Joining of Data-driven Forensics and Multimedia Forensics - Practical Application on DeepFake Image and Video Data

Dennis Siegel
Dept. of Computer Science
Otto-von-Guericke-University
Magdeburg, Germany
dennis.siegel@ovgu.de

Christian Kraetzer
Dept. of Computer Science
Otto-von-Guericke-University
Magdeburg, Germany
christian.kraetzer@ovgu.de

Stefan Seidlitz
Dept. of Computer Science
Otto-von-Guericke-University
Magdeburg, Germany
stefan.seidlitz@ovgu.de

Jana Dittmann
Dept. of Computer Science
Otto-von-Guericke-University
Magdeburg, Germany
jana.dittmann@ovgu.de

***Abstract*—DeepFake technology poses a new challenge to the validation of digital media integrity and authenticity. In contrast to 'traditional' forensic sub-disciplines (for example dactyloscopy), there are no standardized process models for DeepFake detection yet that would enable its usage in court in most countries. In this work, two existing best-practice methodologies (a data-centric model and a set of image authentication procedures) are combined and extended for the application of DeepFake detection. The extension includes aspects required to expand the focus from digital images to videos and enhancements in the quality assurance for methods (here focusing on the peer review aspect). Particular emphasis is put on the different actors involved in the forensic examination process. The new methodology is applied to the example of DeepFake detection in two application scenarios, based on image and video respectively. The process itself is further separated in the initial assessment of the media followed by DeepFake detection. In total 36 features from nine existing and implemented tools are used as methods. In addition, the value types, ranges and their tendency for a DeepFake are determined for each feature. To further diversify the application field, the DeepFake detectors represent both handcrafted and deep learning based feature spaces for Media content analysis. The whole process is then manually evaluated, highlighting potential loss, error and uncertainties within the process and individual tools. With the discussed potential extensions towards video evidence and machine learning involved, we identified additional requirements. These requirements are addressed in this paper as a proposal for an extended methodology to serve as starting point for future research and discussion in this domain.**

Keywords-forensics; media forensics; DeepFake detection; machine learning.

I. Introduction and Motivation

Recent advances in computer vision and deep learning enabled a new digital media manipulation technology called DeepFakes, replacing identities in digital images, videos and audio material. They pose a challenge to the integrity and authenticity of digital media and the trust placed in media objects for forensic science. With the advances in technology and also DeepFake quality, they are no longer easily recognizable as such to the bare eye. For this reason, most existing protection approaches use machine learning algorithms for DeepFake detection. The use of machine learning makes it necessary to fulfil additional requirements for artificial intelligence (AI) systems (i.e., legal regulations). In consequence, DeepFake detectors are still not suitable for court room usage. This is due to aspects such as lack of maturity, including (besides precisely validated error rates) modeling and standardization efforts so that they can be integrated into established forensic procedures.

In this paper, as an extended work of [1], this gap (i.e., the lack of process modeling and investigation steps) is partially addressed by the following contributions:

- conceptional joining of IT and media forensic methodologies on the selected example of the existing *Data-Centric Examination Approach (DCEA)* [2], [3] and the *Best Practice Manual for Digital Image Authentication (BPM-DI)* from the *European Network of Forensic Science Institute (ENFSI)* [4].
- strengthening of the Human-in-the-Loop aspect in the forensic examination by highlighting the human operators involved and usage of algorithms in a decision support system.
- application of our concept to both an image and video DeepFake detection scenario, by utilizing a total of nine tools for general purpose media analysis, image processing and DeepFake detection.

With the focus on process modelling in the context of individual investigations, the prerequisites for the use of the individual tools are not considered in this paper in detail. This includes essential aspects such as initial model training, appropriate benchmarking and certification of the proposed tools. For these aspects the reader is referred to [5].

The paper is structured as follows. First, an overview of the state of the art on digital forensics, standards and regulations as well as the topic of DeepFake and its different types is presented in Section II. Following that, our concept of combining data-driven and media forensics can be found in Section III based on the DCEA [2] and BPM-DI [4]. Additional details on different human operators involved in the process are provided. In Section IV the proposed concept is being used in two application scenarios, both for image and video. This application is divided into two parts: first, an initial assessment is carried out, to validate the suitability of the material and then the DeepFake detection is performed. Finally, conclusions are drawn from the evaluation results presented and future directions are outlined in Section V.

It has to be noted, that this paper is an extended version of our work, presented at the SECURWARE 2023 conference [1]. This paper significantly expands on the aspect of human

operators involved in the forensic process as well as validating the practical applicability of the proposed methodology. To be more precise, it presents an expanded view of the fundamentals in the context of forensics by integrating additional Best Practice Manuals (BPM) of ENFSI (especially [6], presented in Section II-A) as well as DeepFake creation and detection (found in Section II-C). These fundamentals are used to conceptualize human operators involved in the forensic investigation in Section III-A. Furthermore, the practical application of the proposed methodology is expanded by testing on both an image and video DeepFake detection scenario. The Methods used in these scenarios are separated in Initial Assessment (Section IV-A), to validate the suitability of DeepFake detection Methods and the DeepFake detection itself (Section IV-B). To further support the separation, six additional tools are introduced.

II. Forensic Investigations in the Context of DeepFake Detection

With the potential of DeepFake manipulations in digital media it is even more important to validate integrity and authenticity of digital media especially for intended court room usage. The following sections address the current state and challenges in digital forensics, existing and upcoming regulations and the topic of DeepFake. These three aspects state fundamentals for the intended court room usage and while they are established in themselves, they are mostly considered in isolation.

A. Digital Forensics

Digital forensics is a subdomain of forensics, which is defined as "*the use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation, and presentation of digital evidence derived from digital sources [...]*" [7]. In [8] the domain of digital forensics is further divided into computer and multimedia forensics based on their link to the outside world. Computer forensics operates exclusively in the digital domain, whereas multimedia forensics uses sensors to capture and connect with the real world.

In general, the application of media forensics is governed by national legislation. For this reason, our focus will be on European documents and views on media forensics. Here, the European Network of Forensic Science Institutes (ENFSI) provides a broad list of BPM and guidelines in forensics.

All recent BPM share a common structure, governed by a common template: the scope of the BPM, definitions and terms, resources (including personnel), methods, validation and estimation of uncertainty of measurement, quality assurance, handling of items, initial assessment, prioritization and sequence of examinations, evaluation and interpretation as well as presentation of results.

The discussions on the personnel usually include discussions on the separation of duties between different roles as well as aspects of training and proficiency testing. Another item of relevance here is the validation and estimation of uncertainty of measurement. One main goal of the validation considerations is defined in [9] as precisely described, tool driven and repeatable processes: "*For software tools that can be configured in a variety of ways and/or uses a number of different parameters, it is particularly important to document the set-up and individual parameter values in order to produce a process that can be repeated*". These reproducibility requirements are the same for 'manual analysis software' as well as 'automated' (i.e., pattern recognition driven) software solutions.

In [6] extensive considerations are put into the validation and estimation of uncertainty. An important aspect of these discussions lies in the distinguishing between verified or non-verified functions and tools, 'validated processes' and 'trustworthy processes'.

Not only automated processes are within the scope of the verification and re-verification work to be performed. In [6] specifically the human-based methods are also included: "*Human-based functions are the pivotal elements within technical forensic processes, all forensic processes are likely to require user interaction, therefore an evaluation of user capability must be made as part of validated process within the laboratory. Even if an instrument-based function returns a valid result, it may still be reliant on the correct interpretation by the user associating the result.* [...] *Verification of human-based (user) functions are covered within proficiency testing* [...]."

The availability of the required forensic practitioners with sufficient training and currently valid certification (if required) is an important factor in every forensic investigation. In this paper, this is accompanied by the need to ascertain that other relevant types of personnel (e.g., data scientists) are also available to perform tasks that need forms or specific know-how (e.g., the creation/curation/update of trained models for AI-based investigation methods).

The whole issue on the validation of tools and processes is a necessity in the risk assessment required for case handling. In [6] it is stated on that issue: "*For the interpretation of evidential significance in the context of the case, a laboratory should always consider the use of techniques and equipment whose risks have been formally assessed; as part of the required functional verification, in preference to those, which have not. This does not mean that a method or process that has not been formally evaluated cannot be used to aid the analysis; rather it means that if there is a wish to use such a solution, a formal justification as to why it has been chosen in preference to one that is part of a validated process must be made. When designing a validation process, five key elements of a successful validation policy are:*

1) An understanding of known errors and uncertainty
2) The Statement of Requirements;
3) Risk Analysis and Assessments;
4) Effective validation test sets; and
5) Routine verification.*"*

According to this list, the fourth and fifth elements are more or less self-explanatory. Elements 1, 2 and 3 needs additional

explanations, which are given directly in the following for the 'An understanding of known errors and uncertainty', 'Statement of Requirements' as well as the 'Risk Analysis and Assessments': The 'understanding of known errors and uncertainty' needs a closer specification of the term 'uncertainty', which in [6] is specified as: "*is the unknown (random) difference (delta) between the measurement taken and its true value. It can never be completely defined, or eliminated, and is represented as a bounded region, in which the true value exists within its given confidence level.*" In complex systems, uncertainty is aggregating: "*Uncertainty within a system is additive in nature, and generally increases with the number of functions deployed within a process. The decision as to whether the uncertainty should be calculated at the function level or abstracted to the process level is at the discretion of each laboratory.* [...] *Software solutions will also contain additional uncertainty on top of the uncertainty associated with the physical systems, including the operating system, they are running on. This is especially true for software, which relies on functions with no formal specification and/or calibrated standard. As a result, software uncertainty properties will also need to be acknowledged and accounted for.*"

Regarding the considerations on "*uncertainty within image authentication*", BPM-DI [4] identifies three domain specific potential factors as: "*tool inaccuracies*", "*operator inaccuracies*" and "*data inconsistencies*". Also acknowledging that those factors are interlinked, the BPM-DI elaborates: "*Given the intricate dependencies which could exist between uncertainties that arise at various points during the image authentication analysis procedures, the uncertainty attached to a specific measurement cannot always be quantified.*"

The 'Statement of Requirements' is defined in [6] as: "*The statement of requirements defines the problem to be solved by a technical process. It should provide explanatory text to set the scene for a lay reader, summarising the problem, noting the scope and acceptable risks or limits of any solution and acknowledging the relevant stakeholders. It should be created independently of and without regard to any particular implementation or solution.*" Furthermore, the statement of requirements "*provides the interface (or formal bridge) between what the customer believes is achievable (customer requirements), and so desires, and what the laboratory can realistically achieve (laboratory capability) with the available staff, tools and the incurred time costs.*"

Optimally, this statement of requirements is not only a list, which expresses a set of needs and corresponding associated constraints and conditions but also includes a "*list of well-formed, testable requirements*." [6] In the ENFSI BPM FIT, an exemplary list of types for such requirements are presented, including functional and performance requirements as well as requirements focusing on the interfaces for the solution, its compliance with local laws and processes, etc. In addition it is stated that "*If the risks are considered too great then either the statement of requirements will need to be amended, or alternate solutions sought, to reduce the risks to acceptable levels.*". It basically determines, which methods are to be used within a forensic examination to be conducted, based on customer requirements. For the 'Risk Analysis and Assessments' [6] states: "*risk analysis and verification stages are paramount in creating a reliable validation method*", with the BPM providing a very general description how to perform such an risk analysis and how to record/document the risk in a formal assessment process. Different examples for corresponding evaluation questions to be used within such an assessment process are provided, including method-specific questions, implementation specific question as well as questions regarding the labs organizational procedures regarding the usage within a process. Summarising the discussion on risk analysis, [6] states: "*Risk analysis can not only be used to explain why a verified function has been used within a validated process, but also why in certain circumstances a formally unverified function has been chosen in preference.*"

The ENFSI BPM FIT [6] explicitly integrates the competence of the forensic practitioner(s) available to handle a case into the risk analysis: "*The lower the level of knowledge [of the analyst], the greater will be the potential errors and risks.*" But also experienced analysts might encounter challenges when interpreting the output of verified functions. In this case, the escalation procedure recommended is: "*If a new, unknown, discrepancy is detected then the evaluation will need to be highlighted for the peer review, and one or more of the verified tools may need to be reassessed, along with the existing validated process.*"

With regard to the usage of non-verified functions, which is a very likely scenario for certain media forensics investigation that still lack maturity and for which only lower technology readiness level solutions exist so far, the recommendation of [6] with regard to the corresponding risk assessment would be: "*When using a non-verified function during analysis it is important that the analyst is competent enough to research the characteristics of the returned results, and can qualify them against standard validation methods employed within the laboratory* [...]."

In the field of digital imaging, there are currently three Best Practice Manuals existing. The first document addresses the aspect of forensic facial image comparison [10] and formulates the respective investigation steps. At the beginning, competing hypotheses are made, which need to be examined. In the context of comparing facial images, these hypotheses could be whether a subject in an image is a specific person or some other person. The comparison is performed based on the ACE-V methodology, which stands for *Analysis*, *Comparison*, *Evaluation* and *Verification*. ACE-V is a common practice in forensic comparison tasks, such as fingerprint [11] and facial image comparison [10].

In [11] the *Analysis* is described as: "*The examiner makes a determination, based upon previous training, experience, understanding, and judgments, whether the print is sufficient for comparison with another print. If one of the prints is determined to be insufficient, the examination is concluded with a determination that the print is insufficient for comparison purposes.*" This highlights the importance of validating

the suitability of material for a forensic investigation. In the domain of facial images, ENFSI provides a list with potential factors influencing the facial appearance [10]. This potentially non-exhaustive list contains the aspects of "*image resolution/distance from camera*", "*image compression*", "*aspect ratio*", "*lighting*", "*occlusion*", "*camera angle*", "*image/lens distortion*", "*number of available images*" and "*date an image was captured*" [10]. In addition, the National Institute of Standards and Technology (NIST) and German Federal Office for Information Security (BSI; the German national cyber security authority) are collaborating to create the so called Open Source Face Image Quality (OFIQ) metric [12], to estimate the facial image quality. This metric is intended to be derived from the Face Image Quality Assessment (FIQA) discussed by Schlett et al. [13]. If the quality of the media file is not sufficient, it has to be either discarded or enhanced. Procedures on image and video enhancement can be found in [14]. However, it has to be noted that "*Image enhancement processes alter the appearance and content of an image and may distort facial features or introduce artefacts that mislead the comparison examination*." [14]. The *Comparison* component of ACE-V for the domain of facial images is discussed in detail in [10]. This comparison is performed by an examiner on the basis of a so-called facial feature list, including a total of 19 facial components, such as eyes, nose and mouth. In *Evaluation*, the results of the comparison are used to confirm or refute the competing hypotheses. In the end, the examination process has to be repeated independently by another examiner for *Verification* purposes.

The most recent document on image forensics and also the closest to the topic of DeepFake detection, is the *Best Practice Manual for Digital Image Authentication (BPM-DI)* [4]. In its own words it "*aims to provide a framework for procedures, quality principles, training processes and approaches to the forensic examination*" in the context of image authentication. For this purpose it describes a total of four aspects to categorize and structure investigation steps. These aspects consist of two different analysis methods, namely **Auxiliary data analysis** and **Image content analysis**, which are used based on different **Strategies** fulfilling different purposes. The last method class is **Peer review**, enabling the validation, interpretation and evaluation of the individual methods and their outcomes by a forensic human examiners.

At the national level, the German situation is relevant for the authors. Here, the guidelines for IT forensic by BSI [55] are currently relevant. The DCEA is an extension of these guidelines, which has three main components: a model of the *phases* of a forensic process, a classification scheme for *forensic method classes* and *forensically relevant data types*.

The six DCEA *phases* are briefly summarized as: *Strategic preparation (SP), Operational preparation (OP), Data gathering (DG), Data investigation (DI), Data analysis (DA) and Documentation (DO)*. While the first two (*SP and OP*) contain generic (*SP*) and case-specific (*OP*) preparation steps, the three phases *DG*, *DI* and *DA* represent the core of any forensic investigation. At this point it is necessary to emphasize the importance of the *SP*, because it is the phase that also includes all standardization, benchmarking, certification and training activities considered. For details on the phase model the reader is referred, e.g., to [2] or [15].

In terms of data types, the DCEA proposes a total of six for digital forensics and ten for digitized forensics. In [3], the data types are specified in the context of media forensics and are referred to as *media forensic data types (MFDT)*. The resulting eight can be summarized as: digital input data *MFDT1* (the initial media data considered for the investigation), processed media data *MFDT2* (results of transformations to media data), contextual data *MFDT3* (case specific information, e.g., for fairness evaluation), parameter data *MFDT4* (contain settings and other parameter used for acquisition, investigation and analysis), examination data *MFDT5* (including the traces, patterns, anomalies, etc that lead to an examination result), model data *MFDT6* (describe trained model data, e.g., face detection and model classification data), log data *MFDT7* (data, which is relevant for the administration of the system, e.g., system logs), and chain of custody & report data *MFDT8* (describe data used to ensure integrity and authenticity, e.g., hashes and time stamps as well as the accompanying documentation for the final report).

An additional extension is made in the process modeling, in which individual processing steps are represented as atomic black box components. These components are accompanied by a description of the process performed. The individual components have four connectors input, output, parameters and log data. In addition, with the increasing use of machine learning, a fifth connection required for knowledge representation is defined. The labeled model can be found in [3].

B. Standards and Regulations in the Context of Media Forensics

With the intended court room usage of forensic methods, standardization is required in investigation and analysis procedures. One of the more established standards is the United States Federal Rules of Evidence (FRE; especially FRE 702, see [16]) and the Daubert standard in the US. Although these standards only apply in the US, its usage, e.g., in Europe has been discussed in [17]. In this work, the focus is on modelling media forensic methods within an investigation, whereby the following two (of five) Daubert criteria are particularly relevant [17]:

- "*whether the technique or theory has been subject to peer review and publication*";
- "*the existence and maintenance of standards and controls*".

In the context of standards and controls, the European Commission proposed the Artificial Intelligence Act (AIA), addressing the usage of Artificial Intelligence (AI) systems [18]. At the current time, the proposal has been adjusted and approved by the European Parliament [19]. This upcoming regulation places particular emphasis on the human in control aspects (Art. 14). The decisive factor is therefore not only

the decision of the AI system, but the process of decision-making, which must be comprehensible for the human operator and thus enable the decision to be questioned and challenged. In addition, the International Criminal Police Organization (INTERPOL) recently published a document, addressing the usage of AI systems for law enforcement purposes [20]. Furthermore, the National Institute of Standards and Technology (NIST) currently develops a dataset for DeepFake detection for validation of methods [21]. All documents have in common that a human operator should comprehend and oversee the processing and decision-making of the AI system.

C. DeepFakes

With the advances in machine learning and computer vision DeepFake are a recent form of digital media manipulation and generation. In contrast to previous manipulation techniques, DeepFake utilizes deep learning to artificially generate or manipulate existing digital media, such as image, video and audio data. The application of DeepFakes is very versatile and can also be used for positive aspects, as described in [22]. Independently of their intended purpose, DeepFakes have to be identifiable both for integrity and authenticity of digital media and is further enforced by the recently adopted AIA [19].

Just as the created media of the DeepFake manipulations differ (i.e., image, video and audio), so do the creation methods. Mirsky et al. [23] divided the DeepFake generation methods into the four main categories, which are reenactment, replacement, editing and synthesis. Reenactment refers to the controlling of expressions of one person by another person without changing the identity. In contrast, replacement (e.g., face swap) is an attempt of impersonation by replacing the identity. Editing does not require a second identity, instead specific facial traits of the given face are adjusted and changed. Common examples of such forgeries include changing ethnicity, facial hair or age of a face. Akhtar [24] further states the possibility to add injury, effects of drugs or other health-related issues to the image of a person. The last category of face synthesis does not require a particular identity, as it creates new, non-existent persons. In addition, the different approaches of generation have various different generation methods, e.g., encoder-decoder networks or generative adversarial networks (GAN), potential traces of manipulation may vary depending on the generation method.

Li et al. [25] detect DeepFakes by analyzing warping artefacts, which are a common trace in face swap approaches. The DeepFake algorithm generates face images with a fixed size, which are afterwards adjusted by using affine transformation to get the resolution of the face in the target image. This process results in warping artefacts. For synthetic face images mostly GANs, such as StyleGAN [26] or its extensions [27]–[29] are used [24]. Each StyleGAN version introduces its own individual artefacts and fixes issues of its predecessors. In consequence, these architectures leave individual forensic traces, which are comparable to fingerprints. This direction is further explored by Yu et al. [30] and Marra et al. [31]. A more detailed overview of the specific artefacts originating from different generation methods is given in the survey of Akhtar [24]. In terms of DeepFake detection, methods can be divided into spatial and temporal feature spaces [23]. Initially, the focus of detection was solely on the proposal of suitable deep learning based detectors without any form of explanations. More recently publications further prioritise forensic aspects in detection. In [32] DeepFake detection with the consideration of compliance with existing and upcoming regulations are shown.

III. Conceptional Extension and Joining of Data-driven and Media Forensic

For the conceptual connection of data-driven and media forensics, the BPM-DI [4] is considered as a basis and extended for the case of DeepFake detection for video. To classify this further, it should be noted that [4] proposes the application in practice on a specific investigation. According to the phase modeling, this includes the phases *OP*, *DG*, *DI* and *DA*, with *SP* being omitted. In consequence, the tools used for the forensic investigation are assumed to be tested and verified (i.e., its error rates are known by means of benchmarking and also limits of applicability have been identified).

A. Human Operators Involved in the Forensic Investigation

Based on the findings in [1], more attention must be paid to the people involved in the forensic investigation. Table I provides a non-exhaustive list of typical roles of human operators in forensic processes. Note: Here a homogenized version of the ENFSI terminology is used since it slightly varies in terms and definitions between different ENFSI BPM (e.g., 'Case lead' vs. 'Case Leader' vs. 'Section Heads/Operations Managers' in the technical departments of a lab). In this set of typical roles involved, different subsets can be identified: For investigations, the minimal subset involved would be {Customer, Case Leader, System Administrator}. While the roles of the Customer and the Case Leader in a forensic investigation are obvious, the System Administrator is responsible for the availability and technical reliability of the used case management system(s) and the corresponding resources. Since this usually involves a need for elevated access privileges, special care has to be taken to prevent unauthorised access to investigation procedures and results by the System Administrator. In other typical working contexts, e.g., proficiency testing, other roles are involved (in the example of the proficiency testing the minimal subset would contain {Standardization Body, Examiner}).

B. Methods of the Forensic Investigation

In this paper, the considerations on methods used in forensic investigations are based on the categorization provided in [4]. The aspect of **Auxiliary data analysis** (see **Methods** in Figure 1) focuses on all traces of a media file. This includes the **Analysis of external digital context data**, which takes meta data of the file system into account. It can be used to identify potential traces of editing, for example by investigating the modify, access and change (MAC) times. The **File structure**

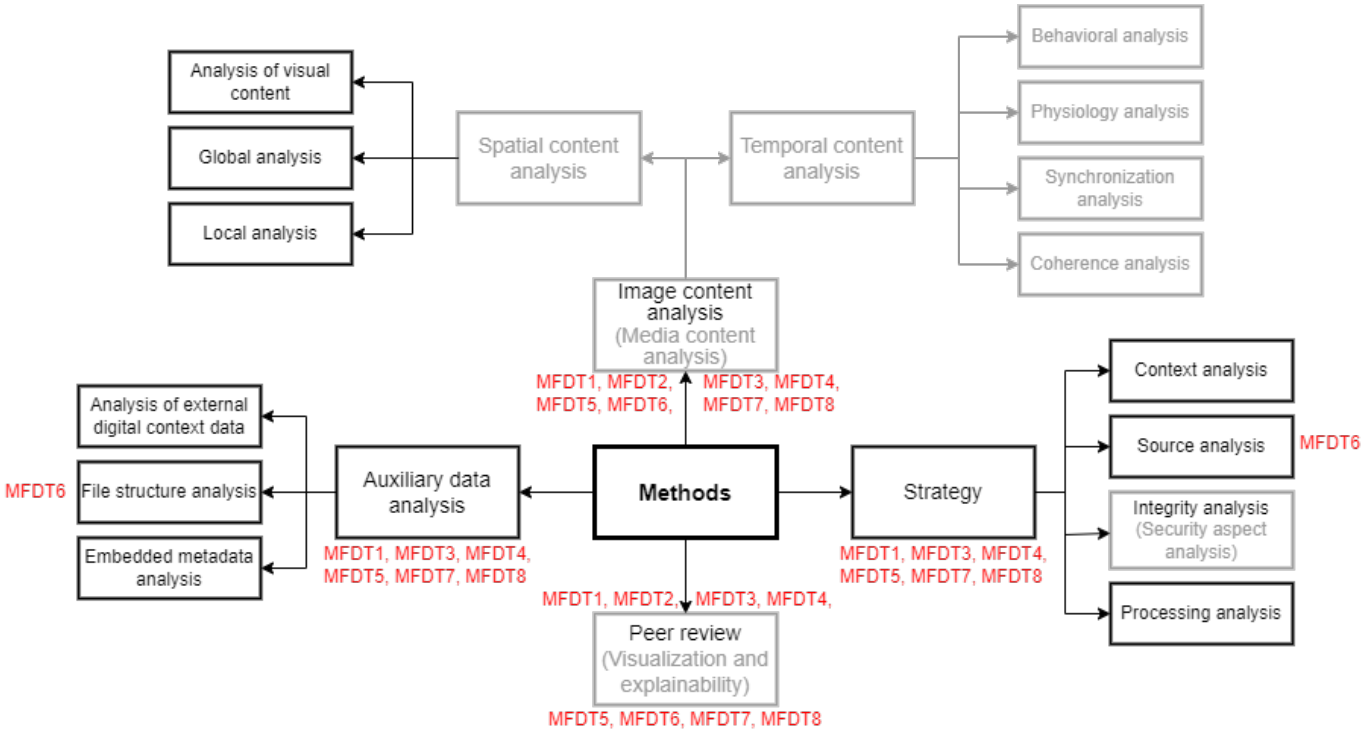

Figure 1. Categorization of forensic methods proposed in [4], extended on the case of media forensics, especially DeepFake detection. Extensions are marked in gray. Integration of media forensic data types (MFDT) can be found in red. Figure redrawn from [1].

TABLE I
TYPICAL ROLES OF HUMAN OPERATORS INVOLVED IN THE PROCESS OF A FORENSIC INVESTIGATION (NON-EXHAUSTIVE LIST). FURTHER SEPARATION BASED ON THE POINT OF TIME THE INVOLVEMENT OCCURS. THE PHASE DO IS OMITTED AS IT IS RELEVANT FOR ALL OPERATIONS PERFORMED.

Role	Phases	Description
System Administrator	SP, OP, DG, DI, DA	Entity responsible for the administration and maintenance of the system. Does not require any case specific information.
Data Scientist	SP	Human operator involved in the training of a machine learning model. The Data Scientist is responsible for managing and curating the datasets.
AI Expert	SP	Performs quality assurance on the feature space. The AI Expert applies explainable AI techniques to the feature space.
Standardization Body	SP	Entity verifying functions and tools. In the context of DeepFake detection this includes performing a benchmark and certifying the trained model.
Customer	OP	Entity requesting the examination of a digital media.
Third Party	OP	Acts as intermediary between Customer and Examiner. Formulates competing hypothesis for the investigation to mitigate potential bias of the Examiner.
Case Leader	OP	Examiner who prioritises the sequence of examination steps to be performed and assigns each to appropriate Examiners.
Examiner	OP, DG, DI, DA	Person(s) performing the forensic investigation of digital media.

analysis covers the examination of the file format. The format found for the examined file is compared with common formats including the specific version number. This can be a clue to the tools used to store the file. For videos, this is also useful to determine the potential origin based on the codec and its version used. **Embedded metadata analysis** takes into account all embedded metadata that can be found in the specific media. These can be used for the two main purposes of identifying the capturing device and gathering more details on the capturing process. For the identification of the capturing device the resolution and corresponding pixel format of images and videos can be used as a first indicator. For audio devices the sampling rate can be used as an equivalent. It is also possible for the device information to be specified in the metadata, but this is optional. For details on the capturing, there are optional metadata regarding the date and time of the recording and the GPS (Global Positioning System) location. In comparison to the BPM-DI [4], no extensions are required so far.

As discussed in Section II-C DeepFakes can occur in image, video as well as audio files. To address this aspect the BPM-DI [4] needs to extend the **Methods** to include spatial and temporal feature spaces in particular. This extension is suggested by a change in two steps, first the **Image content analysis** (see **Methods** Figure 1) has to become broader to also address video files by introducing **Media content analysis**. Second, a further separation of methods is presented, according to the categorization of DeepFake detection methods proposed

in [23] dividing into **Spatial** and **Temporal content analysis**. **Methods** of **Spatial content analysis** correspond to BPM-DI [4] **Image content analysis**, which are **Analysis of visual content**, **Global analysis** (i.e., analysis of the entire image) and **Local analysis** (i.e., analysis of a particular image region). These Methods can be found to the left of **Spatial content analysis** in Figure 1.

In contrast, **Temporal content analysis** is another required modality of DeepFake detection. There the first **Method** utilizes the **Behavioral analysis** shown in video or audio. For example in [33] facial movement is analyzed using facial action units to detect DeepFakes of Barack Obama, which is further enforced by the availability of reference data for this person. **Physiology analysis** relies on the assumption, that DeepFake creation lack physiological signals, e.g., in heart rate [34] or eye blinking behavior [22]. **Methods** for **Synchronization analysis** utilize different types of media to validate their correlation. In most cases this is done by extracting features from both audio and video and comparing them against each other. Previous research has been done for example on emotions [35] or lip synchronization [42]. **Coherence analysis** focuses on the aspect, that DeepFakes are created on a frame by frame basis, which might result in flickers and jitters in the video.

The general purpose of the category **Strategy** (see **Methods** in Figure 1) is to categorize previously mentioned **Methods**, both **Auxiliary data analysis** and **Media content analysis**, based on the specific investigation goal. In this work, we consider three of the investigation goals of BPM-DI [4] as they stand and extend the other. These address the correctness of the context the media is put into (**Context analysis**), identification of the device used to capture the media (**Source analysis**) and which processing steps applied to the media (**Processing analysis**). Extensions are made to the **Integrity analysis**, which initially identifies whether the questioned media was altered after acquisition. The extension aims to take into account all security aspects and additionally leave room for future requirements, (e.g., compliance with the AIA [18]). The existing method of **Integrity analysis** can be seen as method within the category of **Security aspect analysis**.

The **Peer review** (see **Methods** in Figure 1) of the BPM-DI [4] is the integration of a human examiner to analyze and interpret results during the whole process. With the introduction of machine learning techniques, especially for DeepFake detection, an extension of this aspect is proposed by introducing techniques to improve **Visualization and explainability**. Its purpose is therefore to support the human examiner in the process of investigation and decision making. With the introduction of machine learning algorithms, special attention has to be paid to the reproduceability of individual methods, their visualization and the entire examination process.

The application of data types is based on the existing 8 media forensic data types (MFDT) [3] mentioned in Section II-A and can also be seen in Figure 1 in red. Since the individual analysis **Methods** are kept generic our assignment of the data types is based on the higher level categories and is the same for the corresponding subcategories. In general, all **Methods** given require a process-accompanying documentation, which are specified to log data (*MFDT7*) and chain of custody & report data (*MFDT8*). Both **Auxilary data analysis** and **Strategy** work on the initial media representations (*MFDT1*), utilizing case specific information (*MFDT3*) and parameters (*MFDT4*) to yield examination data (*MFDT5*). In addition, model data (*MFDT6*) is required for both **File structure analysis** of and **Source analysis** to have a reference model of file structures or camera models respectively. The same can be said for **Media content analysis**, with the addition of various additional representations of the media (*MFDT2*) specific to the method of analysis and the potential usage of machine learning to introduce model data (*MFDT6*). One difference can be found in **Peer review**, in the initial proposal it suggests the analysis and interpretation of media representations (*MFDT2*) and examination data (*MFDT5*). By extending this category to **Visualization and explainability** and the identification of different human operators [5] it further introduces additional data types to be explained. These human operators include, but are not limited to, the forensic investigator, who requires *MFDT2*, *MFDT3*, and *MFDT5*, and the data scientist, who requires *MFDT3*, *MFDT4*, and *MFDT6*. Independent of the human operator, the data types *MFDT1*, *MFDT7* and *MFDT8* are required. In consequence, all MFDTs must be addressed in the method of **Visualization and explainability**.

To enable a more specific and descriptive assignment of the occurring data types, the individual processing steps have to be known, which is specific to the application used for the analysis. This is shown in more detail in the practical example given in Section IV.

IV. Application of DeepFake Detection on the Extended Modelling

To validate the applicability of the proposed extended **Methods** (see Figure 1), a practical application on the example of DeepFake detection is performed on two scenarios. The first scenario describes the forensic examination of an image. Here, an DeepFake image originating from the OpenForensics [36] dataset is selected, which can be found in Figure 3. The image contains two persons, of which only one (the person on the left) contains DeepFake manipulation. In the second scenario, the forensic examination of a video is performed. For this purpose, the DeepFake video 'id0_id1_0000' of Celeb-DF [37] is selected.

In accordance with the methodology discussed above, a total of *9* existing and implemented tools are considered in order to cover a broad spectrum of methods. This is an extension of [1] by a further 6 tools. Initial steps of the forensic investigation begin with the Customer requesting a forensic investigation for specific media data. This request should be made to an independent Third Party, to minimise potential biases of the Examiner. In this paper, the term Third Party is used as specified in [4] as the intermediary between the Customer and the actual investigation. Other ENFSI BPM use the same term with different meanings (e.g., the more traditional meaning

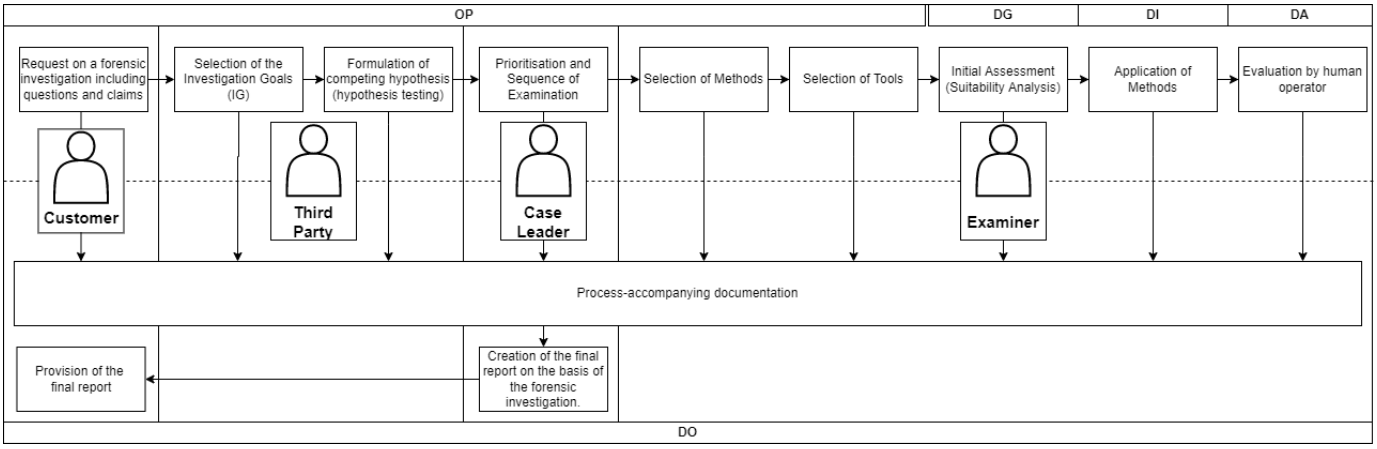

Figure 2. Workflow of an forensic investigation, separation according to the responsible human operator. Corresponding phases of the DCEA [2] are shown at the top and bottom respectively. The dashed line indicates the distinction between phases. For simplicity, log data (MFDT7) and chain of custody & report data (MFDT8) are combined to process-accompanying documentation.

Figure 3. Image selected for application scenario 1. The image shows two persons, of which the left person contains DeepFake manipulation. Image taken from the training set of the OpenForensics [36] dataset (id: 0b02353c85).

of an independent third party subcontractor or an independent lab with similar capabilities for result verification). However, as the focus of this paper is the topic of DeepFake detection, the initial request of a Customer and the request assessment of Third Party are omitted. It further assumes, that the competing hypothesis for the investigation derived are:

- The material under investigation appears to be tampered in a way that indicates a DeepFake manipulation
- There are no traces of post-processing or the identified post-processing seems plausible

It should be noted that the hypotheses considered here are specified for the case of DeepFake detection, as the detectors used are only intended for this purpose. In general, the hypotheses are derived from the request of the Customer and are closely connected to the **Strategies** discussed in Section III-B.

At this point the Case Leader determines and prioritises the investigation steps to be performed and assigns them to the corresponding forensic Examiner. Following the ACE-V methodology, the suitability of the material for the forensic examination has to be validated first. An initial assessment, including exemplary collection of tools for this purpose is discussed in Section IV-A. If none of the material (i.e., individual frames of a video) is found sufficient, no further investigation is performed. As described in [14], it would be possible to use enhancing techniques, but in the image domain these may distort features or introduce artefacts that mislead the examination. It has to be further noted, that the enhancement could possibly also remove traces of DeepFake manipulation. Once the digital media appears suitable for a forensic investigation, the selected Methods are applied. An exemplary application can be found in Section IV-B. The forensic Examiner then evaluates the results gathered for each Method. The entire process is supported by process-accompanying documentation. This documentation consists of log data (MFDT7), which is relevant for the administration of the system, and chain of custody & report data (MFDT8), which is used for forensic investigations reporting. Finally, all reportings are combined in a final report by the Case Leader, which is made available to the Customer. This workflow is further illustrated in Figure 2.

In the following, the individual processing steps and groups of features (hereinafter referred to as PS) as well as individual features (hereinafter referred to as ID) will be labeled and categorized in the extended BPM-DI [4] for **Auxiliary data analysis** (shown in Figure 4), **Media content analysis** (shown in Figure 5) and **Strategies** (shown in Figure 6).

A. Initial Assessment of the Media Under Investigation

To further extend the findings of [1] the initial assessment of the digital media is carried out using the methods of **Auxiliary data analysis** and **Spatial content analysis**. To analyse the metadata of the media both ExifTool [38] (PS-exif) and FFmpeg [39] (PS-ffmpeg) are used. While a variety of entries are available in the metadata, a total of eight features (ID-exif$_n$) are selected from ExifTool for this exemplary

approach and categorized according to the Ext. BPM-DI. Three features are selected from FFmpeg (ID-ffmpeg$_n$), two of which correspond to ExifTool features, so that a direct comparison between these two tools is possible. The first set of three Exiftool features address **Analysis of external digital context data** with the aim of **Processing analysis**. These can give first indications of possible manipulations, for example by validating timestamps for modification, access and creation (ID-exif$_1$), file size (ID-exif$_2$) or system feature flags such as user permissions (ID-exif$_3$). Furthermore, three additional features can be used for **File structure analysis**, by extracting the file format (ID-exif$_4$), its format version (ID-exif$_5$) and in case of a video file the used codec (ID-exif$_6$). The extracted information of **File structure** can then be compared to **Standard formats**, unveiling potential traces for **Processing analysis**. In addition, file formats and codecs can give an indication of the software or device to enable **Source analysis** as well. The third set, consisting of two features, which address **Embedded metadata analysis**, with the aim of **Context analysis**, by extracting the media files width and height (ID-exif$_7$, ID-ffmpeg$_1$) and frame rate if it is a video (ID-exif$_8$, ID-ffmpeg$_2$). The features ID-exif$_4$-ID-exif$_8$ can further be used to validate the suitability of subsequent DeepFake detectors. This refers in particular to media properties such as width and height of an image or frame (ID-exif$_7$), frame rate for videos (ID-exif$_8$) and format (ID-exif$_4$) or codec specific compression (ID-exif$_6$) and frame compression (ID-ffmpeg$_3$).

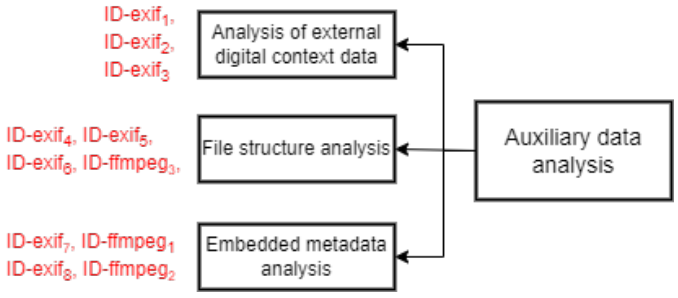

Figure 4. Individual features extracted using ExifTool [38] and FFmpeg [39] (in red) categorized in the extended BPM-DI [4] for the category Auxiliary data analysis. This is an extended version of [1], with the introduction of additional features (ID-ffmpeg).

The second set of Tools utilizes Methods of **Spatial content analysis** to estimate the quality of images, either globally (**Global Analysis**) or in the context of individual faces visible in the images (**Analysis of visual content**). For this exemplary approach, the image quality is estimated using blur (ID-global$_1$) and JPEG compression (ID-global$_2$) detection, which are derived from the frequency domain of the image. Face detection is performed using both MTCNN [40] (ID-face$_1$ and ID-face$_2$) and dlibs 68 landmarks [41] (ID-face$_3$). In addition, the head position (ID-face$_4$) is determined using the roll, yaw and pitch angle on the basis of the landmark positions. As the aforementioned tools and methods are used for the initial assessment, they provide context-specific information (MFDT3). The categorization of Tools in the extended BPM-DI can be found in Table II.

When using the Tools and Methods, it should first be noted, that not every media file contains all the information, as some features are video specific. In consequence, there are no information available regarding video codec (ID-exif$_6$), frame rate (ID-exif$_8$) and frame types (ID-ffmpeg$_3$) in scenario 1. However, FFmpeg provides a frame rate (ID-ffmpeg$_2$), which is due to the fact that it works primarily on video data and thus, incorrectly assumes a motion image. More information can be collected in scenario 2. The frame types in the video show a repeating pattern of an I-frame, followed by 11 P-frames (ID-ffmpeg$_3$) and shows no inconsistencies. Considering the results of both PS-global and PS-face, facial movement can be derived from the value range in ID-face$_4$, which could be the reason for motion blur identified in ID-global$_1$. In the context of the face detection algorithms considered, it should first be noted that only MTCNN provides a confidence measure (ID-face$_2$) for the detected faces. There are also small differences between the MTCNN (ID-face$_1$) and dlib (ID-face$_3$) face detector, as MTCNN did not find a face in two frames. This is a limiting factor for detection approaches based on Methods of Temporal content analysis. However, as the DeepFake detectors considered in Section IV-B utilize the dlib face detector, their suitability is given. In addition, it has to be noted that the *'C:'* in MAC timestamps (ID-exif$_1$) states the time of file change provided by the file system, as the meta data field for creation is empty in both scenarios. A summary of the results collected in this step of initial assessment can be found in Table III.

B. Practical Application of the Extended Methods to DeepFake detection

One of the more promising feature spaces for DeepFake detection utilizes the mouth region, addressing two flaws in DeepFake synthesis. First, the synthesis occurs on a frame-by-frame basis, which results in inconsistencies in the temporal domain, enabling aspects of lip movement analysis. In [42] the detection is performed based on lip synchronization, by considering both audio and video and detecting inconsistencies between phonemes in audio and visemes in video. A similar approach has been taken for the LipForensics detector [43] by identifying unnatural mouth movement. The second aspect utilizes the post processing, especially blurring, performed in DeepFake synthesis. In [44] and [45] texture analysis is performed on the mouth region to identify manipulations. A combination of both approaches is given in [46], where hand-crafted features are used to detect DeepFakes based on mouth movement and teeth texture analysis described as DF_{mouth}.

To evaluate the suitability of the proposed Ext. BPM-DI modeling for DeepFake detection the two detectors DF_{mouth} [46] and LipForensics [43] are selected, representing a hand-crafted as well as deep learning based detector. Both address **Media content analysis**, **Strategies** and **Peer review**. In addition, DF$_{mouth}$ utilizes the features ID-exif$_7$ and ID-exif$_8$ of **Auxiliary data analysis** for internal feature

TABLE II
COLLECTION OF TOOLS AND FEATURES USED FOR THE INITIAL ASSESSMENT OF DIGITAL IMAGE AND VIDEO DATA. CATEGORIZATION BASED ON THE PROPOSED EXTENDED BPM-DI [1]. THE SUITABILITY FOR A FORENSIC EXAMINATION IS HIGHLIGHTED IN BOLD AND ITALIC, WHERE BOLD VALUES INDICATE HIGHER SUITABILITY FOR VALUES CLOSER TO THE UPPER BOUNDARY. IN CONTRAST, ITALIC VALUES INDICATE A HIGHER SUITABILITY CLOSE TO THE LOWER BOUNDARY.

Ext. BPM-DI		feature	description	value	processing step	analysis	strategy	data type
Auxiliary data analysis	Analysis of external digital context data	ID-exif$_1$	MACtime	timestamp	PS-exif	File system metadata	Processing analysis	MFDT3
		ID-exif$_2$	file size	string				
		ID-exif$_3$	system feature flags					
	File structure analysis	ID-exif$_4$	file format	string		File structures	Source & Processing analysis	
		ID-exif$_5$	file format version	version number				
		ID-exif$_6$	video codec	string				
	Embedded meta-data analysis	ID-exif$_7$	file resolution	int [0, ∞]		Additional metadata	Context analysis	
		ID-exif$_8$	file frame rate	real [0, ∞]				
	Embedded meta-data analysis	ID-ffmpeg$_1$	file resolution	int [0, ∞]	PS-ffmpeg	Additional metadata	Context analysis	
		ID-ffmpeg$_2$	file frame rate	int [0, ∞]				
	File structure analysis	ID-ffmpeg$_3$	frame types	string		File structures	Processing analysis	
Media content analysis	Spatial content analysis	ID-global$_1$	image blur estimation	**real [0, 1]**	PS-global	Global analysis	Processing analysis	MFDT3
		ID-global$_2$	JPEG compression estimation	*real [0, 1]*				
		ID-face$_1$	Face detector MTCNN (ROI)	5 landmarks	PS-face	Analysis of visual content	Context analysis	
		ID-face$_2$	Face detector MTCNN (confidence)	**real [0, 1]**				
		ID-face$_3$	face detector dlib (ROI)	68 landmarks				
		ID-face$_4$	face orientation	real [-90, 90]				

TABLE III
RESULTS OF THE INITIAL ASSESSMENT FOR SCENARIO 1 (IMAGE) AND SCENARIO 2 (VIDEO). VALUES IDENTIFIED FOR SCENARIO 2 ARE GIVEN AS RANGES, TO PROVIDE AN OVERVIEW OF ALL FRAMES.

feature	scenario 1	scenario 2
ID-exif$_1$	M: 2021:02:22 22:19:00+01:00 A: 2024:03:14 15:21:16+01:00 C: 2024:03:14 15:21:05+01:00	M: 2019:11:13 14:17:36+01:00 A: 2024:03:14 01:00:00+01:00 C: 2024:03:14 17:02:07+01:00
ID-exif$_2$	385 kB	2.1 mB
ID-exif$_3$	r w - r - - r - -	r w - r - - r - -
ID-exif$_4$	image/jpeg	video/mp4
ID-exif$_5$	1.01	0.2.0
ID-exif$_6$	-	-
ID-exif$_7$	1024x683	944x500
ID-exif$_8$	-	30
ID-ffmpeg$_1$	1024x683	944x500
ID-ffmpeg$_2$	25/1	30/1
ID-ffmpeg$_3$	-	repeating pattern of 11 P-frames between single I-frames
ID-global$_1$	no blur: 0.1160	blur: [0.0163, 0.0232]
ID-global$_2$	not compressed: 0.0699	not compressed: [0, 0.0004]
ID-face$_1$	2 faces found	1 face found in 467 of 469 frames
ID-face$_2$	left face: 1 right face: 0.9999	[0.9838, 0.9999]
ID-face$_3$	2 faces found	1 face found in 469 of 469 frames
ID-face$_4$	left face: {-4.64, 8.34, -24.99} right face: {-2.73, 1.32, -26.94}	Roll: [-21.25, 15.75] Yaw: [-10.84, 40.94] Pitch: [-35.32, 10.00]

normalization. With their intention of identifying DeepFakes the general **Strategy** of application is **Integrity analysis**. Starting with the SP phase for DF$_{mouth}$, the detector is introduced in [46] and trained using the WEKA machine learning toolkit [47]. For the classification the decision tree classifier J48 [48] is used on the datasets DeepfakeTIMIT [49], [50], Celeb-DF [37] and DFD [51]. Detection performance peaks at 96.3% accuracy on a distinct training and test split of DFD. Considering distinct datasets for training and testing, detection performance peaks at 76.4% accuracy trained on DeepfakeTIMIT and tested on DFD. In a later benchmark approach given in [5] DF$_{mouth}$ is applied on a larger variety of DeepFake synthesis methods, including FaceForensics++ [51], DFD [51], Celeb-DF [37] and HiFiFace [52]. With an achieved detection performance of 69.9% accuracy the approaches suitability is identified only for certain DeepFake synthesis methods. With the limitations of DF$_{mouth}$ in mind, it is first split into five processing steps and categorized according the extended model. The individual features are then used for decision support by human operator, using the thresholds provided by the classifier in [46].

1) The video under investigation is first split into individual frames (PS-mouth$_1$) to first focus on **Spatial content analysis**.
2) For each frame a face detection algorithm is applied, in [46] using dlib's 68 landmark detection model [41] to extract the corresponding region for the mouth region (PS-mouth$_2$), which shows a dependency on the underlying model for face detection.
3) Then in PS-mouth$_3$, based on the keypoint geometry, it is determined whether the mouth is open (referred to as "state 1") or closed ("state 0"). Furthermore, the occurrence of teeth (referred to as "state 2") are examined based on texture analysis.
4) Based on the extracted mouth region and the information gathered, a total of 16 features are extracted. The first set of features, ID-mouth$_1$-ID-mouth$_7$ and ID-mouth$_{12}$ refer to **Physiological analysis** by describing mouth movements and the presence of teeth, by embedding individual frame features back into the temporal context of the video (PS-mouth$_4$). With the idea of DeepFakes having fewer mouth movements, values closer to 0 indicate a DeepFake for the features ID-mouth$_1$-ID-mouth$_6$. Features ID-mouth$_7$ and ID-mouth$_{12}$ aim to identify potential post-processing of the media, where lower values in ID-mouth$_{12}$ and higher values in ID-mouth$_7$ indicate a DeepFake. These are used for **Context analysis** to identify temporal inconsistencies. The normalization of features is done based on the frame rate (ID-exif$_8$) identified in **Auxiliary data analysis**.

5) The second group of features (PS-mouth$_5$), which consist of ID-mouth$_8$-ID-mouth$_{11}$ and ID-mouth$_{13}$-ID-mouth$_{16}$, refers to **Local analysis** to describe the sharpness of objects (here mouth and teeth region). In general, higher values for the features addressing state 1 (ID-mouth$_8$-ID-mouth$_{11}$) and lower values for the features addressing state 2 (ID-mouth$_{13}$-ID-mouth$_{16}$) indicate a potential DeepFake. The underlying **Strategy** is **Processing analysis**. The normalization of features is done based on the video frame resolution (ID-exif$_7$) identified in **Auxiliary data analysis**.

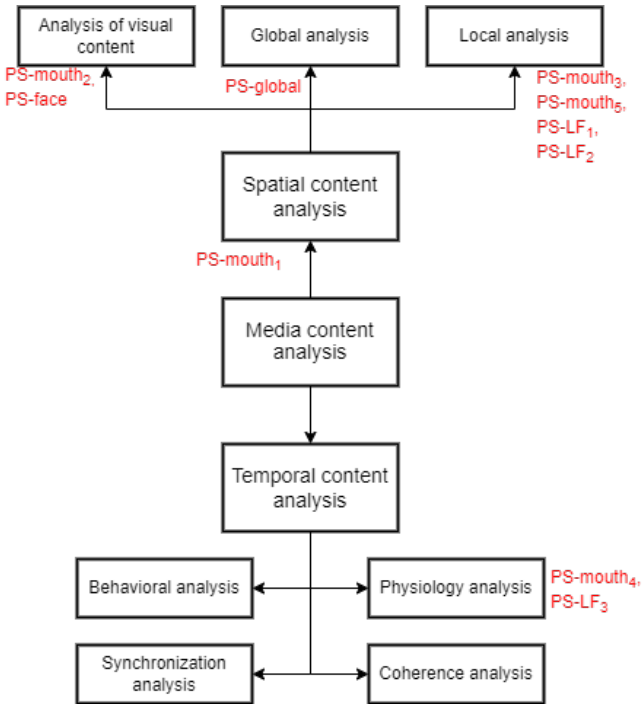

Figure 5. Processing steps (PS, in red) for ExifTool [38] and the DeepFake detectors DF$_{mouth}$ [46] and LipForensics [43] categorized in the extended BPM-DI [4] for the category Media content analysis. This is an extended version of [1], with the introduction of additional features (PS-face and PS-global).

More details on the individual features, their description as well as the categorization in the forensic methods can be found in the upper part of Table IV. Although all features can be categorized as *MFDT5*, the individual processing steps are more complex, containing multiple data types. For a more detailed description, the reader is referred to [22].

The second detector LipForensics [43] (herinafter referred to as LF) is included on a theoretical basis. For LF a total of three PS can be identified.

1) In the first step (PS-LF$_1$) the preprocessing occurs. First, a total of 25 frames are extracted from the video. These frames are converted to grayscale images, cropped to the mouth region and scaled to a resolution of 88x88. The resulting image representation can be categorized as *MFDT2*. With the intend of using only the mouth region, the corresponding method is **Local analysis** and the underlying strategy **Context analysis**.
2) In PS-LF$_2$ the feature extraction is done using a pre-trained ResNet-18 architecture trained on lip reading (*MFDT6*). As the result a feature vector of size 512 is generated (*MFDT3*). Again, the corresponding method is **Local analysis** and the underlying strategy **Context analysis**.
3) The resulting feature vector is used for classification purposes (PS-LF$_3$) using a multiscale temporal convolutional network (MS-TCN). The classification result *MFDT5* contains a classification label and the corresponding probability. With the aim of identifying unnatural behavior in mouth movement the corresponding method is **Physiology analysis** and the strategy of **Processing analysis**.

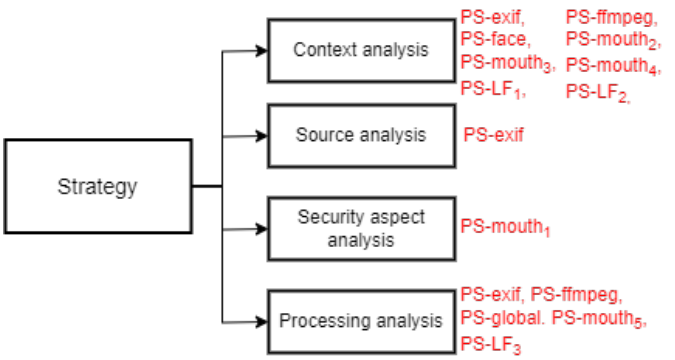

Figure 6. Processing steps (PS, in red) for ExifTool [38] and the DeepFake detectors DF$_{mouth}$ [46] and LipForensics [43] categorized in the extended BPM-DI [4] for the category Strategy. This is an extended version of [1], with the introduction of additional features.

For the practical implementation, starting with scenario 1, it has to be noted that LipForensics is not suitable for image data. In contrast, DF$_{mouth}$ also uses spatial features (PS-mouth$_5$), providing an indication for potential DeepFake manipulation. Only four of the sixteen features can be extracted with a single image, as only one state of the mouth can be given. In addition, when looking at the picture, it can be seen that both people show a closed mouth (state 0), which does not contribute to the decision for DF$_{mouth}$. However, applying the algorithm to the image classifies the left person's mouth as open with visible teeth (state 2) and the right person's mouth as open without visible teeth (state 1). The reason for this is the inaccurate position of the landmarks, particularly in the area of the lower lip margin. The assignment of the left face to mouth state 2, on the other hand, is less plausible, especially as the texture details are very small as shown by ID-mouth$_{15}$ = 0, for example. These findings, are a indication of possible manipulation, but since DF$_{mouth}$ operates outside the boundaries of its intended use, they cannot be directly related to DeepFake manipulation. A more detailed analysis of the features is possible in scenario 2 as it is a video and both LF and DF$_{mouth}$ are applicable. However, as LF requires all videos of the Celeb-DF dataset to evaluate, only the pre-

TABLE IV
CATEGORIZATION OF DF_{mouth} [46] (TOP SECTION) AND LIPFORENSICS [43] (BOTTOM SECTION) IN THE FORENSIC CONTEXT, BASED ON THE PROPOSED EXTENDED BPM-DI. FOR FEATURE VALUES HIGHLIGHTED IN BOLD HIGHER VALUES INDICATE A DEEPFAKE AND FOR ITALIC LOWER VALUES INDICATE A DEEPFAKE, PRESENTED IN [1].

Ext. BPM-DI		feature	description	value	processing step	analysis	strategy	data type
Media content analysis	Temporal content analysis	ID-mouth$_1$	abs max change Y	*real [0, ∞]*	PS-mouth$_4$	Physiology analysis	Context analysis	MFDT5
		ID-mouth$_2$	max change Y	*real [0, ∞]*				
		ID-mouth$_3$	min change Y	**real [-∞, 0]**				
		ID-mouth$_4$	abs max change X	*real [0, ∞]*				
		ID-mouth$_5$	max change X	*real [0, ∞]*				
		ID-mouth$_6$	min change X	**real [-∞, 0]**				
		ID-mouth$_7$	percentage time state 1	**real [0, 1]**				
		ID-mouth$_{12}$	percentage time state 2	*real [0, 1]*				
	Spatial content analysis	ID-mouth$_8$	max regions state 1	*real [0, ∞]*	PS-mouth$_5$	Local analysis	Processing analysis	
		ID-mouth$_9$	max FAST keypoints state 1	*real [0, ∞]*				
		ID-mouth$_{10}$	max SIFT keypoints state 1	*real [0, ∞]*				
		ID-mouth$_{11}$	max sobel pixel state 1	*real [0, ∞]*				
		ID-mouth$_{13}$	min regions state 2	**real [0, ∞]**				
		ID-mouth$_{14}$	min FAST keypoints state 2	**real [0, ∞]**				
		ID-mouth$_{15}$	min SIFT keypoints state 2	**real [0, ∞]**				
		ID-mouth$_{16}$	max sobel pixel state 2	**real [0, ∞]**				
Media content analysis	Spatial content analysis	ID-LF$_1$	extraction of 25 frames, grayscale, crop and align	int [0, 255]	PS-LF$_1$	Local analysis	Context analysis	MFDT2
	Spatial content analysis	ID-LF$_2$	feature extraction utilizing ResNet-18	feature vector of size 512	PS-LF$_2$	Local analysis	Context analysis	MFDT3
	Temporal content analysis	ID-LF$_3$	classification of mouth movement based on MS-TCN	label: {real, fake} probability: real [0, 1]	PS-LF$_3$	Physiology analysis	Processing analysis	MFDT5

processing part of the detector (LF_1) is considered. DF_{mouth} provides more details on discussion and interpretation. Considering the contents of the video in conjunction with the video, the percentage of time the person is showing an open mouth (ID-mouth$_7$ and ID-mouth$_{12}$, adding up to 0.3091) appears to be relatively low, since the person is talking in the video. The texture analysis (ID-mouth$_8$-ID-mouth$_{11}$ and ID-mouth$_{13}$-ID-mouth$_{16}$) does not provide any indications on DeepFake manipulation as the individual features are not that high or low respectively. Also, the DeepFake detector classifies this Video as no DeepFake in this instance. In consequence, there are indications of post-processing in the mouth region, but they cannot be identified as DeepFake manipulation. The full set of features extracted for both scenarios can be found in Table V.

TABLE V
RESULTS OF THE DEEPFAKE DETECTION FOR SCENARIO 1 (IMAGE) AND SCENARIO 2 (VIDEO).

feature	scenario 1	scenario 2
ID-mouth$_1$	-	3.8333
ID-mouth$_2$	-	3.8333
ID-mouth$_3$	-	-3.0804
ID-mouth$_4$	-	1.4904
ID-mouth$_5$	-	1.4904
ID-mouth$_6$	-	-1.2923
ID-mouth$_7$	-	0.2878
ID-mouth$_{12}$	-	0.0213
ID-mouth$_8$	right face: 0.0469	0.1148
ID-mouth$_9$	right face: 0.0781	0.0714
ID-mouth$_{10}$	right face: 0.0313	0.0492
ID-mouth$_{11}$	right face: 6.3438	7.7049
ID-mouth$_{13}$	left face: 0.0556	0.0441
ID-mouth$_{14}$	left face: 0.0741	0.0526
ID-mouth$_{15}$	left face: 0	0.0441
ID-mouth$_{16}$	left face: 6.6852	7.3659
ID-LF$_1$	2 images of mouth	469 images of mouth
ID-LF$_2$	-	-
ID-LF$_3$	-	-

With the introduction of machine learning algorithms in combination with previously discussed aspects of human in control and human oversight, the **Peer review** component becomes even more important. Its aim should be to enable the human operator to validate the results of each machine learning step to reduce the potential for error. Figure 7 demonstrates a potential direction to enhance the Method of **Peer review** on the basis of DF_{mouth} and ExifTool [54]. In general, the aim of this visualization is to remove the decision-making from the detector. Instead, the individual features are displayed and evaluated by the human operator. To enable the advanced methodology and the human operator to make a decision, this first conceptual example consists of four segments.

1) A filter for the forensic Methods of analysis (i.e., Auxiliary data analysis and Media content analysis), Strategy, detector and data type (see the top left box of Figure 7). Based on the selected features only suitable features are shown and selectable for further investigation.
2) The second block (see the top right box of Figure 7) acts as media player. It has different views to either visualize the video, individual frames (including potential

visualizations for explainability) and the metadata.
3) Based on the selected feature, this element shows its categorization in the forensic Methods and visualizes its value for each frame (see the bottom left box of Figure 7).
4) The last block (see the bottom right box of Figure 7) integrates the human operator in the decision-making process. The operator is provided with questions based on specific features and values to identify potential errors of the algorithm. In addition, the detectors thresholds for classification are provided without the decision itself. This is done to reduce the risk of bias by the Examiner based on the decision being provided.

In addition, it should be noted that each step in the pipeline discussed involving machine learning for DF_{mouth} could also have been performed by manually labeling the data to reduce the error susceptibility. However, this would come at the expense of the required review time, especially for long videos with high frame rates.

This potential usage of machine learning indicates the necessity of the *SP* phase within the investigation process. Models have to be benchmarked properly to identify both error rates and potential limitations in their usage, to comply with the Daubert criteria discussed previously [17]. Furthermore, in the context of forensic investigations they have to be certified, so that these are approved for the investigation. These required steps must be performed before the actual investigation in the *SP* phase, which is not considered in the BPM-DI, in contrast to our extended BPM-DI.

V. Conclusion and Future Work

In this work an extension to the ENFSI BPM for digital image authentication is proposed, utilizing data-driven forensics by adding the eight media forensic data types (MFDT) from DCEA [2], [3] in **Methods** of BPM-DI [4]. This makes it possible to establish a connection between existing practices in forensic science and DeepFake detection. On the one hand, the forensic investigataion steps, that are necessary for DeepFake detection are shown. On the other hand, the necessary requirements are outlined, particularly with regard to the provision of information generated by the detector. In addition, extensions are proposed in the **Media content analysis Methods** using **Spatial** and **Temporal content analysis** to reflect the typical analysis domain of DeepFake detection (and other video authentication methods). Furthermore, the extension of the **Peer review** component to address also **Visualization and explainability** was touched upon by introducing a graphical interface that provides further information about the internal processing of the DeepFake detector. Here, the aspects 'human in the loop' and 'human in control' as well as the topic 'explainable AI' represent important foundations for this component as there are different operators involved in the forensic examination process.

The extended BPM-DI model is applied the forensic investigation process of image and video data. By using a total of nine existing and implemented tools as methods the applicability can be shown. Potential limitations and errors have been shown, as the selected DeepFake detectors of DF_{mouth} and LipForensics are intended for the analysis of video data. In addition, it was found that the deep learning based features are too complex to achieve the same granularity as the detector DF_{mouth}. Another limitation resulted from the structuring according to the phases, as suggested in DCEA. By omitting the Strategic Preparation (SP) phase, the detection approaches introduced for investigation have to be trained, benchmarked and certified beforehand. On this basis, the suitability of the individual detectors for the respective investigation must be determined, but this is not possible without prior knowledge of SP. In order to compensate for these disadvantages, the preliminary work of the detectors under consideration was taken as the findings of the SP phase and further verified in the application within this paper. Moreover, the interplay of individual **Methods** have been identified. Starting with the initial assessment of the media provides further insights for the suitability of individual detection approaches.

The evaluation of the proposed methodology on the two examples shows the need to manually verify particular aspects of the algorithms used. In the first scenario, relating to the analysis of image media, both DF_{mouth} and LipForensics act outside their intended use. While Lipforensics cannot be used, DF_{mouth} would only analyse one face in its default configuration. By adding the **Analysis of visual content** in the Initial Assessment the investigation can be extended to both faces shown in the image. This also provides the insight that both faces show a closed mouth, which differs from the results of DF_{mouth}. Consequently, the joining of multiple methods leads to inconsistencies being unveiled, which remain hidden when viewed in isolation. The same applies to the second scenario, in which a video is the subject of the investigation. In particular, information from the **Auxiliary data analysis** as a result of the Initial Assessment can be applied. It includes the usage for feature engineering and normalization as shown in PS-$mouth_4$ and PS-$mouth_5$. Furthermore, PS-$mouth_4$ states (see Table IV and Figure 5), that spatial traces can be utilized in the temporal context as well. In addition, the **Auxiliary data analysis** can be further used as a selection strategy for single-image approaches. This applies in particular to the individual frame types in the video (ID-$ffmpeg_3$). However, the DeepFake could not be identified with certainty in either scenario. Instead, the manual review of the results revealed inconsistencies that indicate changes to the media. This highlights the importance of the transparency and interpretability of the algorithms used, which is required for the forensic examiner to comprehend the results.

In contrast to our previous work in [1], the applicability of the proposed extended BPM-DI has been further improved by adding six tools to a total of nine. However, not all methods of the proposed model could be covered with the selected detectors. This shows that individual tools cannot and should not cover all methods. This is further supported by the findings in [46], where DF_{mouth} is only one of three modalities used for DeepFake detection. In relation to these modalities, ENFSI also provides a list of facial features in their "Best Practice

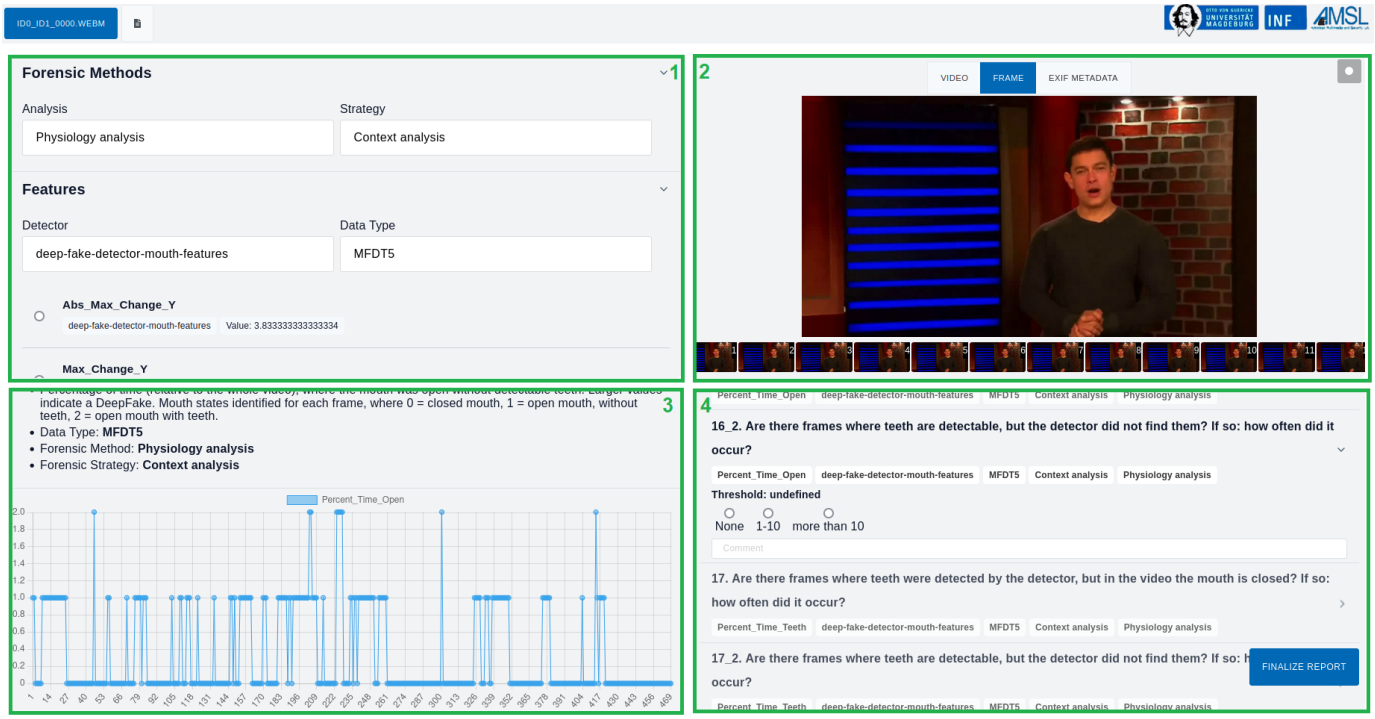

Figure 7. Demonstration of the extended Methods, exemplified on DF_{mouth} for video id0_id1_0000 of the Celeb-DF dataset - from a student project in the context of the lecture "Multimedia and Security", 2023 Department of Computer Science, Otto-von-Guericke-University of Magdeburg, expanded version of [1], further highlighting the individual areas of the interface.

Manual for Facial Image Comparison" [10] that can be used as a reference. A further extension can be derived from the findings in the SP phase, as the detectors can generally only detect certain types of DeepFakes, which is related to the specific traces of manipulation in the media. Lastly, it was also discussed that DeepFakes can occur in audio data, which is not specifically included in the extended model. For this purpose, there is the "Best Practice Manual for Digital Audio Authenticity Analysis" [53], which has to be addressed in the future.

Author Contributions and Acknowledgments

The work in this paper is funded in part by the German Federal Ministry of Education and Research (BMBF) under grant number FKZ: 13N15736 (project "Fake-ID"). Special thanks to team TraceMap, consisting of Stephan Haussmann, Hannes Hinniger, Tjark Homann and Malte Rathjens (a student project in the context of the lecture "Multimedia and Security", 2023 Department of Computer Science, Otto-von-Guericke-University of Magdeburg) for providing an initial demonstrator used as a basis for Figure 7.

Author Contributions: Initial idea & methodology: Jana Dittmann (JD), Christian Kraetzer (CK); Conceptualization: Dennis Siegel (DS); Modeling & application in the context of DeepFake: DS; Writing – original draft: DS; Writing – review & editing: CK, Stefan Seidlitz (StS), JD and DS.

References

[1] D. Siegel, C. Kraetzer, and J. Dittmann. "Joining of Data-driven Forensics and Multimedia Forensics for Deepfake Detection on the Example of Image and Video Data," *Proceedings of the SECURWARE 2023, The Seventeenth International Conference on Emerging Security Information, Systems and Technologies*, IARIA, 2023, pp. 43–51.

[2] S. Kiltz. "Data-centric examination approach (DCEA) for a qualitative determination of error, loss and uncertainty in digital and digitised forensics," *Ph. D. Thesis. Otto-von-Guericke-University Magdeburg, Fakultät für Informatik*, 2020.

[3] D. Siegel, C. Kraetzer, S. Seidlitz, and J. Dittmann. "Forensic data model for artificial intelligence based media forensics - illustrated on the example of DeepFake detection," *Electronic Imaging 34*, 2022, pp. 1–6.

[4] European Network of Forensic Science Institutes (ENFSI). "Best practice manual for digital image authentication," *ENFSI-BPM-DI-03*, 2021.

[5] C. Kraetzer, D. Siegel, S. Seidlitz and J. Dittmann. "Human-in-control and quality assurance aspects for a benchmarking framework for DeepFake detection models," in Electronic Imaging, 2023, pp. 379–1 - 379-6, https://doi.org/10.2352/EI.2023.35.4.MWSF-379.

[6] European Network of Forensic Science Institutes. "Best practice manual for the forensic examination of digital technology," *ENFSI-BPM-FIT-01*, 2015.

[7] M. Reith, C. Carr, and G. H. Gunsch. "An examination of digital forensic models," *Int. J. Digit. EVid. 1, 3*, 2002.

[8] R. Böhme, F. C. Freiling, T. Gloe, and M. Kirchner. "Multimedia forensics is not computer forensics," *Computational Forensics*, Springer, 2009, pp. 90–103.

[9] European Network of Forensic Science Institutes (ENFSI). "Best practice guidelines for ENF analysis in forensic authentication of digital evidence (BPM-ENF-001)," Technical Report, ENFSI, Wiesbaden, Germany, 2009.

[10] European Network of Forensic Science Institutes (ENFSI). "Best practice manual for facial image comparison," *ENFSI-BPM-DI-01*, 2018.

[11] National Institute of Justice. "Fingerprint Sourcebook," 2011.

[12] German Federal Office for Information Security (BSI). "Open source face image quality (OFIQ)," *online available at https://www.bsi.bund.de/dok/OFIQ-e*, last access: 2024-05-31.

[13] T. Schlett, C. Rathgeb, O. Henniger, J. Galbally, J. Fierrez, and C. Busch. "Face image quality assessment: a literature survey," *ACM Computing Surveys, Volume 54, Issue 10s*, 2022, pp. 1-49.

[14] European Network of Forensic Science Institutes (ENFSI). "Best practice manual for forensic image and video enhancement," *ENFSI-BPM-DI-02*, 2018.

[15] R. Altschaffel. "Computer forensics in cyber-physical systems : applying existing forensic knowledge and procedures from classical IT to automation and automotive," *Ph. D. Thesis. Otto-von-Guericke-University Magdeburg, Fakultät für Informatik*, 2020.

[16] Legal Information Institute. "Rule 702. testimony by expert witnesses," 2019 *online available at https://www.law.cornell.edu/rules/fre/rule_702*, last access: 2024-05-31.

[17] C. Champod and J. Vuille. "Scientific evidence in europe - admissibility, evaluation and equality of arms," *International Commentary on Evidence 9, 1*, 2011.

[18] European Commission. "Proposal for a regulation of the european parliament and of the council laying down harmonisd rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts," *COM/2021/206 final*, April,21 2021.

[19] European Parliament. "Amendments adopted by the european parliament on 14 june 2023 on the proposal for a regulation of the european parliament and of the council on laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts," *(COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))*, June, 14 2023.

[20] United Nations Interregional Crime and Justice Research Institute (UNICRI) and International Criminal Police Organization (INTERPOL). "Toolkit for responsible AI innovation in law enforcement: principles for responsible AI innovation," *online available at https://unicri.it/Publication/Toolkit-for-Responsible-AI-Innovation-in-Law-Enforcement-UNICRI-INTERPOL*, last access: 2024-05-31.

[21] National Institute of Standards and Technology (NIST). "Digital and multimedia evidence," *online available at https://www.nist.gov/spo/forensic-science-program/digital-and-multimedia-evidence*, last access: 2024-05-31.

[22] C. Kraetzer, D. Siegel, S. Seidlitz, and J. Dittmann. "Process-driven modelling of media forensic investigations - considerations on the example of deepfake detection," *Sensors 22, 9*, 2022.

[23] Y. Mirsky and W. Lee. "The creation and detection of deepfakes: a survey," *ACM Comput. Surv. 54, 1, Article 7*, 2021.

[24] Z. Akhtar. "Deepfakes generation and detection: a short survey," *Journal of Imaging 9, 1*, 2023

[25] Y. Li and S. Lyu. "Exposing deepfake videos by detecting face warping artifacts," In Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1-7.

[26] T. Karras, S. Laine, and T. Aila. "Training generative adversarial networks with limited data," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[27] T. Karras, S. Laine, T. Aila, J. Hellsten, J. Lethinen, and T. Aila. "Analyzing and improving the image quality of styleGAN," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[28] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lethinen, and T. Aila. "Training generative adversarial networks with limited data," *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020.

[29] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lethinen, and T. Aila. "A style-based generator architecture for generative adversarial networks," *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021.

[30] N. Yu, L. Davis, and M. Fritz. "Attributing fake images to GANs: learning and analyzing GAN fingerprints," Proceedings of the *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[31] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi. "Do GANs leave artificial fingerprints?," *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2019, pp. 506-511.

[32] B. Lorch, N. Scheler, and C. Riess. "Compliance challenges in forensic image analysis under the artificial intelligence act," *30th European Signal Processing Conference (EUSIPCO)*, 2022, pp. 613–617.

[33] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li. "Protecting world leaders against deep fakes," *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019 pp. 38–45.

[34] V. Conotter, E. Bodnari, G. Boato, and H. Farid. "Physiologically-based detection of computer generated faces in video," *IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 248–252.

[35] B. Hosler, D. Salvi, A. Murray, F. Antonacci, P. Bestagini, S. Tubaro, and M. C. Stamm. "Do deepfakes feel emotions? a semantic approach to detecting deepfakes via emotional inconsistencies," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021, pp. 1013–1022.

[36] T.-N. Le, H.H. Nguyen, J. Yamagishi, and I. Echizen. "OpenForensics: large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild," *International Conference on Computer Vision (ICCV)*, 2021.

[37] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu. "Celeb-df: a large-scage challenging dataset for deepfake forensics," *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3204–3213, doi:10.1109/CVPR42600.2020.00327.

[38] P. Harvey. "Exiftool," *online available at https://exiftool.org/*, last access: 2024-05-31.

[39] FFmpeg developers. "FFmpeg," *online available at https://ffmpeg.org/*, last access: 2024-05-31.

[40] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters 23(10)*, 2016, pp. 1499-1503.

[41] D. E. King. "Dlib-ml: a machine learning toolkit," *J. Mach. Learn. Res. 10*, 2009, pp. 1755–1758.

[42] S. Agarwal, H. Farid, O. Fried, and M. Agrawala. "Detecting deep-fake videos from phoneme-viseme mismatches," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2814–2822.

[43] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic. "Lips don't lie: a generalisable and robust approach to face forgery detection," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 5039-5049.

[44] F. Matern, C. Riess, and M. Stamminger. "Exploiting visual artifacts to expose deepfakes and face manipulations," *IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 2021, pp. 83–92.

[45] A. Elhassan, M. Al-Fawa'reh, M. T. Jafar, M. Ababneh, and S. T. Jafar. "DFT-MF: enhanced deepfake detection using mouth movement and transfer learning," *SoftwareX 19*, 2022.

[46] D. Siegel, C. Kraetzer, S. Seidlitz, and J. Dittmann. "Media forensics considerations on deepfake detection with hand-crafted features," *Journal of Imaging 7, 7*, 2021.

[47] M. A. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. "The weka data mining software: an update," SIGKDD Explor., 11(1):10–18, 2009.

[48] J. R. Quinlan. "C4.5: programs for machine learning," *Morgan Kaufmann Publishers Inc.*, San Francisco, CA, USA, 1993.

[49] P. Korshunov and S. Marcel. "Deepfakes: a new threat to face recognition? Assessment and detection," *arXiv preprint arXiv:1812.08685*, 2018.

[50] C. Sanderson and B. Lovell. "Multi-region probabilistic histograms for robust and scalable identity inference," *Lecture Notes in Computer Science (LNCS)*, 2009, pp. 199–208.

[51] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. "Faceforensics++: learning to detect manipulated facial images," *International Conference on Computer Vision (ICCV)*, 2019.

[52] Y. Wang, X. Chen, J. Zhu, W. Chu, Y. Tai, C. Wang, J. Li, Y. Wu, F. Huang, and R. Ji. "Hififace: 3d shape and semantic prior guided high fidelity face swapping," *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 2021, pp. 1136-1142.

[53] European Network of Forensic Science Institutes. "Best practice manual for digital audio authenticity analysis". *ENFSI-FSA-BPM-002*, 2018.

[54] D. Siegel and J. Dittmann. "TraceMap," Student project within the lecture of Multimedia and Security [MMSEC], Otto-von-Guericke-University Magdeburg, 2023, unpublished.

[55] German Federal Office for Information Security (BSI). "Leitfaden IT-forensik," *online available at https://www.bsi.bund.de/dok/6620610*, last access: 2024-05-31, available in German only, 2011.