https://doi.org/10.2352/EI.2023.35.4.MWSF-379

Human-in-control and quality assurance aspects for a benchmarking framework for DeepFake detection models

Christian Kraetzer[1]***, Dennis Siegel***[1]***, Stefan Seidlitz***[1]***, Jana Dittmann***[1]

[1] ***Otto-von-Guericke University, Magdeburg, Germany***

Abstract

Human-in-control is a principle that has long been established in forensics as a strict requirement and is nowadays also receiving more and more attention in many other fields of application where artificial intelligence (AI) is used. This renewed interest is due to the fact that many regulations (among others the the EU Artificial Intelligence Act (AIA)) emphasize it as a necessity for any critical AI application scenario. In this paper, human-in-control and quality assurance aspects for a benchmarking framework to be used in media forensics are discussed and their usage is illustrated in the context of the media forensics sub-discipline of DeepFake detection.

Introduction and Motivation

Intended courtroom usage of forensic methods requires standardized investigation and analysis procedures that underwent quality assurance as well as standardization prior to application to case work. Internationally accepted best practices governing this field are e.g. the United States Federal Rules of Evidence (FRE; esp. FRE 702, see [18]) and the Daubert standard in the US.
Authors like Champod et al. point out that, even if the Daubert standard is only directly legally binding for court proceedings on US federal level, they are also in many other countries worldwide considered as a best practice for evaluation of the degree of maturity of forensic methods as basis for expert testimonies intended to be used in court (see e.g. [3], where the influence of the Daubert standard on the evaluation and admissibility of scientific evidence in Europe is discussed).
Within this paper focusing on the benchmarking of media forensic methods, especially the following three (out of five) Daubert criteria are relevant ([3]):

- "*whether the expert's technique or theory can be or has been tested – that is, whether the expert's theory can be challenged in some objective sense, or whether it is instead simply a subjective, conclusory approach that cannot reasonably be assessed for reliability*"
- "*the known or potential rate of error of the technique or theory when applied*"
- "*the existence and maintenance of standards and controls*"

Especially the second and the last of the criteria quoted above are of importance within this context, because they imply on one hand a strong need for process modeling as foundation of work in standardization and on the other hand require benchmarking work to allow to suitably measure or estimate the potential rate of error of the method when applied in practice.

Many process models for forensic processes exist for 'traditional' forensic sub-disciplines (e.g. dactyloskopy), with the intended purpose of making corresponding investigations fit for courtroom usage. What they usually have in common is that they define standards for application of methods and requirements for the certification of practitioners, strictly putting an expert operator in control of the investigation, leading to an expert testimony in court. Most media forensic approaches today still lack maturity in this regard because the focus here currently lies mostly only on proposing AI detectors for specific forensic tasks, like image manipulation detection or DeepFake detection, neglecting most of the necessary modeling, benchmarking and standardization work required to make such approached mature enough for court room appearance.
This gap (i.e., the lack of required domain specific process modeling and benchmarking work) is addressed in this paper in part by the following contributions in this paper:

- An extension of existing modeling work on domain specific process models for media forensic investigations (here illustrated on the example of DeepFake detection), to include human-in-the-loop and human-in-control aspects as requested by changing requirements/legislation worldwide, esp. the upcoming EU Artificial Inteligence Act (AIA).
- An empirical estimation of the generalization power (or lack there-off) of pre-existing DeepFake detectors in intra and inter data set benchmarking, using different data selection strategies and classifiers.
- Initial tests on 2- vs. multi-class modeling of the decision problem, showing interesting results for the potential attribution / identification of the used DeepFake synthesis method.

The rest of the paper is structured as follows: First, a very brief overview over the current state of the art on domain specific process modeling for media forensics in Europe and Germany is given. The following section presents the modeling work in this paper, extending an existing Data-Centric Examination Approach for Incident Response- and Forensics Process Modeling (DCEA) by including quality assurance aspects for a benchmarking framework for DeepFake detection models. Based on this modeling work, the core part of this paper presents empirical benchmarking activities on the example case of DeepFake detection, describing the setup and results for performance benchmarking for various DeepFake detection models compared in the same framework. The paper closes with conclusions and a summary of perspectives for potential future work.

Domain specific process modeling for media forensics in Europe and Germany

The most recent best practice document for media forensics in Europe is, at the time of writing this paper, the European Network of Forensic Science Institutes (ENFSI) Best Practice Manual (BPM) for Digital Image Authentication [8]. In its own words it "*aims to provide a framework for procedures, quality principles, training processes and approaches to the forensic examination*" and is intended "*to establish and maintain working practices in the field of forensic Image Authentication* (IA) *that will: deliver reliable results, maximize the quality of the information obtained and produce robust evidence. The use of consistent methodology and the production of more comparable results will facilitate interchange of data between laboratories*." It generalizes a workflow for an image authentications examination and provides a classification scheme for methods for digital image authentication but insists that it "*is not a standard operating procedure (SOP) and addresses the requirements of the judicial systems in general terms only*" [8].
The reason why the ENFSI BPM does not intend to be a standard operating procedure or a forensic process model as basis for standardization purposes is, that such processes are governed by national law and ENFSI has no directive authority in Europe. Here, national regulation would be required to define the precise legal context for any media forensic investigation and the usage of the corresponding results in court.

Regarding the German situation, which is relevant for the authors of this paper, the most relevant best practice document regarding IT forensics in general (incl. media forensics) is the BSI (German Federal Office for Information Security) guide on IT forensics [2] (German: "*Leitfaden IT-Forensik*"). It provides various means for modeling forensic processes, including the definition of a generic phase-driven investigation & reporting model, a basic data model and a classification of methods and tools. Like many other best practice documents in this field it covers basic investigation principles, process models, forensic data types, etc. but does not provide domain specific process models and guidelines for specific media forensic investigations such as DeepFake detection. Here, existing research, such as the latest extension to the BSI guidelines [2] described as the Data-Centric Examination Approach for Incident Response- and Forensics Process Modeling (DCEA) summarized in [14] and [25], is used as basis for extending the scope of these guidelines to achieve a higher degree of maturity for the state of the art in taylor-made models for media forensics (incl. DeepFake detection).

The core of 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 the importance of the SP has to be pointed out, since it is the phase that also includes all standardization, benchmarking, certification and training activities considered in this paper. For details on the phase model the reader is referred, e.g. to [14] or [1].

The second core aspect of DCEA is the definition of *forensic method classes* as presented in [14]. The third aspect is the specification of *forensically relevant data types*. They can be summarized as: MFDT1 "digital input data" (the initial media data considered for the investigation), MFDT2 "processed media data" (results of transformations to media data), MFDT3 "contextual data" (case specific information (e.g. for fairness evaluation)), MFDT4 "parameter data" (contain settings and other parameter used for acquisition, investigation and analysis), MFDT5 "examination data" (including the traces, patterns, anomalies, etc that lead to an examination result), MFDT6 "model data" (describe trained model data (e.g. face detection and model classification data)), MFDT7 "log data" (data, which is relevant for the administration of the system (e.g. system logs)), and MFDT8 "chain of custody & report data" (describe data used to ensure integrity and authenticity (e.g. hashes and time stamps) as well as the accompanying documentation for the final report).

In general, each processing operation (or operator) in an DCEA process pipeline is considered here as an atomar processing black box component with an identifier and (usually) a description of the processing performed in this operation. Each component has four well defined connectors: *input*, *output*, *parameters* and *log data* (see figure 1). To pay respects to the particularities of this field and make the following modeling task easier, a fifth connector is defined within this paper for a specific type of operator which requires a knowledge representation or a model for its processing operation. In that case, this fifth connector is labeled *model*. A detailed description of the modeling of these operators is given in [25].

The focus of the proposed extensions of the DCEA lies in this paper on the integration of the human operator into the procedures. Human-in-control is an principle that has long been established in forensics as a strict requirement and is nowadays also receiving more and more attention in any field of application where artificial intelligence (AI) is used. Among other regulations, the EU Artificial Intelligence Act (AIA, [7]) emphasizes it as a necessity for any critical application scenario. This extension is shown in figure 1 where two human operators are added to the component: One (labeled 'HO') as operator in control of the functionalities of the component and another one (labeled 'Sys admin') in the loop on the infrastructure, analyzing the system logs (MFDT7) and reacting to potential technical events such as an hard disc failure, etc.

Example case: Quality assurance aspects for a benchmarking framework for DeepFake detection models

Depending on the actual position of the component in a forensic investigation pipeline, the human operator (HO) in control could be a someone defining in-house quality assurance strategies (e.g. human operator 'HO1 ' in figure 2), a media forensics expert performing explainable AI (xAI) tasks in the used feature space ('HO 2' in figure 2) or a data scientist at a standardiza-

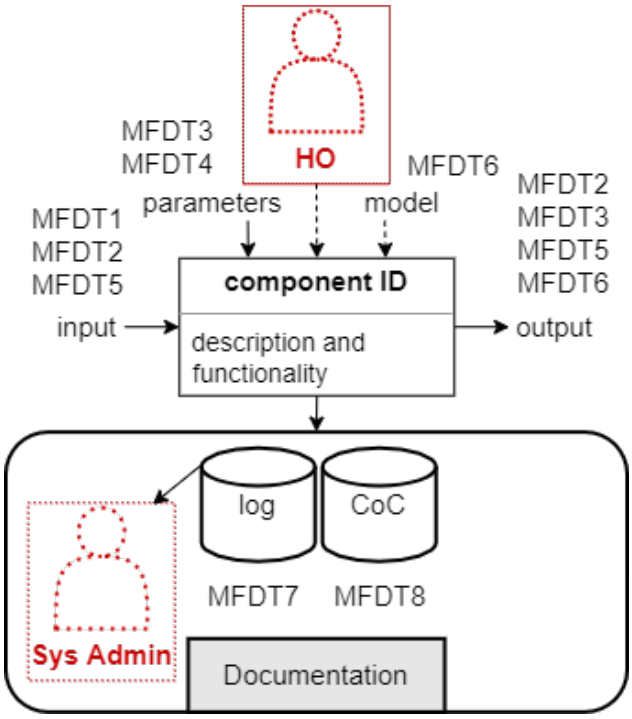

Figure 1. *Template structure integrating the human operator(s) (HO) highlighted in red.*

tion body like NIST running a benchmark and performing certification of the model trained ('HO 3' in figure 2). Obviously, all these different example HO would need different expertise and might have conflicting intentions.

The empirical evaluations performed in this paper focus on the interplay between 'HO 1', 'HO 2' and 'HO 3' in figure 2. Their interaction represents the cycle of decision model development (or training), its benchmarking and reasoning on the obtained results. For the performed evaluations in DeepFake detection, the following evaluation goals are defined:

- Estimation of generalization power (or lack there-off) in intra and inter data set evaluations, using different data selection strategies and classifiers
- Initial discussion on 2- vs. n-class classification (where n is the number of DeepFake synthesis methods plus one class for original, non-modified videos)
- Impact of data augmentation in training (model robustness)
- First considerations on video post-processing operations as potential counter-forensics

Evaluation setup

The evaluation setup is build according to the process model and the evaluation goals discussed in the previous section. Its general purpose is to provide an evaluation framework for DeepFake detection models, based on suitable DeepFake data sets. The video selection is done for each data set, where the selected number of videos corresponds to the minimal size of all data sets given. The extracted source data is augmented using different augmentation methods. All videos are processed in feature extractors introduced in [24] to classify DeepFakes based on eye (DF_{eye}), mouth (DF_{mouth}) and image foreground ($DF_{foreground}$) regions respectively. In addition, meta data is gathered to enable a human operator (here 'HO 1') to further curate the data. The extracted feature lists are then split into distinct training and test data for all model generation and benchmark strategies. This separation is further used to enable different evaluation scenarios, such as intra and inter data set evaluations.

Benchmarking data set selection

Previous experiments given in [16] have shown that early DeepFake video data sets, such as TIMIT-DF [23, 15], show visible flaws in the videos, making them unsuitable for a fair benchmarking of detectors. Therefore, a manual curation and evaluation of data sets to be used is performed. FaceForensics++ [21, 22] is another early data set, that includes various DeepFake synthesis methods, but also got a recent extension in HiFiFace [27]. Initially, DeeperForensics [11] was included into the data pool to be used in this paper as an augmented data set based on FaceForensics++, but was then replaced by in-house augmentation for comparability reasons. The DeepFake Detection data set (DFD) by Google and JigSaw [6] is available as a part of FaceForensics++, providing both additional real videos as well as the output of a DeepFake synthesis method. Celeb-DF [19] is large data set, using an Autoencoder for synthesis. Furthermore, FakeAVCeleb [13] was originally considered for usage in this paper, due to the fact that it also includes audio data and a labeling of ethical background and gender, but it was dropped due to a low resolution of 224x224. In table 1 a summary of selected data sets can be found.

Data augmentation

Based on the selected data sets an equal amount of 363[1] videos per subset of each data set are taken for evaluation. The selection occurs pseudo-random based on a seed (here, the randomly chosen value 7 is taken as seed). To further augment the data sets and simulate a less optimal training scenario, the selected videos undergo two different post-processing operations: One additional data set is generated by re-sampling the videos to 15 frames per second, a second data set is created by resizing them to a width of 480 pixels while keeping the aspect ratio to prevent distortion. This augmentation is done using the FFmpeg library [9]. In total, 7986 videos (363 + 7*3*363) are used in this benchmark.

For classification, a total of five different classifiers from WEKA [10] are selected to represent a variety of different algorithms. These are NaiveBayes [12], LibSVM [4], Simple Logistics [17, 26], JRIP [5] and J48 [20].

To ensure the distinct split of training and test data two different approaches are taken. The first one utilizes methods built into WEKA, which includes a 66% training 34% testing percentage split, as well as 3-, 5- and 10-fold stratified cross-validation. The second approach involves manual pre-processing and dividing of the samples in fixed splits. This allows for more precise grouping of the data and thus enables addressing of specific evaluation questions. For reproducibility, the splits occur pseudo-randomly by using a deterministic script with a seed (again the value 7 is

[1]The number of files in the smallest set used (here 'DFD-actors') defines the size of the subsets drawn from all other data sets used, to ensure equally sized representations in training and testing.

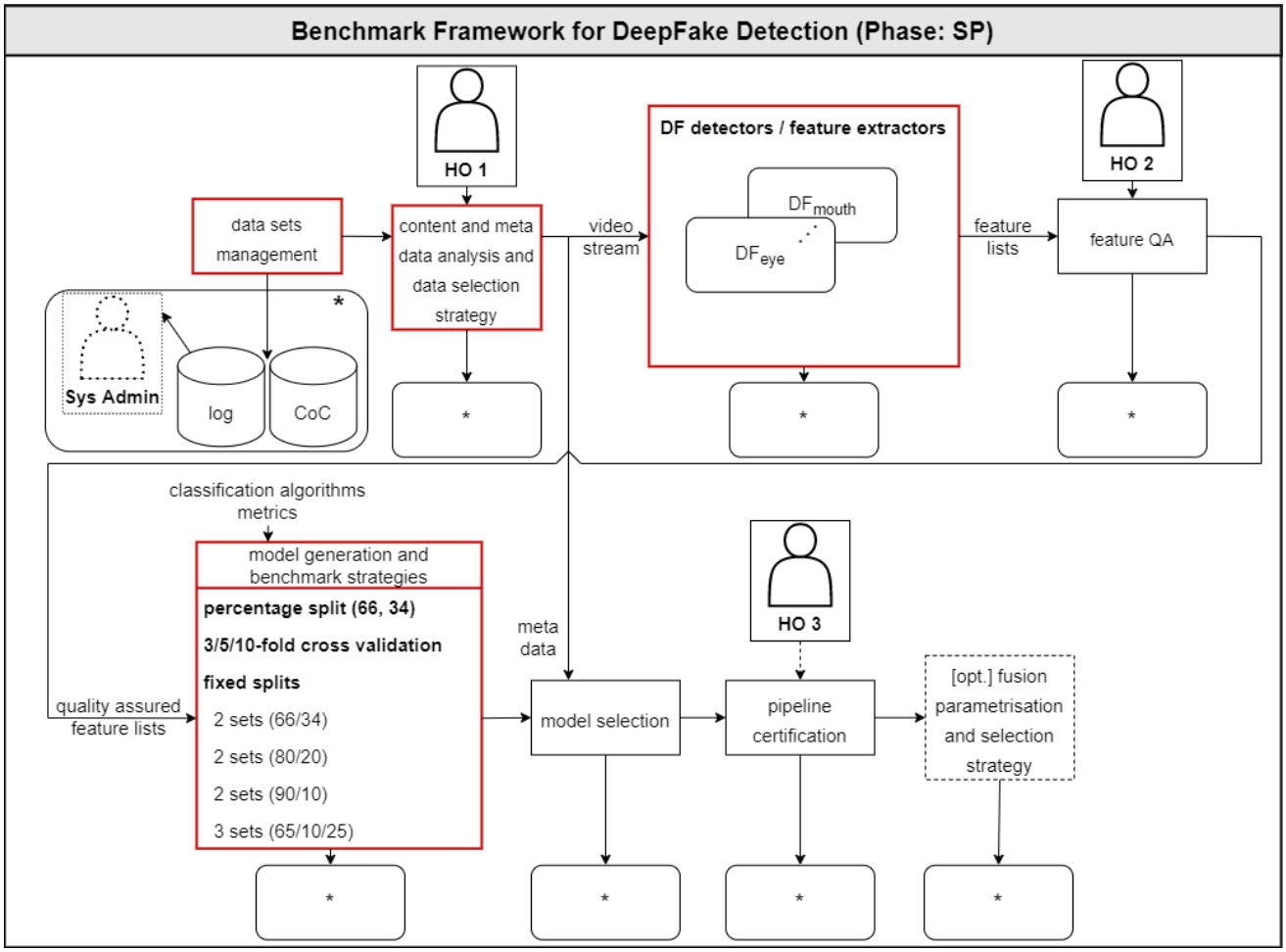

Figure 2. *Illustration of the DeepFake detection pipeline created as a template in the forensic process model phase of Strategic Preparation (SP), with the inclusion of human operators (HO) implementing human-in-control as well as human-in-the-loop (for 'Sys Admin'). Contribution is highlighted in red.*

data set	# individuals	# real video	# DeepFake video	subset	# selected videos
FaceForensics++ [21, 22]	?[2]	1 000	4 000	youtube-real	363
				Face2Face	363
				FaceShifter	363
				NeuralTexture	363
DFD [6]	28	363	3 068	DFD-actors (real)	363
HiFiFace [27][1]	?[2]	0	1 000	FaceSwap	363
Celeb-DF [19]	59	890	5 639	Celeb-real	363
				Celeb-synthesis	363

Overview of the data sets used in this paper for benchmarking of DeepFake detection models.

[1]: Based on the youtube-real subset of FaceForensics++.

[2]: Numbers correspond, but unfortunately the exact number have not been disclosed by the original authors.

taken). Using this script, disjointed training and testing splits of 66%/34%, 80%/20% and 90%/10% are generated automatically.

Evaluation results

As discussed previously, the evaluation is done in multiple individual experiments. In the first experiment the evaluation aims at different model generation and benchmark strategies, using the non-augmented data for evaluation. With the consideration of all three detectors DF_{eye}, DF_{mouth} and $DF_{foreground}$ the same tendencies of classification can be found, with some small exceptions. Figure 3 displays the results on the example of DF_{eye}. In general, it can be said, that there are almost no differences between 3-, 5- and 10-fold cross validation in this benchmark. In terms of predefined splits, an increase in detection performances can be found with increasing training data set size. This comes with an exception for the J48 classifier on the detectors DF_{eye}, where smaller training splits yield higher detection performance on the test set, indicating generalization problems (here in the handling of outliers in the test data) for this setup. Besides this mall glitch in the performance of J48, none of the tested classifiers is signifi-

cantly better than the others. Each of the detection approaches had a different classifier scoring best, in all cases achieved on the 90/10 fixed split. LibSVM for DF_{eye} (Kappa=0.4991), J48 for DF_{mouth} (0.4113) and Simple Logistic for $DF_{foreground}$ (0.3620). The Kappa statistics of DF_{mouth} are in the range of 0.2544-0.4113, showing a significant drop in performance compared to previous results in [24]. This might suggest that anomalies in the mouth region are data set specific and do not occur to the same extent as in the previous experiment. Same can be said for $DF_{foreground}$ ranging Kappa values from 0.1810 to 0.3620, showing lower but less fluctuating performances.

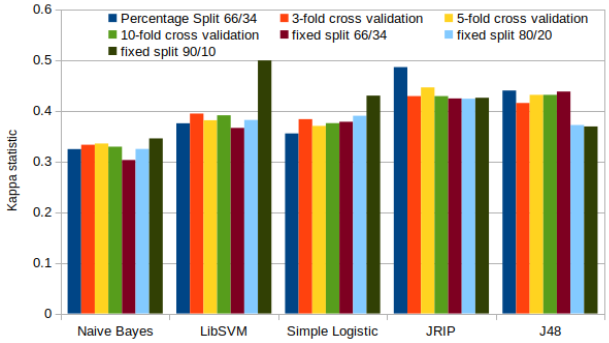

Figure 3. *Detection performances (Kappa values) for different model generation and benchmark strategies, on the example of* DF_{eye}

The second experiment addresses the usage of augmentation strategies in both training and testing. For this purpose, the data set is divided into native and augmented videos. Independently of the detector, augmentation usage solely for training or testing, results in a drop of detection performance. But it also has to be noted, that the integration of augmentation strategies in both training and testing did not impact the detectors negatively, and even increased the performance of DF_{mouth} and $DF_{foreground}$ (see the corresponding table).

The third experiment focuses on the impact of different DeepFake synthesis methods and considers every method as an individual class. Based on the considered data sets this results in a 6-class classification problem, which is then back projected to 2-class ('original' vs. 'DeepFake') for direct comparison. In terms of individual synthesis methods, it turned out that HiFiFace is clearly different from the others, especially for DF_{eye}. Here, none of the other types is classified as HiFiFace and also the HiFiFace subset is solely classified as 'real' or 'HiFiFace'. This suggests that more recent DeepFakes show less flaws in creation, here on the case of eye region and blinking specifically. This distinction is not found for DF_{mouth} and $DF_{foreground}$. However, considering the results, the separation does not show an improvement in detection performance in any detector compared to a 2-class classification. Nevertheless, it allows for an attribution / identification of the used synthesis method and therefore for a better justification of the decision made by using this model.

detector	2-class	6-class
DF_{eye}	73.55% (0.4312)	72.73% (0.4348)
DF_{mouth}	69.90% (0.3347)	67.60% (0.3186)
$DF_{foreground}$	71.83% (0.3199)	62.19% (0.2228)

Comparison of 2- and 6-class DeepFake detection.

Summary, Conclusions and Future Work

Summarizing the empirical results presented in this paper, it is shown that the promising results previously shown in [24] are not reliable (i.e., not generalizing well) when properly benchmarked: The extension of the data considered (in different evaluation scenarios) shows challenges in generalization power, an important lesson learned regarding human-in-control and QA aspects, highlighting the relevance of benchmarking for data selection as well as feature and decision model quality assurance.
First tests on 2- vs. multi-class modeling of the decision problem show interesting initial results for the potential attribution / identification of the used DeepFake synthesis method.
Important future work would be to extend the introduced benchmarking framework to include additional datasets to cover an even wider range of DeepFake synthesis methods and also more different sets of 'genuine' (non-DeepFake) samples with different preprocessing histories. In this regard, the first data augmentation tests discussed here could be a suitable starting point for creating more robust detector models. Extensions along this line could e.g. use the DeeperForensics data set (with its augmentations) as an extension of FaceForensics++.
Besides the generalization issue, also the closely related question of training bias and fairness has to be considered in future work, potentially with evaluations using the FakeAVCeleb data set with its metadata annotations (incl. among other characteristics an indication on the ethical background of the person in the video).
From the perspective of potential courtroom fitness, an important future step would be to find a independent and trustworthy third party like NIST in the US or the BSI in Germany that could be motivated to perform independent benchmarking (and potentially also certification) of methods and trained models.

Acknowledgements

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").
Author Contributions: Initial idea & methodology: Jana Dittmann (JD) and Christian Kraetzer (CK); Conceptualization: Christian Kraetzer (CK), Dennis Siegel (DS), Stefan Seidlitz (StS) and JD; Re-modelling of the process model components: CK, DS, StS; Empirical evaluations: DS; Writing – original draft: CK; Writing – review & editing: DS, StS and JD.

training data set →	no augmentation (no aug)		with augmentation (w aug)		combination of both
↓ detector test data set →	no aug	w aug	no aug	w aug	for train and test
DF_{eye}	73.55% (0.4312)	57.72% (0.1155)	65.25% (0.1931)	70.42% (0.1919)	72.38% (0.3643)
DF_{mouth}	69.90% (0.3347)	69.86% (0.1089)	71.25% (0.3239)	70.09% (0.1558)	70.94% (0.2489)
$DF_{foreground}$	71.83% (0.3199)	70.58% (0.1109)	71.11% (0.2814)	70.11% (0.0611)	71.50% (0.2080)

Evaluation results for augmentation strategies. All values are determined using J48 under default parameters.

References

[1] Robert Altschaffel. *Computer forensics in cyber-physical systems : applying existing forensic knowledge and procedures from classical IT to automation and automotive*. PhD thesis, Otto-von-Guericke-Universität Magdeburg, Fakultät für Informatik, 2020.

[2] BSI. *Leitfaden IT-Forensik*. German Federal Office for Information Security, 2011.

[3] Christophe Champod and Joëlle Vuille. Scientific evidence in europe - admissibility, evaluation and equality of arms. *International Commentary on Evidence*, 9(1), 2011.

[4] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[5] William W. Cohen. Fast effective rule induction. In *Twelfth International Conference on Machine Learning*, pages 115–123. Morgan Kaufmann, 1995.

[6] Nick Dufour and Andrew Gully. Contributing Data to Deepfake Detection Research. `https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html`, September, 24 2019. Accessed: 09/09/2021.

[7] European Commission. Proposal for a Regulation of the european parliament and of the council Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. *COM(2021) 206 final*, April, 21 2021. [Online]. Available at: `https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX\%3A52021PC0206` [Last retrieved: 14.09.2021].

[8] European Network of Forensic Science Institutes. Best practice manual for digital image authentication. *ENFSI-BPM-DI-03*, October 2021. [Online]. Available at: `https://enfsi.eu/wp-content/uploads/2022/12/1.-BPM_Image-Authentication_ENFSI-BPM-DI-03-1.pdf` [Last retrieved: 12.01.2023].

[9] FFmpeg. FFmpeg, 2018. [Online]. Available at: `https://ffmpeg.org/` [Last retrieved: 12.01.2023].

[10] Mark A. Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *SIGKDD Explor.*, 11(1):10–18, 2009.

[11] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 2886–2895. Computer Vision Foundation / IEEE, 2020.

[12] George H. John and Pat Langley. Estimating continuous distributions in bayesian classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, San Mateo, 1995. Morgan Kaufmann.

[13] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset, 2021.

[14] Stefan Kiltz. *Data-Centric Examination Approach (DCEA) for a qualitative determination of error, loss and uncertainty in digital and digitised forensics*. PhD thesis, Otto-von-Guericke-Universität Magdeburg, Fakultät für Informatik, 2020.

[15] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *CoRR*, abs/1812.08685, 2018.

[16] Christian Kraetzer, Dennis Siegel, Stefan Seidlitz, and Jana Dittmann. Process-driven modelling of media forensic investigations-considerations on the example of deepfake detection. *Sensors*, 22(9), 2022.

[17] Niels Landwehr, Mark Hall, and Eibe Frank. Logistic model trees. 95(1-2):161–205, 2005.

[18] Legal Information Institute. Rule 702. testimony by expert witnesses, Dec 2019.

[19] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 3204–3213. IEEE, 2020.

[20] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[21] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics: A large-scale video dataset for forgery detection in human faces. *arXiv*, 2018.

[22] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 1–11. IEEE, 2019.

[23] Conrad Sanderson and Brian Lovell. Multi-region probabilistic histograms for robust and scalable identity inference. *LNCS*, 5558:199–208, 2009.

[24] Dennis Siegel, Christian Kraetzer, Stefan Seidlitz, and Jana Dittmann. Media forensics considerations on deepfake detection with hand-crafted features. *Journal of Imaging*, 7(7), 2021.

[25] Dennis Siegel, Christian Krätzer, Stefan Seidlitz, and Jana Dittmann. Forensic data model for artificial intelligence based media forensics-illustrated on the example of deepfake detection. *Electronic Imaging*, 34:1–6, 2022.

[26] Marc Sumner, Eibe Frank, and Mark Hall. Speeding up logistic model tree induction. In *9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 675–683. Springer, 2005.

[27] Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Hififace: 3d shape and semantic prior guided high fidelity face swapping. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 1136–1142. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track.

Author Biography

Jana Dittmann *is a Professor on multimedia and security at the University of Otto-von-Guericke University Magdeburg (OvGU). She is the leader of the Advanced Multimedia and Security Lab (AMSL) at OvGU, which is partner in national and international research projects and has a wide variety of well recognized publications in IT security.* ***Christian Kraetzer*** *is a post-doc researcher and* ***Dennis Siegel*** *as well as* ***Stefan Seidlitz*** *are PhD students at AMSL.*