., De Santis, A., Costabile, G., & Epifani, M. (2012). The Forensic Analysis of a False Digital Alibi. *6th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing* (pp. 114-121). IEEE.

Caviglione, L., Webdzel, S., & Mazurczyk, W. (2017). The Future of Digital Forensics: Challenges and the Road Ahead. *IEEE Security & Privacy, 15*(6), 12-17.

Chabot, Y., Bertaux, A., Nicolle, C., & Kechadi, T. (2014). A Complete Formalized Knowledge Representation Model for Advanced Digital Forensics Timeline Analysis. *Digital Investigation, 11*(2), s95-s105.

Chabot, Y., Bertaux, A., Nicolle, C., & Kechadi, T. (2015). An Ontology-based Approach for the Reconstruction and the Analysis of Digital Timelines. *Digital Investigation, 15*(C), 83-100.

Chai, T., & Draxler, R. R. (2014). Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? - Arguments against avoiding RMSE in the literature. *Geoscientific Model Development, 7*(3), 1247-1250.

Champod, C., & Vuille, J. (2011). Scientific Evidence in Europe - Admissibility, Evaluation and Equality of Arms. *International Commentary of Evidence, 9*(1).

Chari, S., Gruen, D. M., Senevirante, O., & McGuiness, D. L. (2020). Foundation of Explainable Knowledge-Enabled System. *arXiv preprint arXiv:2003.07520*.

Chaski, C. E. (2005). Who's at the Keyboard? Authorship Attribution in Digital Evidence Investigation. *International Journal of Digital Evidence, 4*(1), 1-13.

Chau, M., Xu, J. J., & Chen, H. (2002). Extracting Meaningful Entities from Police Narrative Reports. *Proceedings of the 2002 annual national conference on digital government research*, (pp. 1-5).

Che, Z., Purushotham, S., & Khemani, R. L. (2015). Distilling Knowledge from Deep Networks with Applications to Healthcare Domain. *arXiv preprint arXiv:1512.03542*.

Chen, J., Song, L., Wainwright, M., & Jordan, M. (2018). An Information-Theoretic Perspective on Model Interpretation. *Proceedings of the 35th Intl. Conf. on Machine Learning (ICML)*, *80*, pp. 882-891.

Chen, K., Clark, A. J., De Vel, O., & Mohay, G. (2003). ECF-Event Correlation for Forensics. *Australian Computer, Networks & Information Forensics Conference*.

Chen, Y., Zhao, X., & Jia, X. (2015). Spectral-spatial classification of hyperspectral data based on deep belief network. *IEEE Journal of Selected Topics in Applied Earth Observation and Remote Sensing, 8*(6), 2381-2392.

Chen, Z. (2020). Graph Convolutional Networks for Graph with Multi-Dimensionally Weighted Edges. *arXiv preprint arXiv:1808.06099*.

Chiadighikaobi, I. R., & Abdullah, J. (2017). Malicious Code Intrusion Detection using Machine Learning and Indicators of Compromise. *International Journal of Computer Science and Information Security, 15*(9).

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science, 7*, e623.

Chirath, D. A. (2019). *Email Forensics: Investigation Techniques*. Retrieved from Forensic Focus Article: https://www.forensicfocus.com/articles/email-forensics-investigation-techniques/

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representation using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv preprint arXiv:1406.1078*.

Choi, E., Bahadori, M. T., Kulas, J. A., Schuetz, A., Stewart, W. F., & Sun, J. (2016). RETAIN: an Interpretable Predictive Model for Healthcare Using Reverse Time Attention Mechanism. *NIPS*, (pp. 3512-3520).

Chowdhury, G. G. (2003). Natural Language Processing. *Annual Review of Information Science and Technology, 37*(1), 51-89.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modelling. *arXiv preprint arXiv:1412.3555*.

Chung, M. H., & Gray, P. (1999). Data Mining. *Journal of Management Information Systems, 16*(1), 11-16.

Ciresan, D., Ueli, M., Jonathan, M., Gambardella, L. M., & Jurgen, S. (2011). Flexible, High Performance Convolutional Neural Networks for Image Classification. *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, *2*, pp. 1237-1242.

Clancey, W. J. (1983). The epistemology of a rule-based expert system - a framework for explanation. *Artificial Intelligence, 20*(3), 215-251.

Cole, A. K., Gupta, S., Gurugubelli, D., & Rogers, K. M. (2015). A Review of Recent Case Law Related to Digital Forensics: the Current Issues. *Conference of Digital Forensics, Security and Law*, *2*, pp. 95-103.

Cole, L., Austin, D., & Cole, L. (2004). Visual Object Recognition Using Template Matching. *Austalasian Conference on Robotics and Automation .*

Collier, P. A., & Spaul, B. J. (1992). A forensic methodology for countering computer crime. *Artificial Intelligence Review, 6*(2), 203-215.

Constantini, S., Gasperis, G. D., & Olivieri, R. (2019). Digital Forensics and Investigation Meets Artificial Intelligence. *Annals of Mathematics and Artificial Intelligence, 86*, 193-229.

Constantini, S., Giovanni, D. G., & Olivieri, R. (2019). Digital Forensics Analysis: An Answer Set Programming Approach for Generating Investigagtion Hypothesis. *Annals of Mathematics and Artificial Intelligence, 86*(1-3), 193-229.

Cook, D. R., & Weisberg, S. (1982). Criticism and Influence Analysis in Regression. *Sociological Methodology, 13*, 313-361.

Cortes, C., & Vapnik, V. N. (1995). Support-vector Networks. *Machine Learning, 20*(3), 273-279.

Cortez, P., & Embrechts, M. J. (2013). Using Sensitivity Analysis and Visualization Techniques to Open Black Box Data Mining Models. *Information Sciences, 225*, 1-17.

Cortez, P., & Emrechts, M. J. (2011). Opening black box data mining models using sensitivity analysis. *IEEE Symposium on Computational Intelligence and Data Mining (CIDM).*

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current directions in psychological science, 19*(1), 51-57.

Coyle, D., & Weller, A. (2020). Explaining machine learning reveals policy challenges. *Science, 368*(6498), 1433-1434.

Craven, M. W. (1996). *Extracting comprehensible models from trained neural networks.* Ph.D. Dissertation, The University of Wisconsin-Madison, Department of Computer Science.

da Silva, D. B., Schmidt, D., da Costa, C. A., da Rosa, R. R., & Eskofier, B. (2021). DeepSigns: A predictive model based on Deep Learning for the early detection of patient health deterioration. *Expert Systems with Applications, 165*, 113905.

Dabkowski, P., & Gal, Y. (2017). Real time image saliency for black box classifiers. *31st Intl. Conf. on Neural Information Processing Systems*, (pp. 6970-6979).

Dahl, G. E., Sainath, T. N., & Hinton, G. E. (2013). Improving deep neural networks for LVCSR using rectfied linear units and dropout. *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, (pp. 8609-8613).

Dai, B., Ding, S., & Wahba, G. (2013). Multivariate Bernoulli Distribution. *Bernoulli, 19*(4), 1465-1483.

Dan L., Jitender D., Spaulding W. & Shuart B. (2004). Towards missing data imputation: a study of fuzzy k-means clustering method. *In International conference on rough sets and current trends in computing* (pp. 573-579).

Dasklis, T. K., & Arakelian, V. (2021). Special Issue on Financial Forensics and Fraud Investigation in the Era of Industry 4.0. *Digital Finance, 3*, 299-300.

David, H., Padhraic, S., & Heikki, M. (2001). *Principles of data Mining.* The MIT Press.

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1*(2), 224-227.

Davis, B., Bhatt, U., Bhardwaj, K., Marculescu, R., & Moura, J. M. (2020). On Netwrok Science and Mutual Information for Explainable Deep Neural networks. *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, (pp. 8399-8403).

Davis, R., & Shrobe, H. S. (1993). What is a Knowledge Representation? *AI Magazine, 14*(1), 17-33.

De Fauw, J., Ledsman, H. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., & Ronneberger, O. (2018). CLinically applicable deep learning for diagnosis and referral in retina disease. *nature Medicine, 24*(9), 1342-1350.

De Myttenaere, A., Golden, B., Le Grand, B., & Rossi, F. (2016). Mean absolute percentage error for regression models. *Neurocomputing, 192*, 38-48.

De Vel, O., Anderson, A., Corney, M., & Mohay, G. (2001). Mining E-Mail Content for Author Identification Forensics. *SIGMOD Record, 30*(4), 55-64.

Delua, J. (2021). *Supervised Vs. Unsupervised Learning: What's the Difference?* Retrieved from IBM: https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning.

Desai, D. R., & Kroll, J. A. (2017). Trust But Verify: A Guide to Algorithms and the law. *Havard Journal of Law & Technology, 31*(1).

DFRWS Technical Committee. (2001). *A Roadmap for Digital Forensic Research.* DFRWS.

Dhurandhar, A., Chen, P., Luss, R., TU, C., Ting, P., Shanmugam, K., & Das, P. (2018). Explanations based on the missing: Towards contrastive explanations with pertinent negatives. *Proceedings of the Advances in Neural Information Processing Systems*, (pp. 592-603).

Di Bucchianico, A. (2008). Coefficient of Determination (R2). *Encyclopedia of Statistics in Quality and Reliability, 1*.

Donadello, I. (2018). *Semantic Image Representation - Integration of Numerical Data and Logical Knowledge for Cognitive Vision.* Doctoral Thesis, University of Trento.

Donadello, I., Serafini, L., & Garcez, A. D. (2017). Logic Tensor Networks for Semantic Image Representation. *IJCAI*, (pp. 1596-1602).

Dong, Y., Su, H., Zhu, J., & Zhang, B. (2017). Improving Interpretability of Deep Neural Networks with Semantic Information. *IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, (pp. 975-983).

Dongare, A. D., Kharde, R. R., & Kachare, A. D. (2012). Introduction to Artificial Neural network. *International Journal of Engineering and Innovative Technology (IJEIT), 2*(1), 189-194.

Doowon, J. (2020). Artificial Intelligence Security Threat, Crime, and Forensics: Taxonomy and Open Issues. *IEEE Access*, 184560-184574.

Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspective. *Proceedings of the 1st Intl. Workshop on Comprehensibility and Explanation in AI and ML, Co-located with the 16th Intl. Conf. of the Italian Assoc. for AI.* Ceur.

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., . . . Wood, A. (2017). Accountability of AI under the Law: The Role of Explanation. *arXiv preprint arXiv:1711.01134*.

Dror, I. E., & Mnooking, J. L. (2010). The use of technology in human expert domains: challenges and risks arising from the use of automated fingerprint identification systems in forensic science. *Law, Probability and Risk, 9*(1), 47-67.

Du, X., Le-Khac, N.-A., & Scalon, M. (2017). Evaluation of Digital Forensics Process Models with Respect to Digital Forensics as a Services. *arXiv preprint arXiv:1708.01730*.

Dumais, S. T. (1994). Latent Semantic Indexing (LSI) and TREC-2. *NIST Special Publication*, 105-105.

Dunn, J. C. (1973). A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics, 3*(3), 32-57.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *3rd Innovations in Theoretical Computer Science Conference*, (pp. 214-226).

Ebrahimi, J., Rao, A., Lowd, D., & Dou, D. (2018). HotFlip: White-box Example for Text Classification. *Proceedings of the Association of Computational Linguistics*, *2*, pp. 31-36.

Eggensperger, K., Feurer, M., Hutter, F., Bergstra, J., Snoek, J., Hoos, H. H., & Leyton-Brown, K. (2013). Towards an empirical foundation for assessing Bayesian optimization of hyperparameters. *NIPS Workshop on Bayesian Optimization in Theory and Practice Work*, (pp. 1-5).

Elshawi, R., Maher, M., & Sakr, S. (2019). Automated mahine learning state-of-the-art and open challenges. *arXiv preprint arXiv:1906.02287*.

Emad, A., Alaa, E. A., Bsoul, M., Essam, A. D., & Otoom, A. F. (2019). Simplified Features for Email Autorship Identification. *Internantional Journal of Security and Networks, 8*(2), 72-81.

European Network of Forensic Science Institute (ENFSI). (2015). *Best Practice Manual for the Forensic Examination of Digital Technology (Version 01).* ENFSI. Retrieved from https://enfsi.eu/wp-content/uploads/2016/09/1._forensic_examination_of_digital_technology_0.pdf

European Network of Forensic Science Institutes (ENFSI). (2015b). *Guideline for evaluative reporting in forensic science: strenghtening the evaluation of forensic results across Europe.* Retrieved from http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf

European Union Agency for Cybersecurity (ENISA). (2015). *Electronic Evidence - A Basic Guide for First Responders.* ENISA. Retrieved from https://www.enisa.europa.eu/publications/electronic-evidence-a-basic-guide-for-first-responders

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., . . . Song, D. (2017). Robust Physical-World Attacks on Deep Learning Models. *arXiv preprint arXiv:1707.08945*.

Faggella, D. (2020). *What is Machine Learning? - An Informed Definition*. Retrieved from Emerj: https://emerj.com/ai-glossary-terms/what-is-machine-learning/

Falkner, S., Klein, A., & Hutter, F. (2018). BOHB: robust and efficient hyperparameter optimization at scale. *35th Intl. Conf. on Machine Learning (ICML)*, *4*, pp. 2323-2341.

Farid, D. M., & Rahman, M. Z. (2010). Anomaly Network Intrusion Detection Based on Improved Self Adaptive Bayesian Algorithm. *JCP, 5*(1), 23-31.

Farid, H. (2019). Digital Forensics in a Post-Truth Age. *Forensic Science International, 289*, 268-269.

Farkhund, I., Rachid, H., Benjamin, F. C., & Mourad, D. (2008). A Novel Approach of Mining Write-Prints for Authorship Attribution in E-mail Forensics. *Digital Investigation, 5*, 42-51.

Faye, M. (2010). The Use of Artificial Intelligence in Digital Forensics: An Introduction. *Digital Evidence and Electronic Signature Law Review, 7*, 35-41.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine, 17*(3), 37.

Ferreira, S., Antunes, M., & Correia, M. (2021). Exposing Manipulated Photos and Videos in Digital Forensics Analysi. *Journal of Imaging, 7*(7), 102.

Ferreira, S., Antunes, M., & Correira, M. E. (2021). A Dataset of Photos and Videos for Digital Forensics Analysis Using Machine Learning Processing. *Data, 6*(8), 87.

Florea, M., Potlog, C., Pollner, P., D, A., Garcia, O., Bar, S., . . . Asif, W. (2019). Complex Project to Develop Real Tools for Identifying and Countering Terrorism: Real-time Early Detection and Alert System for Online Terrorist Content Based on Natural Language Processing, Social Network Analysis, AI & Complex Event Processing. *Open Access Repository, Birmigham University*.

Forensic Science Regulator (FSR). (2016). *Draft Guidance: Digital Forensics Method Validation.* Crown Prosecution Service.

Fowlkes, E. B., & Mallows, C. L. (1983). A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association, 78*(383), 553-569.

Framling, K. (2020). Explainable AI without Interpretable Model. *arXiv preprint arXiv:2009.13996v1*.

Framling, K. (2022). Contextual Importance and Utility: A Theoretical Foundation. In G. Long, X. Yu, & S. Wang (Eds.), *AI 2021: Advances in Artificial Intelligence. AI 2022. Lecture Notes in Computer Science* (Vol. 13151, pp. 117-128). Springer, Cham.

Frawley, W. J., Piatetsky-Shapiro, G., & J, M. C. (1992). Knowledge Discovery in Databases: An Overview. *AI Magazine, 13*(3), 57.

Freedman, D. (2009). *Statistical Models: Theory and Practice* (2nd ed.). Cambridge University Press.

Freire-Obregon, D., Narducci, F., Barra, S., & Castrillon-Santana, M. (2017). Deep Learning for Source Camera Identification on Mobile Devices. *arXiv preprint arXiv:1710.01257*.

Furnell, S. (2003). Cybercrime: Vandalizing the Information Society. In J. M. Lovelle, B. M. Rodriguez, J. E. Gayo, P. R. del Puerto, & L. J. Aguilar, *Lecture Notes in Computer Science* (Vol. 2722, pp. 8-16). Springer, Berlin.

Gaby, D., & Benjamin, F. C. (2013). Subject-based semantic document clustering for digital forensics investigation. *Data and Knowledge Engineering, 86*, 224-241.

Gandotra, E., Bansal, D., & Sofat, S. (2014). Malware Analysis Classification: A Survey. *Journal of Information Security, 5*, 56-64.

Gani, K., Hacid, H., & Skraba, R. (2012). Towards multiple identify detection in social networks. *Proceedings of the 21st International Conference on World Wide Web* (pp. 503-504). ACM Digital Library.

Gao, J., Lanchantin, J., Soffa, M., & Qi, Y. (2018). Black-box generation of adversarial text sequence to evade deep learning classifiers. *Proceedings of IEEE Security and Privacy Workshop*, (pp. 50-56).

Garamvolgyi, B., Ligeti, K., Ondrejova, A., & von Galen, M. (2021). Admissibility of Evidence in Criminal Proceedings in the EU. *EUCRIM Article, 2020*(3), 201-208.

Garcez, A. D., & Lamb, L. C. (2020). Neurosymbolic AI: The 3rd Wave. *arXiv preprint arXiv:2012.05876*.

Garcez, A., Gori, M., Lamb, L., Serafini, L., Spranger, M., & Tran, S. N. (2019). Neural-Symbolic Computing: An Effective Methodology for Principled Integration of Machine Learning and Reasoning. *arXiv preprint arXiv:1905.06088*.

Garfinkel, S. L. (1999). *USB deserves more support*. Retrieved from Business, The Boston Globe: https://simson.net/clips/1999/99.Globe.05-20.USB_deserves_more_support+.shtml

Garfinkel, S. L. (2010). Digital Forensic Research: The Next 10 Years. *Digital Investigation, 7*, 64-73.

Garofalakis, M., & Kumar, A. (2004). Deterministic Wavelet Thresholding for Maximum-Error Metrics. *Proccedings of the 23rd ACM SIGMOD-SIGACT Symposium on Principles of Databse System*, (pp. 166-176).

Gary, B.-C. (2014). *Community Discussion of the Definition of Digital Object*. Retrieved from Data Foundation and Terminology: https://www.rd-alliance.org/group/data-foundation-and-terminology-wg/post/community-discussion-definition-digital-object.html

Gazzaniga, M. S. (2015). *Tales from both sides of the brain: A life in neuroscience.* Ecco/HarperCollins .

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and Tensorflow.* Canada: O'Reilly Media, Inc.

Gers, F., Schmidhuber, J., & Cummins, F. (1999). Learning to Forget: Continual Prediction with LSTM. *Proceedings of the 9th Intl. Conf. on Artificial Neural Networks* (pp. 850-855). IEEE, London.

Gidudu, A., Hulley, G., & Tshilidzi, M. (2007). Image Classification using SVMs: one-against-one vs one-against-all. *arXiv preprint arXiv:0711.2914*.

Giova, G. (2011). Improving Chain of Custody in Forensic Investigation of Electronic Digital Systems. *International Journal of Computer Science and Network Security, 11*(1), 1-9.

Gladyshev, P., & Patel, A. (2004). Finite State Machine Approach to Digital Event Reconstruction. *Digital Investigation, 1*(2), 130-149.

Gleicher, M. (2016). A Framework for Considering Comprehensibility in Modelling. *Big Data, 4*(2), 75-88.

Gobel, T., Schafer, T., Hachenberger, J., Turr, J., & Herald, B. (2020). A Novel Approach for Generating Synthetic Datasets for Digital Forensics. *Advances in Digital Forensics XVI, IFIP ACT, 589*, 73-63.

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2013). Peeking inside the black box: Visualizing statistical learning with plot of individual conditional expectation. *arXIv preprint arXiv:1309.6392*.

Gonzalez-Cuautle, D., Corral-Salinas, U., Sanchez-Perez, G., Perez-Meana, H., Toscano-Medina, K., & Hernandez-Suarez, A. (2019). An Efficient Botnet Detection Methodology using Hyper-parameter Optimization Through Grid Search Techniques. *Intl. Workshop on Biometric and Forensics (IWBF).*

Goode, S. (2009). The admissibility of electronic evidence. (29), 1.

Goodfellow, I. J., Mirza, M., Xiao, D., Courville, A., & Bengio, Y. (2014). An empirical investigation of catastrophic forgetting in gradient-based neural networks. *Proceedings of ICLR.*

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, d., Ozair, S., . . . Bengio, Y. (2014b). Generative Adversarial Networks. *arXiv preprint arxiv:1406.2661*.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2020). Generative Adversarial Networks. *Communication of the ACM, 63*(11), 139-144.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014a). Explaining and Harnessing Adversarial Examples. *arXiv preprint arXiv:1412.6572*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* MIT Press.

Goodison, S. E., Robert, C. D., & Brian, A. J. (2015). *Digital Evidence and the U.S. Criminal Justice System: Identifying and Other Need to More Effectively Acquire and Utilize Digital Evidence*. Retrieved from RAND Corporation: https://www.rand.org/pubs/research_reports/RR890.html.

Goodman, B., & Flaxman, S. (2016). European Union Regulations on Algorithmic Decision-Making and A 'Right to Explanation'. *arXiv preprint arXiv:1606.088813*.

Gordon, S., & Ford, R. (2002). Cyberterrorism? *Computer & Security, 21*(7), 636-647.

Graham, J., Jones, S., Booth, G., Champod, C., & Evett, I. W. (2006). The Nature of Forensic Science Opinion - a Possible Framework to Guide Thinking and Practice in Investiagtion and in COurt Proceedings. *Science & Justice, 46*(1), 33-44.

Grajeda, C., Breitinger, F., & Baggili, I. (2017). Availability of datasets for digital forensics: and what is missing. *Digital Investigation, 22*, S94-S105.

Graycar, A., & Rusell, S. G. (2002). *Identifying and responding to electronic frauds risks*. Retrieved from Handle: hdl.handle.net/2328/38668

Griffin, L. (2018). *Issues in Digital Evidence: Rules & Types*. Retrieved from Study.com: https://study.com/academy/lesson/issues-in-digital-evidence-rules-types.html

Grimm, P. W., Capra, D. J., & Joseph, G. P. (2017). Authenticating Digital Evidence. *Baylor Law Review* (69), 1.

Gross-Brown, R., Ficek, M., Agundez, J. L., Dressler, P., & Laoutaris, N. (2015). Data Transparency Lab Kick Off Workshop (DTL 2014) Report. *ACM SIGCOMM Computer Communication Review, 45*(2), 44-48.

Gruber, N., & Jockisch, A. (2020). Are GRU Cells More Specific and LSTM Cells More Sensitive in Motive Classification of Text? *Frontiers in Artificial Intelligence, 3*(40).

Guarino, N., Oberle, D., & Staab, S. (2009). What is an Ontology? In S. Staab, & R. Studer (Eds.), *Handbook on Ontologies. International Handbooks on Information Systems.* Springer, Berlin.

Guera, D., & Delp, E. J. (2018). Deepfake Video Detection using Recurrent Neural Networks. *15th IEEE Intl. Conf. on Advanced Video and Signal Based Surveillance (AVSS)*, (pp. 1-6).

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., & Giannotti, F. (2019). A Survey of Methods for Explaining Black Box Models. *ACM Computing Survey, 51*(5), 1-42.

Gunning, D. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine, 40*(2).

Guo, H., Jin, B., & Qian, W. (2013). Analysis of Email Header for Forensics Purpose. *Proceedings of the Intl. Conf. on Communication Systems and Network technologies.* IEEE.

Hajiramezanali, E., Hasanzadeh, A., Duffield, N., Narayanan, K. Z., & Qian, X. (2020). Variational Graph Recurrent Neural Networks. *arXiv preprint arXiv:1908.09710*.

Hall, M., Frank, E., Homles, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Exploratons Newsletter, 11*(1), 10-18.

Hall, P. (2018). On the Art and Science of Machine Learning Explanations. *arXiv preprint arXiv:1810.02909*.

Hall, S. W., Sakzad, A., & Choo, K.-K. R. (2021). Explainable artificial intelligence for digital forensics. *Wiley Interdisciplinary Reviews: Forensic Science*, e1434.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a reciever operating characteristic (ROC) curve. *Radiology*, 29-36.

Harrington, P. (2012). *Machine Learning in Action.* Simon and Schuster.

Harrison, R. L. (2010). Introduction to monte carlo simulation. *AIP Conference Proceedings*, *1204*, pp. 17-21.

Hartigan, J. A. & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. J*ournal of the royal statistical society. Series C (Applied Statistics), 28(1), 100-108*.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical learning* (2nd ed.). Springer.

Hauck, R. V., Atabakhsh, H., Ongvasith, P., Gupta, H., & Chen, H. (2002). Using Coplink to Analyze Criminal-Justice Data. *Computer, 35*(3), 30-37.

Haugeland, J. (1989). *Artificial Intelligence: The Very Idea.* MIT Press.

Hayes, A. (2020). *Wearable Technology*. Retrieved from Investopedia: https://www.investopedia.com/terms/w/wearable-technology.asp

Haykin, S. O. (2008). *Neural Networks and Learning Machines* (3rd ed.). Pearson.

Heaton, J. B., Polson, N. G., & Witte, J. H. (2017). Deep Learning for Finance: Deep Portfolio. *Applied Stochastic Models in Business and Industry, 33*(1), 3-12.

Hebb, D. (1949). *The Organization of Behavior.* New York: Wiley.

Henelius, A., Puolamaki, K., & Ukkonen, A. (2017). Interpreting Classifiers through Attribute Interactions in Datasets. *arXiv preprint arXiv:1707.07576*.

Henelius, A., Puolamaki, K., Bostrom, H., Asker, L., & Papapetrou, P. (2014). A peek into the black box: exploring classifiers by randomization. *Data Mining and Knowledge Discovery, 28*(5-6), 1503-1529.

Hewling, M. O. (2013). Digital Forensics: An Integrated Approach for the Investigation of Cyber/Computer Related Crimes. *University of Bedfordshire*.

Heyburn, R., Bond, R., Michaela, B., Mulvenna, M., Wallace, J., Rankin, D., & Cleland, B. (2018). Machine learning using synthetic and real data: similarity of evaluation metrics for different healthcare datasets and for different algorithms. *Proceedings of the 13th International FLNS Conference*, (pp. 1281-1291).

Hijazi, S., Kumar, R., & Rowen, C. (2015). Using Convolutional Neural Networks for Image Recognition. *Cadence Design Systems Inc*, 1-12.

Hilderman, R. J., & Hamilton, H. J. (1999). Knowledge Discovery and Interestingness Measure: A Survey. *Citseer*.

Himal, L. (2010). *E-mail Forensic Authroship Attribution (Doctoral Dissertation).* University of Fort Hare. Retrieved from http://libdspace.ufh.ac.za/handle/20.500.11837/708

Hinton, G. E., Krizhevsky, A., & Sida, W. D. (2011). Transforming auto-encoders. *Intl. Conf. on Artificial Neural Networks* (pp. 44-51). Speringer, Berlin.

Ho, T. K. (1995). Random Decision Forests. *Proceedings of the 3rd Intl. Conf. on Document Analysis and Recognition, Montreal, QC.*, (pp. 278-282).

Ho, T. K. (1998). The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8), 832-844.

Hoar, S. B. (2001). *Identity Theft: The Crime of the New Millenium*. Retrieved from United States Department of Justice: https://www.hsdl.org/?view&did=439991

Hofmann, T. (1999). Probabilistic latent Semantic Indexing. *Intl. ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 50-57).

Hong, J.-H., & Cho, S.-B. (2008). A probabilistic multi-class strategy of one-vs-rest support vector machines for cancer classification. *Neurocomputing, 71*(16-18), 3275-3281.

Hopkins, B., & Skellam, J. G. (1954). A new method for determining the type of distribution of plant individuals. *Annals of Botany, 18*(2), 213-227.

Horsman, G. (2019). Tool Testing and Reliability Issues in the Fiels of Digital Forensics. *Digital Investigation, 28*, 163-175.

Horsman, G., & Lyle, j. R. (2021). Dataset construction challenges for digital forensics. *Foresic Science International: Digital Investigation, 38*, 301264.

Hua, Y., Guo, J., & Zhao, H. (2015). Deep belief networks and deep learning. *Proceedings of 2015 International Conference on Intelligent Computing and Internet of Things* (pp. 1-4). IEEE.

Huang, J., Chai, J., & Cho, S. (2020). Deep Learning in Finance and Banking: A Literature Review and Classification. *Frontiers of Business Research in China, 14*, 1-24.

Hussain, F. (2017). Internet of Everything. *Internet of Things*, 1-11.

Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2011). Sequential model-based optimization for general algorithm configuration. *Proc. LION*, *5*, pp. 507-523.

Hutter, F., Kotthoff, L., & Vanschoren, J. (2019). In *Automatic Machine Learning: Methods, Systems, Challenges.* Springer.

Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., & Baesens, B. (2011). An Empirical Evaluation of the Comprhensibility of Decision table, Tree and Rule based Predictive models. *Decision Supoort Systems, 51*(1), 141-154.

Hyeok, K., Cholyong, J., & Ryang, U. (2016). Rare Association Rule Mining for Network Intrusion Detection. *arXiv preprint arXiv:1610.04306*.

Hyndman, R. J., & Koehler, A. B. (2006). Another Look at Measures of Forecast Accuracy. *International Journal of Forecasting, 22*(4), 679-688.

Ikuesan, A. R., & Venter, H. S. (2019). Digital Behavioural-fingerprint for User Attribution in Digital Forensics: Are we there yet? *Digital Investigation, 30*, 73-89.

Ilkou, E., & Maria, K. (2020). Symbolic Vs Sub-symbolic AI Methods: Frinds or Enemies? *CIKM (2020).*

Injadat, M., Moubayed, A., Nassif, A. B., & Shami, A. (2020). Systematic ensemble model selection approach for educational data mining. *Knowledge-based Systems, 200*, 105992.

Injadat, M., Salo, F., Nassif, A. B., Essex, A., & Shami, A. (2018). Bayesian optimization with machine learning algorithms towards anomaly detection. *IEEE Global Communication Conference*, 1-6.

Jackson, P. (1998). *Introduction to Expert Systems* (3rd ed.). Addison Wesley.

James, A. N., Hsien, W. H., & Matthew, A. B. (2019). Application of Machine Learning to Imaging and Diagnosis. *Biophysical Reviews, 28*(1), 111-118.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning.* Springer.

James, J. I., & Gladyshev, P. (2018). Challenges with Automation in Digital Forensic Investigation. *arXiv Preprint arXiv:1303.4498*.

Jamieson, k., & Talwalkar, A. (2015). Non-Stochastic best arm identification and hyper-parameter optimization. *Artificial Intelligence and Statistics*, 240-248.

Javed, A. R., Jalil, Z., Zehra, W., Gadekallu, T. R., Young, D. S., & Piran, J. (2021). A Comprehensive Survey on Digital Video Forensics: Taxonomy, Challeneges, and Future Directions. *Engineering Applications of Artificial Intelligence, 6*.

Jianyu, X., Shancang, L., & Qinglian, X. (2019). Video-Based Evidence Analysis and Extraction in Digital Forensic Investigation. *Special Section on Deep learning, Security, and Forensic Research Advances and Challenges, 7*, 55432.

Jiawei, H., Micheline, K., & Jian, P. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufman.

Joachims, T. (1999). *Svmlight: Support Vector Machine*. Retrieved from SVM-Light Support Vector Machine: http://svmlight.joachims.org

Johansson, U., konig, R., & Niklasson, L. (2004a). The truth is there - Rule extraction from opaque models using genetic programming. *Proceedings of FLAIRS conference*, (pp. 658-663).

Johansson, U., Niklasson, L., & Konig, R. (2004b). Accuracy vs. Comprehensibility in Data Mining Models. *Proceeding of the 7th Intl. Conf. on Information Fusion*, (pp. 295-300).

Johnson, P., & Mead, D. (1991). Legislative knowledge base systems for public administration: some practical issues. *Proceedings of the 3rdIntl. Conf. on Ai and Law (ICAIL)*, (pp. 108-117).

Jones, D. R., Schonlau, M., & Welch, W. J. (1998). Efficient Global Optimization of Expensive Black-box Functions. *Journal of Global Optimization, 13*(4), 455-492.

Jordan, M. I., & Mitchell, T. M. (2015). Machine Learning: Trends, Perspectives, and Prospects. *Science, 349*(6245), 255-260.

Juan, R. (2003). Using tf-idfto determine word relevance in document queries. *Proceedings of the first instructional conference on machine learning*, *242*, pp. 29-48.

Jure, Z. (1994). Introduction to Artificial Neural Network (ANN) Methods: What they are and how to use them. *Acta Chimica Slovenica, 41*, 327.

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in My Hand: Who's the Fairest in the Land? On the Interpretations, Illustrations, and Implications of Artificial Intelligence. *Business Horizon, 62*(1), 15-25.

Karimpanal, T., & Bouffanais, R. (2018). Self-organizing maps for storage and transfer of knowledge in reinforcement learning. *Adaptive Behavior, 27*(2), 111-126.

Karnin, Z., Koren, T., & Somekh, O. (2013). Almost optimal exploration in multi-armed bandits. *30th Intl. COnf. on Machine Learning (ICML)*, *28*, pp. 2275-2283.

Karpathy, A., & Johnson, J. F.-F. (2016). Visualizing and Understanding Recurrent Networks. *ICLR Workshop Track.* Retrieved from http://vision.standford.edu/pdf/KarpathyICLR2016.pdf

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukhthankar, R., & Fei-Fei, L. (2014). Large-Scale Video Classification with Convolutional Neural Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1725-1732). IEEE.

Kazemi, S. M., Rishab, G., Kshitij, J., Ivan, K., Akshay, S., Peter, F., & Pascal, P. (2020). Representation Learning for Dynamic Graphs: A Survey. *Journal of Machine Learning Research, 21*, 1-73.

Kelly, L., Sachan, S., Ni, L., Almaghrabi, F., Allmendinger, R., & Chen, Y.-W. (2020). Explainable Artificial Intelligence for Digital Forensics: Opportunities, Challenges and a Drug Testing Case Study. *Digital Forensic Science*.

Kenneally, E. E. (2001). Gatekeeping out of the box: Open source software as a mechanism to assess reliability for digital evidence. *Virginia Journal of Law & Technology* (6), 1.

Kerkhoff, W., Stoel, R., Mattijssen, E., & Hermsen, R. (2013). The likelihood ration approach in cartridge case and bullet comparison. *AFTE J, 45*(3), 284-289.

Kerr, O. S. (2011). *Computer Crime Law* (2nd ed.). West Acadmic Publishing.

Khan, H., Hanif, S., & Muhammad, B. (2021). A Survey of Machine Learning Applications in Digital Forensics. *Trends in Computer Science and Information Technology, 6*, 20-24.

Khan, M. N. (2012). Performance Analysis of Bayesian Networks and Neural Networks in Classification of File System Activities. *Computer & Security, 31*(4), P391-401.

Khan, M. N., Chatwin, C. R., & Young, R. C. (2007). A Framework for Post-Event Timeline Reconstruction Using Neural Networks. *Journal of Digital Forensics and Incident Response, 4*(3-4), 146-157.

Kim, B., Wattenberg, G. J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2017). Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV) . *arXiv Preprint arxiv:1711.11279*.

Kingma, D. (2014). Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes. *Intl. Conf. on Learning Representation (ICLR).*

Kipf, T. N., & Welling, M. (2016). Semi-supervised Classification with Graph Convolutional Networks. *arXiv preprint arXiv:1609.02907*.

Kipf, T. N., & Welling, M. (2016). Variational Graph Auto-encoder. *Neural Information Processing Systems (NeurIPS).* arXiv preprint arXiv:1312.6114.

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of cluster in k-means clustering. *International Journal, 1*(6), 90-95.

Koehrsen, W. (2018). *Comparison of activation functions for deep neural networks*. Retrieved from https: //towardsdatascience.com/bayes-rule-applied-75965e4482ff

Koh, P. W., & Liang, P. (2017). Understanding Black-box Predictions via Influence Functions. *Proceedings of the 34th Intl. Conf. on Machine Learning*, *70*, pp. 1885-1894.

Konig, R., Johansson, U., & Niklasson, L. (2008). G-Rex: A Versatile Framework for Evolutionary Data Mining. *IEEE Intl Conf. on Data Mining*, (pp. 971-974).

Koroniotis, N., Moustafa, N., & Stinikova, E. (2020). A new network fornsic framework based on deep learning for Internet of Things networks: A particle deep framework. *Future Generation Computer Systems, 110*, 91-106.

Koutsoukas, A., Monaghan, K. J., Li, X., & Huan, J. (2017). Deep-Learning: investigating deep neural networks hyper-parameters and comparison of performanc to shallow methods for modelling bioactivity data. *Journal of Chemoinformatics, 9*(42), 1-13.

Kpalma, K., & Ronsin, J. (2007). An Overview of Advances of Pattern recognition and Computer Vision. In *Vision Systems: Segmentation and Pattern Recognition* (p. 26).

Krakovna, V., & Doshi-Velez, F. (2016). Increasing the Interpretability of Recurrent Neural Networks Using Hidden Markov Models. *arXIv preprint arXiv:1606.05320*.

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal, 37*(2), 233-243.

Krivulin, N., Dennis, G., & Charles, H. (2005). Parallel implementation of a random search procedure: an experimental study. *5th WSEAS Intl. Conf. on Simulation, Modelling and Optimization (SMO'05).*

Kuchta, K. J. (2000). Computer Forensics Today. *Information Systems Security, 9*(1), 1-5.

Kuhn, M., & Kjell, J. (2013). *Applied Predictive Modelling.* Springer.

Kumar, K., Sofat, S., Aggarwal, N., & Jain, S. K. (2012). Identification of User Ownership in Digital Forensic Using Data Mining Techniques. *International Journal of Computer Applications, 50*(4), 1-5.

Kunang, Y. N., Nurmaini, S., Stiawan, D., & Yudho, B. (2020). Improving Classification Attacks in IoT Intrusion Detection Sytem Using Bayesian Hyperparameter Optimization. *3rd Intl. Seminar on Research of Information Technology and Intelligent System (ISRITI).*

Kuncheva, L. I. (2008). Fuzzy Classifiers. *Scholarpedia, 3*(1), 2925.

Kurosawa, K., & Kuroki, K. A. (2009). Individual Camera Identification using Correlation of Fixed Pattern Noise in Image Sensors. *Journal of Forensic Sciences, 54*(3), 639-641.

Kwak, N. (2008). Principal Component Analysis Based on L1-norm Maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30*(9), 1672-1680.

Kwon, H., Lee, S., & Jeong, D. (2021). User Profiling via Application Usage Pattern on Digital Devices for Digital Forensics. *Expert Systems with Application, 168*.

Langley, P. (2011). The Changing Science of Machine Learning. *Machine Learning, 82*, 275-279.

Larry, E. D., & Lars, E. D. (2012). *Digital Forensics for Legal Professionals: Understanding Digital Evidence from the Warrant to the Courtroom.* Science Direct.

Lau, S. (2017). *A Walkthrough of Convolutional Neural Network - Hyperparameter Tuning*. Retrieved from Medium: https://towardsdatascience.com/a-walkthrough-of-convolutional-neural-network-7f474f91d7bd

Lau, T., & Bidermann, A. (2020). Assessing AI Output in Legal Decision-Making with Nearest Neighbors. *Penn State Law Review, 24*(3).

Laurens, v. d., Eric, P., & Herik, J. v. (2009). Dimensionality Reduction: A Comparative Review. *Journal of Machine learning Research, 10*(1), 66-71.

Lazic, L., & Bogdanoski, M. (2018). E-mail Forensics: Techniques and Tools for Forensic Investigation. *10th Intl. Conference on Business Information Security*, (p. 25).

LeCun, Y. (1988). A theoretical framework for back-propagation. *Proceedings of the 1998 connectionist models summer school.*

LeCun, Y., Yoshua, B., & Geoffrey, H. (2015). Deep Learning. *Nature, 521*(7553), 436-444.

Lee, B. C., Downey, D., Lo, K., & Weld, D. S. (2020). LIMEADE: A General Framework for Explanation-Based Human Tuning of Opaque Machine Learners. *arXiv preprint arXiv:2003.04315*.

Lee, S. I., & Yoo, S. J. (2020). Multimodal deep learning for finance: integrating and forecasting international stock markets. *The Journal of Supercomputing, 76*(10), 8294-8312.

Lee, W., Stolfo, S. J., & Mok, W. (1999). A Data Mining Framework for Building Intrusion Detection Models. *IEEE Symposium Security and Privacy* (pp. 12-132). IEEE CS Press.

Lehman, E. L., & casella, G. (1998). *Theory of Point Estimation* (2nd ed.). Springer, New York.

Leslie, D. (2019). *Understanding artificial intelligence and safety: A guide for the responsible design and implementation of AAI systems in the public sector.* The Alan Turin Institute.

Li, D., Jitender, D., Spaulding, W., & Shuart, B. (2004). Towards missing data imputation: a study of fuzzy k-means clustering method. *Intl. Conf. on rough sets and current trends in computing* (pp. 573-579). Springer, Berlin.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2017). Hyperband: a novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning research, 18*(1), 1-52.

Li, Y., Fu, Y., Li, H., & Zhang, S.-W. (2009). The Improved Training Algorithm of Back Propagation Neural Network with Self-adaptive Learning Rate. *Intl. Conf. on Computational Intelligence and Natural Computing*, *1*, pp. 73-76.

Li, Z. (2002). A saliency map in primamry visual corts. *Trends in cognitive sciences, 6*(1), 9-16.

Liddy, E. D. (2001). Natural Language Processing. In *Encyclopedia of Library and Information Science* (2nd ed.). Marcel Decker, Inc.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy, 23*(1), 18.

Lior, R., & Oded, M. (2008). *Data mining with decision trees: theory and applications.* World Scientific Pub Co Inc.

Liou, C.-Y., Cheng, W.-C., & Liou, J.-W. L.-R. (2014). Autoencoder for words. *Neurocomputing, 139*, 84-96.

Lipton, Z. C. (2018). The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is bothe Important and Slippery. *Queue, 16*(3), 31-57.

Little, R. J. (1993). Statistical Analysis of Masked Data. *Journal of Official Statistics, 9*(2), 407.

Liu, F., Zhang, X., Ye, Y., Zhao, Y., & Li, Y. (2015). MLRF: multi-label classification through random forest with label-set partition. *Intl. Conf. on Intelligent Computing* (pp. 407-418). Springer, CHam.

Liu, V. (2016). *Metasploit Anti-Forensics Project (MAFIA)*. Retrieved from Bishopfox: https://resources.bishopfox.com/resources/tools/other-free-tools/mafia/

Lobo, F. G., Goldberg, D. E., & Pelikan, M. (2000). Time complexity of genetic algorithms on exponentially scaled problems. *Proceedings on Genetic Evolutionary Computation Conference*, (pp. 151-158).

Lone, A. H., & Mir, R. N. (2019). Forensic-chain: Blockchain based digital forensics chain of custody with PoC in Hyperledger Composer. *Digital Investigation, 28*, 44-55.

Lord, N. (2020). *What are Memory Forensics? A Definition of Memory Forensics*. Retrieved from Data Insider Blog: https://digitalguardian.com/blog/what-are-memory-forensics-definition-memory-forensics

Lorenzo, P. R., Nalepa, J., Kawulok, M., Ramos, L. S., & Pastor, J. R. (2017). Particle Swarm Optimization for Hyper-Parameter Selection in Deep Neural Networks. *Proceedings of the Genetic and Evolutionary Computation Conference*, (pp. 481-488).

Ludmila, P., Frantisek, B., & Jan, P. (2021). Semi-Automatic Adaptation of Diagnostic Ruls in the Case-Base Reasoning Process. *Applied Sciences, 11*(1), 292.

Lukas, J., Fridrich, J., & Goljan, M. (2006). Digital Camera Identification from Sensor Pattern Noise. *IEEE Transactions on Information Forensics and Security, 1*(2), 205-214.

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, (pp. 4768-4777).

Lynch, C. (2000). Authenticity and Integrity in the Digital Environment: an Exploratory Analysis of the Central Role of Trust. *Museums in a Digital Age*, 314-332.

Ma, X., & Hovy, E. (2016). End-to-end Sequence Labelling via Bi-directional LSTM-CNNs-CRF. *arXiv preprint arxiv:1603.01354*.

Mabu, S., Chen, C., Lu, N., Shimada, K., & Hirasawa, K. (2010). An intrusion-detection model based on fuzzy class-association-rule mining using genetic network programming. *IEEE transactions on system, amn, and cybernetics, part C (Applications and Reviews), 41*(1), 130-139.

MacDonell, S. G., Buckingham, D., Gray, A., & Sallis, P. (2002). Software Forensics: Extending Authorship Analysi Techniques to Computer Programs. *3rd Biannual Conference of the International Association of Forensic Linguistic (AIFL)*, (pp. 1-8).

Maclaurin, D., Duvenaud, D., & Adams, R. P. (2015). Gradient-based Hyperparameter Optimization through Reversible Learning. *Proceedings of the 32nd Intl. COnf. on Machine Learning*, *37*, pp. 2113-2122.

Mahdi, H. M., Maaruf, A., Peter, S. E., & Rich, P. (2015). A Review on Internet of Things (IoT), Internet of Everythin (IoE) and Internet of Nano Things (IoNT). *Internet Tecnologies and Application (ITA)*, (pp. 219-224).

Maimon, O., & Rokach, L. (2005). *Data Mining and Knowledge Discovery Handbook.*

Manhaeve, R., Dumnacic, S., Kimming, A., Demeester, T., & De Raedt, L. (2021). Neural Probabilistic Logic Programming in DeepProbLog. *Artificial Intelligence, 298*, 103504.

Manning, C. D., Raghavan, P., & Schutze, H. (2008). *Introduction to Information Retrieval.* Cambridge University Press.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, (pp. 55-60).

Manning, C., & Schutze, H. (1999). *Foundations of Statistical Natural Language Processing.* MIT Press.

Mao, J., Gan, C., Kohli, P., Tenebaum, J. B., & Wu, J. (2019). The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences from Natural Supervision. *arXiv preprint arxiv:1904.12584*.

Marcinowski, M. (2021). Deep Learning v. Human Rights. *Proceedings of the 1st Intl. Workshop on Bias, Ethics and Fairness in Artificial Intelligence: Representation and Reasoning (Befair 2021).*

Marija, L. (2021). *39 worrying cybercrime statistics.* Retrieved 2021, from Legal Jobs blog post: https://legaljobs.io/blog/cyber-crime-statistics

McCarthy, J. (2004). *What is Artificial Intelligence?* Retrieved from https://www.csee.umbc.edu/courses/471/papers/turing.pdf

McCulloch, W., & Walter, P. (1943). A Logical Calculus of Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics, 5*(4), 115-133.

McHugh, J. (2001). Testing intrusion detection systems: A critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory. *ACM Transactions on Information and System Security, 3*(4), 262-294.

McKemmish, R. (2008). When is Digital Evidence Forensically Sound? *IFIP International Conference on Digital Forensics* (pp. 3-15). Springer, Boston.

McMenamin, G. R. (2020). Forensic Stylistics: The Theory and Practice of Forensic Stylistics. In *The Routledge Handbook of Forensic Linguistics* (2nd ed.).

MEAN ABSOLUTE PERCENTAGE ERROR (MAPE). (2000). In P. M. Swamidass (Ed.), *Encyclopedia of Production and Manufacturing Management.* Springer, Boston.

Mehryay, M., Afshin, R., & Ameet, T. (2018). *Foundations of Machine Learning* (2nd ed.). MIT Press.

Michalski, R. S. (1983). A theory and methodology of inductive learning. *Artificial Intelligence, 20*(2), 111-161.

Mikolov, T., Karafiat, M., Burget, L., Cernosky, J., & Khudanpur, S. (2010). Recurrent Neural Network Based Language Model. *Interspeech, 2*(3), 1045-1048.

Miljanovic, M. (2012). Comparative Analysis of Recurrent and Finite Impulse Response Neural Networks in Time Series Prediction. *Indian Journal of Computer and Engineering, 3*(1).

Miller, G. A. (1956). The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychology Review, 63*(2), 81-97.

Miller, T. (2019). Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, 1-38.

Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of Inmates Running the Asylum. *IJCAI Workshop on Explainable AI, 36*, 36-40.

Minsky, M., & Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry.* MIT Press, Cambridge.

Miyamoto, D., Hazeyama, H., & Kadobayashi, Y. (2008). Detecting Methods of Virus Email Based on Mail Header and Encoding Anomaly. *Advances in Neuro-infromation Processing*.

Mohammad, R. A. (2021). Analyzing Cyber-Attack Intention for Digital Forensics Using case-Based Reasoning. *arXiv preprint arXiv:2101.01395*.

Mohammad, R. M. (2019). A Neural Network Based Digital Forensic Classification. *15th Intl. Conf. on Computer Systems and Application (AICCSA).*

Molnar, C. (2019). Interpretable machine learning. https://christophm.github.io/interpretable-ml-book/

Montasari, R., & Hill, R. (2019). Next Generartion Digital Forensics: Challenges and Future Paradigms. *12th IEEE International Conference on Global Security, Safety and Sustainability (ICGS3)*, (pp. 205-212).

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Muller, K.-R. (2017). Explaining nonlinear classifiction decisions with deep Taylor decomposition. *Pattern Recognition, 65*, 211-222.

Montavon, G., Samek, W., & Muller, K.-R. (2018). Methods for Interpreting and Understanding Deep Neural Networks. *Digital Signal Processing, 73*, 1-15.

Morgan, E. M. (1948). Hearsay Dangers and the Application of the Hearsay Concept. *Havard Law Reviews, 6*(2), 177-219.

Morovati, K., & Kadam, S. S. (2019). Detection of Phishing Emails with Email Forensic Analysis and Machine Learning Techniques. *International Journal of Cyber-Security and Digital Forensics (IJCSDF), 8*(2), 98-107.

Morrison, G. S. (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. *Science and Justice, 51*, 91-98.

Morrison, G. S., & Thompson, W. C. (2017). Assessing the Admissibility of a New Generation of Forensic Voice Comparison Testimony. *SCI. & TECH. Law Review, 18*(2).

Mrityunjay, C. U., & Gupta, S. (2017). Novel Approach for Email Forensics. *International Journal of Engineering Research & Technology (IJERT), 5*(10), 1-6.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective.* Cambridge: MIT Press.

Murrie, D. C., Boccaccini, M. T., Turner, D. B., Meeks, M., Woods, C., & Tussey, C. (2009). Rater (dis) agreement on risk assessment measures in sexually violent predator proceedings: evidence of adversarial allegiance in forensic evaluation? *Psychology, Public Policy, and Law, 15*(1), 19-53.

Murtaz, M., Azwar, H., Ali, S. B., & Rehman, S. (2018). A Framework for Android Malware Detection and Classification. *5th International Conference on Engineering Technologies and Applied Sciences (ICETAS).* IEEE.

Nair, V., & Hinton, G. E. (2010). Rectified Linear Units Improve Restricted Boltzman Machines. *Proceedings of the 27th International Conference on Machine Learning*, (pp. 807-814).

National Institute of Justice. (2008). *Electronic Crime Scene Investigation: A Guide for First Responders* (2nd ed.). U.S. Department of Justice, Washington D. C.

Nevitt, J., & Hancock, G. R. (2000). Improving the root mean square error of approximation for nonnormal conditions in structural equation modelling. *The Journal of Experimental Education, 68*(3), 251-268.

Ng, A. (2012). Clustering with the K-means Algorithm. *Machine Learning*.

Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., & Clune, J. (2016). Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *Proceedings of the NIPS*, (pp. 3395-3403).

Nirkhi, S., Dharaskar, R. V., & Thakare, V. M. (2012). Data Mining: A Prospective Approach for Digital Forensics. *Intl. Journal on Data Mining and Knowledge Management Process, 2*(6), 41-48.

Noble, W. S. (2006). What is a support vector machine? *Nature biotechnology, 24*(12), 1565-1567.

Noever, D. (2020). The Enron Corpus: Where the email bodies are buried? *arXiv preprint arXiv:2001.10374*.

Novak, M., Grier, J., & Gonzalez, D. (2018). New Approaches to Digital Evidence Acquisition and Analysis. *NIJ Journal* (208).

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdi, W., Vida, M.-E., . . . Papadopoulos, S. (2020). Bias in data-driven artificial intelligence systems - an introductory survey. *Wiley Interdisciplinary Reviews - Wires Data Mining and Knowledge Discovery, 10*(3), e1356.

Ogilvie, E. (2002). Cyberstalking. *Trends and Issues in Crime and Criminal Justice* (166), 1-6.

Orujov, F., Maskeliunas, R., R, D., & Wei, W. (2020). Fuzzy based image edge detection algorithm for blood vessel detection in retinal images. *Applied Soft Computing, 94*, 106452.

O'shea, K., & Nash, R. (2015). An Introduction to Convolutional Neural Networks. *arXiv preprint arXiv:1511.08458*.

Overill, R., & Collie, J. (2020). DEEP: Extending the Digital Forensics Process Model to Criminal Investigation. *Athens Journal of Science, 7*(4), 225-240.

Palmer, G. (2001). A road map for digital forensic research. *Proceedings of the 1st Digital Forensic Research Workshop, Utica, NY*, (pp. 27-30).

Panagiotis, S., Theodoros, S., & Petros, D. (2018). Examining Deep Learning Architecture for Crime Classification and Prediction. *arXiv preprint arXiv:1812:00602*.

Parth, B., Ilya, S., Nidhal, B., Robi, P., & Dimah, D. (2017). Machine Learning in Transportation Data Analytics. *Data Analystics for Intelligent Transportation Systems*, (pp. 283-307).

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms that Control Money and Information* (Vol. 320). Cambridge: Havard University Press.

Pathak, A. R., Pandey, M., & Rautaray, S. (2021). Topic-level sentiment analysis of social media data using deep learning. *Applied Soft Computing, 108*, 107440.

Pathak, J., Vidyarthi, N., & Summer, S. L. (2005). A fuzzy-based algorithm for auditors to detect element of fraud in settled insurance claim. *Managerial Auditing Journal, 20*(6), 0268-6902.

Pavel, P., & Jana, N. (1998). Novel Methods for Feature Subset Selection with Respect to Problem Knowledge. In H. Liu, & H. Motoda (Eds.), *Feature Extraction, Construction and Selection* (p. 101).

Pavlo, R. (2014). Impact of Training Set batch Size on the Performance of Convolutional Neural Networks for Diverse Datasets. *Information Technology and Management Science, 20*(1), 20-24.

Pearl, J. (2009). *Causality* (2nd ed.). Cambridge University Press.

Pedreschi, D., Giannotti, F., Guidotti, R., Monraele, A., Pappalardo, L., Salvatore, R., & Turinin, F. (2018). Open the Black Box Data-Driven Explanation of Black Box Decision Systems. *arXiv preprint arXiv:1806.09936*.

Peilun, W., Fan, Y., & Hui, G. (2021). Holmes: AN Efficient and Lightweight Semantic Based Anomalous Email Detector. *arXiv preprint arXiv:2104.08044*.

Peter, S. (1987). *Parts: A Study in Ontology* (Vol. 0199241465). Oxford University Press.

Piatetsky-Shapiro, G. (1991). Knowledge Discovery in Real Databases: A report. *AI Magazine*, 68-70.

Pica, A. (2013). An Overview on Image Forensics. *International Scholarly Research Notices, 2013*.

Piryonesi, S. M., & El-Diraby, T. E. (2021). Using Machine Learning to Examine Impact of Type of performance Indicator on Flexible Pavement Deterioration Modeling. *Journal of Infrastructure Systems, 27*(2), 04021005.

Pollit, M., Casey, E., Jaquet-Chiffelle, D. O., & Gladyshev, P. (2018). *A framework for harmonizing forensic science practices and digital/multimedia evidence.* Organization of Scientific Area Committees for Forensic Science (OSAC).

Pour, M. S., Bou-Harb, E., Varma, K., Neshenko, N., Pados, D. A., & Choo, K.-K. R. (2019). Comprehending the IoT Cyber Threat Landscape: A Data Dimensionality Reduction Technique to Infer and Characterize Internet-Scale IoT Probing Campaigns. *Digital Investigation, 28*, s40-s49.

Prasse, p., Knaebel, R., Machlica, L., Pevný, T., & Scheffer, T. (2019). Joint Detection of Malicious Domains and Infected Clients. *Machine Learning*, 1352-1368.

Preece, A., Ashelford, R., Armstrong, H., & Braines, D. (2018). Hows and Whys of Artificial Intelligence for Public Sector Decisions: Explanation and Evaluation. *arXiv preprint arXiv:1810.02689*.

President's Council of Advisors on Science and Technology (PCAST). (2016). *Forensic science in criminal courts: ensuring scientific validity of feature-comparison methods.* Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf

Qadir, S., & Noor, B. (2021). Applications of Machine Learning in Digital Forensics. *Intl. Conf. on Digital Futures and Transformative Technologies (ICoDT2)* (pp. 1-8). IEEE.

Quinlan, J. R. (1986). Introduction of decision trees. *Machine Learning, 1*, 81-106.

Quoc, V. L., Marc'Aurelio, R., Rajat, M., Matthieu, D., Kai, C. G., Jeff, D., & Andrew, Y. N. (2012). Building High-level Features Using Large Scale Unsupervised Learning. *29th Intl. Conf. on International Conference on Machine Learning* (pp. 507-514). arXiv preprint arXiv:1112.6209.

Rami, M. A., & Mohammed, A. (2019). A comparison of machine learning techniques for file system forensic analysis. *Journal of Information Security and Applications, 46*(1), 53-6.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66*(336), 846-850.

Rani, P., Liu, C., Sarkar, N., & Vanman, E. J. (2006). An empirical study of machine learning techniques for affect recognition in human-robot interaction. *Pattern Analysis and Applications, 9*(1), 58-69.

Rania, A., Tet, Y., & Morad, B. (2020). Using Topic Modelling Methods for Short-Text Data: A Comparative Analysis. *Frontier in Artificial Intelligence, 3*(42).

Rankin, D., Black, M., Bond, R., Wallace, J. M., & Epelde, G. (2020). Reliability of Supervised Machine Learning Using Synthetic Data in Health Care. *JMIR Med Inform, 8*(7), e18910.

Reith, M., Carr, C., & Gunsch, G. (2002). An Examination of Digital Forensic Models. *International Journal of Digital Evidence, 13*(2), 1-2.

Ren, L., & Glasure, Y. (2009). Applicability of the Revised Mean Absolute Percentage Error (MAPE) Approach to some Popular Normal and Non-normal Independent Time Series. *International Advances in Economic Research, 15*(4), 409-420.

Rezaeianjouybari, B., & Shang, Y. (2020). Deep learning for prognostics and health management: state of the art, challenges, and opportunities. *Measurement, 163*, 107929.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 1135-1144).

Rissland, E. L., & Ashley, K. (1987). A case-based system for trade secrets law. *Proceedings of the 1st Intl. Conf. on AI and Law*, (pp. 60-66).

Rivest, R., & Dusse, S. (1992). The MD5 Message-Digest Algorithm. *MIT Laboratory for Computer Science and RSA Data Security*, 330-344.

Robertson, S. E. (2004). Understanding Inverse Document Frequency: On Theoretical Argument for IDF. *Journal of Documentation, 60*(5), 503-520.

Roder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of topic Coherence Measures. *Proceedings of the 8th ACM Intl. Conf. on Web Search and Data Mining* (pp. 399-408). ACM Digital Library.

Rogers, D. M. (2005). Anti-forensic presentation given to Lockheed Martin.

Rosenblatt, F. (1957). *The Perceptron - a perceiving and recognizing automation.* Cornell Aeronautical Laboratory.

Rosenblatt, F. (1958). The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review, 65*(6), 386-408.

Rosenblum, N., Zhu, X., & Miller, B. P. (2011). Who Wrote this Code? Identifying the Authors of Program Binaries. In V. Atluri, & C. Diaz (Eds.), *Computer Security* (Vol. 6879). Springer, Berlin.

Roth, A. (2015). Trial by machine. *Geo. law Journal, 104*, 1245.

Roth, A. (2017). Machine Testimony. *Yale Law Journal, 126*(7), 1972-2053.

Rouhollahi, Z. (2021). Towards Artificial Intelligence Enabled Financial Crime Detection. *arXiv preprint arXiv:2105.10866*.

Rousseeuw, P. J. (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of CLuster Analysis. *Computational and Applied Mathematics, 20*, 53-65.

Rubin, D. B. (1993). Statistical Disclosure Limitation. *Journal of Official Statistics, 9*(2), 461-468.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decision and use interpretable models instead. *Nature Machine Learning, 1*(5), 205-215.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533-536.

Ruping, S. (2006). *Learning Interpretable Models.* Ph.D. Dissertation, University of Dortmund. Retrieved from https://d-nb.info/997491736/34

Rusell, S. J., & Norvig, P. (2009). *Artificial Intelligence: A Modern Approach* (3rd ed.). Upper Saddle River, New Jersey: Prentice Hall.

Ryan, D. J., & Shpantzer, G. (2010). *Legal Aspect of Digital Forensics.* Retrieved from http://euro.ecom.cmu.edu/program/law/08-732/Evidence/RyanShpantzer.pdf

Safara, F., Souri, A., & Serrizadeh, M. (2020). Improved Intrusion Detection Method for Communication Networks Using Association Rule Mining and Artificial Neural Networks. *IET Communications, 14*(7), 1192-1197.

Sahoo, A. K., Pradhan, C., & Das, H. (2020). Performance evaluation of different machine learning methods and deep-learning based convolutional neural network for health decision making. *Nature inspired computing for data science*, 201-212.

Salih, A., Zeebaree, S. T., Ameen, S., Alkhyyat, A., & Shukur, H. M. (2021). Survey on the role of artificial intelligence, machine learning and deep learning for cybersecurity attack detection. *7th Intl. Engibeering Conference on Research and Innovation amid Global Pandemic* (pp. 61-66). IEEE.

Sally, M. C., & Terence, L. L. (1999). Maintenance and Limitations Issues of Case-Based Reasoning Technology in a Manufacturing Application. *Proceedings of AAAI Technical Report.*

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Muller, K.-R. (2019). Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. *Lecture Notes in Computer Science*, 11700.

Samek, W., Weigand, T., & Muller, K.-R. (2017). Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *arXiv preprint arXiv:1708.08296*.

Sammut, C., & Webb, G. I. (2010). Mean Absolute error. *Encyclopedia of Machine Learning, 652*.

Samuel, A. L. (2000). Some studeis in machine learning using the game of checkers. *IBM Journal of Research and Development, 44*(1.2), 206-226.

Sanger, T. D. (1989). Optimal unsupervised learing in a single-layer linear feed-forward neural network. *Neural networks, 2*(6), 459-473.

Santos, I., Laorden, C., Xabier, U.-P., Sanz, B., & Bringas, P. G. (2012). Spam Filtering through Anomaly Detection. *Communications in Computer and Information Intl. Conf. on E-Business and Telecommunication. 314*, pp. 203-216. Springer.

Saporta, A., Gui, X., Agrawal, A., Pareek, A., Truong, Q., Nguyen, C. D., . . . Rajpukar, P. (2021). Benchmarking saliency methods for chest X-ray interpretation. *medRxiv preprint*.

Sarunas, G., & Jevgenijus, T. (2020). Habits attribution and digital evidence object modes based tool for cybercrime investigation. *Baltic Journal of Modern Computing, 8*(2), 275-292.

Sato, M., & Tsuikimoto, H. (2001). Rule Extraction from Neural Networks via Decision Tree Induction. *Intl. Conf. on Neural Networks.*

Scarcelli, F., Gori, M., Tsoi, A. C., Hegenbuchner, M., & Monfardini, G. (2008). The Graph Neural Network Model. *IEEE Transactions on Neural Network, 20*(1), 61-80.

Schatz, B. (2007). *Digital Evidence: Representation and Assurance (Ph.D. Thesis).* Queensland University of Technology.

Schlobohm, D. A., & Waterman, D. A. (1987). Explanation for an expert system that performs estate planning. *Proceedings of the 1st Intl. Conf. on AI and Law*, (pp. 18-27).

Scientific Working Group on Digital Evidence (SWGDE). (2018). *Establishing Confidence in Digital and Multimedia Evidence Forensics Results by Error Mitigation Analysis (Version 2.0).* Retrieved from https://drive.google.com/file/d/1pK_6eveU8Wb9TC9DvVpw1XNKPndwokKk/view

Seeger, M. (2004). Gaussian processes for machine learning. *International Neural Systems, 14*(2), 69-106.

Selvaruju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from Deep Networks via Gradient-Based Localization. *IEEE Intl. Conf. on Computer Vision*, (pp. 618-626).

Senator, T. E., Goldberg, H. G., Wooton, J., Cottini, M. A., Khan, U. A., Klinger, C. D., . . . Wong, R. (1995). The FinCEN Artificial Intelligence System: Identifying Potential Money Laundering from Reports of Large Cash Transactions. *Proceedings of AAAI.*

Sencar, H. T., & Memon, N. (Eds.). (2012). *Digital Image Forensics: There ia More to a Picture Than Meets the Eye.* Springer, New York.

Seo, Y., Defferard, M., Vandergheynst, P., & Bresson, X. (2018). Structured Sequence Modelling with Graph Convolutional Recurrent Networks. *Intl. Conf. on Neural Information Processing* (pp. 362-373). Springer, Cham.

Shai, S.-S., & Shai, B.-D. (2014). *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press.

Shanks, D. R. (2006). Complex choices better made unconsciously? *Science, 313*(5788), 760-1.

Shao, S., Tunc, C., Al-Shawi, A., & Hariri, S. (2019). Automated Twitter Author Clustering with Unsupervised Learning for Social Media Forensics. *16th Intl. Conf. on Computer Systems and Applications (AICCSA)*, (pp. 1-8).

Shapiro, J. (2001). Genetic Algorithms in Machine Learning. In Karkaletsis, & C. D. Spyropoulos (Eds.), *ACAI, LNAI* (Vol. 2049, pp. 146-168). Springer-Verlag berlin.

Sherman, D. M. (1989). Expert systems and ICAI in tax law: Killing two birds with one AI stone. *In Proceedings of the 2nd international conference on Artificial intelligence and law*, (pp. 74-80)

Shi, Y., & Eberhart, R. C. (1998). Parameter Selection in Particle Swarm Optimization. *Evolutionary Programming VII*, 591-600.

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning Important Features Through Propagating Activation Differences. *Proceedings of the 34th Intl. Conf. on Machine Learning*, *70*, pp. 3145-3153.

Shukla, S., Misra, M., & Varshney, G. (2020). Identification of Spoofed Emails by Applying Email Forensics and Memorty Forensics. *Proceedings of the 10th Intl. Conf. on Communication and Network Security* (pp. 109-111). ACM Digital Library.

Silberschartz, A., & Tuzhilin, A. (1995). On Subjective Measures of Interestingness in Knowledge Discovery. *Proceedings of the First Intl. COnf. on Knowledge Discovery and Data Minin (KDD '95)*, (pp. 271-285).

Siljak, D. D. (2008). Dynamic Graphs. *Nonlinear Analysis: Hybrid Systems, 2*(2), 544-567.

Singh, B., & Sharma, D. K. (2021). Image Forgery Over Social Media Platforms - A Deep Learning Approach for its Detection and Localization. *8th Intl. Conf. on Computing for Sustainable Global Developemnt (INDIACom)* (pp. 705-709). IEEE.

Singh, M., Singh, M., & Kaur, S. (2019). Detect Bot-Infected Machines using DNS Fingerprint. *Digital Investigation, 28*, 14-33.

Smit, N. M., Morgan, R. M., & Lagnado, D. A. (2018). A Systematic Analysis of Misleading Evidence in Unsafe Rulings in England and Wales. *Science & Justice, 8*(2), 128-137.

Snoek, J., Larochlle, H., & Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. *Neural Information Processing Systems*, 2951-2959.

Solanke, A. A., Chen, X., & Ramírez-Cruz, Y. (2021). Pattern Recognitiona and Reconstruction: Detecting Malicious Deletions in Textual Communications. *IEEE Intl. Conf. on Big Data (Big Data)*, (pp. 2574-2585).

Soltani, S., & Seno, S. A. (2017). A Survey of Digital Evidence Collection and Analysis. *7th Intl. Conf. on Computer and Knowledge Engineering (ICCKE)*, (pp. 247-253).

Sommer, P. (2018). Accrediting Digital Forensics: What are the Choices? *Digital Investigation, 25*, 116-120.

Spafford, E. H., & Weeber, S. A. (1993). Software Forensics: Can we track code to its author? *Computers & Security, 12*(6), 585-595.

Spencer, J. R. (1992). Court Experts and Expert Witnesses: Have we a Lesson to Learn from the French? *Current Legal Problems, 45*(2), 213-236.

Spencer, J. R. (2010). The green paper on obtaining evidence from one Member State to another and securing its admissibility: the reaction of one British Lawyer. *Zeitschrift für die internationale Strafrechtsdogmatik, 9*, 602-606.

International Organization for Standardization, Information Technology - Vocabulary - Part 37 (2017). *Biometrics (ISO/IEC 2382-37-2017(E)).* ISO. Retrieved from https://standards.iso.org/ittf/PubliclyAvailableStandards/

Steinholtz, O. S. (2018). *A Comparative Study of Black-box Optimization Algorithms for Tuning Hyperparameters in Deep Neural Networks.* M.S. Thesis, Dept. Elect. Eng., Lulea Univ. Technology.

Steve, M. (2021). *Special Report: Cyberwarfare in the C-Suite.* Retrieved 2020, from Cybersecurity Ventures: https://cybersecurityventures.com/wp-content/uploads/2021/01/Cyberwarfare-2021-Report.pdf

Studiawan, H., Sohel, F., & Payne, C. (2020). Sentiment Analysis in a Forensic Timeline with Deep Learning. *IEEE Access, 8*, 60664-60675.

Su, G., Wei, D., Kush, R., & Malioutov, D. M. (2016). Interpretable Two-level Boolean Rule Learning for Classification. *arXiv preprint arXiv:1606.05798.*

Sublime, J., & Kalinicheva, E. (2019). Automatic Post-Disaster Damage Mapping Using Deep-Learning Techniques for Change Setection: Case Study of the Tohoku Tsunami. *Remote Sensing, 11*(9), 1123.

Sun, S., Cao, Z., Zhu, H., & Zhao, J. (2020). A Survey of Optimization Methods from Machine Learning Perspective. *IEEE Transactions on Cybernetics, 50*(8), 3668-3681.

Sunde, N., & Dror, I. E. (2019). Cognitive and Human Factors in Digital Forensics: Problems, Challenges, and the Way Forward. *Digital Investigation, 29*, 101-108.

Suri, P., & Roy, N. R. (2017). COmparison between LDA & MNF for event-detection from large text stream data. *3rd International Conference on Computational Intelligence and Communication Technology*, (pp. 1-5).

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing Propoerties of Neural Networks. *arXiv preprint arXiv:1312.6199*.

Tamas, A. (2006). Event Sequence Mining to Develop Profiles for Computer Forensic Investigation Purpose. *Australian Workshops on Grid Computing and e-Research*, *54*, pp. 145-153.

Tan, A.-H. (1999). Text Mining: The State of the Art and the Challenges. *Proceedings of the Pakdd 1999 Workshop on Knowledge Discovery from Advanced Databases*, *8*, pp. 65-70.

Tang, Y., Zhang, Y.-Q., Chawla, N. V., & Krasser, S. (2008). SVMs Modeling for Highly Imbalanced Classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39*(1), 281-288.

Taylor, M., & Quayle, E. (2003). *Child Pornography: An Internet Crime.* Brunner-Routledge.

The Consultative Committee for Space Data System (CCSDS). (2012). *Reference Model for an Open Archival Information System.* Magenta Book.

Thiagarajan, J. J., Kailkhura, B., Sattigeri, P., & Ramamurthy, K. N. (2016). TreeView: Peeking into Deep Neural Networks via Feature-Space Partitioning. *arXiv preprint arXiv:1611.07429*.

Thompson, K. (1958). Programming Techniques: Regular Expression Search Algorithm. *Communication of the ACM, 116*, 419-422.

Thompson, W. C. (2017). How should forensic scientists present source conclusions? *Seton Hall Law Review, 48*, 773.

Thornton, C., Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2013). Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms. *2013 ACM SIGKDD*, (pp. 847-855).

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63*(2), 411-423.

Tieleman, T., & Hinton, G. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning, 4*(2), 26-31.

Tim, J. M. (2017). *Models for Machine Learning*. Retrieved from IBM Developer: https://developer.ibm.com/articles/cc-models-machine-learning/#reinforcement-learning

Timmer, S. T., Meyer, J.-J. C., Prakken, H., Renooij, S., & Verheij, B. (2015). A structure-guided approach to capturing Bayesian reasoning about legal evidence in argumentation. *Proceedings of the 15th Intl. Conf. on AI and Law (ICAIL)*, (pp. 109-118).

Toro-Vizcarrondo, C., & Wallace, T. D. (1968). A test of the mean square error criterion for restrictions in linear regression. *Journal of the American Statistical Association, 63*(322), 558-572.

Trivedi, R., Farajtabar, M., Biswal, P., & Zha, H. (2018). Representation Learning Over Dynamic Graphs. *arXiv preprint arXiv:1803.04051*.

Tsai, M.-J., Lai, C.-L., & Liu, J. (2007). Camera/Mobile Phone Source Identification for Digital Forensics. *Intl. Conf. on Acoustic, Speech and Signal Processing.* IEEE.

Tsoumakas, G., & Katakis, I. (2007). Multi-label Classification: An Overview. *International Journal of Data Warehousing and Mining (IJDWM), 3*(3), 1-13.

Turin, A. M. (1950). Computing Machinery and Intelligence. *Mind, 59*(236), 433. Retrieved from https://www.csee.umbc.edu/courses/471/papers/turing.pdf

Turnbull, B., & Randhawa, S. (2015). Automated Event and Social Network Extraction from Digital Evidence Sources with Ontological Mapping. *Digital Investigation, 13*, 94-106.

Turvey, B. E. (2012). An Introduction to Behavioural Evidence Analysis. In *Criminal Profiling: An Introduction to Behavioural Evidence Analysis* (4th ed., pp. 121-140).

U.S. Department of Homeland Security. (2015). *Best Practices for Seizing Electronic Evidence: A Pocket Guide for First Responders V.4.2.* Retrieved from https://www.crime-scene-investigator.net/PDF/best-practices-for-seizing-electronic-evidence-v4.pdf

Uma, M., & Nikkath, B. S. (2021). Machine Learning Forensics to Guage the Likelihood of Fraud in Emails. *Intl. Conf. on COnference on Communication and Electronics System (ICCE)* (pp. 1567-1572). IEEE.

Underwood, G., Foulsman, T., van Loon, E., Humphreys, L., & Bloyce, J. (2006). Eye movements during scence inspection: A test of the saliency map hypothesis. *European Journal of Cognitive Psychology, 18*(3), 321-342.

van Barr, R. B., van Beek, H. M., & van Ejik, E. J. (2014). Digital Forensics as a Service: A Game Changer. *Digital Investigation, 11*(1), s54-62.

van Beek, H. M., van den, B. J., Boztas, A., van Eijk, J., Schramp, R., & Ugen, M. (2020). Digital forensics as a service: Stepping up the game. *Forensic Science International: Digital Investigation, 35*, 301021.

van Beek, H. M., Van Eijk, E. J., Van Baar, R. B., Ugen, M., Bodde, J. N., & Siemelink, A. J. (2015). Digital Forensic as a Service: game On. *Digital Investigation, 15*, 20-38.

Varun, C., Arindam, B., & Vipin, K. (2009). Anomaly Detection: A Survey. *ACM Computing Survey (CSUR), 41*(3), 1-58.

Venguerov, M., & Cunningham, P. (1998). Generalised Syntactic Pattern Recognition as a Unifying Approach in Image Analysis. *LNCS, 1451*, 913-920.

Vens, C., Struyf, J., Schietgat, L., Dzeroski, S., & Blockeel, H. (2008). Decision trees for hierarchical multi-label classification. *Machine Learning, 73*(2), 185.

Vieira, S., Pinaya, W. H., & Mechelli, A. (2017). Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and application. *Neuroscience & Biobehavioural Reviews, 74*(Part A), 58-70.

Vincent, P., & Larochelle, H. (2010). Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *Journal of Machine Learning Research, 11*, 3371-3408.

Vlek, C., Prakken, H., Renooij, S., & Verheij, B. (2016). A Method for Explaining Bayesian Networks for Legal Evidence with Scenarios. *Artificial Intelligence Law, 24*, 285-324.

Wachter, S., Mittelstadt, B., & Rusesell, C. (2018). Counterfactual Explanation without Opening the Black Box: Automated Decision and the GDPR. *Havard Journal of Law & Technology, 31*(2), 841.

Walker, I. R. (2011). *Reliability in scientific research: improving the dependability of measurements, calculations, equipment, and software.* Cambridge University Press.

Wang, G., Chen, H., & Atabakhsh, H. (2004). Automatically Detecting Deceptive Criminal Identities. *Communication of the ACM, 47*(3), 70-76.

Wang, S. C. (2003). Artificial Neural Network. *Interdisciplinary Computing in Java Programming* (pp. 81-100). Springer, Boston.

Wang, T., Rudin, C., Wagnner, D., & Rich, S. (2013). Learning to Detect Patterns of Crime. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 515-530). Springer, Berlin.

Watanabe, S. (1985). *Pattern Recognition: Human and Mechanical.* John Wiley.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law, 7*(2), 76-99.

Waweru, A. (2021). *Understanding Pattern Recognition in Machine Learning*. Retrieved from https://www.section.io/engineering-education/understanding-pattern-recognition-in-machine-learning/

Webb, L., Craissati, J., & Keen, S. (2007). Characteristics of Internet Child Pornography Offenders: A Comparison with Child Molesters. *SAGE Journals of Sexual Abuse, 19*(4), 449-465.

Welch, T. (1997). Computer Crime Investigation and Computer Forensics. *Information Systems Security, 6*(2), 56-80.

Welling, M., & Kingma, D. P. (2019). An Introduction to Variational Autoencoders. *Foundations and Trends in Machine Learning, 12*(4), 307-392.

Wenzl, R. (2014). *The Wichita Eagle*. Retrieved from https://www.kansas.com/news/special-reports/btk/article2188657.html

Westerlund, M. (2019). The emergence of Deepfake Technology: A Review. *Technology Innovation Management Review*.

Westin, A. F. (1952). The Wire-Tapping Problem: An Analysis and a Legislative Proposal. *Columbia Law Review , 52*(2), 165-208.

White, P. (2010). In *Crime Scene to Court: The Essentials of Forensic Science.* Royal Society of Chemistry.

Whitley, D. (1994). A genetic algorithm tutorial. *Statistics and Computing, 4*(2), 65-85.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research, 30*(1), 79-82.

Willmott, C. J., & Matsuura, K. R. (2009). Ambiguities inherent in sums-of-Squares-based error statistics. *Atmospheric Environment, 43*(3), 749-752.

Wolfson, A. (2005). Electronic Fingerprints: doing away with the conception of computer-generated records as hearsay. *Michigan Law Review, 104*(1).

Wong, C.-I., Wong, K.-Y., Ng, K.-W., Fan, W., & Yeung, K.-H. (2014). Design of a Crawler for Online Social Networks Analysis. *WSEAS Transactions on Communications, 13*, 263-274.

Wright, S. (1921). *Correlation and Causation.*

Wu, T.-F., Lin, C.-J., & Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research, 5*, 975-1005.

Xia, F., Tang, L., Wang, L., & Vinel, A. (2012). Internet of things. *International Journal of Communication Systems, 25*(9), 1101-1102.

Xiaoyu, D., Hargreaves, C., Sheppard, J., Anda, F., Sayakkara, A., Le-Khac, N.-A., & Scalon, M. (2020). Sok: Exploring the State of the Art and the Future Potential of Artificial Intelligence in Digital Forensics Investigation. *Intl. Conf. on Availability, Reliability and Security. 46*, pp. 1-10. ACM Digital Library.

Xu, K., Ba, J. L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., . . . Bengio, Y. (2015). Show, attend and tell: neural image caption generation with visual attention. *Proceedings of ICML*, *37*, pp. 2048-2057.

Xu, W., Liu, X., & Gong, Y. (2003). Document Clustering Based on Non-negative Matrix Factorization. *Procceding of the 26th Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, (pp. 267-273).

Yampolski, R. V. (2020). Unexplainability and Incomprehensibility of AI. *Journal of AI & Consciousness, 7*(2), 277-291.

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing, 415*, 295-316.

Yannikos, Y., Graner, L., Steinebach, M., Winter, C. (2014). Data corpora for digital forensics education and research. *In Intl. Conf. on Digital Forensics,* (pp. 309-325)

Yeh, C.-K., Kim, J. S., Yen, I. E., & Ravikumar, P. (2018). Representer point selection for explaining deep neural networks. *Advances on Neural Information Processing Systems*, 9311-9321.

Yong, C., Hui, Z., Rui, L., Zhiwen, Y., & Lin, J. (2019). Experimental Explorations on Short Text Topic Mining Between LDA and NMF Based Schemes. *Knowledge-Based Systems, 163*, 1-13.

Yu, T., & Zhu, H. (2020). Hyper-Parameter Optimization: A Review of Algorithm and Application. *arXiv preprint arxiv:2003.05689*.

Yunyao, L., Rajasekar, K., Sriram, R., & Shivakumar, V. (2008). Regular Expression Learning for Information Extraction. *Conference on Empirical Methods in Natural Language Processing* (pp. 21-30). ACL Anthology.

Zeiler, M. D., & Fergus, R. (2013). Visualizing and Understanding Convolution Networks. *arXiv preprint arXiv:1311.2901*.

Zeiler, M. D., Krishnan, D., Taylor, G. W., & Fergus, R. (2010). Deconvolution networks. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *10*, p. 7.

Zeiler, M. D., Taylor, G. W., & Fergus, R. (2011). Adaptive deconvolutional networks for mid and high level feature learning. *Intl. Conf. on Computer Vision*, *1*, p. 6.

Zeleznikow, J., & Stranieri, A. (1995). The split-up system: integrating neural networks and rule-based reasoning in the legal domain. *Proceedings of the 5th ICAIL*, (pp. 185-194).

Zell, A. (1994). *Simulation Neuronaler Netze [Simulation of Neural Networks]* (1st ed.). Addison-Wesley.

Zhan, Y., Chen, Y., Zhang, Q., & Kang, X. (2017). Image forensics based on transfer learning and convolutional neural network. *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*, (pp. 165-170).

Zhang, E., & Yi, Z. (2009). Average Precision. *Encyclopedia of Database Systems*.

Zhang, S., Xu, J., Huang, E., & Chen, C.-H. (2016). A new optimal sampling rule for multi-fidelity optimization via ordinal transformation. *IEEE Intl. Conf. on Automation Science and Engineering*, (pp. 670-674).

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2015). Learning Deep Feature for Discriminative Localization. *arXiv preprint arXiv:1512.04150*.

Zia, T., Liu, P., & Han, W. (2017). Application-Specific Digital Forensics Investigative Model in Internet of Things (IoT). *12th Intl. Conf. on Availability, Reliability and Security*, (pp. 1-7).

Zilke, J. R., Mencía, E. L., & Janssen, F. (2016). DeepRED - Rule Extraction from Deep Neural Networks. *Intl. Conf. on Neural Networks* (pp. 457-473). Springer.

Zou, Q., Xie, S., Lin, Z., Wu, M., & Ju, Y. (2016). FInding the best classification threshold in imbalanced classification. *Big Data Research, 5*, 2-8.