Fool me once: A systematic review of techniques to authenticate digital artefacts

Forensic Science International: Digital Investigation 45 (2023) 301516

Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

journal homepage: www.elsevier.com/locate/fsidi

Fool me once: A systematic review of techniques to authenticate digital artefacts

Christopher Neale

The Open University, Milton Keynes, MK7 6AA, United Kingdom

ARTICLE INFO

Article history:
Received 19 November 2022
Received in revised form
1 February 2023
Accepted 2 February 2023
Available online xxx

Keywords:
Anti-forensics
Authentication
Digital artefacts
Reliability
Tampering

ABSTRACT

When conducting digital forensic investigations, practitioners are concerned with understanding whether the digital artefacts they encounter are authentic and have not been the subject of tampering activity. This is one factor of investigations which could potentially impact of the reliability of any subsequent findings. Some research into this problem has already been undertaken, however there is currently very little understanding of how effective current technique are. In this paper, a Systematic Review (SR) of existing literature will be undertaken to identify the techniques that currently exist to authenticate digital artefacts. Furthermore, consideration will be given to understanding whether existing techniques are effective in solving the problem of digital artefact authentication and whether they are accessible by the practitioner community. The results of the SR will show that while research effort has been devoted to this problem, there are relatively few techniques which can be generally applied. Additionally, very little effort has been devoted to understanding the effectiveness of these techniques. Furthermore, the lack of standardised datasets for evaluation makes comparison between techniques impossible and none of the identified papers provided publicly available implementations. The shortcomings identified in this SR show that further research effort in this area could benefit the community in its aim to produce more reliable findings in forensic investigations.

1. Introduction

The prevalence of digital devices in society means that there is an inevitability that they will be present at many crime scenes. Therefore the data held on them is a crucial aspect of many investigations (Horsman, 2021). However, some such as Stoykova (2021) take a pessimistic view on how digital evidence is increasingly presented and accepted in courts. Specifically some are of the view that scientific validation of the digital forensics methodology or tools, and many digital investigations still lack quality assurance and accountability (Stoykova, 2021). As an example, in the UK legal system, there exists a presumption that digital evidence is reliable, unless proven otherwise. This has been the default legal position since the repeal of section 69(1) (b) of the Police and Criminal Evidence Act (PACE, 1984) with s.60 of the Youth and Criminal Evidence Act (1999). According to the UK Crown Prosecution Service, the impact is that "*Computer evidence must now follow the Common Law rule.*" This rule states "*that a presumption will exist that the computer producing the evidential record was working properly […]the record is therefore admissible as real evidence*" (Crown Prosecution Service, 2017). Whilst this clearly makes the presentation of digital evidence easier, there must also be a consideration given to the reliability of this evidence.

In this context, many factors can influence reliability, with Neale et al. (2022) proposing that a Zero Trust Digital Forensics approach could in theory be applied to investigations as a way of increasing reliability. The approach was defined in this paper as 'a strategy adopted by investigators whereby each aspect of an investigation is considered to be unreliable until verified' but consideration of how it could be applied in practice was left for future research. Therefore, this paper will continue the work and focus on one aspect that was highlighted in Neale et al. (2022), namely the authenticity of digital artefacts encountered during investigations. A cursory study of existing literature shows that there has been some existing work to study how the problem of artefact authentication can be achieved. For example Shanmugam et al. (2011) use attack trees to identify anti-forensic activity. Alternatively, Horsman and Errickson (2019) propose the creation of a database to record digital tool marks (DTMs) of specific applications which are used to tamper

E-mail address: cjn.correspondence@protonmail.com.

https://doi.org/10.1016/j.fsidi.2023.301516

with digital data. However, several such techniques appear to be presented as proof-of-concept ideas, rather than as working implementations which could be used in a realistic practical context. Understanding whether such existing research effectively answers questions regarding the existing or otherwise of tampering activity is also non-trivial. Therefore, this work will focus on understanding the current state of research into verifying the authenticity of digital artefacts through the use of a Systematic Review (SR). This is a technique especially suited to summarising existing research and identifying any gaps which are suited for further research (Kitchenham, 2004). By way of contribution, this paper will identify such existing techniques and analyse what types of artefacts these techniques can be applied to. Additionally, analysis of the different ways in which the effectiveness of such techniques can be understood provides insight into whether this body of work is accessible to the practitioner community.

The rest of this paper is structured as follows. In section 2, the specific research question used and the methodology undertaken will be discussed before the details of the review conducted are given in section 3. Section 4 presents the results of the review and these are then discussed in section 5. The limitations of this work are outlined in section 6 and recommendations made in section 7 before section 8 concludes the paper.

2. Research question and methodology

The research question used as a basis for this work is given as.

RQ: To what extent do current techniques identify digital artefact tampering?

For the sake of clarity, some of the terms used in the research question are discussed here. First, we use the definition of 'digital artefact' as described by Horsman and Errickson (2019b). This is stated as "*a digital object containing data which may describe the past, present or future use or function of a piece of software, application or device for which it is attributable to*." Examples include files containing content (such as pictures, documents log files etc.) as well as metadata, such as timestamps and author information. The digital forensics literature uses a variety of terms to describe this concept, such as 'digital evidence' (e.g. Bednar-Schadle, 2018; Casey, 2019; Kirmani and Banday, 2020), 'digital traces' (e.g. Overill and Silomon, 2010; Reddy et al., 2020) or 'digital data' (e.g. Horsman, 2020; Iqbal and Alharbi, 2019) as well as 'digital artefacts' (e.g. Carrier, 2006; Johnson and Davies, 2020; Lees, 2013; Horsman and Errickson, 2019a). Generally, each term broadly describes the same idea, however it is noted that 'digital evidence' could potentially be seen as problematic. This is due to the fact that the term evidence suggests that it is actively being used to support or refute some hypothesis (as explained in Stoykova, 2021). This is not always be the case, given that it can exist before an investigator has had time to formulate any such hypothesis.

We also use the term 'authentication' throughout as a synonym for process of identifying tampering activity. We take the definition of authentication from the Scientific Working Group on Digital Evidence (SWGDE) as being '*The process of substantiating that the data is an accurate representation of what is purports to be*' (SWGDE 2016).

In terms of research methodology, a Systematic Review (SR) was chosen as a suitable method to answer the stated research question as it is a useful method for summarising existing evidence concerning a technology (Kitchenham, 2004). Additionally, SRs are effective at identifying gaps in current research which can be used to provide a framework or background in order to appropriately position new research activities (Kitchenham, 2004). This particular review was primarily conducted by a single researcher between October 2021 and March 2022 in the context of PhD study and is therefore more indicative of a rapid review, or scoping review, both of which are cited in the above as valid approaches to conducting such research.

In order to answer the research question, the following sub-questions were devised.

RQ-1.1: What techniques exist to identify digital artefact tampering?
RQ-1.2: What types of artefacts can be authenticated using these techniques?
RQ-1.3: Which of these techniques are effective in identifying tampering?
RQ-1.4: Are these techniques available for digital forensic practitioners?

RQ-1.1 is concerned with finding relevant peer-reviewed work to identify an appropriate corpus of research work on the identification of artefact tampering. RQ-1.2 is concerned with understanding the range of artefact types that have previously been studied in order to identify potential gaps. RQ-1.3 aims to study whether the techniques from RQ-1.1 have been subjected to evaluation procedures in order to determine effectiveness, both for individual techniques and as a comparator between techniques.

It could be argued that RQ-1.4 goes beyond the bounds of the primary research question by attempting to provide insight into the availability of the techniques from RQ-1.1 to the forensic community. However, we argue that this plays a role in the extent to which current techniques can be applied to solve the problem.

3. Conducting the review

The process used in undertaking the SR was based on the recommendations made in Kitchenham (2004), including the updated recommendations provided in the tertiary review Kitchenham and Brereton (2013). A visual representation is given in Fig. 1.

In step 1., searches were conducted across a number of relevant databases, including SCOPUS, Google Scholar, Science Direct and IEEE Xplore to identify examples of research relevant to the question. This was done to pilot the search strategy, including identification of a suitable search string and database sources, and suitable inclusion and exclusion criteria and data extraction criteria. This information was then used to finalise the research protocol described in this section and which was agreed with the research supervisors (step 2.).

In step 3., the search string constructed from the research pilot was used to identify an initial dataset by searching a single database, SCOPUS. Previous SRs conducted within digital forensics, such as Coronel et al. (2018) and Edward and Ojeniyi (2019) had used multiple databases to identify an initial dataset of papers. However, the pilot research conducted in step 1 found that the many of the identified papers of likely relevance to the research question were present within SCOPUS search results. These same results were only occasionally present in the results of the other databases: Science Direct, IEEE Xplore, SpringerLink and ACM Digital Library. Furthermore, with limited research capacity and time constraints on the work, the use of a single database eliminates the need to manually remove duplicates of results (Kitchenham, 2004).

The search string was constructed through experimentation in step 1. It was developed to accommodate for variations in terminology cross the digital forensic literature, where terms such as 'digital forensics' and 'cyber forensics' can sometimes be used interchangeably. Many of these synonyms were included through

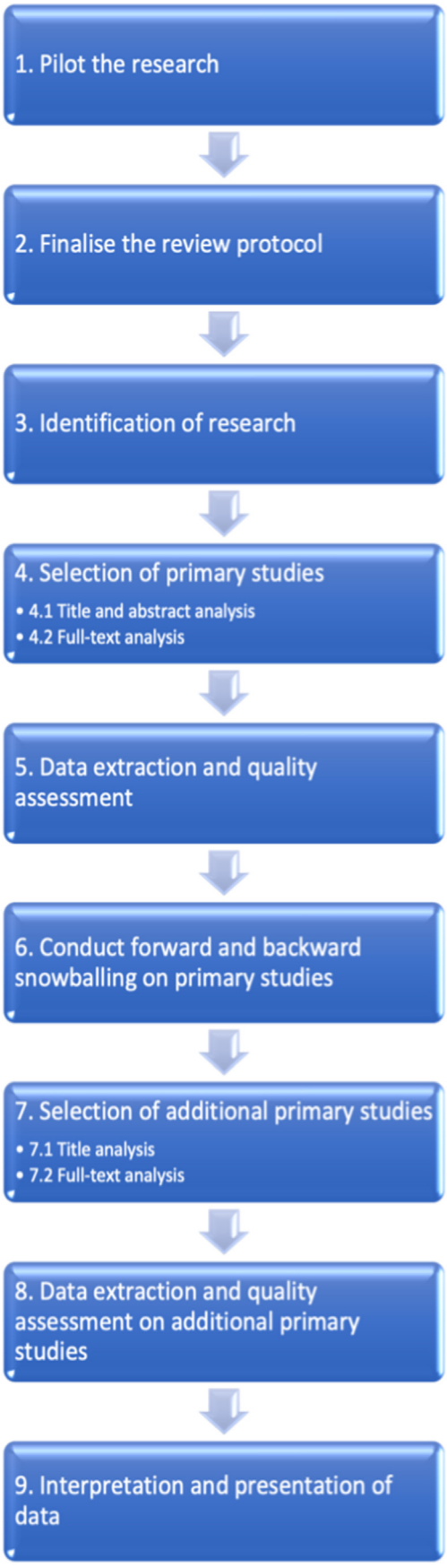

Fig. 1. Systematic Review process undertaken.

the use of Boolean 'OR' operators in the search string in order to gain the broadest dataset possible. The final search string that was used was:

> ((“digital forensics” OR “cyber forensics” OR “computer forensics”) AND (“evidence” OR “artefacts” OR “traces”)) AND (“tampering” OR “manipulation” OR “anti forensics” OR “counter forensics” OR “forgery” OR “verification” OR “validation”)

Once an initial dataset was created in step 3, step 4 involved evaluating each result against the inclusion and exclusion criteria which are more fully described later in this section. The title of the paper, as well as the text abstract were initially used in step 4.1 to conduct this evaluation, and those papers selected were then retrieved in full in order to conduct a full-text analysis in step 4.2. Data extraction and quality assessment was then performed on this smaller subset of papers in step 5. in order to ascertain properties as outlined in Table 1. This step was moderated by an independent researcher.

Step 6. followed the recommendations set out in Kitchenham and Brereton (2013) to improve the results generated by SRs by conducting forward and backward snowballing on the primary studies. This is used to identify further relevant work of interest. For each primary study selected in step 4., the provided reference list was added as an additional dataset (backward snowballing). Additionally, Google Scholar was used to identify which papers cited those in the primary dataset (forward snowballing) and these were also added to this additional dataset. Papers included here were then subject to the same selection criteria (step 7.). and papers which were selected were then subjected to the same data extraction methods and moderation (step 8.). Finally, in step 9., the results from the data extraction of both datasets were interpreted and presented in this document.

The inclusion and exclusion criteria that were used in the SR are provided in Table 2.

4. Results

The search string was used to query the SCOPUS database in December 2021, returning 1257 results. These were subjected to screening in step 4.1 as previously described and this returned 72 results. These papers were then subjected to full-text analysis in step 4.2 and after this was completed, 23 were selected to form the initial dataset. 49 papers were therefore excluded at this stage and the reasons given, along with frequency are shown in Table 3.

Data extraction was performed on the 23 selected papers (step 5) before forward and backward snowballing was performed on the papers, identifying 12 papers suitable for further screening (step 7.1). 8 of these papers were excluded during the second screening (step 7.2) with reasons recorded in Table 4. Therefore 4 papers were added to the initial dataset to create a final dataset of 27 papers.

Fig. 2 shows the process undertaken as a PRISMA diagram generated using the online tool provided by Haddaway et al. (2022)

4.1. Moderation of results

In steps 5 and 8, moderation of a sample of the papers was provided by an independent researcher. In total, 5 of the 27 papers were selected at random for moderation (19%). For each of these, the independent researcher was asked to conduct the same data extraction and quality assessment as set out previously in Table 1. One this had been conducted, the primary and independent researchers then discussed the results to determine agreement. Fleiss kappa (Fleiss et al., 1979) was used as a measure of agreement

Table 1
Data extraction conducted in the SR.

Data Extracted	Sub-question data relates to
Title	RQ-1.1
Authors	RQ-1.1
Year	RQ-1.1
Location of publication	RQ-1.1
Short Description	RQ-1.1
Quality (as assessed against criteria set out by Kitchenham (2004))	RQ-1.3
Artefacts technique applies to	RQ-1.2
Category of technique	RQ-1.1
Whether the technique has been evaluated	RQ-1.3
Stated measures of effectiveness claimed by the study	RQ-1.3
The dataset used in any evaluation work	RQ-1.3
Whether the technique is available for practitioners (with at least a 'proof-of-concept' implementation)	RQ-1.4

Table 2
Inclusion and Exclusion criteria used in the SR.

Inclusion Criteria		
Identifier	Description	Justification
I-1	Papers presenting a technique or methodology to detect whether digital artefacts have been tampered with	These are the papers of most relevance to the overall research question considered by the SR
I-2	Papers whose primary focus is to present a tampering technique to be applied to digital artefacts, where there is enough detail included to describe how to detect such activity	These papers could provide some insight in helping to answer the various SQ's considered in the review as they describe the methodology to detect tampering activity despite it not being a primary focus
Exclusion Criteria		
Identifier	**Description**	**Justification**
E−1	Papers primarily focussed on multimedia forensics	The pilot review conducted in step 1. identified a large amount of research into multimedia forensics (e.g. the tampering of image files/video files etc.). It is therefore unrealistic to include these papers due to the time and resource constraints
E−2	Papers where the primary application of the technique is a 'small' number of artefacts	The pilot review conducted in step 1. identified several papers which were concerned with tampering activity against specific artefact types. However, the number of these was vast and it was not possible to include each of them in the SR due to time and resource constraints. Additionally, it was unclear whether these techniques could be generalised in each case and therefore they were excluded from this review.
E−3	Papers not in the English language	English is the only language understood by the primary researcher and it was not possible to acquire translations of non-English papers within the required timeframe.
E−4	Papers which have not been subject to peer review	Peer-review represents the gold-standard of scientific research as it has been subject to more rigorous types of review than other types

Table 3
Reasons for excluded papers in step 4.2 of the review process.

Reason for exclusion	Frequency
Doesn't propose a new tampering detection technique (not I-1)	28
Doesn't propose a generalised tampering detection technique (E−2)	6
Duplicate of another result	3
Focus on prevention through design rather than detection (not I-1)	3
Paper content not retrievable (e.g. due to licencing and copyright restrictions)	9

Table 4
Reasons for excluded papers in step 7.2 of the review process.

Reason for exclusion	Frequency
Doesn't propose a new tampering detection technique (not I-1)	6
Doesn't propose a generalised tampering detection technique (E−2)	2

before and after this discussion. Table 1 is expanded below as Table 5 to show the level of agreement reached between researchers. For categories where factual data was extracted (e.g. Title/Authors etc.), a value of 'Not Applicable' (N/A) is given.

Using a standard interpretation of Fleiss kappa (Landis and Koch, 1977), substantial agreement or better was reached in each category after discussion. Initially, only fair agreement was reached in regard to the artefacts the techniques applied to due to the moderation form not being completely filled in by the independent researcher. However, after clarification on this specific result,

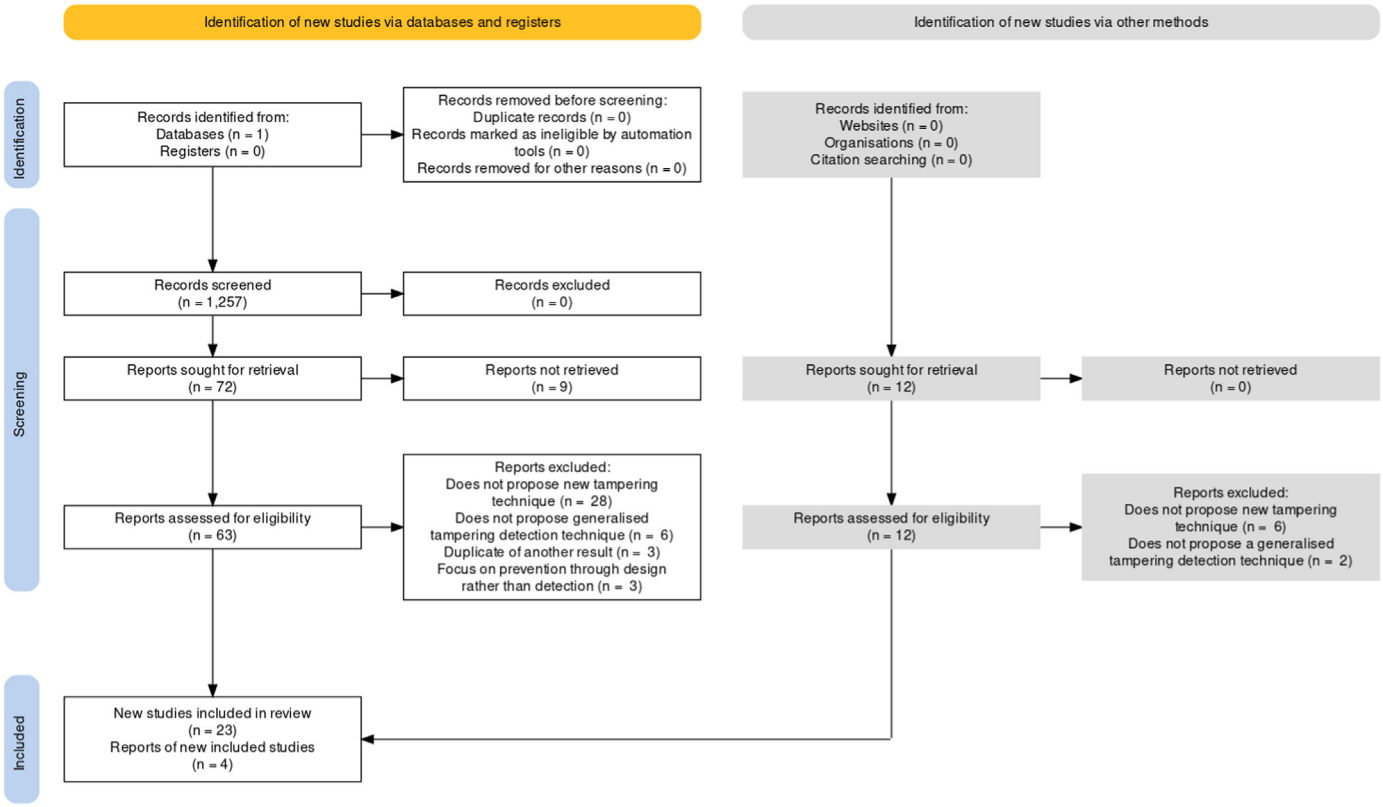

Fig. 2. PRISMA diagram of Systematic review process.

Table 5
Results of moderation and measures of agreement between researchers.

Data Extracted	Sub-question data relates to	Fleiss kappa score before moderation discussion	Fleiss kappa score after moderation discussion
Title	RQ-1.1	N/A	N/A
Authors	RQ-1.1	N/A	N/A
Year	RQ-1.1	N/A	N/A
Location of publication	RQ-1.1	N/A	N/A
Short Description	RQ-1.1	N/A	N/A
Quality (as assessed against criteria set out by Kitchenham (2004))	RQ-1.3	0.61	1
Artefacts technique applies to	RQ-1.2	0.32	0.61
Category of technique	RQ-1.1	1	1
Whether the technique has been evaluated	RQ-1.3	−0.43	0.61
Stated measures of effectiveness claimed by the study	RQ-1.3	N/A	N/A
The dataset used in any evaluation work	RQ-1.3	1	1
Whether the technique is available for practitioners (with at least a 'proof-of-concept' implementation)	RQ-1.4	1	1

substantial agreement was reached.

The greatest change in agreement can be found in an assessment of whether the technique had been evaluated. Initially, only 'poor agreement' was achieved. However during the moderation discussion, further clarification between the researchers on the criteria required to describe whether a paper had 'evaluated' a technique resulted in a significant increase.

The level of agreement reached during the moderation of the results from steps 5 and 8 further substantiate the findings presented in the remainder of this section.

The results presented in the remainder of this section are taken from the dataset after moderation had been completed and are provided in relation to the research sub-question they address.

4.2. RQ-1.1: what techniques exist to identify digital artefact tampering?

In total, 27 papers were identified which were primarily concerned with proposing techniques to verify the authenticity of digital artefacts. These are summarised in Table 6. The techniques have also had a category applied to provide a description of the type of approach used.

The categories represented in this initial dataset are shown in Table 7.

4.3. RQ-1.2: what types of artefacts can be authenticated using these techniques?

The types of digital artefacts that are the focus of the papers

Table 6
Papers concerned with presenting a technique to verify the authenticity of digital artefacts.

Paper	Short description	Category of technique
Nisioti et al. (2021)	Use of game theory methods to guide investigator actions when detecting attacks in a corporate network.	Game theory
Dija et al. (2020)	Manual inspection of specific RAM structures in order to indicate tampering via malware on Windows devices	Manual inspection
Shafiee Hasanabadi et al. (2020)	Game theory approach to model the best way for investigators to act in response to the presence of rootkits.	Game theory
Al-Sharif et al. (2020)	Technique to check for specific strings in the RAM dump of running Java programs to determine how they have been used when compared to assumed source code.	Feature extraction
Schneider et al. (2020)	Description of various manual detection techniques used by participants in an experimental setting to detect inconsistencies	Manual inspection
Mothi et al. (2020)	Evaluates each phase of a digital forensic investigation by modelling as vectors and using custom transformations to determine whether anti-forensic activity might be present	Model checking
Horsman and Errickson (2019a)	Uses 'Digital Tool Marks' to detect when tools have been used in an anti-forensic manner on a machine.	Signature-based detection
Pieterse et al. (2019b)	Use of modelling of potential tampering activities through the creation of 'attack trees' to which a model-checking algorithm is applied to determine whether such activity is present.	Model checking
Pieterse et al. (2019a)	Uses a classification model to evaluate whether data from smartphones is authentic by calculating an authenticity score based on several 'assessment points'	Model checking
Pieterse et al. (2019b)	Use of modelling of potential tampering activities through the creation of 'attack trees' to which a model-checking algorithm is applied to determine whether such activity is present.	Model checking
Freiling and Hösch (2018)	Description of various manual detection techniques used by participants in an experimental setting to detect inconsistencies	Manual inspection
Pieterse et al. (2018)	Use of a Smartphone Data Evaluation Model to assess whether digital evidence retrieved from a smartphone device is authentic based on whether it meets pre-determined 'requirements'	Model checking
Rani and Kumari (2017)	Presents a high-level process for identifying 'anti-forensics' in cloud environments	High-level process
Hoelz and Maues (2017)	Use of threat modelling to assess the likelihood and impact of 'anti-forensics' by evaluating the cost of risk mitigation and the selection of tools and techniques as countermeasures to it.	Threat modelling
Pieterse et al. (2017)	Uses a reference architecture to model components of a smartphone and then compares evidence to 'theories of normalities' in order to detect inconsistencies and deviations from these theories in order to evaluate the authenticity of the evidence	Model checking
Singhal et al. (2015)	Builds a network forensic model (an extension of a previous logic-based reasoning framework) to automate causality correlation of evidentiary data collected after a security event to construct evidence graphs for further analysis.	Model checking
Day (2014)	Manual inspection of 'obscurely formatted areas of media', which are either immune to anti-forensics or which are often overlooked and which could uncover inconsistencies to suggest tampering has taken place.	Manual inspection
Rekhis and Boudriga (2012b)	Model checking method using TLA + to model a system under attack and then uses reasoning logic to determine whether tampering has occurred	Model checking
Khan (2012)	Compares two classification approaches, Bayesian networks and Neural Networks, to classify whether file system activity is authentic or tampered.	Classification algorithm via a learning network model
Rekhis and Boudriga (2012a)	Model checking method using TLA + to model a system under attack and then uses reasoning logic to determine whether tampering has occurred	Model checking
Liu et al. (2012)	Uses attack graphs to model how anti-forensics might have been used when investigating an attack on a corporate network.	Model checking
Shanmugam et al. (2011)	Validates digital artefacts by modelling the system as a decision tree and then using an inductive reasoning system to examine the tree for anti-forensic activity	Model checking
Yusoff et al. (2010)	Examination of digital artefacts for properties relating to their reliability, taken from centuries old methods for validating Islamic hadith in order to generate a 'trust value' for each artefact	Manual inspection
Wojcik et al. (2006)	Describes how to build a 'trust model' of digital evidence in order to assess its authenticity.	Model checking
Geiger (2005)	Provides signatures to enable manual inspection of the use of specific anti-forensic tools	Manual inspection
Stallard and Levitt (2003)	Uses an expert system and decision tree to detect semantic incongruities which occur when 'invariant relationships' are violated	Model checking
Helman and Liepins (1993)	Uses a probability model to distinguish between normal and anomalous transactions on a computer system in order to detect misuse.	Model checking

Table 7
Categories of techniques for verifying the authenticity of digital artefacts.

Category	Frequency
Model checking	14
Game theory	2
Manual inspection	6
Feature extraction	1
Signature-based detection	1
High-level process	1
Threat modelling	1
Classification algorithm via a learning network model	1

Table 8
Artefacts covered by the papers in the final dataset.

Artefact	Number of papers
Any	13
Any on Windows Operating System	1
Any on iOS/Android Operating System	5
Corporate networks	1
Cloud systems	1
Disk images	1
RAM images	1
RAM images of Windows systems	1
RAM images of Java programs	1
Rootkits on any system	1
Artefacts created by a range of anti-forensic tools	1

included in the final dataset are stated in Table 8.

Note that there were several other studies where the primary focus was on a narrow range of artefacts, however these papers had been excluded based on the criterion E−2. Therefore, the table above covers the range of artefacts to which a generalised technique can apply. Techniques against specific types of artefacts are

out of scope for this review.

It is worth also noting that while there are 5 results for 'Any of iOS/Android operating system', all of these studies were conducted by the same authors (Pieterse et al.), with the results published between 2017 and 2019. As each is a distinct study, none have been excluded based on duplication, however there is considerable overlap between them, with the later studies representing refinements over their previously proposed techniques, rather than significantly new ideas.

4.4. RQ-1.3: which of these techniques are effective in identifying tampering?

Of the 27 studies included in the final dataset, 17 included some form of evaluation of the proposed technique. For these papers, an attempt was made to extract results relating to the effectiveness of the technique, such as accuracy, number of false positives etc. using the metrics used by the authors. In every case, the dataset used for evaluation was generated by the authors themselves for use in that specific study. The evaluation metrics used in these 17 papers are recorded in Table 9.

Finally, each paper was evaluated against the standard quality criteria used to assess primary studies in a systematic review (Kitchenham, 2004). These results are shown in Table 10.

4.5. RQ-1.4: are these techniques available for digital forensic practitioners?

Of the 27 primary studies, 6 were based on manual inspection techniques, meaning that the study provided the necessary information to allow a practitioner to repeat the various steps required in order to apply the technique. Of the 21 other studies, the studies were examined for information relating to whether the proposed technique had been made publicly available for use by the digital forensics community. Where this information could not be obtained within the study itself, additional internet searches were made to attempt to locate related software. These searches included the names of the authors of each paper and any technique names provided in the paper in the Google search engine. Additionally, contact was made with the corresponding email address provided each paper in order to seek such an implementation directly from the authors.

Of these 21 studies, no examples were found where the proposed technique had been made available to the wider community in any form. No papers provided a direct link to a publicly available implementation of any technique and no implementations were found through internet searches by either the primary researcher or the moderator. Furthermore, no responses were received from the emails sent to the authors corresponding addresses.

All of the data used in this research, including the original SCOPUS export of original results and the results of the data extraction processes in steps 5 and 8 can be found online at https://github.com/cn3759/ArtefactAuthenticationSystematicReview.

5. Discussion

A discussion of the results described in the previous section is now provided.

5.1. RQ-1.1: what techniques exist to identify digital artefact tampering?

With respect to the first sub-question, the techniques found during the SR have previously been outlined in Table 6 and for which substantive further discussion is not required. The technique approaches were also categorised, details of which were given in Table 7. These show that while several approaches have been suggested by digital forensics researchers, the main types are model checking and manual inspection. Techniques have been categorised as 'model checking' when a method of abstraction of a system under investigation is performed to create a model of its behaviour. For example by modelling it as an attack tree (Shanmugam et al., 2011). Typically, this model is subjected to some form of scrutiny to make assertions about the system it represents. For example, through applying some form of rules-based system in order to ascertain the authenticity of the system execution (Rekhis and Boudriga, 2012b). This approach forms 52% of the proposed techniques that were included in the final dataset of this SR. The only other significant approach taken by researchers was 'manual inspection'. These techniques involved scrutiny of specific artefacts in order to make a determination of authenticity in relation to detecting artefact tampering. Typically, the researchers were able to determine specific properties of an artefact which indicated the potential that tampering had occurred, either on the artefact itself or an additional one.

5.2. RQ-1.2: what types of artefacts can be authenticated using these techniques?

RQ-1.2 was concerned with understanding the types of artefacts that could be authenticated using the techniques in the final dataset. These were given in Table 8. The largest proportion were potentially applicable to any artefact on any system, however this still only applied to less than half the techniques in the final dataset. Therefore, the majority of techniques for identifying artefact tampering included in this SR do have some kind of caveat as to when they can be applied, even if they are general enough to avoid being excluded as per E−1 or E−2. This is a potential concern, as it means that there are fewer techniques available that can easily be applied to systems which have been subject to less research.

Table 9
Evaluation metrics used.

Evaluation metric	Papers used
Accuracy/false positives	Khan (2012)
Accumulated payoff	Nisioti et al. (2021)
Attack path	Liu et al. (2012)
Detection of specific traces	Dija et al. (2020); Pieterse et al. (2019a); Pieterse et al. (2019b); Pieterse et al. (2018); Pieterse et al. (2017); Shanmugam et al. (2011); Geiger (2005)
Game theory metrics (Stability/Applicability/Efficiency/Feasibility/Scalability/Individual Rationality/Robustness/Profitability/Iterations)	Shafiee Hasanabadi et al. (2020)
Sensitivity/false alarm	Helman and Liepins (1993)
Time taken	Schneider et al. (2020); Freiling and Hösch (2018); Stallard and Levitt (2003)
Time taken and accuracy of string comparison	Al-Sharif et al. (2020)
Trust values	Yusoff et al. (2010)

Table 10
Quality of primary studies based on evaluation against Kitchenham (2004) criteria.

Value	Description	Number of papers
1	Evidence from at least one properly-designed randomised controlled trial	0
2	Evidence obtained from well-designed pseudo-randomised controlled trials (i.e. non-random allocation to treatment)	0
3−1	Evidence obtained from comparative studies with concurrent controls and allocation not randomised, cohort studies, case−control studies or interrupted time series with a control group	0
3−2	Evidence obtained from comparative studies with historical control, two or more single arm studies, or interrupted time series without a parallel control group	0
4−1	Evidence obtained from a randomised experiment performed in an artificial setting	5
4−2	Evidence obtained from case series, either post-test or pre-test/post-test	0
4−3	Evidence obtained from a quasi-random experiment performed in an artificial setting	13
5	Evidence obtained from expert opinion based on theory or consensus	9

Furthermore, there were a large number of papers in the initial dataset that were excluded based on either of E−1 or E−2 exclusion criteria. This suggests that a much larger body of research has focused on identifying artefact tampering in specific scenarios or on specific artefact types. It could be that the aggregation of this works allows for the identification of artefact tampering across a wide enough range of artefact types as to render the generalised techniques studied here unnecessary in practice. Evaluating the level of truth in this is beyond the scope of this SR, however even if this were the case, there is the distinct disadvantage to practitioners in having to wait for peer-reviewed research to be published that applies to each specific investigative scenario encountered before being able to apply such techniques to cases.

5.3. RQ-1.3: which of these techniques are effective in identifying tampering?

That 17 of the 27 papers included some form of evaluation of the proposed technique shows that there is some understanding of the importance of this within the discipline. The most common metrics used were the time taken to complete the technique (4 papers) and whether specific traces were detected by the technique (7 papers). The use of metrics related to detecting specific traces suggest that these authors have been primarily motivated to understand whether the tampering activities under scrutiny have produced the hypothesised properties. However, this does not provide any information as to whether the proposed technique is effective in conducting this detection.

The use of temporal metrics in 4 papers suggests that these authors have been concerned with one aspect of the potential applicability of their approach, namely time taken. This however presents a narrow view of effectiveness and disregards any understanding of whether the techniques produce accurate and reliable results. In fact, only 2 of the papers included these types of measurements for effectiveness that might have been expected (precision of the technique, the number of false positives and false negatives experienced etc.). That only 2 of the 27 papers in the primary dataset (7.4%) considers these types of metrics demonstrates the immaturity of the approaches. Additionally, we note that the most of recent of these is over 10 years old, showing that recent efforts to explore techniques for identifying tampering have not considered such attributes in their work. The variety of metrics used in the evaluation of these techniques also indicates that there is no standardised measure upon which techniques can be properly compared.

All 17 papers which included evaluation created a new dataset in order to do so, with none of these datasets being made available to the wider research community. As a result, external validation of the techniques would be extremely difficult to undertake and new research is unable to benefit from the data, creating additional difficulty in comparing techniques. Furthermore, it is difficult to evaluate whether the creation of individual datasets introduces any bias which may have had an impact on any evaluation of effectiveness that has been conducted. Creation of suitable datasets for use in digital forensics has been cited as a challenge to the field more generally (Arshad et al., 2018; Horsman and Lyle, 2021). Nevertheless, we argue here that without such efforts, it is impossible to be able to make satisfactory conclusions about the effectiveness of any of the techniques.

Additionally, the evaluation of the 27 papers against the quality criteria set out by Kitchenham (2004) demonstrates the quality of the evidence generated by the research in the final dataset. This was generally low, with 22 of the 27 papers being in the bottom two categories. This does not allow for conclusions to be drawn into the quality of the techniques themselves. It does show that the research conducted in these studies is not of sufficient robustness to draw conclusions about their effectiveness.

In summary, the research studied as part of this SR does not allow for any significant conclusions to be drawn into the effectiveness of any of the techniques. This is perhaps a surprising result, but it is the only position that can be taken given the issues described above.

5.4. RQ-1.4: are these techniques available for digital forensic practitioners?

Given that none of the techniques was available for use to the wider community, the answer to RQ-1.4 was rather more straightforward than had been anticipated. Not a single instance was identified where the research work had been made accessible to the practitioner community. This is a concerning observation which suggests that the researcher community is unable to have the impact on practice with regards to this issue.

6. Limitations

Several limitations were experienced when conducting this review. First, only one researcher was assigned to the SR and therefore there were significant resource and experience constraints. Further research collaboration would have likely provided more robust outcomes at all stages, including the evaluation of the primary studies in stages 4 and 7 as well as the data extraction in steps 5 and 8.

Additionally, the research was subject to significant time constraints as it was conducted to allow for completion as part of a PhD pilot study. This meant that the scope was tightened to only include generalised techniques, resulting in the introduction of additional exclusion criteria E−1 and E−2.

7. Recommendations

While the work outlined in this paper was somewhat limited in scope due to the practical reasons outlined in section 6, it is still possible to provide some recommendations for future work.

First, this review has considered generalised artefact detection techniques through the stated exclusion criteria. However, there are many techniques which relate to specific but widely used artefact types such as SQLite databases and images files. A similar study which focussed on these would enable to community to determine whether techniques for identifying tampering of these specific types are effective and practical.

Secondly, new papers which propose techniques for identifying tampering should consider the metrics used to evaluate the techniques. We recommend that appropriate metrics would include (but are not limited to) accuracy, precision, recall, false positives, false negatives and time taken. Ideally these would become standardised in the field over time such that comparisons between techniques can be more easily made.

Finally, authors should consider the datasets used in any such evaluations. Requirements for such datasets can be found in Garfinkel et al. (2009). Not only should these requirements be followed where possible, but we recommend that authors consider making their datasets publicly available. This has the advantage that multiple techniques can be evaluated against the same data, further enabling comparison between techniques and in time allowing practitioners to make more informed choices about which they apply to their cases. This is especially true when combined with the second recommendation relating to evaluation metrics. An example of such a dataset for SQLite artefacts can be found in Nemetz et al. (2018).

8. Conclusion

In conclusion, this Systematic Review identified 27 pieces of primary research concerned with proposing a generalised technique for identifying artefact tampering. The 'model checking' category of techniques was most common, followed by manual inspection of specific artefact properties. Reasons for the apparent popularity of this choice of approach are unclear and understand the potential advantages could be used to inform the initial design of further techniques in future work. However, the primary research included in this review did not provide any consistent means for evaluating the effectiveness of the proposed techniques. The evaluations conducted were also of generally low quality when assessed against criteria set out by Kitchenham (2004). Therefore, no conclusions can be drawn as to whether any of the identified research can be classified as effective in achieving the claimed objective of identifying tampering. There is also no way to be able to make comparisons between techniques with regards to their effectiveness, or indeed demonstrate improvement with future research. This is due to the use of individual datasets used in evaluation, none of which has been made publicly available. In addition, none of the techniques themselves are available to the practitioner community, severely limiting the impact that this type of research can have for the discipline as a whole.

Therefore, while the papers included in this SR do demonstrate the need for techniques to be able to identify tampering, existing research in this area has so far failed to provide any solution which can be shown to be effective. Furthermore, none can be meaningfully compared against other past or future techniques, nor are available to digital forensics practitioners for use in real cases.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data is available at https://github.com/cn3759/ArtefactAuthenticationSystematicReview

Acknowledgements

Thanks to my colleague who wishes to remain anonymous but who provided excellent moderation during the SR process. I owe you a pint or two of a certain continental lager, I would also like to thank Ian Kennedy, Bashar Nuseibeh, Yijun Yu and Blaine Price for their sage wisdom and advice throughout.

References

Al-Sharif, Z.A., Al-Saleh, M.I., Alawneh, L.M., Jararweh, Y.I., Gupta, B., 2020. Live forensics of software attacks on cyber–physical systems. Future Generat. Comput. Syst. 108, 1217–1229. https://doi.org/10.1016/j.future.2018.07.028.

Arshad, H., Jantan, A. Bin, Abiodun, O.I., 2018. Digital forensics: review of issues in scientific validation of digital evidence. J. Inf. Process. Syst. 14, 346–376. https://doi.org/10.3745/JIPS.03.0095.

Bednar-Schadle, T., 2018. MISINTERPRETATION OF DIGITAL EVIDENCE: RECOMMENDATIONS TO IMPROVE DATA INTEGRITY.

Carrier, B., 2006. A Hypothesis-Based Approach to Digital Forensic Investigations. NA. https://doi.org/10.1017/CBO9781107415324.004.

Casey, E., 2019. Forensic science international : digital investigation trust in digital evidence. Forensic Sci. Int. Digit. Investig. 31, 200898. https://doi.org/10.1016/j.fsidi.2019.200898.

Coronel, B., Cedillo, P., Campos, K., Camacho, J., Bermeo, A., 2018. A systematic literature review in cyber forensics: current trends from the client perspective. In: 2018 IEEE 3rd Ecuador Tech. Chapters Meet. https://doi.org/10.1109/ETCM.2018.8580266. ETCM 2018.

Crown Prosecution Service, 2017. Computer records evidence. accessed 6.27.22. https://www.cps.gov.uk/legal-guidance/computer-records-evidence.

Day, D., 2014. Seizing, imaging, and analyzing digital evidence: step-by-step guidelines. Cyber Crime Cyber Terror. Investig. Handb. 71–89. https://doi.org/10.1016/B978-0-12-800743-3.00007-4.

Dija, S., Ajana, J., Indu, V., Sabarinath, M., 2020. Cyber forensics: discovering traces of malware on windows systems. 2020. In: IEEE Recent Adv. Intell. Comput. Syst. RAICS, pp. 141–146. https://doi.org/10.1109/RAICS51191.2020.9332496, 2020.

Edward, E.O., Ojeniyi, J.A., 2019. A systematic literature review on digital evidence admissibility: methodologies, challenges and research directions. In: 2019 15th Int. Conf. Electron. Comput. Comput. ICECCO. https://doi.org/10.1109/ICECCO48375.2019.9043250, 2019.

Fleiss, J.L., Nee, J.C., Landis, J.R., 1979. Large sample variance of kappa in the case of different sets of raters. Psychol. Bull. 86, 974–977. https://doi.org/10.1037/0033-2909.86.5.974.

Freiling, F., Hösch, L., 2018. Controlled experiments in digital evidence tampering. DFRWS 2018 EU - Proc. 5th Annu. DFRWS Eur. S83. https://doi.org/10.1016/j.diin.2018.01.011. –S92.

Garfinkel, S., Farrell, P., Roussev, V., Dinolt, G., 2009. Bringing science to digital forensics with standardized forensic corpora. Digit. Invest. 6, S2–S11. https://doi.org/10.1016/j.diin.2009.06.016.

Geiger, M., 2005. Evaluating commercial counter-forensic tools. Digit. Forensic Res. Work. 1–12, 2005.

Haddaway, N.R., Page, M.J., Pritchard, C.C., McGuinness, L.A., 2022. PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis. Campbell Syst. Rev. 18, e1230. https://doi.org/10.1002/cl2.1230.

Helman, P., Liepins, G., 1993. Statistical foundations of audit trail analysis for the detection of computer misuse. IEEE Trans. Software Eng. https://doi.org/10.1109/32.241771.

Hoelz, B., Maues, M., 2017. Anti-forensic threat modeling. IFIP Adv. Inf. Commun. Technol. 511, 169–183. https://doi.org/10.1007/978-3-319-67208-3_10.

Horsman, G., 2021. Digital evidence and the crime scene. Sci. Justice 61, 761–770. https://doi.org/10.1016/j.scijus.2021.10.003.

Horsman, G., 2020. Digital evidence certainty descriptors (DECDs). Forensic Sci. Int. Digit. Investig. 32, 200896. https://doi.org/10.1016/j.fsidi.2019.200896.

Horsman, G., Errickson, D., 2019a. When finding nothing may be evidence of something: anti-forensics and digital tool marks. Sci. Justice 59, 565–572. https://doi.org/10.1016/j.scijus.2019.06.004.
Horsman, G., Errickson, D., 2019b. When finding nothing may be evidence of something. Anti-forensics and digital tool marks 59, 565–572. https://doi.org/10.1016/j.scijus.2019.06.004.
Horsman, G., Lyle, J.R., 2021. Dataset construction challenges for digital forensics. Forensic Sci. Int. Digit. Investig. 38, 301264. https://doi.org/10.1016/j.fsidi.2021.301264.
Iqbal, S., Alharbi, S.A., 2019. Advancing automation in digital forensic investigations using machine learning forensics. Digit. Forensic Sci.
Johnson, C., Davies, R., 2020. Using digital forensic techniques to identify contract cheating. A Case Study 105–113.
Khan, M.N.A., 2012. Performance analysis of Bayesian networks and neural networks in classification of file system activities. Comput. Secur. 31, 391–401. https://doi.org/10.1016/j.cose.2012.03.003.
Kirmani, M.S., Banday, M.T., 2020. Digital forensics in the context of the internet of things. Cyber Warf. https://doi.org/10.4018/978-1-7998-2466-4.ch069. Terror. 1178–1200.
Kitchenham, B., 2004. Procedures for Performing Systematic Reviews NICTA Technical Report 0400011T, vol. 1.
Kitchenham, B., Brereton, P., 2013. A systematic review of systematic review process research in software engineering. Inf. Software Technol. 55, 2049–2075. https://doi.org/10.1016/j.infsof.2013.07.010.
Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159. https://doi.org/10.2307/2529310.
Lees, C., 2013. Determining removal of forensic artefacts using the USN change journal. Digit. Invest. 10, 300–310. https://doi.org/10.1016/j.diin.2013.10.002.
Liu, C., Singhal, A., Wijesekera, D., 2012. Using attack graphs in forensic examinations. Proc. - 2012 7th Int. Conf. Availability, Reliab. Secur. ARES 2012 596–603. https://doi.org/10.1109/ARES.2012.58.
Mothi, D., Janicke, H., Wagner, I., 2020. A novel principle to validate digital forensic models. Forensic Sci. Int. Digit. Investig. 33. https://doi.org/10.1016/j.fsidi.2020.200904.
Neale, C., Kennedy, I., Price, B., Yu, Y., Nuseibeh, B., 2022. Forensic science international : digital investigation the case for Zero trust digital forensics. Forensic Sci. Int. Digit. Investig. 40, 301352. https://doi.org/10.1016/j.fsidi.2022.301352.
Nemetz, S., Schmitt, S., Freiling, F., 2018. A standardized corpus for SQLite database forensics. DFRWS 2018 EU - Proc. 5th Annu. DFRWS Eur. 24, S121. https://doi.org/10.1016/j.diin.2018.01.015. –S130.
Nisioti, A., Loukas, G., Rass, S., Panaousis, E., 2021. Game-theoretic decision support for cyber forensic investigations. Sensors 16.
Overill, R.E., Silomon, J.A.M., 2010. Digital meta-forensics: quantifying the investigation. In: Proc. 4th Int. Conf. Cybercrime Forensics Educ. Train. (CFET 2010). Canterbury, UK (September 2010).
Pieterse, H., Olivier, M., van Heerden, R., 2019a. Classifying the authenticity of evaluated smartphone data. IFIP Adv. Inf. Commun. Technol. 569, 39–57. https://doi.org/10.1007/978-3-030-28752-8_3.
Pieterse, H., Olivier, M., van Heerden, R., 2019b. Detecting manipulated smartphone data on android and iOS devices. Commun. Comput. Inf. Sci. 973, 89–103. https://doi.org/10.1007/978-3-030-11407-7_7.
Pieterse, H., Olivier, M., van Heerden, R., 2018. Smartphone data evaluation model: identifying authentic smartphone data. Digit. Invest. 24, 11–24. https://doi.org/10.1016/j.diin.2018.01.017.
Pieterse, H., Olivier, M., van Heerden, R., 2017. Evaluating the authenticity of smartphone evidence. IFIP Adv. Inf. Commun. Technol. 511, 41–61. https://doi.org/10.1007/978-3-319-67208-3_3.
Pieterse, H., Olivier, M., Van Heerden, R., 2019c. Evaluation framework for detecting manipulated smartphone data. SAIEE Africa Res. J. 110, 67–76. https://doi.org/10.23919/SAIEE.2019.8732797.
Rani, D.R., Kumari, G.G., 2017. A framework for detecting anti-forensics in cloud environment. In: Proceeding - IEEE Int. Conf. Comput. Commun. Autom. ICCCA, pp. 1277–1280. https://doi.org/10.1109/CCAA.2016.7813913, 2016.
Reddy, G.U., Madhu Bala, M., Padmaja, B., 2020. An overview on digital forensics tools used in crime investigation for forgery detection. 2020 Int. Conf. Comput. Sci. Eng. Appl. ICCSEA. https://doi.org/10.1109/ICCSEA49143.2020.9132965, 2020.
Rekhis, S., Boudriga, N., 2012a. A Hierarchical Visibility theory for formal digital investigation of anti-forensic attacks. Comput. Secur. 31, 967–982. https://doi.org/10.1016/j.cose.2012.06.009.
Rekhis, S., Boudriga, N., 2012b. A system for formal digital forensic investigation aware of anti-forensic attacks. IEEE Trans. Inf. Forensics Secur. 7, 635–650. https://doi.org/10.1109/TIFS.2011.2176117.
Schneider, J., Wolf, J., Freiling, F., 2020. Tampering with digital evidence is hard: the case of main memory images. Forensic Sci. Int. Digit. Investig. 32, 300924. https://doi.org/10.1016/j.fsidi.2020.300924.
Shafiee Hasanabadi, S., Habibi Lashkari, A., Ghorbani, A.A., 2020. A survey and research challenges of anti-forensics: evaluation of game-theoretic models in simulation of forensic agents' behaviour. Forensic Sci. Int. Digit. Investig. 35. https://doi.org/10.1016/j.fsidi.2020.301024.
Shanmugam, K., Powell, R., Owens, T., 2011. An approach for validation of digital anti-forensic evidence. Inf. Secur. J. 20, 219–230. https://doi.org/10.1080/19393555.2011.604667.
Singhal, A., Liu, C., Wijesekera, D., 2015. A logic based network forensics model for evidence analysis. Proc. ACM Conf. Comput. Commun. Secur. 2015-Octob 1677. https://doi.org/10.1145/2810103.2810106.
Stallard, T., Levitt, K., 2003. Automated analysis for digital forensic science: semantic integrity checking. Proc. - Annu. Comput. Secur. Appl. Conf. ACSAC 2003-Janua 160. https://doi.org/10.1109/CSAC.2003.1254321. –167.
Stoykova, R., 2021. Digital evidence: unaddressed threats to fairness and the presumption of innocence. Comput. Law Secur. Rep. 42. https://doi.org/10.1016/j.clsr.2021.105575.
SWGDE, 2016. 2016-06-23 SWGDE Digital and Multimedia Evidence Glossary V3.
Wojcik, M., Venter, H., Eloff, J., Olivier, M., 2006. Applying machine trust models to forensic investigations. IFIP Int. Fed. Inf. Process. 222, 55–65. https://doi.org/10.1007/0-387-36891-4_5.
Yusoff, Y., Ismail, R., Zainuddin, H., 2010. Adopting hadith verification techniques in to digital evidence authentication. J. Comput. Sci. 6, 591–596.