# Joining of Data-driven Forensics and Multimedia Forensics for Deepfake Detection on the Example of 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
kraetzer@iti.cs.uni-magdeburg.de

Jana Dittmann
*Dept. of Computer Science*
*Otto-von-Guericke-University*
Magdeburg, Germany
jana.dittmann@iti.cs.uni-magdeburg.de

***Abstract*—DeepFake technology poses a new challenge to the validation of digital media integrity and authenticity. In contrast to 'traditional' forensic sub-disciplines (e.g., 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). The new methodology is applied to the example of DeepFake detection, utilizing three existing tools as methods. One for the Auxiliary data analysis and two DeepFake detectors based on hand-crafted and deep learning based feature spaces for Media content analysis are used. A total of 27 features were considered. In addition, the value types, ranges and their tendency for a DeepFake are determined for each feature. With the discussed potential extensions towards video evidence and machine learning, 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.**

***Index Terms*—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, 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)* [1], [2] and the *Best Practice Manual for Digital Image Authentication (BPM-DI)* from the *European Network of Forensic Science Institute (ENFSI)* [3].
- illustration of applicability and benefits of our concept on the example of three existing applications ExifTool [4], the hand-crafted DeepFake detector DF$_{mouth}$ [5] as well as the deep learning based DeepFake detector LipForensics [6].

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. 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 [7].

The paper is structured as follows. First, a brief overview of the state of the art on digital forensics, standards and regulations as well as implications on the topic of DeepFake 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 [1] and BPM-DI [3]. Which is then applied to the practical example of detecting DeepFakes using three different features spaces [5] as methods in Section IV. Finally, conclusions are drawn from the evaluation results presented and future directions are outlined in Section V.

## 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 there 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 [...]*" [8]. In [9] 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 *Best Practice Manuals (BPM)* and guidelines in forensics. In the field of digital imaging, there are three Best Practice Manuals. The first document addresses the aspect of forensic facial image comparison [10] and formulates the respective investigation steps. This comparison is conducted 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. The second BPM focuses on best practices of enhancement techniques for images and videos [11]. Here, approaches for enhancing image and video as well as strategies for selecting suitable frames are presented. This is also done on the basis of a human operator. 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)* [3]. 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 the German Federal Office for Information Security (BSI; the national cyber security authority) [38] are currently relevant. The data-centric examination approach (DCEA) is an extension of these guidelines. The DCEA 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 [1] or [12].

In terms of data types, the DCEA proposes a total of six for digital forensics and ten for digitized forensics. In [2], 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 [2].

### 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 [13]) 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 [14]. 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 [14]:

- "*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 [15]. At the current time, the proposal has been adjusted and approved by the European Parliament [16]. 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 [17]. Furthermore, the National Institute of Standards and Technology (NIST) currently develops a data set for DeepFake detection for validation of methods [18]. 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 [19]. 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 [16]. DeepFake detection methods can be divided into methods utilizing spatial and temporal feature spaces [20]. This classification goes hand in hand with the diversified creation of DeepFakes, which can take place on image, video and audio files. 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 prioritize forensic aspects in detection. In [21] 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 [3] is considered as a basis and extended for the case of DeepFake detection for video. To classify this further, it should be noted that [3] 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. An overview of the proposed extended BPM-DI can be found in Figure 1.

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 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 [3], 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 [3] 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 [20] dividing into **Spatial** and **Temporal content analysis**. **Methods** of **Spatial content analysis** correspond to BPM-DI [3] **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 [22] 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 [23] or eye blinking behavior [19]. **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 [24] or lip synchronization [25]. **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**,

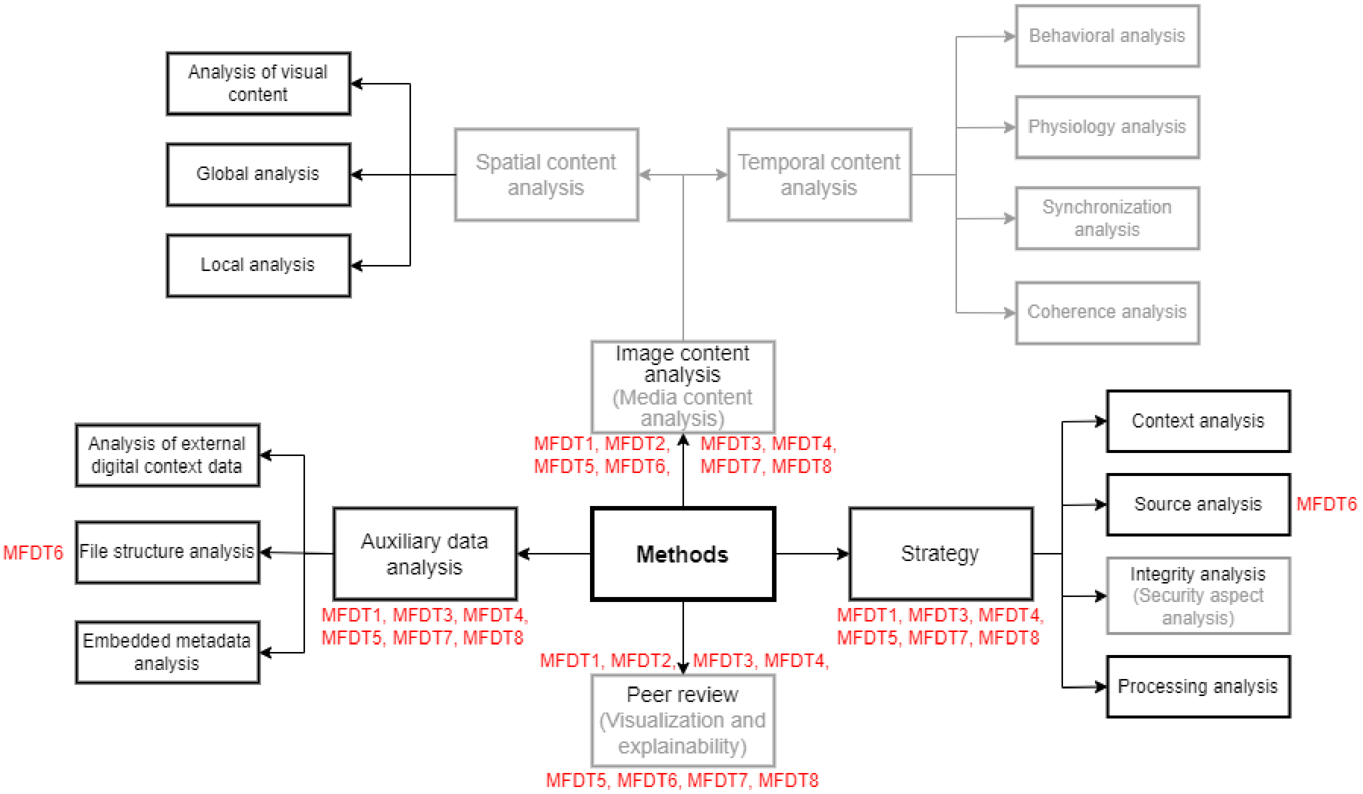

Behavioral analysis
Analysis of visual content
Physiology analysis
Global analysis
Spatial content analysis
Temporal content analysis
Synchronization analysis
Local analysis
Coherence analysis
Image content analysis (Media content analysis)
Context analysis
Analysis of external digital context data
MFDT1, MFDT2, MFDT5, MFDT6,
MFDT3, MFDT4, MFDT7, MFDT8
Source analysis
MFDT6
MFDT6
File structure analysis
Auxiliary data analysis
Methods
Strategy
Integrity analysis (Security aspect analysis)
Embedded metadata analysis
MFDT1, MFDT3, MFDT4, MFDT5, MFDT7, MFDT8
MFDT1, MFDT3, MFDT4, MFDT5, MFDT7, MFDT8
MFDT1, MFDT2, MFDT3, MFDT4,
Processing analysis
Peer review (Visualization and explainability)
MFDT5, MFDT6, MFDT7, MFDT8

Fig. 1. Categorization of forensic methods proposed in [3], 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.

based on the specific investigation goal. In this work, we consider three of the investigation goals of BPM-DI [3] 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 [15]). 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 [3] 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) [2] 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 [7] 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-B.

## 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. To cover a wide range of applications three existing tools of different categories are used. The first tool is ExifTool [4], which does not use any form of machine learning. ExifTool is an open source tool, which is able to read, write and edit metadata for a wide range of image and video formats. In addition, two existing machine learning based DeepFake detectors are used. To diversify the approaches, one based on hand-crafted features and one based on Deep Learning were chosen.

### *A. Semantics-driven DeepFake Detection Approaches as Methods in the Context of the Best Practice Manual*

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 [25] 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 [6] by identifying unnatural mouth movement. The second aspect utilizes the post processing, especially blurring, performed in DeepFake synthesis. In [26] and [27] texture analysis is performed on the mouth region to identify manipulations. A combination of both approaches is given in [5], 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 $DF_{mouth}$ [5] and LipForensics [6] are selected, representing both a hand-crafted as well as deep learning based detector.

### *B. Practical Application of the Extended Methods*

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 [3] for **Auxiliary data analysis** (shown in Figure 2), **Media content analysis** (shown in Figure 3) and **Strategies** (shown in Figure 4). The first step in verifying the authenticity of the media content under study is carried out using the methods of **Auxiliary data analysis**. For this purpose the open source tool ExifTool [4] is used. It is able to read, write and edit metadata for a wide range of image and video formats. In the context of this work, it is used for extracting the metadata (PS-exif). While a variety of entries are available in the metadata, a total of eight features (ID-exif$_n$) are selected for this exemplary approach and categorized according to the Ext. BPM-DI. These can be found in the top part of Table I. The first set of three 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$_7$) and frame rate if it is a video (ID-exif$_8$). 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$_6$).

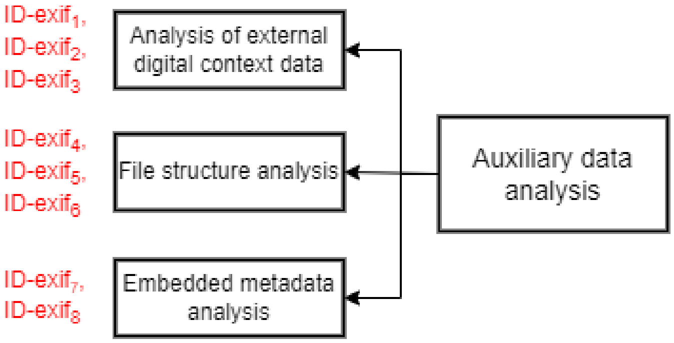


Fig. 2. Individual features extracted using ExifTool [4] (in red) categorized in the extended BPM-DI [3] for the category Auxiliary data analysis.

In terms of DeepFake detectors, 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 normalization. With their intention of identifying DeepFakes the general **Strategy** of application is **Integrity analysis**. Starting with DF$_{mouth}$, the detector is introduced in [5] and trained using the WEKA machine learning toolkit [28]. For the classification the decision tree classifier J48 [29] is used on the datasets Deepfake-TIMIT [30], [31], Celeb-DF [32] and DFD [33]. 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 [7] DF$_{mouth}$ is applied on a larger variety of DeepFake synthesis methods, including FaceForensics++ [33], DFD [33], Celeb-DF [32] and HiFiFace [34]. 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 [5].

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 [5] using dlib's 68 landmark detection model [35] 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**.

More details on the individual features, their description as well as the categorization in the forensic methods can be found in the middle part of Table I. 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 [19].

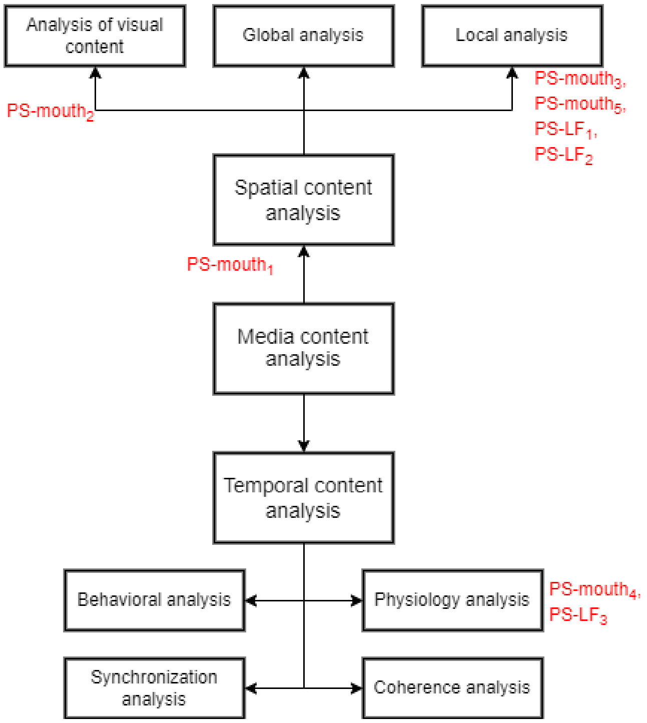


Fig. 3. Processing steps (PS, in red) for ExifTool [4] and the DeepFake detectors DF$_{mouth}$ [5] and LipForensics [6] categorized in the extended BPM-DI [3] for the category Media content analysis.

The second detector LipForensics [6] (herinafter refered 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**.

With the introduction of machine learning algorithms in

TABLE I

CATEGORIZATION OF EXIFTOOL [4] (TOP SECTION), DF$_{mouth}$ [5] (MIDDLE SECTION) AND LIPFORENSICS [6] (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.

| 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, ∞] | | | | |
| 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 |
| | | 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}<br>probability: real [0, 1] | PS-LF$_3$ | Physiology analysis | Processing analysis | MFDT5 |

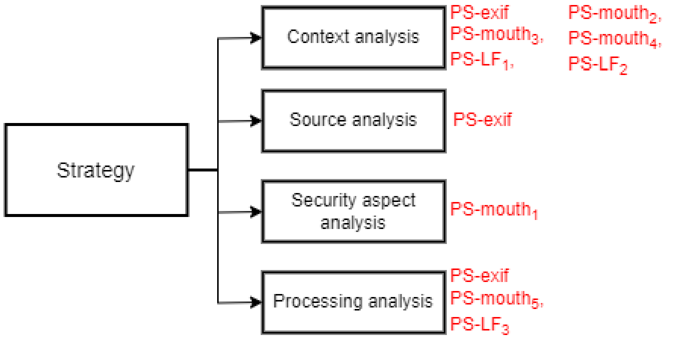
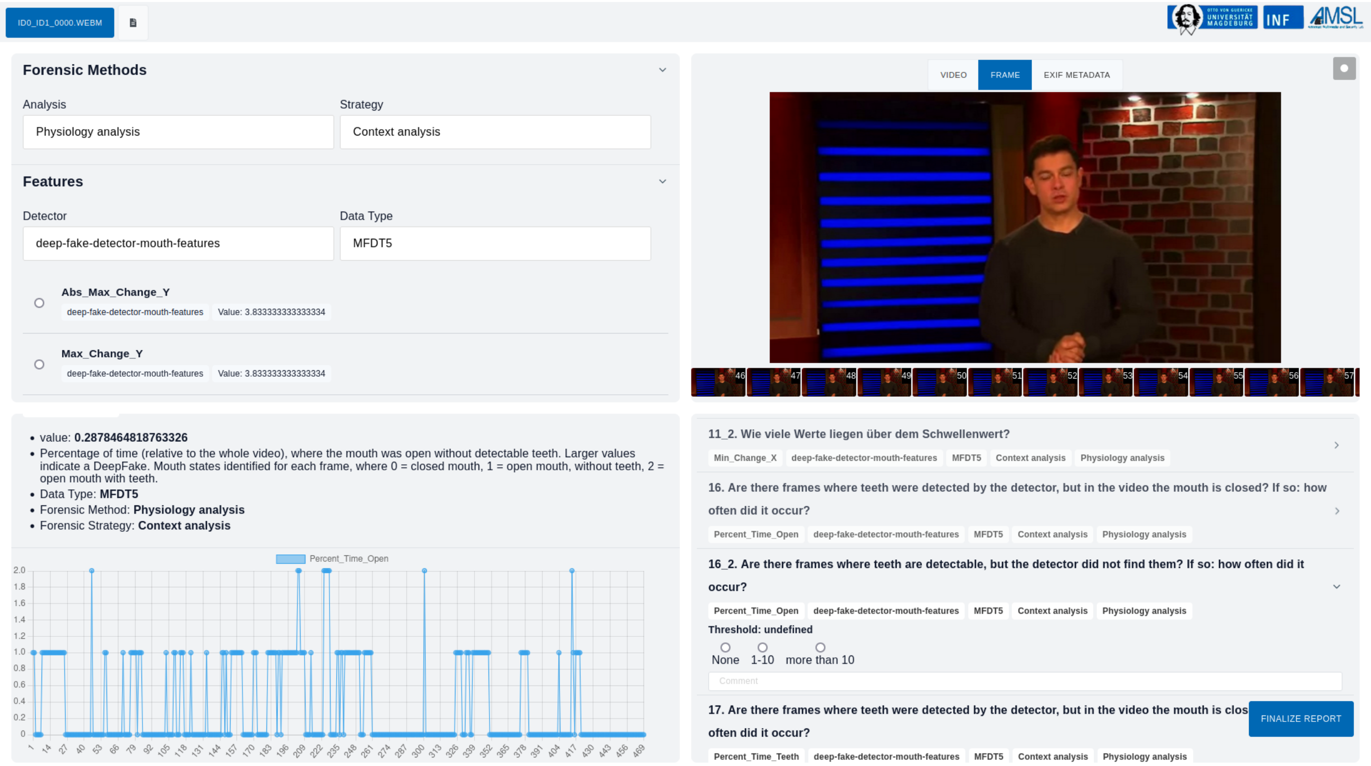


Fig. 4. Processing steps (PS, in red) for ExifTool [4] and the DeepFake detectors DF$_{mouth}$ [5] and LipForensics [6] categorized in the extended BPM-DI [3] for the category Strategy.

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 5 demonstrates a potential direction to enhance the Method of **Peer review** on the basis of DF$_{mouth}$ and ExifTool [37]. 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 5). 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 5) 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 5).
4) The last block (see the bottom right box of Figure 5) 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.

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



**Forensic Methods**

Analysis: Physiology analysis

Strategy: Context analysis

**Features**

Detector: deep-fake-detector-mouth-features

Data Type: MFDT5

○ **Abs_Max_Change_Y**
deep-fake-detector-mouth-features Value: 3.833333333333334

○ **Max_Change_Y**
deep-fake-detector-mouth-features Value: 3.833333333333334

- value: **0.2878464818763326**
- Percentage of time (relative to the whole video), where the mouth was open without detectable teeth. Larger values indicate a DeepFake. Mouth states identified for each frame, where 0 = closed mouth, 1 = open mouth, without teeth, 2 = open mouth with teeth.
- Data Type: **MFDT5**
- Forensic Method: **Physiology analysis**
- Forensic Strategy: **Context analysis**

**11_2. Wie viele Werte liegen über dem Schwellenwert?**
Min_Change_X deep-fake-detector-mouth-features MFDT5 Context analysis Physiology analysis

**16. Are there frames where teeth were detected by the detector, but in the video the mouth is closed? If so: how often did it occur?**
Percent_Time_Open deep-fake-detector-mouth-features MFDT5 Context analysis Physiology analysis

**16_2. Are there frames where teeth are detectable, but the detector did not find them? If so: how often did it occur?**
Percent_Time_Open deep-fake-detector-mouth-features MFDT5 Context analysis Physiology analysis
**Threshold: undefined**
○ None ○ 1-10 ○ more than 10
Comment

**17. Are there frames where teeth were detected by the detector, but in the video the mouth is clos often did it occur?**
Percent_Time_Teeth deep-fake-detector-mouth-features MFDT5 Context analysis Physiology analysis

Fig. 5. Demonstration of the extended Methods, exemplified on $\mathrm{DF}_{mouth}$ 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.

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 [14]. 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 [1], [2] in **Methods** of BPM-DI [3]. 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. Here, the aspects ‘human in the loop’ and ‘human in control’ as well as the topic ‘explainable AI’ represent important foundations for this component and will be further elaborated in a future paper.

The extended BPM-DI model is applied to the three existing applications ExifTool, the hand-crafted DeepFake detector $\mathrm{DF}_{mouth}$ as well as the deep learning based DeepFake detector LipForensics, showing its applicability to a wide range of approaches. In addition, it was found that the deep learning based features are too complex to achieve the same granularity as the detector $\mathrm{DF}_{mouth}$. 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. Moreover, the interplay of individual **Methods** have been identified. This includes the use of **Auxiliary data analysis** for feature engineering and normalization as shown in PS-$\mathrm{mouth}_4$ and PS-$\mathrm{mouth}_5$. Furthermore, PS-$\mathrm{mouth}_4$ states (see Table I and Figure 3), that spatial traces can be utilized in the temporal context as well.

Beyond that, 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 indicates that additional tools are needed for integration. Lastly, an even more detailed categorisation of methods can be explored. With regard to the ENFSI “Best Practice Manual for Facial Image Comparison”, the method of Local analysis could be split according the facial feature list [10]. 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" [36], 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 5.

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, JD and DS.

## References

[1] 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.

[2] D. Siegel, C. Kraetzer, S. Seidlitz, 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.

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

[4] P. Harvey, "Exiftool," https://exiftool.org/, 2016.

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

[6] A. Haliassos, K. Vougioukas, S. Petridis, 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.

[7] C. Kraetzer, D. Siegel, S. Seidlitz, 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.

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

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

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

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

[12] 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.

[13] Legal Information Institute, "Rule 702. testimony by expert witnesses," 2019.

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

[15] 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.

[16] 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.

[17] UNICRI, INTERPOL, "Toolkit for responsible AI innovation in law enforcement: principles for responsible AI innovation," June 2023.

[18] National Institute of Standards and Technology (NIST), "Digital and multimedia evidence,", https://www.nist.gov/spo/forensic-science-program/digital-and- multimedia-evidence, 2022.

[19] 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.

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

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

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

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

[24] B. Hosler, D. Salvi, A. Murray, F. Antonacci, P. Bestagini, S. Tubaro, 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.

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

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

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

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

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

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

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

[32] Y. Li, X. Yang, P. Sun, H. Qi, 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.

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

[34] Y. Wang, X. Chen, J. Zhu, W. Chu, Y. Tai, C. Wang, J. Li, Y. Wu, F. Huang, 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.

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

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

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

[38] German Federal Office for Information Security (BSI), "Leitfaden IT-forensik," https://www.bsi.bund.de/dok/6620610 available in German only, 2011.