

D4.5

Intermediate toolset for robust, explainable, fair, and privacy-preserving AI

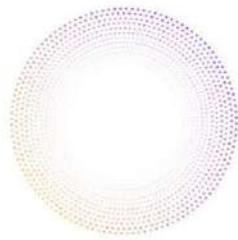
Project Title	AI4Media – A European Excellence Centre for Media, Society and Democracy
Contract No.	951911
Instrument	Research and Innovation Action
Thematic Priority	H2020-EU.2.1.1. - INDUSTRIAL LEADERSHIP - Leadership in enabling and industrial technologies - Information and Communication Technologies (ICT) / ICT-48-2020 - Towards a vibrant European network of AI excellence centres
Start of Project	1 September 2020
Duration	48 months



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

info@ai4media.eu

www.ai4media.eu



Deliverable title	Intermediate toolset for robust, explainable, fair, and privacy-preserving AI
Deliverable number	D4.5
Deliverable version	1.0
Previous version(s)	N/A
Contractual date of delivery	August 31st 2023
Actual date of delivery	September 13th 2023
Deliverable filename	AI4Media Deliverable D4.5
Nature of deliverable	Report
Dissemination level	Public
Number of pages	151
Work Package	WP4
Task(s)	T4.2, T4.3, T4.4, T4.5
Partner responsible	IBM
Author(s)	Anisa Halimi, Naoise Holohan, Ambrish Rawat, Stefano Braghin, Kieran Fraser (IBM), Nicu Sebe (UNITN), Daniel Gatica-Perez, Sina Sajadmanesh, Maya Guido (IDIAP), Hervé Le Borgne (CEA), Vasileios Mygdalis (AUTH), Thomas Köllmer (FhG-IDMT), Christoforos Papastergiopoulos, Vasileios Mezaris (CERTH), Mara Graziani (HES-SO), Lorenzo Seidenari (UNIFI), Gianluigi Lopardo, Gabriele Ciravegna, Frederic Precioso, Damien Garreau (3IA-UCA)
Editor	Anisa Halimi and Naoise Holohan (IBM)
Officer	Evangelia Markidou

Abstract	This deliverable presents the second collection of work and outcomes from WP4 in AI4Media, focusing on Trustworthy AI. The document describes the continuing investigations and results of our work targeting four dimensions namely, (i) AI Robustness, (ii) Explainable AI, (iii) AI Privacy, and (iv) AI Fairness, each respectively corresponding to tasks T4.2, T4.3, T4.4, and T4.5. For each dimension, we present an overview of each partner's contribution and the methodology used, along with results, relevant publications, and software if available.
Keywords	AI, Media, AI Robustness, Explainable AI, AI Privacy, AI Fairness

Copyright

© Copyright 2023 AI4Media Consortium

This document may not be copied, reproduced, or modified in whole or in part for any purpose without written permission from the AI4Media Consortium. In addition to such written permission to copy, reproduce, or modify this document in whole or part, an acknowledgement of the authors of the document and all applicable portions of the copyright notice must be clearly referenced. All rights reserved.





Contributors

NAME	ORGANIZATION
Anisa Halimi	IBM
Naoise Holohan	IBM
Ambrish Rawat	IBM
Stefano Braghin	IBM
Kieran Fraser	IBM
Nicu Sebe	UNITN
Daniel Gatica-Perez	IDIAP
Sina Sajadmanesh	IDIAP
Maya Guido	IDIAP
Hervé Le Borgne	CEA
Vasileios Mygdalis	AUTH
Thomas Köllmer	FhG-IDMT
Christoforos Papastergiopoulos	CERTH
Vasileios Mezaris	CERTH
Mara Graziani	HES-SO
Lorenzo Seidenari	UNIFI
Gianluigi Lopardo	3IA-UCA
Gabriele Ciravegna	3IA-UCA
Frederic Precioso	3IA-UCA
Damien Garreau	3IA-UCA

Peer Reviews

NAME	ORGANIZATION
Adrian Popescu	CEA
Nicu Sebe	UNITN





Revision History

Version	Date	Reviewer	Modifications
0.1	May 12th 2023	Anisa Halimi	First draft sent to partners for contributions
0.2	July 17th 2023	Anisa Halimi	Updated version with contributions from partners
0.4	July 25th 2023	Adrian Popescu	Updated version with review from Adrian Popescu
0.5	July 27th 2023	Nicu Sebe	Updated version with review from Nicu Sebe
0.6	July 28th 2023	Filareti Tsalakanido	Updated version with review from Filareti Tsalakanido
0.7	August 24th 2023	Naoise Holohan	Updated version with contributions from partners
0.8	September 8th 2023	Anisa Halimi, Naoise Holohan	Updated version regarding Use Cases
1.0	September 13th 2023	Anisa Halimi, Naoise Holohan	Final version

The information and views set out in this report are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf.





Table of Abbreviations and Acronyms

Abbreviation	Meaning
AD	Average Drop
AI	Artificial Intelligence
AIF360	AI Fairness 360
AOD	Average Odds Difference
ART	Adversarial Robustness Toolbox
AT	Adversarial Training
AWP	Adversarial Weight Perturbation
CAM	Class Activation Mapping
CE	Cross Entropy
CEO	Calibrated Equality of Odds
CNN	Convolutional Neural Network
DCNN	Deep Convolutional Neural Network
DFD	DeepFake Detection
DI	Disparate Impact
DNN	Deep Neural Network
DP	Differential Privacy
EOD	Equal Opportunity Difference
EOO	Equality of Odds
EU	European Union
FHE	Fully Homomorphic Encryption
FL	Federated Learning
FoR	Fake-or-Real
GAN	Generative Adversarial Network
GAP	Global Average Pooling
GDPR	General Data Protection Regulation
GNN	Graph Neural Network
Grad-CAM	Gradient-weighted Class Activation Mapping
HCP	Hyperspherical Class Prototypes
HDT	Heuristic Decision Tree
HSJ	HopSkipJump
IC	Increase in Confidence
LLM	Large Language Model
ML	Machine Learning
MSE	Mean Square Error
PET	Privacy Enhancement Technologies
PGD	Projected Gradient Descent
RNG	Random Number Generator





ROC	Reject Option Classification
RPN	Region Proposal Network
SM	Saliency Map
SMPC	Secure Multiparty Computation
SPD	Statistical Parity Difference
SPSM	Spatial Smoothing
SRC	Self-Residual-Calibration
STN	Spatial Transformer Network
TTS	Text-to-Speech





Contents

1	Executive Summary	17
2	Introduction	19
2.1	Trustworthy AI Overview	19
2.2	WP4 Timeline	19
2.3	Document Organisation	21
3	Robust AI (Task 4.2)	22
3.1	Exploring Robustness of an AI-based Deepfake Detection Service	22
3.1.1	Introduction	22
3.1.2	Adversarial attack and defense of a Deepfake Detector	23
3.1.3	Results & Discussion	24
3.1.4	Relevance to AI4Media use cases and media industry applications	27
3.2	Federated Model Fusion and Robustness	27
3.2.1	Relevance to AI4Media use cases and media industry applications	28
3.3	Mitigating Robust Overfitting via Self-Residual-Calibration Regularization	29
3.3.1	Experiments	30
3.3.2	Conclusions	33
3.3.3	Relevant Resources and Publications	34
3.3.4	Relevance to AI4Media use cases and media industry applications	34
3.4	Geometrically-inspired training scheme for adversarial robustness	34
3.4.1	Overview	35
3.4.2	Robust One-class Classification-based training loss	35
3.4.3	Experiments	37
3.4.4	Conclusions	38
3.4.5	Relevant publications	38
3.4.6	Relevance to AI4Media use cases and media industry applications	38
3.5	Matching Pairs: Attributing Fine-Tuned Models to their Pre-Trained Large Language Models	38
3.5.1	Overview	39
3.5.2	Experiments	39
3.5.3	Conclusion	40
3.5.4	Relevant Resources and Publications	41
3.5.5	Relevance to AI4Media use cases and media industry applications	41
4	Explainable AI (Task 4.3)	42
4.1	Deep Learning Insights into Synthetic Audio Detection: An Interpretable Approach Using Saliency Maps	42
4.1.1	Overview	42
4.1.2	Methodology	43
4.1.3	Results	43
4.1.4	Relevance to AI4Media use cases and media industry applications	45
4.2	Learning Visual Explanations for DCNN-Based Image Classifiers Using an Attention Mechanism	46
4.2.1	Overview	46
4.2.2	Methodology	47
4.2.3	Results	49





4.2.4	Relevant Resources and Publications	52
4.2.5	Relevance to AI4Media use cases and media industry applications	52
4.3	Wasserstein loss for Semantic Editing with GANs	52
4.3.1	Overview	52
4.3.2	Methodology	54
4.3.3	Results	55
4.3.4	Relevant Resources and Publications	56
4.3.5	Relevance to AI4Media use cases and media industry applications	56
4.4	Disentangling Neuron Representations with Concept Vectors	56
4.4.1	Approach	57
4.4.2	Contribution	58
4.4.3	Experiments	58
4.4.4	Relevant Resources and Publications	59
4.4.5	Relevance to AI4Media use cases and media industry applications	59
4.5	Multitask-Adversarial Learning Architecture	59
4.5.1	Approach	59
4.5.2	Contribution	60
4.5.3	Relevant Resources and Publications	60
4.5.4	Relevance to AI4Media use cases and media industry applications	60
4.6	Explaining Autonomous Driving with Visual Attention and End-to-End Trainable Region Proposals	61
4.6.1	Approach	61
4.6.2	Results	61
4.6.3	Relevant Resources and Publications	62
4.6.4	Relevance to AI4Media use cases and media industry applications	63
4.7	SMACE: Semi-Model-Agnostic Contextual Explainer	63
4.7.1	Overview	63
4.7.2	Approach	64
4.7.3	Results	66
4.7.4	Relevant Resources and Publications	68
4.7.5	Relevance to AI4Media use cases and media industry applications	68
4.8	A Sea of Words: An In-Depth Analysis of Anchors for Text Data	69
4.8.1	Overview	69
4.8.2	Methodology	69
4.8.3	Results	72
4.8.4	Relevant Resources and Publications	72
4.8.5	Relevance to AI4Media use cases and media industry applications	72
4.9	Interpretable Neural-Symbolic Concept Reasoning	72
4.9.1	Overview	73
4.9.2	Methodology	73
4.9.3	Results	74
4.9.4	Relevant Resources and Publications	75
4.9.5	Relevance to AI4Media use cases and media industry applications	76
4.10	Concept Embedding Models: Beyond the Accuracy-Explainability Trade-Off	77
4.10.1	Overview	77
4.10.2	Methodology	77
4.10.3	Results	79
4.10.4	Relevant Resources and Publications	82
4.10.5	Relevance to AI4Media use cases and media industry applications	82





4.11	First Nice Workshop on Interpretability (NWI)	82
4.11.1	Summary	83
4.11.2	List of invited talks	83
5	Privacy-Enhancing AI (Task 4.4)	85
5.1	Federated Unlearning: How to Efficiently Erase a Client in FL?	85
5.1.1	Overview	85
5.1.2	Methodology	85
5.1.3	Results	87
5.1.4	Relevant Resources and Publications	88
5.1.5	Relevance to AI4Media use cases and media industry applications	88
5.2	Diffprivlib – A General-Purpose Differential Privacy Library	88
5.2.1	Overview	88
5.2.2	Methodology	89
5.2.3	Results	90
5.2.4	Relevant Resources and Publications	91
5.2.5	Relevance to AI4Media use cases and media industry applications	92
5.3	A Utility-Preserving De-Identification Approach with Relation Extraction Filtering	92
5.3.1	Method	93
5.3.2	Experimental Evaluation	94
5.3.3	Relevant Resources and Publications	95
5.3.4	Relevance to AI4Media use cases and media industry applications	95
5.4	Secure Federated Learning	95
5.4.1	Methodology	96
5.4.2	Results and Updates	97
5.4.3	Relevance to AI4Media use cases and media industry applications	97
5.5	Differentially Private Graph Learning	97
5.5.1	Overview	97
5.5.2	GAP: Differentially Private GNNs with Aggregation Perturbation	98
5.5.3	ProGAP: Progressive GNNs with Aggregation Perturbation	100
5.5.4	Relevant Resources and Publications	102
5.5.5	Relevance to AI4Media use cases and media industry applications	102
5.6	Reversible adversarial attacks for privacy protection	103
5.6.1	Overview	103
5.6.2	Transformation based adversarial attacks	103
5.6.3	The linear case	104
5.6.4	The non-linear case	105
5.6.5	Experiments	106
5.6.6	Conclusion	106
5.6.7	Relevant Resources and Publications	107
5.6.8	Relevance to AI4Media use cases and media industry applications	108
6	Fair AI (Task 4.5)	109
6.1	Datasheet for the Dataset on European Press Coverage of Covid-19 Vaccination News	109
6.1.1	Goals	109
6.1.2	Method	110
6.1.3	Discussion	110
6.1.4	Relevance to AI4Media use cases and media industry applications	111
6.2	Exploring Fairness of an AI-based Deepfake Detection Service	111





6.2.1	Introduction	111
6.2.2	Identifying Bias in the DeepFake Detection Service	112
6.2.3	Toward Fair Deepfake Detection	115
6.2.4	Relevance to AI4Media use cases and media industry applications	118
6.3	Debiasing Neural Models Using Explainable Artificial Intelligence	118
6.3.1	Introduction	118
6.3.2	Debiasing Neural Models Using XAI	118
6.3.3	Relevant Resources and Publications	119
6.3.4	Relevance to AI4Media use cases and media industry applications	120
7	Contributions to the AI4Media WP8 Use Cases and Media Industry Applications	121
8	Ongoing Work and Conclusions	124
8.1	Ongoing Work	124
8.1.1	AI Robustness (Task 4.2)	124
8.1.2	Explainable AI (Task 4.3)	124
8.1.3	Privacy-Enhancing AI (Task 4.4)	125
8.1.4	AI Fairness (Task 4.5)	126
8.2	Conclusions	126
A	Appendix	141
A.1	Datasheet for the dataset on European press coverage of Covid-19 vaccination news	141
A.2	Abstracts from the invited talks at the First Nice Workshop on Interpretability (NWI)	147





List of Tables

2	DFD service performance when not under attack.	24
3	Average frame (image) prediction time with and without defence applied.	25
4	DFD service (video) performance when under attack	25
5	DFD service (frame) performance under attack	25
6	Test robust accuracy (%) on three benchmarks. “Natural” is the natural accuracy. “BC” is the highest test accuracy observed during training. “FC” is test accuracy on the last epoch. “D” indicates the difference between the highest and final checkpoints. The symbol ↓ means the lower score is better, and the symbol ↑ means the higher score is better.	31
7	Test accuracy on CIFAR-10. “+D” means semi-supervised data augmentation is used. “WRN” means WideResNet is used. “+FL” means Flooding is used. “+AWP” means the adversarial weight perturbation [49] is used. PGD-AT is the recent SOTA model with considering many useful tricks [37]. “ES” means the early-stopping is used.	33
8	Classification accuracy of the competing methods.	37
9	Robustness (classification accuracy) in PGD black-box attack, by using the Vanilla ResNet architecture as attack model.	37
10	Cross-method black-box PGD attacks in CIFAR-10.	38
11	Model Attributions on $m_{\#}$ from the different methods. Dashes (–) are used when multiple models (m_f) are attributed to m_b . TP denotes True Positives	40
12	Evaluation results for a VGG-16 (upper half) and ResNet-50 (lower half) backbone classifier using 2,000 randomly-selected testing images of ImageNet. The best and 2nd-best performance for a given evaluation measure are shown in bold and underline, respectively.	51
13	Ablation study. We vary the number of proposals produced by STN and RPN. Both STN and RPN perform better using a number of boxes around 100. In general, STN can obtain higher driving accuracy even with a low number of proposals, compared to RPN.	62
14	Task accuracy on the <i>MNIST-addition</i> dataset. The neural-symbolic baselines use the knowledge of the symbolic task to distantly supervise the image recognition task. DCR achieves similar performances even though it learns the rules from scratch.	76
15	Error rate of Booleanised DCR rules w.r.t. ground truth rules. Error rate represents how often the label predicted by a Booleanised rule differs from the fuzzy rule generated by our model. The error rate is reported with the mean and standard error of the mean.	76
16	Performance measured on the TAB dataset.	95
17	Comparison results on MNIST dataset	106
18	Fairness assessment of DFD service. Values of 0 (for Statistical Parity difference, Average Odds difference and Equal Opportunity difference) and 1 (for Disparate Impact) indicate a fair deepfake detector.	113
19	Results for fairness metrics using predictions from DFD service and post-processed using Reject-Option Classification, optimized for Statistical Parity Difference.	116
20	Balanced accuracy scores for post-processed (fair) predictions of the DFD service and the associated adjustments made to the classification threshold for each attribute to achieve fairness.	117





List of Figures

1	WP4 four-year timeline, showing the position of the present deliverable (D4.5) with reference to the lifetime of the AI4Media project.	20
2	WP4 Tasks, comprising four vertical tasks (technical) and two horizontal tasks. . .	20
3	Success rate vs L2 distance threshold. Successful attacks are those which changed the DFD prediction.	26
4	Time taken to generate batches of adversarial frames (batch size = 32).	27
5	Model convergence under the presence of poisoned training data with different aggregation methods. Federated Averaging (FedAvg) shows oscillation while other approaches show a more stable (and robust) convergence.	28
6	Learning curves on CIFAR-100 and SVHN for different models, i.e., PGD-AT, TRADES, AT-FL, and Ours. (Best view in color.)	32
7	Visualization of the spectrogram, saliency map, and superimposed image of an audio sample. (a) The spectrogram shows the frequency components of the audio sample over time. (b) The saliency map shows the areas of the audio sample that are most salient, or attention-grabbing. (c) The superimposed image shows the audio sample with the saliency map overlaid.	44
8	Average saliency maps for the synthetic (top) and real (bottom) audio classes. The x -axis represents the time domain from 0 to 251, and the y -axis shows the Mel frequency bins from 0 to 128. The color scheme, depicted by the legend, illustrates the activation strength from low (blue) to high (red) according to the Jet color map.	45
9	Distribution of confidence scores, calculated as the absolute difference between cosine similarities of new instances with the average saliency maps of synthetic and real audio classes. The median confidence score (depicted by the dashed red line) is approximately 0.1088, indicating that for a typical instance, the difference in cosine similarities is about this value.	46
10	The network architectures of the developed approaches: (a) L-CAM-Fm training, (b) L-CAM-Img training, (c) L-CAM-Fm/-Img inference.	48
11	Visualization of SMs from various XAI methods superimposed on the original input image to produce class-specific visual explanations for the VGG-16 (columns 1 to 3) and ResNet-50 (columns 4 to 6) backbones.	50
12	Two examples of using class-specific SMs (superimposed on the input image) produced by L-CAM-Img [†] on VGG-16 for classes “pug” and “tiger cat” (left) and classes “soccer ball” and “Maltese” (right).	51
13	CEA Method overview. For each semantic attribute (<i>e.g.</i> , “Glasses”) we learn a mapping \mathbf{H}_k that moves the distribution of latent codes lacking the attribute to the distribution of codes having that attribute. We enforce that each latent code is moved near a point that shares similar semantics, thus only changing that attribute. To preserve identity, the resulting distribution does not entirely match the target distribution.	53





14	Failure cases of a classifier-based method. Latent transformer (LT) [102] learns edits in latent space under the guidance of a latent classifier. (left) On FFHQ [106] for “ <i>Male</i> ’ → ‘ <i>Female</i> ”’: without L_2 -regularization on the edited codes, the edited images are unrealistic (as shown in the qualitative result on the left) before reaching the desired editing. The classifier leads to out-of-distribution regions as it allocates high confidence to regions larger than that of the training samples [103]. The quantitative analysis on attribute and identity preservation shows highly degraded results. (right) On MultiMNIST [104] “ <i>1 digit</i> ’ → ‘ <i>2 digits</i> ”’: the edited images remain unchanged (no digit is being added) while the classifier indicates the opposite (predicts 2 digits with high confidence). The classifier leads to regions close to the decision boundaries where there are adversarial samples. The quantitative analysis shows that only 32% of images are correctly edited.	53
15	(left) Qualitative results for facial attribute editing. We report the editing results for $\alpha = \pm 2$. We observe that our approach better preserves identity and some facial attributes (<i>e.g</i> expression, absence of makeup) compared to LT. (right) Qualitative comparison between classifier-based edits (2^{nd} col.) and our Wasserstein-based edits w/o any constraint (3^{rd} col.) and w/ disentanglement constraint (4^{th} col.).	56
16	Step 1. A set of images that maximally activate a neuron in a model layer is taken. Step 2. The Euclidean distance between images in activation space is used as the similarity space on which the clustering is performed. This returns the appropriate number of clusters for a given distance threshold. Step 3. K-means clustering computes the cluster membership. Step 4. From the images in each cluster, a concept vector is calculated, which points toward the non-neuron aligned direction in activation space.	57
17	UMAP of the maximally activating images kept after k-means clusters and outlier removal in latent space: (left) separate clusters for polysemantic neuron 35 (right) a single cluster for monosemantic neuron 16.	58
18	Multi-task adversarial architecture	59
19	A convolutional backbone generates a feature map. Then, a region proposal function extracts RoIs that are pooled and weighed by an attention layer. Separate region proposal and attention modules are trained for each high level command in order to focus on different regions and output the appropriate steering angle.	62
20	Precision-Recall Curves for detecting failed episodes.	63
21	Structure of a composite decision system with D input features x_1, \dots, x_D , and N models m_1, \dots, m_N . A decision policy P (<i>i.e.</i> , a set of decision rules) is finally applied to produce an outcome O . Note that in general both the models and the rules take a subset of input features as input, though not necessarily the same. . . .	64
22	Comparison of SMACE, SHAP, and LIME on the ability to identify the set of features contributing negatively to a decision, regardless of individual attribution. Correctly identifying negative features is a desirable property: to change the decision, each of them must be moved. When the conditions are not met, the three methods are used to extract the negative features, and we generate perturbed samples around the original values. We then compare the average decision made on the samples. . . .	67
23	Anchors explaining the positive prediction of a black-box model f on an example ξ from the Restaurant review dataset. The anchor $A = \{great, not, bad, fine\}$ (in blue), having length $ A = 4$ is selected. Intuitively, these four words together ensure a positive prediction by f with high probability (precision : 0.97), while being not too uncommon (coverage : 0.12).	70





24	An illustration of Algorithm 3 with evaluation function $p = \text{Prec}$. Each blue dot is an anchor, with x coordinate its length and y coordinate its value for p . Here, $\epsilon = 0.2$ and the maximal length of an anchor is $b = 10$ (the length of ξ). In the end, the anchor A such that $ A = 3$ and $p(A) = 0.9$ is selected (red circle).	71
25	(left) Deep Concept Reasoner (DCR) generates fuzzy logic rules using neural models on concept embeddings. Then DCR executes the rule using the concept truth degrees to evaluate the rule symbolically. (right) Schema of DCR modules: first neural models ϕ and ψ generate the rule, and then the rule is executed symbolically. . . .	73
26	Mean ROC AUC for task predictions for all baselines across all tasks (the higher the better). DCR often outperforms interpretable concept-based models. <i>CE</i> stands for concept embeddings, while <i>CT</i> for concept truth degrees. Models trained on concept embeddings are not interpretable as concept embeddings lack a clear semantic for individual embedding dimensions.	75
27	Sensitivity of model explanation when changing the radius of the input perturbation. The lower, the better. DCR explanations engender trust as they are stable under small perturbations of the input. The same does not hold generally for LIME explanations of XGBoost or Relu Net decision rules.	75
28	Model confidence as a function of the number of perturbed features on counterfactual examples. The lower, the better. Similarly to interpretable methods, DCR prediction confidence quickly drops after inverting the truth degree of a small set of relevant concepts, facilitating the discovery of counterfactual examples.	75
29	Concept Embedding Model: from an intermediate latent code \mathbf{h} , we learn two embeddings per concept, one for when it is active (i.e., $\hat{\mathbf{c}}_i^+$), and another when it is inactive (i.e., $\hat{\mathbf{c}}_i^-$). Each concept embedding (shown in this example as a vector with $m = 2$ activations) is then aligned to its corresponding ground truth concept through the scoring function $s(\cdot)$, which learns to assign activation probabilities \hat{p}_i for each concept. These probabilities are used to output an embedding for each concept via a weighted mixture of each concept’s positive and negative embedding.	78
30	Accuracy-vs-interpretability trade-off in terms of task accuracy and concept alignment score for different concept bottleneck models. In CelebA, our most constrained task, we show the top-1 accuracy for consistency with other datasets. .	79
31	Qualitative results for our CEM with t-SNE visualisations of “has white wings” concept embedding learnt in CUB with sample points coloured red if the concept is active in that sample	80
32	Qualitative results for hybrid with t-SNE visualisations of “has white wings” concept embedding learnt in CUB with sample points coloured red if the concept is active in that sample	80
33	top-5 test neighbours of CEM’s embedding for the concept “has white wings” across 5 random test samples.	81
34	Mutual Information (MI) of concept representations (\hat{C}) w.r.t. input distribution (X) and ground truth labels (Y) during training. The size of the points is proportional to the training epoch.	81
35	Effects of performing positive random concept interventions (left and center left) and incorrect random interventions (center right and right) for different models in CUB and CelebA. As in [141], when intervening in CUB we jointly set groups of mutually exclusive concepts.	82
36	Pictures from NWI: (left) captivated audience; (middle:) discussions during the coffee break; (right:) The speakers and the organization team at the restaurant. . .	84





37	Phases of Federated Unlearning: (a) First, clients and the server participate in a federated learning process to train a global model. (b) One of the clients wants to opt out of federation and wants to unlearn their data. The target client i locally runs Projected Gradient Descent to obtain model \mathbf{w}_i^u . (c) The server and the remaining clients perform a few steps of federated learning with \mathbf{w}_i^u as the initial point to obtain the final ‘unlearned’ model.	86
38	Backdoor accuracy (efficacy) of the fully retrained and the PGD-based unlearned model in each dataset, and their comparison with the FedAvg model before unlearning.	87
39	Clean accuracy (fidelity) of the fully retrained and the PGD-based unlearned model in each dataset.	87
40	Communication costs (efficiency) of the proposed unlearning method and the baseline approach with respect to the clean accuracy (fidelity) in each dataset for $N = 5$	88
41	Example code showing the difference of seeding a Diffprivlib function with an integer, and seeding with a <code>RandomState</code> instance. In the former, repeated execution returns the same “random” value. In the latter, different values are returned, but repeated execution of the same script will give the same overall output. Typically, the second behaviour is what is desired.	90
42	Example performance of the Random Forest classifier with differential privacy across various epsilon (privacy loss) values. The simulations were completed using Scikit-Learn’s <code>make_blobs</code> data generator, with 10,000 samples generated over 3 centres, with a 80/20 train/test split. As can be seen from the plot, maximum performance is achieved approximately when $\epsilon = 0.02$	91
43	Computation time for secure sampling versus naive (insecure) sampling in blue. The increased sampling time for using more uniform variates is offset by the (exponentially) increased security and attack resistance.	92
44	Proposed method diagram.	93
45	Example of entities and relationships	94
46	Overview of GAP’s architecture: (1) The encoder is trained using only node features (\mathbf{X}) and labels (\mathbf{Y}). (2) The encoded features are given to the aggregation module to compute private K -hop aggregations (here, $K = 2$) using the graph’s adjacency matrix (\mathbf{A}). (3) The classification module is trained over the private aggregations for label prediction.	99
47	Accuracy vs. privacy cost (ϵ) of edge-level private algorithms (top) and node-level private methods (bottom). $-\infty$, -EDP, and -NDP suffixes correspond to non-private, edge-level and node-level DP, respectively.	100
48	An example PROGAP architecture with three stages. MLP and JK represent multi-layer perceptron and Jumping Knowledge [186] modules, respectively. NAP denotes the normalize-aggregate-perturb module used to ensure the privacy of the adjacency matrix, with its output cached immediately after computation to save privacy budget. Training is done progressively, starting with the first stage and then expanding to the second and third stages, each using its own head MLP. The final prediction is obtained by the head MLP of the last stage.	101
49	Accuracy-privacy trade-off of edge-level (top) and node-level (bottom) private methods. The dotted line represents the accuracy of the non-private PROGAP.	102
50	Architecture of the proposed AdvRevGAN model.	105





51	Adversarial examples and reconstructed images on MNIST Dataset. The first column depicts original images \mathbf{x}_i , the next three columns are the corresponding adversarial examples \mathbf{y}_i^{adv} generated by the proposed method, MUAT and UAP respectively while above them demonstrated the wrong class that predicted by the model. In the last two columns are demonstrated the reconstructed images \mathbf{x}_i^{rec} derived by MUAT and our proposed method respectively.	107
52	Relative Performance vs Corrected Relative Performance plots for predictions of the DFD service on videos of Celeb-DF (left) and FF++ (right) illustrating bias toward (green) and against (red) attributes.	114
53	Deepfake Data Relative Performance vs Pristine Data Relative Performance for predictions of the DFD service on videos from Celeb-DF (left) and FF++ (right).	115
54	RP-vs-CRP plots for predictions made by the DFD service and post-processed using ROC-SP on videos from Celeb-DF (left) and FF++ (right).	117
55	Hierarchy of technical Work Packages, showing the flow of work from top to bottom. WP4 is highlighted in green.	121
56	Full listing of WP4 components, compiled by WP4 partners.	123





1 Executive Summary

This deliverable presents the research carried out as part of technical tasks of Work Package 4 of the AI4Media project, entitled *Explainability, Robustness and Privacy in AI*. These tasks, T4.2, T4.3, T4.4 and T4.5, cover the areas of AI Robustness, Explainability, Privacy, and Fairness respectively, and are accompanied by Tasks 4.1 and 4.6 which cover legal and benchmarking aspects of the Work Package that are not part of this deliverable. For each contribution in this report, we provide an overview of the work carried out, as well as references to the publications and software released by each partner.

This deliverable covers work carried out after the submission of the first deliverable, D4.1, and includes outcomes produced in the intervening two years, from M13 (September 2021) to M36 (August 2023). A considerable volume of work has been completed during this time, including contributions from partners (in alphabetical order) **AUTH**, **CEA**, **CERTH**, **FhG-IDMT**, **HESO**, **IBM**, **IDIAP**, **3IA-UCA**, **UNIFI**, and **UNITN**. Tasks 4.4 and 4.5 were due to come to completion with this deliverable, but both will be extended to the end of the project to allow for additional research to be completed, and delivered as part of D4.7 in August 2024.

Introductory remarks are given in Section 2, covering an overview of the Trustworthy AI field (Section 2.1), the timeline of Work Package 4 (Section 2.2) and the structure of this document (Section 2.3).

Contributions towards the **Robust AI** task (T4.2) are detailed in Section 3. This includes work on (i) exploring the robustness of an AI-based detection of deep learning manipulations (known as deepfakes) in images and videos (Section 3.1), (ii) robustness of federated learning model training, by analysing the impact of different aggregation algorithms (Section 3.2), (iii) addressing the problem of robust overfitting in adversarial training and designing a novel regularization scheme to overcome it (Section 3.3), (iv) a new, geometrically-inspired, approach to adversarial training to improve robustness in neural networks (Section 3.4), and (v) matching a fine-tuned model to its pre-trained parent/root (Section 3.5).

Contributions towards the **Explainable AI** task (T4.3) are detailed in Section 4. This includes work on (i) using visualisations to explain deep learning insights when detecting synthetic audio (Section 4.1), (ii) designing a new attention mechanism for visual explanations of image classifiers (Section 4.2), (iii) new learning methods for semantic editing of GANs for generating images (Section 4.3), (iv) disentangling neuron representations with concept vectors (Section 4.4), (v) a novel architecture for combining multitask learning and adversarial training (Section 4.5), (vi) explainability in autonomous driving systems (Section 4.6), (vii) explainability in multi-model AI systems (Section 4.7), (viii) an analysis of *Anchors* in text classification systems (Section 4.8), (ix) explainability through concept-based models (Section 4.9), and (x) an extension of concept bottleneck models (Section 4.10). An outline of the first Nice Workshop on Interpretability (NWI) is given in Section 4.11.

Contributions towards the **Privacy-enhancing AI** task (T4.4) are detailed in Section 5. This includes work on (i) unlearning in the federated learning setting, a new field of work (Section 5.1), (ii) continuing work on *diffprivlib*, a general-purpose library for differential privacy computations in Python (Section 5.2), (iii) a utility-preserving de-identification approach for data publication using relation extraction filtering (Section 5.3), (iv) a tool for combining differential privacy, homomorphic encryption and multiparty computation for secure federated learning (Section 5.4), (v) a graph neural network with differentially private learning guarantees (Section 5.5), and (vi) the use of a reversible transformation to create adversarial examples for training (Section 5.6).

Finally, contributions towards the **Fair AI** task (T4.5) are detailed in Section 6. This includes work on (i) datasheets that can be passed along with data and machine learning models to highlight potential risks and recommend appropriate uses (Section 6.1), (ii) fairness of deepfake detection





systems (Section 6.2), and (iii) debiasing neural networks using explainable AI (Section 6.3).

Ongoing work towards the integration of WP4 components within WP8 Use Cases is outlined in Section 7.

In summary, the work presented in this deliverable has resulted in:

- 18 conference and workshop papers (ICASSP ‘23, ACL ‘23, ECCV ‘22, CBMI ‘23, CVPR ‘23, ICML ‘22, ESORICS ‘21, PETS ‘23, UESNIX ‘23, MLSP ‘22, MLSP ‘23, ECML-PKDD ‘23, IJCAI ‘20, ECML PKDD ‘22, AISTATS ‘23, ICML ‘23, NeurIPS ‘22, IJCAI ‘20) and three journal articles (Artificial Intelligence ‘23, Ambient Intelligence and Humanized Computing ‘23, Journal of Ambient Intelligence and Humanized Computing ‘23);
- Three technical reports;
- 13 open-source software and tools that are openly shared (e.g., in GitHub).

In addition, four secondments have been carried out in the context of the AI4Media Junior Fellows Exchange Program, focusing on WP4-related topics and resulting in important outcomes for trusted AI.





2 Introduction

2.1 Trustworthy AI Overview

Artificial Intelligence (AI) holds significant importance in the European Union (EU) due to its potential to foster innovation, drive economic growth, improve public services, and shape social development. While AI offers immense opportunities and numerous benefits, there are also potential risks associated with its development such as security vulnerabilities, lack of transparency, privacy concerns, and bias and discrimination.

Trustworthy AI aims at developing and deploying Machine Learning (ML) technologies that are reliable, transparent, accountable, and aligned with the democratic and ethical values shared in our society. Trustworthy AI is typically divided into four broad dimensions: (i) **AI robustness**, (ii) **Explainable AI**, (iii) **AI Privacy**, and (iv) **AI fairness**.

AI Robustness focuses on detecting and mitigating adversarial attempts such as the introduction of misleading or malicious input to push an ML model towards making incorrect decisions or predictions. These attacks can be achieved through the use of adversarial samples in various data types (e.g., images, text, etc.) and across a broad range of model architectures.

Traditional ML models, such as deep neural networks, are inherently black boxes or operate in a black-box setting¹ so their decision-making processes are difficult to explain. The lack of interpretability and transparency in these models can lead to distrust and reluctance to adopt them, especially in critical applications (e.g., healthcare, finance) where decisions may have a significant impact on individuals. *Explainable AI* aims to provide users with transparency and understanding of how decisions are made by ML models.

AI Privacy focuses on designing and developing techniques to protect individuals' personal information including their sensitive information by maintaining its confidentiality and privacy. It also aims to prevent unauthorized access and misuse as improper handling of personal information can result on unintended parties accessing individuals' sensitive information. Such sensitive information can then be used against the individuals for discrimination or blackmailing. AI models are typically trained on a large amount of data which in many cases contains sensitive information. Thus, AI Privacy aims to produce reliable ML models while ensuring that individuals' privacy is enhanced.

Finally, AI models can inadvertently learn biases from the data that are trained on, reflecting and preserving biases and prejudice already present in our society. This can result in discriminatory treatment in various domains where AI is used such as mortgage lending, hiring, and criminal justice. *AI Fairness* aims to address these issues by developing AI models that treat individuals/groups fairly without favoring or disadvantaging any specific group/individual.

2.2 WP4 Timeline

This work package (WP4) is dedicated to Trustworthy AI. It involves 12 partner institutions, namely - IBM, IDIAP, UPB, FhG, HES-SO, AUTH, CERTH, UCA, UNITN, CEA, KUL, and UNIFI - and runs throughout the entire duration of the AI4Media project (Figure 1). WP4 consists of 6 tasks organized as 4 vertical tasks, i.e., Robust AI (Task 4.2), Explainable AI (Task 4.3), Private AI (Task 4.4), and Fair AI (Task 4.5) and 2 horizontal tasks focusing on the ethical and legal dimensions of AI within the European Union (Task 4.1) and benchmarking of AI systems (Task 4.6) (see Figure 2).

¹A black-box setting refers to a scenario where the user has limited or no access to the internal workings of a model.





Figure 1. WP4 four-year timeline, showing the position of the present deliverable (D4.5) with reference to the lifetime of the AI4Media project.

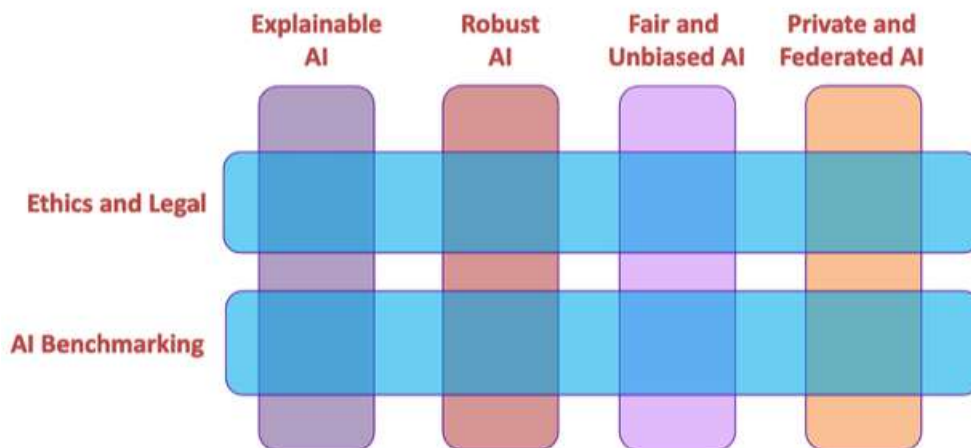


Figure 2. WP4 Tasks, comprising four vertical tasks (technical) and two horizontal tasks.

During the course of the project, this work package will produce 3 types of deliverables: (i) toolset where the technical research output produced from the four vertical tasks will be reported, (ii) legal where the output of the corresponding horizontal task will be reported, and (iii) benchmark for the other horizontal task. This document consists of the second iteration of the toolset deliverable. Each iteration of this deliverable will report the contributions of the partners ranging from new algorithms accompanied by experimental results to toolset modules. In each iteration, we expect individual contributions to be at various stages of this pipeline as investigations mature.

The second iteration of this deliverable is an extension of the first one (D4.1). In this iteration, we present the research outputs that each partner achieved, as well as the secondments outcomes conducted between M13 and M36 of the AI4Media project. This deliverable contains contributions within the dimensions of AI Robustness, Explainable AI, AI Privacy, and AI Fairness. The last iteration of this deliverable will be submitted in month 48 and will present updates on the progress



achieved in each dimension in the last 12 months of the AI4Media Project.

2.3 Document Organisation

This deliverable follows the same structure as D4.1, with a similar structure for all the tasks to ensure a harmonized presentation of the algorithms/tools that were developed since D4.1. Sections 3- 6 describe the contributions towards each vertical task in Figure 2 (i.e., Robust AI, Explainable AI, AI privacy, and AI Fairness), respectively. All sections follow the same structure. We briefly introduce each task, followed by a description of the methods/tools developed by WP4 partners during this period. Section 7 presents steps taken to integrate WP4 technical outputs with AI4Media Use Cases and media industry use cases more generally. Finally, Section 8 concludes the deliverable summarizing the current progress achieved as part of WP4, including an outline of future work from all partners for the final year of the project.





3 Robust AI (Task 4.2)

Machine Learning (ML) models are vulnerable to a variety of threat models [1], [2] in which adversarial samples play a critical role. Adversarial samples consist of inputs (images, texts, tabular data, etc.) deliberately crafted by an attacker in order to produce a desired response by the ML model unintended by the model creators.

There exist four broad types of adversarial threat models depending on how an attacker decides to exploit potential vulnerabilities in an ML model. Poisoning attacks focus on the insertion of malicious data within the datasets used to train a model while inference attacks intend to infer private information about a target model or the data used to train it. Evasion attacks, on the other hand, attempt to modify legitimate input samples in a manner that leads a model to misclassify it, while extraction attacks aim at extracting the parameters of a third party ML model so as to clone it.

In the following, we present new contributions to the **Robust AI** task, which include work on (i) exploring the robustness of an AI-based detection of deep learning manipulations (known as deepfakes) in images and videos (Section 3.1), (ii) robustness of federated learning model training, by analysing the impact of different aggregation algorithms (Section 3.2), (iii) addressing the problem of robust overfitting in adversarial training and designing a novel regularization scheme to overcome it (Section 3.3), (iv) a new, geometrically-inspired, approach to adversarial training to improve robustness in neural networks (Section 3.4), and (v) matching a fine-tuned model to its pre-trained parent/root (Section 3.5).

3.1 Exploring Robustness of an AI-based Deepfake Detection Service

Contributing partners: IBM, CERTH

3.1.1 Introduction

This work was completed as part of a virtual Junior Fellow Exchange between IBM and CERTH and is the first of two evaluations of a Deepfake Detection Service created by CERTH - a second evaluation on fairness is detailed in Section 6.2. The MeVer DeepFake Detection (DFD) service was created to detect deep learning manipulations in images and videos. The system comprises a pre-processing pipeline and model ensemble scheme which is used to obtain a DeepFake probability score indicating whether an input image or video has been manipulated, such as with FaceSwap or similar tools. The DFD service was previously evaluated using three benchmark DeepFake datasets (FaceForensics++ [3], CelebDF-V2 [4], and WildDeepFake [5]), and outperformed a publicly available DeepFake detection model, DeepWare², proving the design robust.

In addition to proving the efficacy of the DFD service on benchmark datasets, the service was also exposed to an adversarial attack to evaluate its performance under hostile conditions. Malicious actors could, for instance, bid to influence the result of the model to allow their deepfake images and videos to escape detection or similarly influence the model to incorrectly classify innocent content as deepfakes with the intent to have the content removed or accounts uploading content banned. The white-box attack chosen for evaluation was Projected Gradient Descent (PGD) [6]. For this attack, it was assumed that the malicious actor had access to all weights of the ensemble models of the DFD service i.e., a worst-case scenario. The attacks demonstrated that PGD was successful at decreasing model performance whilst also leaving little to no indication that an attack had been carried out.

²<https://deepware.ai/>





A key assumption of the PGD adversarial attack experiment was that the attackers had access to all weights of the ensemble models. For proprietary models, this is unlikely to occur as the model weights would not be directly exposed to the public to protect intellectual property. Whilst the previous evaluation [7] illustrated service performance degradation in the worst-case scenario, it did not cover scenarios involving black box attacks, which are far more likely to occur in scenarios where model predictions are exposed via queries from users. In addition, the previous work did not evaluate potential defences that could be implemented within the design of both the ensemble models and DFD service architecture that could aid in mitigating the impact of the attack.

To address these limitations, a black-box attack was selected to evaluate the performance of the DFD service under conditions where a malicious actor does not have access to the internal ensemble models but only has access to the predictions made by the service. This scenario has a much lower barrier-to-entry for malicious actors as they do not require knowledge regarding the model itself and simply require access to the service. Therefore the focus of this work was to:

- Deploy one of the IBM Adversarial Robustness Toolbox (ART)³ black-box attacks against the DFD service;
- Integrate defences from ART to mitigate black-box attacks on the DFD service,

and subsequently measure:

- the performance of the DFD service under black-box attack conditions;
- the performance of the DFD service with a defence mechanism applied in normal and attack scenarios;
- the prediction latency of the DFD service with a defence mechanism applied in normal and attack scenarios.

3.1.2 Adversarial attack and defense of a Deepfake Detector

The driving motivation of this work was to determine whether black-box attacks have a significant impact on the performance of the DFD service, as these types of attacks are more likely to occur, and whether implemented defences can mitigate these attacks whilst also maintaining model performance under normal conditions.

An engineering solution to black-box attacks would naturally be to define rate-limits, query quotas or to only expose the model behind an authentication layer; however, even with these security precautions, determined attackers could yet gain access to the model predictions if they assumed a benign identity. Therefore, in the spirit of trustworthy AI, it is prudent to evaluate the robustness of the DFD service under black-box attacks.

3.1.2.1 Black-box attack The black-box attack chosen in this instance was HopSkipJump (HSJ) [8], a decision-based adversarial attack. As it has shown to require significantly fewer model queries than other attacks to influence model performance and is also robust in the face of a number of defences, it was argued to be a good candidate to evaluate the DFD service.

To set up the attack, IBM's ART [9] was utilized as it is a widely accepted open-source state-of-the-art tool for evaluating the robustness of machine learning models. The toolbox abstracts the complexities of executing adversarial attacks by providing a simple interface through which the model is provided, and adversarial samples are returned.

³<https://github.com/Trusted-AI/adversarial-robustness-toolbox>





Table 2. DFD service performance when not under attack.

Dataset	No Attack					
	Balanced Accuracy			AUC		
	Baseline	JPEG-C	SPSM	Baseline	JPEG-C	SPSM
CelebDF-V2	80.63%	81.55%	69.85%	91.66%	91.36%	88.40%
FaceForensics++	69.64%	66.87%	60.62%	74.64%	73.62%	68.03%
WildDeepFake	81.37%	81.48%	-	90.77%	90.72%	-

In this case, as the DeepFake Detection ensemble model provides a single output probability, and the HSJ class depends on at least two output values representing the probability of both classes (DeepFake and not DeepFake), a thin wrapper was used to override the forward pass of the model and transform the output to a 2-dimensional array representing the probabilities of both classes. This wrapped model was then passed to ART’s PyTorchClassifier class, which was in turn passed to the HSJ class to facilitate the adversarial attack.

To carry out the robustness evaluation, each video of each dataset was first fed to the PyTorchClassifier unaltered to ascertain the prediction score under normal conditions. Then each video frame was passed to HSJ to generate an adversarial frame, which was then passed to PyTorchClassifier to ascertain the prediction score under black-box attack conditions. The L2 distance between the unaltered and adversarial images was calculated to determine the amount of perturbation applied in the adversarial image which could prove useful to the DFD service for identifying a black-box attack by comparing the similarity of queries.

3.1.2.2 Adversarial Defenses As outlined above, an adversarial attack in this case involves adding perturbations to the frames of the video sent to the DFD service such that it elicits an incorrect prediction for deepfake classification. The perturbations are designed to be small to escape detection by human intervention, but large enough to fool the DFD service. Dziugaite et al. [10] hypothesized an approach to remove the adversarial perturbations from images generated by attacks, such as HSJ, by applying a widely used image encoding and compression technique, JPEG Compression, which uses Discrete Cosine Transform to suppress “sharp transitions in intensity and colour hue” which are imperceivable to the human eye, but influential to trained models. JPEG compression can be added as a pre-processing step in the DFD pipeline. Spatial Smoothing (SPSM) [11], proposed by Xu et al., attempts to remove adversarial perturbations by substituting image colour values with median values computed in a sliding window thus, smoothing local colour variance (also known as blur). This defence was selected as an alternative pre-processing defence to JPEG compression to test whether it could yield better performance as it has been shown to be inexpensive and has high detection rates against state-of-the-art attacks.

3.1.3 Results & Discussion

The baseline performance of the DFD service is illustrated in Table 2 for all three benchmark DeepFake datasets. The baseline performance is the performance of the service without a defence or attack applied. The service achieves approx. 80% balanced accuracy and 90% AUC (Area under the Curve) on two of the three datasets, with FaceForensics++ proving slightly more challenging, achieving approx. 69% and 74% respectively. Table 3 illustrates the average frame prediction time per dataset, the maximum being CelebDF-V2 with a latency of 13.56ms. These baseline results facilitate a comparative analysis of the DFD service under attack and defence conditions.





Table 3. Average frame (image) prediction time with and without defence applied.

Dataset	Avg. DFD Frame Prediction Latency (ms)		
	Baseline	JPEG-C	SPSM
CelebDF-V2	13.56	16.41	14.88
FaceForensics++	9.528	16.88	15.34
WildDeepFake	9.398	14.95	-

Table 4. DFD service (video) performance when under attack

Dataset	Under Attack (video)					
	Balanced Accuracy			AUC		
	HSJ	JPEG-C	SPSM	HSJ	JPEG-C	SPSM
CelebDF-V2	25.34%	50%	50%	21.16%	88.58%	88.08%
FaceForensics++	32.59%	50%	50%	28.30%	72.66%	67.95%
WildDeepFake	24.76%	50%	-	13.00%	88.70%	-

On application of the HSJ black-box attack, the performance of the DFD service drops substantially. The HSJ columns of Table 4 illustrate the Balanced Accuracy and AUC achieved across all datasets when under attack. This indicates the vulnerability of the service to black-box attacks. Adversarial attacks evaluated with the CelebDF-V2 and WildDeepFake datasets in particular had the largest drop in performance, with balanced accuracy reducing from approx. 80% to approx. 25%. To mitigate the attack, two defences were applied. The **JPEG-C** and **SPSM** columns of Table 2 illustrate the performance of the DFD service when pre-processing defences were applied and no attack was implemented. Ideally in this scenario, the pre-processing defence should not negatively impact the performance of the ensemble model. The JPEG Compression pre-processing defence had a relatively small impact on the performance of the model, slightly improving performance on CelebDF-V2 and WildDeepFake datasets and slightly reducing performance on FaceForensics++. In contrast, SPSM had a much larger negative effect on performance in the attack-free scenario, reducing balanced accuracy on CelebDF-V2 and FaceForensics++ by approx. 10%. For this reason and due to computational resources, SPSM was not evaluated on the larger WildDeepFake dataset.

Table 3 also highlights the impact of the pre-processing defences on prediction time. Application of JPEG compression resulted in increases between 2.8 to 7.4 milliseconds. Spatial Smoothing resulted in increases of between 1.3 to 5.8 ms. Both defences improve the performance of the DFD service over the attack scenario but fail to restore performance to an attack-free level. Table 4 illustrates the difference in Balanced Accuracy between a standalone HSJ attack (i.e., no defence

Table 5. DFD service (frame) performance under attack

Dataset	Under Attack (frame)					
	Balanced Accuracy			AUC		
	HSJ	JPEG-C	SPSM	HSJ	JPEG-C	SPSM
CelebDF-V2	35.41%	51.34%	52.21%	25.95%	80.57%	83.12%
FaceForensics++	38.88%	50.57%	51.7%	35.04%	66.89%	65.46%
WildDeepFake	36.18%	51.2%	-	26.38%	74.3%	-



applied) and those with a pre-processing defence applied. Further investigation uncovered that, when pre-processing defences are applied in an attack scenario, the DFD service classifies fewer frames as deepfake and as such, on aggregation, no video is ultimately classified as containing a deepfake, resulting in a Balanced Accuracy of 50% across all data sets (for both JPEG compression and Spatial Smoothing). Table 5 illustrates that, whilst defence performance was similar for each dataset, SPSM had a higher accuracy frame-by-frame. Whilst the pre-processing defences do have some mitigating impact, at least at frame-level prediction 5, this analysis indicates that key perturbations added by the HSJ attack are escaping suppression and removal by both defence methods under current implementation and that the untuned pre-processor could be removing key features that the DFD service depends on to identify DeepFakes.

A successful attack against the DFD service, a binary classifier, can be defined as an attack which flips the original prediction regardless of whether the DFD service prediction was correct. Figure 3 illustrates the success rate for different L2 distance thresholds when the DFD service is under attack with and without defences applied. The figure highlights that with an L2 distance budget of 120, the HSJ attack is successful on over 50% of images (grey line), whilst adding the JPEG compression defence, reduces the success rate to slightly below 50% of images at the same L2 distance threshold (red line).

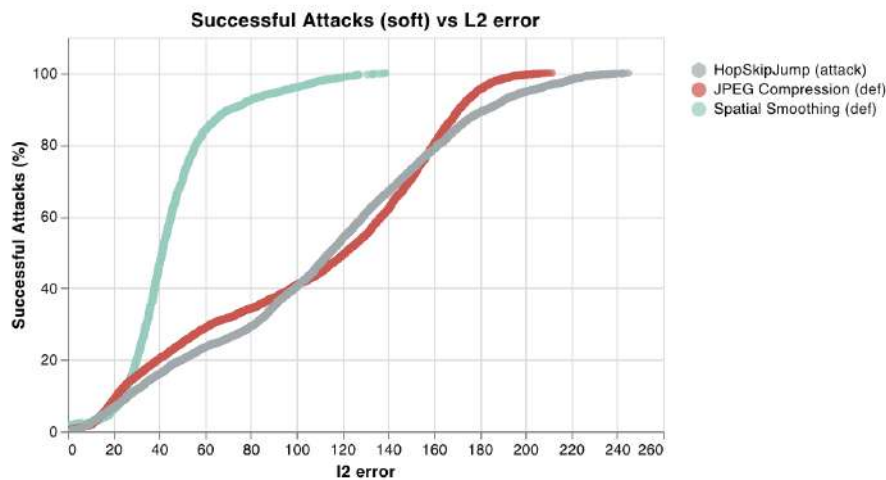


Figure 3. Success rate vs L2 distance threshold. Successful attacks are those which changed the DFD prediction.

In a black-box attack scenario such as HSJ, an adversary needs to query the DFD service multiple times (depending on the max. number of iterations selected to traverse the boundary) when generating an adversarial frame with sufficiently low L2 distance between the original and adversarial frame. Therefore, a potential defence could be to identify queries that are both similar (based on an L2 distance threshold) and converging. As an extra layer of defence, a monitoring system could be set up on the DFD service server to monitor consecutive queries from a user. If multiple queries are submitted which contain frames within a L2 distance threshold, it could signify a black-box attack. Figure 3 aids in selecting a threshold in this instance. For example, if all consecutive queries under 120 were filtered out, approx. 50% of successful attacks could potentially be avoided. The caveat in this case is that legitimate queries may also be included, and so future work should attempt to identify the optimal threshold at which to filter or flag potentially malicious queries. One such similar method, proposed by Li et al. [12], named Blacklight,





relies upon identifying similar queries using probabilistic content fingerprints and has shown to identify black-box attacks after a relatively low number of model queries. Figure 4 also illustrates the latency impact an applied pre-processor defence has on HSJ identifying an adversarial frame. With no defence applied, a batch of adversarial frames took approx. 80s to be generated. When a pre-processing defence was applied, generation latency rose significantly. In particular, adversarial batches of images with lower L2 distances took longest to generate (up to 500s).

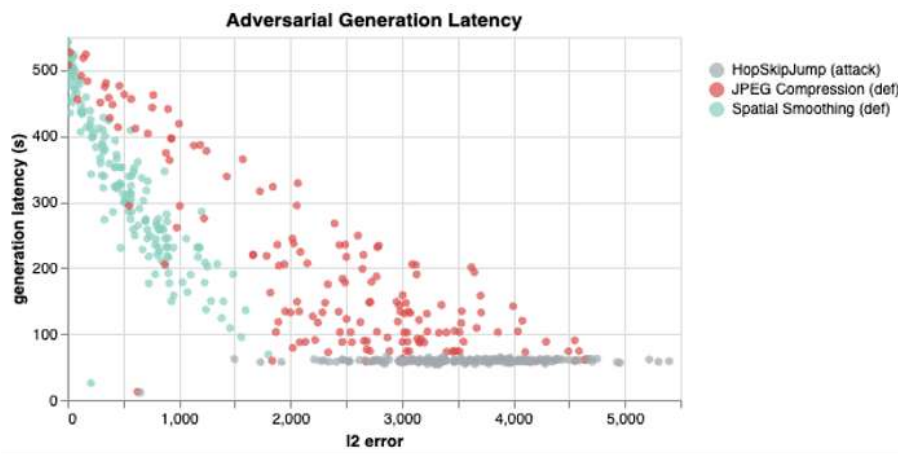


Figure 4. Time taken to generate batches of adversarial frames (batch size = 32).

3.1.4 Relevance to AI4Media use cases and media industry applications

The robustness evaluation of the MeVer Deepfake Detection Service is relevant to UC1 as it aids both the creators and users of the AI model better understand its limitations whilst attempting to detect disinformation. It provides additional transparency over the capability of the model which can facilitate more informed decisions regarding the deployment of the service and help users better interpret results and the associated limitations. For example, understanding that the model is vulnerable to certain black-box attacks would encourage additional scrutiny of requests sent to a model via an API and understanding that the model is susceptible to white-box attacks may initiate stricter developer access rules within the company deploying the model. This work is also relevant to UC2 “AI for News”, as deepfake content becomes more prevalent and journalists must differentiate between content that is authentic and content which has been created to misinform - trusting tools which can help discern real from misleading content, such as the MeVer Deepfake Detection Service, is essential and a robustness evaluation can contribute toward facilitating greater trust in these tools.

3.2 Federated Model Fusion and Robustness

Contributing partners: FHG-IDMT

The default model fusion algorithm for Federated Learning is embarrassingly simple, basically it averages the model weights of all clients and continues with that aggregate. Surprisingly, this works pretty good for a lot of cases where the data distribution is homogeneous between clients and there are no attackers in the system.



We performed research on how different aggregation algorithms perform with respect to certain evaluation criteria and found that large differences in system performance occurs between them, when one client (or more) sends wrong (“poisoned”) data. This was originally not planned to be part of Task 4.2, but makes a good case for AI robustness and is therefore mentioned here. A publication is planned and a more thorough discussion will be put in the next deliverable. For a quick look at the data, Figure 5 gives a first impression on how different aggregation algorithms influence the convergence of the trained model. For example, the plots in the upper right corner show a clear difference between qFedAvg and TWFedAvg, which behave relatively stable in the presence of poisoned data, while FedAvg does not converge at all. A more thorough analysis will be presented in the next deliverable and a respective publication.

Tangentially, this hints at the relationship between *fairness* and *robustness* of AI systems. While both are really broad categories (What is fair? What is robust?), a system that lacks strong fairness properties might also make it easier for an attacker to have a big influence on the model. A fair training process that tries to compensate for differences in the data distribution might be automatically more robust against faulty data – whether they come deliberately from an attacker or are accidental.

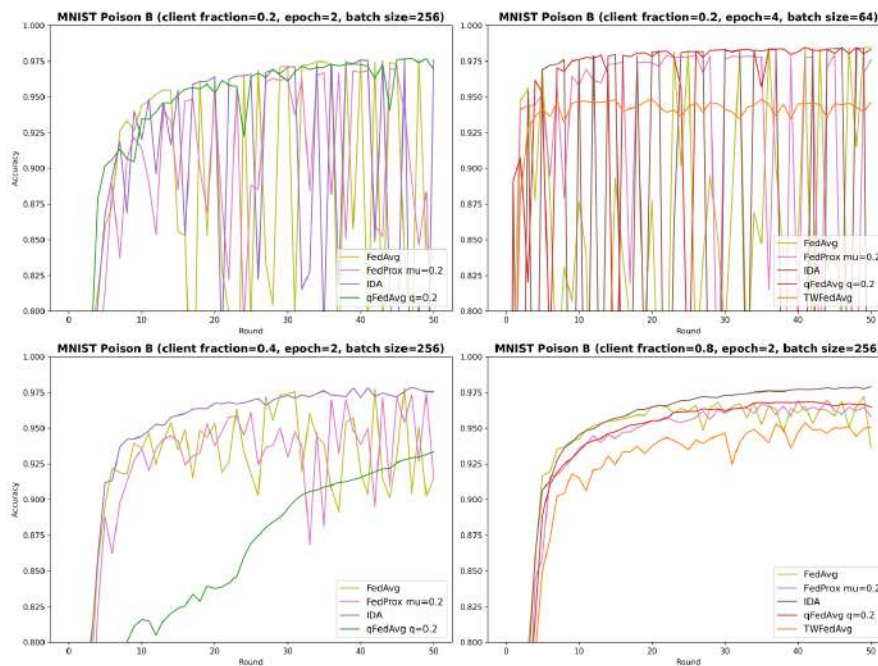


Figure 5. Model convergence under the presence of poisoned training data with different aggregation methods. Federated Averaging (FedAvg) shows oscillation while other approaches show a more stable (and robust) convergence.

3.2.1 Relevance to AI4Media use cases and media industry applications

While robustness is a feature of AI applications that only a few use cases will go without, the proposed approach deals with robustness within Federated Learning systems. Regarding AI4Media, there is no Use Case directly dealing with Federated Learning, yet. Regarding the broader media industry, the outlook of not having to share private data (being it usage, user, or content data), and



therefore avoiding all the practical hassles of data exchange (usage rights, data exchange contracts, data privacy laws, ...) is so promising, that there will be real industry applications for Federated Learning. On that premise, applications that improve the robustness and fairness (for privacy, see Section 5.4) of Federated Learning are worth researching on and will be relevant as in other non-media domains such as medicine or industrial applications.

3.3 Mitigating Robust Overfitting via Self-Residual-Calibration Regularization

Contributing partners: UNITN

Deep Neural Networks (DNNs) are very susceptible to adversarial examples which have been demonstrated to be threatening in various domains [13], including computer vision [14], [15], natural language processing [16], [17], and speech recognition [18]. These specific examples may be generated using various adversarial attack methods [19]–[22], causing the DNNs to behave incorrectly. We usually categorize adversarial attacks as either white-box or black-box attacks, both of which have been widely studied and shown in real-world scenarios [15], [23], [24]. It is remarkable that AutoAttack [25], a recently suggested ensemble of white-box and black-box attacks, has successfully broken several promising robust models. As a result, enhancing the adversarial robustness of DNNs⁴ has become essential for the community.

To this end, various defense methods [22], [26]–[33] have been proposed to defend against different kinds of adversarial attacks [19], [21], [22], [25], which can be mainly categorized into three groups. The first is input-transformation defenses, aiming to remove the adversarial noise from the inputs as studied in [10], [30], [34], [35]. The second belongs to certified defenses [27], [33], [36], which typically provides a tight robustness guarantee. The third promising solution is the Adversarial Training (AT) methods [21], [22], whose principle is to conduct a min-max optimization. The vanilla AT is the projected gradient descent based adversarial training (PGD-AT) [22], [29]. Through careful designs of different training schemes, recent works [37] have empirically verified that PGD-AT is robust to a number of adversarial attacks [19], [21], [22], [25], [38].

Although AT methods are commonly easy to implement and can achieve satisfactory defense results, they are prone to the risk of overfitting [29]. Briefly speaking, the best performance is achieved at a specific intermediate checkpoint, but further training will continue to decrease the robust training loss while increasing the robust test loss. This phenomenon is so-called *robust overfitting*. Several works [29], [31] adopted the early stopping strategy to avoid robust overfitting, in which they usually select the checkpoint by stopping at the first drop of learning rate⁵. However, whether the selected checkpoint is the optimal one remains an open problem. Other methods for solving robust overfitting and improving generalization are using regularization techniques, such as gradient penalties [37], [39], data augmentation [37], semi-supervised learning [40], and implicit regularizer [15], [31], [41], [42]. However, Rice *et al.* [29] observed that most commonly-used regularizations have a limited effect on addressing robust overfitting and improving generalization.

The purpose of this research is to address the robust overfitting problem in AT and to design a novel regularization scheme from a new perspective. We begin by an in-depth examination of the relationship between model calibration and robust overfitting, revealing two intriguing observations that provide new insights for solving robust overfitting. First, we find that robust overfitting is associated with confidence level, in which overconfidence on adversarial samples will easily lead to robust overfitting. Second, we find that there is a trade-off between the confidence of adversarial

⁴Here, adversarial robustness mainly refers to model’s robustness to the perturbations of input data. Briefly speaking, a model is robustness when it can defend against most kinds of adversarial attackers.

⁵The drop of learning rate indicates a special training epoch that requires the reduction of learning rate.





and natural images. This echoes the existing belief that a trade-off exists between natural and robust accuracy [31], [43], [44]. It is worth noting that confidence is not equal to accuracy, which reflects the statement of model calibration.

In response to our new observations, we propose a newly defined regularizer, called **Self-Residual-Calibration (SRC)**. Specifically, the proposed SRC is defined as the absolute residual between the natural and adversarial logit features, which has two advantages. First, by minimizing SRC, the model is encouraged to maintain the trade-off between the confidences of natural and adversarial samples. Second, SRC can be used to determine if a model is well calibrated during training. In most cases, the confidence of natural samples is higher than that of adversarial ones, which causes the SRC to omit the cases with higher confidence of adversarial samples, resulting in an imbalanced problem. To overcome this drawback of imbalanced training, we introduce a weighting strategy for adjusting the weights of samples that can satisfy different conditions. This strategy can be simply formulated by the pinball loss that computes the quantile residual between the logit features of natural and adversarial images.

Finally, we evaluate different robust models against PGD-attack and AutoAttack on three benchmarks, including CIFAR-10, CIFAR-100, and SVHN. The experimental results validate the merits of the proposed SRC over the state-of-the-art (SOTA) methods [31], [37], [42], [45], [46]. Moreover, our SRC is compatible with many regularizations, *e.g.*, CutOut [47], [48], Adversarial Weight Perturbation (AWP) [49], Semi-supervised Learning [40], Flooding [46], *etc.* By combining our SRC with them, the performance can be further improved. For example, with the help of AWP, our method can obtain about 55.0% adversarial accuracy on CIFAR-10 without using additional training data.

3.3.1 Experiments

Dataset: We conduct experiments on four datasets, including CIFAR-10 [50], CIFAR-100 [50], SVHN [51], and Tiny-ImageNet [52]. The results on Tiny-ImageNet can be found in the published article [53].

Baselines: We mainly compare the proposed SRC⁶ method with 6 baselines, including PGD-AT [22]⁷, ALP [45]⁸, TRADES [31]⁹, MART [42]¹⁰, AT with Flooding (AT-FL) [46]¹¹, and normal training on natural images (NT).

Evaluation Protocol: We train robust models using different methods. During testing, we evaluate the accuracies on (train/test) natural and adversarial samples. In addition, we also evaluate different model calibrations: reliability diagram and expected calibration error (ECE) [54], [55].

Network Setup: For CIFAR-10/100, we use pre-activation ResNet-18 [56] and WideResNet-34-20 [57] as the backbones. For SVHN, we use SmallCNN [22] as the backbone, which has three convolutional layers, followed by two fully-connected layers.

3.3.1.1 Evaluation on Robust Overfitting In this section, we analyze the problem of robust overfitting for different methods. The comparison results are shown in Table 6 and Figure 6. Specifically, in Table 6, the “D” columns indicate the difference between the highest and final checkpoints, reflecting the robust overfitting level. That is, a higher number of “D” means more serious robust overfitting level. In Figure 6 we show the adversarial loss curves and adversarial

⁶Note that, “SRC” refers to pinball-based SRC throughout this section.

⁷https://github.com/locuslab/robust_overfitting

⁸<https://github.com/labsix/adversarial-logit-pairing-analysis>

⁹<https://github.com/yaodongyu/TRADES>

¹⁰<https://github.com/YisenWang/MART>

¹¹<https://github.com/takashiishida/flooding>



	CIFAR-10									CIFAR-100			SVHN		
	Natural			PGD-10			AutoAttack			PGD-10			PGD-10		
	FC(↑)	BC(↑)	D(↓)	FC(↑)	BC(↑)	D(↓)	FC(↑)	BC(↑)	D(↓)	FC(↑)	BC(↑)	D(↓)	FC(↑)	BC(↑)	D(↓)
PGD-AT	84.00	82.73	1.27	45.54	53.43	7.89	42.23	48.30	6.07	21.48	28.30	6.82	57.10	58.17	1.07
ALP	82.86	80.59	2.27	51.55	54.71	3.16	47.87	49.82	1.95	-	-	-	-	-	-
TRADES	82.83	81.32	1.51	52.78	56.16	3.38	45.79	48.94	3.18	26.86	27.87	1.01	57.25	57.56	0.31
MART	82.10	77.69	4.41	51.62	57.91	6.29	46.01	48.98	2.97	-	-	-	-	-	-
AT-FL	82.37	82.52	0.15	56.76	57.83	1.19	44.13	45.41	1.28	27.59	28.29	0.7	58.14	58.43	0.29
SRC	81.76	80.58	1.18	54.59	55.58	0.99	50.00	50.39	0.39	27.08	29.14	1.80	57.89	58.07	0.18
SRC ($\tau=0.5$)	83.54	82.75	0.79	53.38	55.30	1.92	49.47	50.86	1.39	-	-	-	-	-	-
SRC w/ Softmax	82.80	82.29	0.51	54.10	56.70	2.60	49.81	50.93	1.12	-	-	-	-	-	-
SRC+FL	80.30	80.70	0.40	57.55	57.90	0.35	50.38	50.35	0.03	29.53	29.77	0.24	57.64	57.95	0.31

Table 6. Test robust accuracy (%) on three benchmarks. “Natural” is the natural accuracy. “BC” is the highest test accuracy observed during training. “FC” is test accuracy on the last epoch. “D” indicates the difference between the highest and final checkpoints. The symbol ↓ means the lower score is better, and the symbol ↑ means the higher score is better.

error curves for different methods, which enable us to investigate the problem of robust overfitting in more detail.

Effectiveness of SRC. As shown in Table 6, the proposed SRC can achieve a smaller degradation in robust accuracy, indicating that SRC can mitigate robust overfitting. For example, there is a 0.99 of performance degradation at the last checkpoint on CIFAR-10, which is on par with the best checkpoint. However, in the results of CIFAR-100 reported in Figure 6(a), we find that SRC begins to decrease after the second drop of learning rate, where robust overfitting still exists. Fortunately, we observe that combining Flooding and SRC helps address this problem, while increasing the robust accuracy (see the black curves shown in Figure 6(a)).

The results in Table 6 indicate that the original SRC can also help improve the robustness at the best checkpoint, but it will achieve higher “D” scores. Finally, because both logit and softmax features can reflect the model’s confidence, we use softmax features instead of logit features. We report the corresponding results in Table 6 and Figure 6. The experimental results show that using the softmax features helps improve the robustness of the model at the best checkpoint. It can achieve the best robust accuracy against strong AutoAttack. However, using the softmax features results in a more oscillating test curve compared to using logit features (see the orange curves in Figure 6(a)). These results verify the advantage of using logit features for SRC.

3.3.1.2 Evaluation on Robust Performance In this part, we mainly evaluate the robust performance of our SRC and the baseline methods. Specifically, the natural accuracy and adversary accuracy of these methods are reported in Tables 6 and 7.

Effectiveness of SRC. From Tables 6 and 7, we observe that the most commonly used regularization methods help improve the performance against PGD-attack with 10 iterations (PGD-10). With the help of many valuable tricks [37], the original PGD-AT can still achieve better performance, especially defending the AutoAttack. Compared to early stopping (ES) [29], three regularizations, i.e., TRADES, MART, and our SRC, have a performance degeneration on natural accuracy. Meanwhile, TRADES and MART have a slight performance improvement on adversary accuracy against AutoAttack, but our SRC can indeed improve the robustness against the AutoAttack, achieving at least 1.18% improvement.

Results of SRC with Flooding. We can find that AT-FL can reduce the degradation of the robust performance between the best and last checkpoints, as shown in Table 6. AT-FL helps finding a more robust model against the PGD-attack, which reaches averaged performance gains of 2.6%, 1.51%, and 1.52% on three datasets *i.e.*, CIFAR-10, CIFAR-100, SVHN compared to recent TRADES. However, AT-FL is still sensitive to the recent strong AutoAttack, which has an 11.93% degradation in robust performance on CIFAR-10. Interestingly, combining Flooding

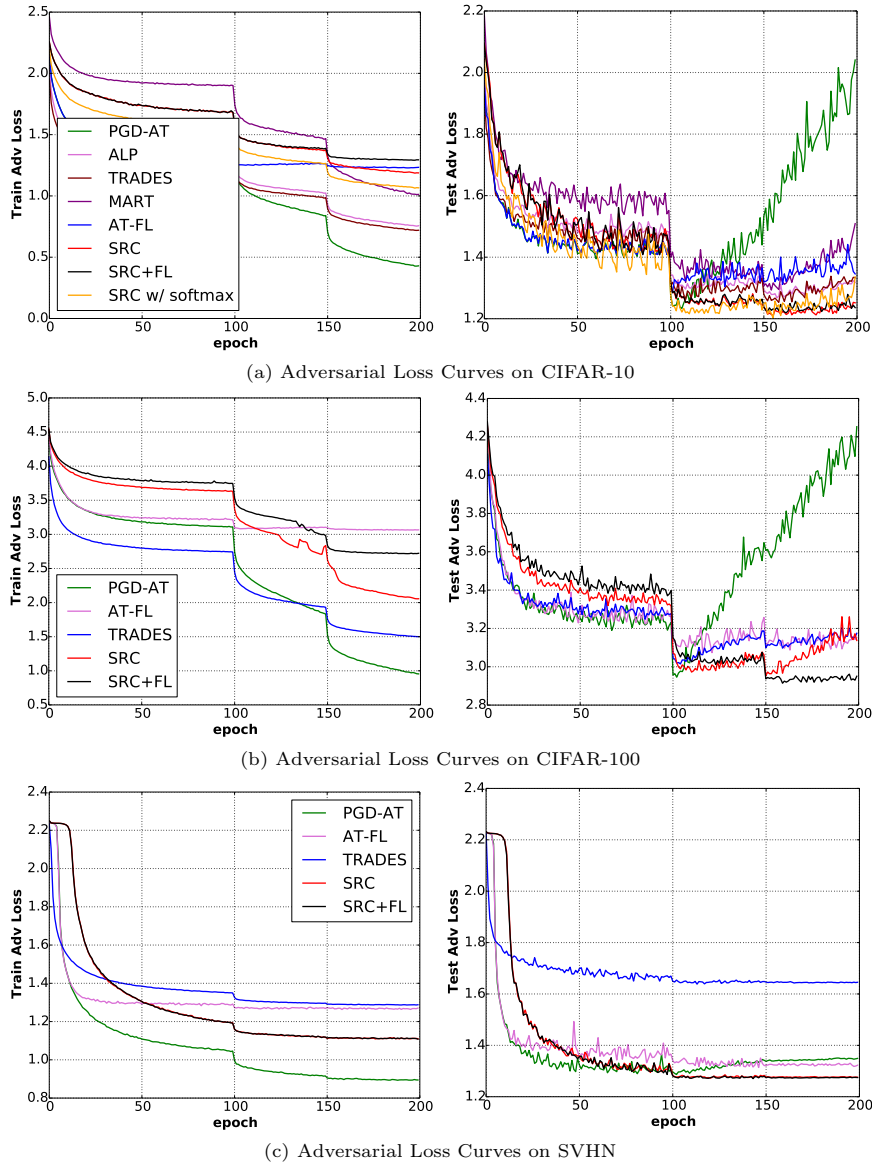


Figure 6. Learning curves on CIFAR-100 and SVHN for different models, i.e., PGD-AT, TRADES, AT-FL, and Ours. (Best view in color.)

and the most widely used regularizations like TRADES and MART hurts the performance against AutoAttack. Instead, combining Flooding and our SRC can achieve satisfactory results under different adversarial attacks. This verifies the effectiveness of the proposed SRC.

Effectiveness of SRC with AWP. Adversarial Weight Perturbation (AWP) [49] is a recent SOTA model which can achieve top performance on the leaderboard¹² of AutoAttack. Following the setting in [49], we combine our SRC with AWP, and we report the new combination method (SRC+AWP) in Table 7. We find that SRC+AWP has an average performance gain of 3.02%

¹²<https://robustbench.github.io/>



Method	Natural	PGD-10	AutoAttack
AT with Tricks [37]	83.40%	54.53%	49.80%
ES [29]	82.73%	53.43%	48.30%
TRADES	81.32%	56.13%	48.94%
MART	77.69%	57.91%	48.98%
SRC	80.58%	55.58%	50.39%
SRC (WRN)	85.84%	58.37%	53.94%
AT-FL	82.34%	57.82%	45.41%
TRADES+FL	-	59.16%	46.58%
MART+FL	-	58.63%	45.56%
SRC+FL ($\beta = 5.0$)	82.29%	57.90%	50.35%
SRC+FL ($\beta = 6.0$)	79.27%	57.95%	51.05%
AWP	81.26%	55.76%	50.34%
SRC+AWP	83.95%	57.31%	51.86%
SRC+AWP+FL	77.36%	58.93%	50.94%
SRC+AWP (WRN)	87.19%	59.83%	55.00%
SRC+D	86.13%	58.93%	54.07%
SRC+D+FL	83.40%	59.28%	51.67%
SRC+D (WRN)	90.12%	63.73%	60.01%
FAST [58]	83.34%	47.37%	42.53%
FreeAT [59]	79.31%	46.49%	41.37%
SRC+FGSM	79.03%	50.05%	46.39%

Table 7. Test accuracy on CIFAR-10. “+D” means semi-supervised data augmentation is used. “WRN” means WideResNet is used. “+FL” means Flooding is used. “+AWP” means the adversarial weight perturbation [49] is used. PGD-AT is the recent SOTA model with considering many useful tricks [37]. “ES” means the early-stopping is used.

over SRC, under two adversarial attacks, showing that AWP can benefit the model robustness. Meanwhile, we further evaluate the adversary accuracy against AutoAttack, when the robust model is WideResNet-34-20. Under this setting, we notice that SRC+AWP can achieve satisfactory results on CIFAR-10, and these results are competitive on the leaderboard, without using any extra or synthesized dataset.

Effectiveness of SRC with Semi-Supervised Learning. Semi-supervised learning, like the prior works [40], [60], can further improve the robust performance, and it helps defend against the AutoAttack. We use the same framework as in [40] and add the SRC during training, which is called “SRC+D”. We find that unlabeled data also has a positive impact on improving the robustness, which achieves at least 6.03% performance improvement. Furthermore, when replacing the backbone with WideResNet-34-20, the best robust performance can be achieved. Specifically, as shown in Table 7, SRC+D (WRN) can reach 60.01% test robust accuracy against AutoAttack, which can achieve at least top-13 on the leaderboard of AutoAttack. We believe that using other data augmentation techniques like [61] may help improve the performance further, and we leave it for the future work.

3.3.2 Conclusions

Overall, in this research, we focus on mitigating the problem of robust overfitting. Our main innovations are as follows:

1. We observe two intriguing properties through the analysis of different model calibrations. These properties reflect an important relationship between robust overfitting and model calibration, motivating us to overcome robust overfitting from a new perspective.





2. We draw inspiration from our experimental observations and introduce a new regularizer for AT, which can effectively avoid robust overfitting and consistently improve defense performance.
3. Experiments show that the proposed SRC can achieve SOTA adversarial accuracy against both PGD-attack and AutoAttack on three commonly used benchmarks.

Our experiments verify that the proposed method can help better defend the cutting-edge adversarial attack methods. Recent studies show that the multi-modal models like CLIP [62] can substantially improve the robustness of a model [63]. This advantage is mainly benefited from the large-scale, diverse, multi-modal data. In our future work, we will also focus on how to handle robust overfitting and enhance adversarial accuracy against adversarial attacks with the help of multi-modal models.

3.3.3 Relevant Resources and Publications

Relevant publications:

- H. Liu, Z. Zhong, N. Sebe, and S. Satoh, “Mitigating Robust Overfitting via Self-Residual-Calibration Regularization”, *Artificial Intelligence*, vol. 137, Article 103877, April 2023. [53]. Zenodo record: <https://zenodo.org/record/7858712>.

Relevant resources:

- The Pytorch implementation can be found in <https://github.com/LynnHongLiu/AIJ2023-SRC>.

3.3.4 Relevance to AI4Media use cases and media industry applications

Adversarial training is one of the methods used to defend against the threat of adversarial attacks but it is prone to overfitting. Briefly speaking, the best performance is achieved at a specific intermediate checkpoint, but further training will continue to decrease the robust training loss while increasing the robust test loss. This phenomenon is so-called robust overfitting. Our solutions to avoid robust overfitting are of utmost relevance to all the use cases in which a deep learning model needs to be learnt. We have discussed in the section the application on image analysis so the approach could be directly relevant to use cases (a) 3A3 (archive exploration), specifically 3A3-11 Visual indexing and search and (b) 7A3 (Re)organisation of visual content by supporting the efficient training and organization of image and video collections. However, the approach can also be applied when other modalities are involved, e.g., 4C3 (audio analysis).

3.4 Geometrically-inspired training scheme for adversarial robustness

Contributing partners: AUTH

A classification system that operates for some specific use case, e.g., a biometric authentication system, could be vulnerable to adversarial attacks, especially if it is based on a neural network. Nevertheless, even when the system vulnerabilities are known, in many cases, it is still very difficult to completely replace the classification system/model that is already installed and running. Therefore, adversarial robustness methods that can be used in existing neural network architectures that are relatively easy to implement, e.g., re-train the neural network with different optimization criteria, are very useful. For this reason, we worked on devising a neural network optimization method that does not add significant computational overhead to the standard training procedure,





which is based on geometric criteria. In order to devise those geometric optimization criteria, we performed a deep dive into one-class classification methods, which were used to define objective functions that aim to minimize the representation variance between items belonging to the same class, for instance, vectorized representations of images that are used to identify a user. As will be shown below, the proposed optimization criteria leads some well known adversarial attacks to fail more often in transferability attack settings, when compared to standard training or adversarial training principles.

3.4.1 Overview

In a more technical fashion, adversarial defenses in classification systems aim to increase their ability to withstand or overcome input perturbation, generated by adversarial attacks. Let a classification system $y = f(\mathbf{x}; \boldsymbol{\theta})$, where f is the model decision function parametrized by $\boldsymbol{\theta}$, \mathbf{x} are the model inputs and y is the model prediction. Robustness is quantified by determining its tolerance to perturbation $\|\mathbf{p}\| < \epsilon$ per se, i.e., $f(\mathbf{x}; \boldsymbol{\theta}) = f(\mathbf{x} + \mathbf{p}; \boldsymbol{\theta})$. Other definitions of adversarial robustness have been proposed in the past, that focus on altering the classification architecture, e.g., input filtering [64], Generative methods [65]. Using the above definition of robustness, we consider such methods irrelevant to the proposed one.

Our work focuses on adversarial defenses that modify the training process of a neural network, while maintaining the same neural network architecture, only by trying to derive in different parameters i.e., $f(\mathbf{x}; \hat{\boldsymbol{\theta}})$. One approach to this end is to fine-tune or re-train the model by exploiting adversarial samples, derived by employing one or more adversarial attack methods, calculated implicitly or explicitly [66], [67]. The main disadvantages of these approaches are the introduced workflow for calculating the adversarial examples, while at the same time, model classification accuracy in clean data is negatively affected [68]. Moreover, due to the adversarial attack-specific nature, there is no guarantee that such defenses remain effective against different types of adversarial defense. Ultimately, the effectiveness of adversarial defense methods that fall into the above category seems to rely on achieving the production of as similar intermediate data representations as possible for both clean and adversarial images belonging to the same specific class. Recently proposed adversarial defenses showed that incorporating distance-based optimization criteria might achieve this goal, without requiring re-training the model with adversarial examples [69], [70]. The second advantage of such methods is that they might employ adversarial training as a complementary step, providing increased robustness to specific adversarial attacks.

This work extends the recently proposed Hyperspherical Class Prototypes (HCP) method [70], by incorporating novel optimization terms inspired by the present state-of-the-art in deep neural network-based one-class classification problems. The proposed method does not imply modifications to the deep neural architectures or the creation of adversarial examples for training purposes. It is deployed in the form of alternative loss functions that supervise the distribution of final and intermediate layer activation values. It is shown that the proposed method increases (or at least does not hinder) the classification accuracy in clean examples, while it provides increased robustness to adversarial attacks at the same time. The proposed method is evaluated in black-box/transferability-based adversarial attack settings in image classification tasks.

3.4.2 Robust One-class Classification-based training loss

The developed method alters the training procedure of a standard neural network architecture, by training in-parallel, additional layer(s) that learn prototype vector centers in the feature space. By minimizing the variance between various class items representations with their corresponding prototype vector, we argue that adversarial attacks require to add more noise to the representation,





in order to be successful. Probably the best way to formulate and learn class prototypes, is by devising one-class classification methods, since the focus of such methods is to learn the optimal way to separate each single class from the rest of the dataset.

Let \mathcal{K} be the set of layers on which the proposed objectives will be applied to, where $\mathbf{g}_k(\mathbf{x}; \boldsymbol{\theta})$ is k -th layer representation of some input \mathbf{x} . This method aims to learn hyperspherical prototypes in the k -th layer defined by the prototype matrices $\mathbf{A}^{(k)} \in \mathbb{R}^{C \times L_k}$, where L_k is the dimensionality of the k -th layer, and radii $\mathbf{R}^{|\mathcal{K}| \times C}$ that will act as one-class classifiers, verifying data sample activations belonging to the j -th class. To this end, the optimization problem for each sample \mathbf{x}_i is the following:

$$\begin{aligned} \min_{\mathbf{R}, \boldsymbol{\Xi}, \mathbf{A}^{(k)}} \quad & \sum_{k \in \mathcal{K}} \sum_{j=1}^C r_{kj}^2 + \sum_{k \in \mathcal{K}} c_k \sum_{i=1}^N \xi_{ki} \\ \text{s.t.} \quad & \sum_{k \in \mathcal{K}} \sum_{j=1}^C \left(-y_{ij} \left(r_{kj}^2 - \|\mathbf{g}_k(\mathbf{x}_i; \boldsymbol{\theta}) - \mathbf{a}_j^{(k)}\|^2 \right) \leq \xi_{ki} \right), \\ & \xi_{ki} \geq 0 \end{aligned} \quad (1)$$

where $\mathbf{a}_j^{(k)}$ is the prototype center for class j , $y_{ij} = 1$ if sample \mathbf{x}_i belongs to class j , or $y_{ij} = -1$, otherwise, ξ_{ki} are the slack variables and $c_k \geq 0$ is a hyperparameter that allows training error (i.e., soft margin formulation) relaxing the optimization constraints. The constraints of the above optimization problem can be optimized by applying the following hinge loss function in every layer selected in \mathcal{K} :

$$\mathcal{L}_M = \sum_j^C \max \left(c_k, -y_{ij} \left(r_{kj}^2 - \|\mathbf{g}_k(\mathbf{x}_i; \boldsymbol{\theta}) - \mathbf{a}_j^{(k)}\|^2 \right) \right). \quad (2)$$

Both the feature vectors and the prototype vectors are trainable parameters. We employ a value of $c_k = 0$. The loss value is $\mathcal{L}_M > 0$ if and only if the one-class classifier decision function misclassifies \mathbf{x}_i . The compactness of the derived class representations is proportional to the learned value of the corresponding radius r_{kj} .

The above function does not produce loss values for marginal data items, i.e., items lying close to the hypersphere boundaries. To this end, we employ a contrastive loss term for items belonging to the same class. We consider a mini-batch of size N is randomly sampled and the contrastive prediction task is defined on pairs of data representations derived from the mini-batch, resulting in $2N$ data points. For a pair of data representations $\mathbf{z}_1 = \mathbf{g}_k(\mathbf{x}_1, \boldsymbol{\theta}) - \mathbf{a}_j^{(k)}$, $\mathbf{z}_2 = \mathbf{g}_k(\mathbf{x}_2, \boldsymbol{\theta}) - \mathbf{a}_j^{(k)}$ belonging to the j -th same class, the loss function is defined as follows:

$$\mathcal{L}_C(\mathbf{z}_1, \mathbf{z}_2) = -\log \left(\frac{\exp(\mathbf{z}_1^T \mathbf{z}_2 / T)}{\exp(\mathbf{z}_1^T \mathbf{z}_2 / T) + \sum_{i=2}^{2N} \exp(\mathbf{z}_1^T \mathbf{z}_i / T)} \right) \quad (3)$$

where \mathbf{z}_i are the remainder mini batch representations and T is the so-called temperature hyperparameter (a value of $T = 0.25$ was used in all our experiments). The introduction of the above loss term promotes the derivation of similar representations in the feature space, without minimizing their Euclidean distance.

However, the \mathcal{L}_C might indirectly increase the Euclidean distance, especially if it is very small, which is something that is contradicting to adversarial robustness. Therefore, we follow the same practice and also employ an Angular loss term to complement this contrastive loss:

$$\mathcal{L}_A(\mathbf{z}_1, \mathbf{z}_2) = \|\mathbf{z}_1^T \mathbf{z}_2\|^2. \quad (4)$$





Table 8. Classification accuracy of the competing methods.

Method/Dataset	CIFAR-10	CIFAR-100	SVHN
Vanilla	93.36	74.04	96.23
Center Loss	93.77	69.75	95.90
PCL [69]	92.30	68.19	95.37
HCP [70]	93.31	72.83	95.85
ROCC	94.46	73.62	96.31

Table 9. Robustness (classification accuracy) in PGD black-box attack, by using the Vanilla ResNet architecture as attack model.

Method/Dataset	CIFAR-10	CIFAR-100	SVHN
Center Loss	57.60	40.40	86.59
PCL [69]	61.61	42.55	84.94
HCP [70]	60.67	46.92	86.50
ROCC	65.09	44.97	86.92

Finally, we formulate the proposed learning procedure called Robust One-class Classification (ROCC) loss function as the combination of the constraints of the abovementioned optimization terms, as follows:

$$\mathcal{L}_{ROCC} = \mathcal{L}_M + \mathcal{L}_C + \mathcal{L}_A. \quad (5)$$

3.4.3 Experiments

ResNet-101 [71] was employed as the baseline architecture. We have employed the publicly available CIFAR-10, CIFAR-100 [50] and SVHN [72] datasets. In our first set of experiments, we compare the classification accuracy of various defences. Table 8 reports the obtained classification accuracy in the respective datasets. As can be observed, the proposed method outperforms all other adversarial robustness methods in every case while it even outperformed the vanilla softmax optimization function in two cases. This can be attributed to the fact that the proposed optimization functions only consider how to obtain better representations for each class, thus being compatible with any standard classification loss function.

In our second set of experiments, we evaluate the Robustness of the competing methods to the iterative projected gradient descent (PGD) [66] attack, with a corresponding parameter $e = 0.1$. To this end, we employed the Vanilla ResNet architecture for generating adversarial samples and inferred their labels by the respective robust models trained using the competing methods. Here it should be noted that this attack is the strongest form of transferability attacks, since the only difference between the attack and target architecture are the network parameters. The results are reported in Table 9. As can be observed, in the 10-class datasets (CIFAR-10, SVHN) the proposed ROCC method outperformed the competition, except for the CIFAR-100 case.

Finally, in our third set of experiments, we employed the competing architectures to attack each other, as "host" and target architectures. We again used the PGD attack with $e = 0.1$. Here, it should be expected that the most robust architectures are supposed to a) remain robust in transferability attacks and b) create strong adversarial samples that are able to fool the other defenses. As can be observed in Table 10, the proposed ROCC method produces the strongest transferability attacks among the competition (red), while at the same time, it remains the most



Table 10. Cross-method black-box PGD attacks in CIFAR-10.

Attack Method/Robust Method	Center Loss	PCL [69]	HCP [70]	ROCC
Center Loss	-	73.46	75.51	80.53
PCL [69]	69.83	-	75.21	78.90
HCP [70]	78.17	79.47	-	83.34
ROCC	64.01	65.16	67.23	-

robust in the opposite scenario (bold).

3.4.4 Conclusions

This work described an adversarial robustness method by exploiting and re-formulating one-class classification inspired optimization criteria. The proposed optimization scheme increases adversarial robustness in black-box adversarial attacks without negative effects on classification accuracy. An interesting link was found, between one-class classification and adversarial robustness. The proposed criteria should also be studied in other forms of computer vision problems, e.g., regression-based problems such as object detection/tracking.

3.4.5 Relevant publications

- V. Mygdalis and I. Pitas, “Exploiting One-Class Classification optimization objectives for increasing Adversarial Robustness”, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023.
Zenodo record: <https://zenodo.org/record/8276389>
- V. Mygdalis and I. Pitas, “Hyperspherical class prototypes for adversarial robustness”, Elsevier Pattern Recognition, vol 125, pp 108527, 2022.
Zenodo record: <https://zenodo.org/record/5137295>

3.4.6 Relevance to AI4Media use cases and media industry applications

This technology provides new optimization objectives for general purpose deep neural network training that can strengthen and robustify them against adversarial threats. Compatible neural networks with the developed optimization objectives can be found in several AI4Media use cases. Such special focus can be found in AI4Media UC1: “AI for Social Media and Against Disinformation”, in neural networks that are fighting disinformation by detecting deep fakes. More concretely, some formats of deep fakes that can be produced in an adversarial manner, e.g., by optimizing for fooling a disinformation detector, will be hindered if this deep fake detector is trained with the developed technology.

3.5 Matching Pairs: Attributing Fine-Tuned Models to their Pre-Trained Large Language Models

Contributing partners: IBM





3.5.1 Overview

Large Language Models (LLMs) or more generally Foundation Models are an emerging technology of general-purpose AI models trained on large volumes of data which can be fine-tuned for a wide range of downstream tasks. LLMs, in particular, can generate novel high-quality text and help drive many downstream applications including machine translation, question answering, and text summarization systems.

However, training these models is challenging as it requires access to vast amounts of data (text corpus) and large compute. This has led to a market where developers, who often don't have access to such resources, source LLMs from third-parties (which we refer to as base models) and fine-tune them for specific domain/tasks. With about 450 start-ups working on generative AI¹³ and over 100,000 models hosted in some repositories¹⁴, there are growing threats like violation of model licenses, model theft, and copyright infringement. Moreover, recent advances have shown that generative AI is also capable of producing harmful content [73] which only exacerbates the problems of accountability within ML supply chains. As approaches like watermarking are shown to be easily bypassed, there's a need for developing general purpose solutions to help with forensics. This work takes the first step to tackle these open challenges by developing defense methods that can attribute a fine-tuned language model to its base model. Establishing this attribution relationship is the first line of defence for an AI forensic investigation.

This work formalizes the role of attribution within the supply chain and presents heuristic and ML based approaches for attribution under different knowledge levels. Furthermore, it shows how such methods can be made more efficient for constrained settings where an attributor may have limited access to the model and/or its API. Attribution is an important step in making the AI model supply chains more robust.

3.5.2 Experiments

This work considers two collections of LLMs — the first one is a set B of pre-trained base LLMs, and the second one is a collection F of fine-tuned LLMs. It assumes that every model $m_f \in F$ was effectively obtained by fine-tuning a model $m_b \in B$. The goal of LLM attribution is to design a function $f : F \rightarrow B$ that maps a given fine-tuned model $m_f \in F$ back to its corresponding base model $m_b \in B$. This work trains a classifier to model this function. The classifier captures the correlations between an arbitrary response and the base model m_b . For example, with a prompt p , this could capture the relationship between a response $m_b(p)$ and m_b . Similarly, one can capture the relationship between a response $m_f(p)$ and m_b where m_f is obtained by fine-tuning m_b . Assuming that such correlations are preserved in a base model and fine-tuned model pair, the classifier can determine the attribution of a fine-tuned LLM.

Given a set of prompts p_1, \dots, p_K , there are multiple ways to prepare them for the classifier. One can apply the target base model, or fine-tuned model to get the responses, and concatenate the prompt and its response. Specifically, this work considers the following input representations (here, SEP refers to commonly used separators like comma or colon):

- Base model only (\mathbf{I}_B): “ $p_i m_b(p_i)$ ”
- Fine-tuned model only (\mathbf{I}_F): “ $p_i m_f(p_i)$ ”
- Base model + fine-tuned model (\mathbf{I}_{B+F}): “ $p_i m_b(p_i) \text{ SEP } p_i m_f(p_i)$ ”
- Separate embeddings for the base model and fine-tuned model.

¹³<https://www.nfx.com/post/generative-ai-tech-market-map>

¹⁴<https://huggingface.co/>





Attribution Method	K	$m_{\#}$										TP
		0	1	2	3	4	5	6	7	8	9	
HDT	K_U	✓	✓	✗	✗	✓	✗	✗	✓	✗	✓	5
Perplexity	K_U	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	1
TripletNet + P1	K_U	✗	✗	✓	✓	✗	✗	✗	✗	✗	✓	3
BERT + I_F + P1	K_U	✗	✓	✓	✓	✗	✗	✗	✓	✓	✓	6
BERT + I_{B+F} + P1	K_U	✗	✓	✓	✗	✗	✗	✓	✓	✓	✓	6
Exact matching	K_R	✓	✓	✓	✗	✗	✗	✓	✗	✗	✓	5
BERT + I_B + P1	K_R	✓	-	✓	-	✗	✗	✓	✓	✓	✓	6
BERT + I_B + P3	K_R	✓	✗	-	✗	✗	✓	✓	✓	-	✓	5
BERT + $I_B+P1+P2$	K_R	✓	✓	✓	-	✗	✓	✓	✓	✓	✓	8

Table 11. Model Attributions on $m_{\#}$ from the different methods. Dashes (-) are used when multiple models (m_f) are attributed to m_b . TP denotes True Positives

The second important bit in the design of the attributor is the choice of prompt set. More specifically, this work considers three approaches: a small set (**P1**) of *edge cases* that are distinct to each corpus, a naive collection (**P2**) of prompts, and reinforcement learning to select a subset (**P3**) from the edge cases.

A summary of the attribution approaches is provided in Table 11 where the approach was tested on 10 pre-trained models labelled 0-9. The Heuristic Decision Tree (HDT) and perplexity based solutions provide baseline approaches. HDT uses a series of discriminative heuristics to categorise F . Similarly, perplexity can be leveraged for measuring attribution by computing the perplexity of m_b relative to the response of m_f to prompt p . A lower perplexity would be indicative of an existing attribution relationship between m_b and m_f . And finally, the exact match as the name suggests looks at responses from m_b and m_f for the same prompts.

Under K_U conditions (described in [74]) the baselines of Perplexity and HDT are only able to correctly attribute 1 and 5 models respectively. Perplexity fails to capture the subtlety of attribution, as repetitive responses lead to lower perplexity and so incorrect attribution. The HDT particularly fails to account for overlap in pre-training and fine-tuning. For instance, DialoGPT-Large and m_{f3} (fine-tuned version of distilgpt2) respond in similar short sentences that leads to incorrect attribution. The TripletNet baseline performs poorly, only correctly attributing 3 of the models. Both BERT based attributors are able to attribute more models correctly in comparison to the baselines.

Examining the models at K_R (described in [74]) shows similar performance. The exact match correctly attributes 5 models and BERT+ I_B identifies 6 models. BERT+ $I_B+P1+P2$ attributor is the most successful by correctly attributing 8 models. Note that this model is the most expensive to train as we have to query a large number of prompts.

3.5.3 Conclusion

This work took initial steps in the LLM attribution problem. It studied LLM attribution in different settings which limit access to B and F to different levels and provides an interesting and realistic study of LLM attribution. It considered a variety of different LLMs that were trained on different datasets, and for different purposes. It postulated that the 10 different LLMs provide a didactic range of models for LLM attribution. In the experiments, it used pre-existing LLMs that were fine-tuned by the open-source community to demonstrate the applicability of our methodology. Overall, our work contributes to the growing understanding of LLM attribution, laying the foundation for



future advancements and developments in this domain.

3.5.4 Relevant Resources and Publications

Relevant publications:

- Myles Foley, Ambrish Rawat, Taesung Lee, Yufang Hou, Gabriele Picco, and Giulio Zizzo. 2023. Matching Pairs: Attributing Fine-Tuned Models to their Pre-Trained Large Language Models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7423–7442, Toronto, Canada. Association for Computational Linguistics. [74].
Zenodo record: <https://zenodo.org/record/8281959>.

Relevant software and/or external resources:

- Model Attribution in Machine Learning, Github repository: <https://github.com/IBM/model-attribution-in-machine-learning>

3.5.5 Relevance to AI4Media use cases and media industry applications

Rapid adoption of machine learning across all industries has raised challenges on model ownership and traceability. For instance, one is likely to face issues on model theft or copyright infringement. This is particularly relevant for sectors like media industry where generative models are being used to create content. Given an access to an API for generating content, how can one trust the source of the API? In this work, we provided a method to establish this relationship and make ML supply chains more robust.



4 Explainable AI (Task 4.3)

The last decade has seen a tremendous adoption of AI technology across a wide range of industries. AI has now become an indispensable part of our society. Accompanying this adoption however is an increasing concern about the opacity of such systems to human scrutiny. The reasons why such systems arrive at specific decisions are in most cases unknown to their users. In many cases, this opacity exists as well for the designers of such systems. This situation is thus one of the main obstacles that prevent the further adoption of AI technology across society today.

Explainable AI hence attempts to provide tools which enable the generation of explanations clarifying how a given model reached a decision and are understandable by humans. The methodologies and tools presented in this section hence addresses the need in the industry and society at large for AI models that can provide human understandable explanations of their underlying mechanisms.

Contributions towards the **Explainable AI** task (T4.3) include work on (i) using visualisations to explain deep learning insights when detecting synthetic audio (Section 4.1), (ii) designing a new attention mechanism for visual explanations of image classifiers (Section 4.2), (iii) new learning methods for semantic editing of GANs for generating images (Section 4.3), (iv) disentangling neuron representations with concept vectors (Section 4.4), (v) a novel architecture for combining multitask learning and adversarial training (Section 4.5), (vi) explainability in autonomous driving systems (Section 4.6), (vii) explainability in multi-model AI systems (Section 4.7), (viii) an analysis of *Anchors* in text classification systems (Section 4.8), (ix) explainability through concept-based models (Section 4.9), and (x) an extension of concept bottleneck models (Section 4.10). An outline of the first Nice Workshop on Interpretability (NWI) is given in Section 4.11.

4.1 Deep Learning Insights into Synthetic Audio Detection: An Interpretable Approach Using Saliency Maps

Contributing partners: CERTH

4.1.1 Overview

The rapid proliferation of digital technology has catalyzed an upsurge in the production of high-quality synthetic audio. While this technological breakthrough offers a plethora of benefits across numerous sectors, it simultaneously introduces a formidable challenge, that is, distinguishing synthetic audio from real audio. Consequently, this has prompted the need for robust detection systems, with deep learning models featuring prominently due to their superior performance.

Nevertheless, these deep learning models often invite criticism for their “black box” nature, characterized by decision-making processes that hinge on intricate internal operations that are neither immediately transparent nor understandable to users [75]. This lack of clarity can erect significant barriers to the practical deployment and user trust in the system’s verdicts.

In order to counteract this limitation, there has been a surge in research endeavors aimed at augmenting the interpretability of deep learning models. This necessitates making the decision-making processes of these models more explicit and intelligible [76]. In the context of synthetic audio detection, enhanced interpretability can provide insights into the specific features the model leverages to classify an audio sample as synthetic or real.

In the presented approach, the principal aim is to augment the accessibility of synthetic audio detection models for non-technical users. By leveraging Gradient-weighted Class Activation Mapping (Grad-CAM) [77], the model’s decision-making process is visualized, thereby transforming





an otherwise opaque computational process into an understandable and interpretable output. This strategy is purposely designed to circumvent the need for expertise in spectrogram interpretation, thereby making the complexity of synthetic audio detection more approachable for users of varying technical backgrounds.

4.1.2 Methodology

Our interpretative AI solution for synthetic audio detection is primarily designed around the Grad-CAM [78]. Grad-CAM, a visualization technique developed by Selvaraju et al. [78], is used for identifying the significant areas within an input that a Convolutional Neural Network (CNN) utilizes for class distinction. Here, we employed the deep CNN model, VGG16, due to its well-acknowledged proficiency in image recognition tasks [79].

The methodology begins with the generation of saliency maps from Mel spectrograms of the audio samples, as shown in Figure 7. Mel spectrograms are a specific kind of spectrograms, which scale frequency in a way that is intended to mimic the human ear's response to different frequencies [80]. The audio data we used is sourced from the Fake-or-Real (FoR) dataset [81]. This dataset comprises more than 117,000 real speech utterances and 87,000 synthetic phrases, collected from various open-source datasets and generated using commercial and open-source Text-to-Speech (TTS) systems. The dataset is gender-balanced and encompasses a variety of voices and recording devices to avoid overfitting.

The saliency maps are then converted into frequency vectors by summing the pixel intensities across the time axis for each frequency bin. This process enables us to concentrate on the frequency regions the model finds critical, effectively eliminating the temporal factor that could introduce significant variability.

Average saliency maps are computed for each class, synthetic and real, from the training dataset. These averages serve as a benchmark when comparing the saliency map of a new audio sample. Here, the comparison is conducted using cosine similarity [82], which quantitatively encapsulates how similar the pattern of frequency importance of a new sample is to the established patterns for each class. The use of cosine similarity, a measure of the cosine of the angle between two vectors, provides an easily interpretable metric for both technical and non-technical users.

To counteract the dominance of less significant areas in the new sample, a thresholding mechanism is utilized. In this context, a threshold is set for the pixel intensities, where only those exceeding this threshold are considered 'important', while others are disregarded. This approach allows the model to focus only on the highly significant regions within the spectrogram.

By employing this methodology, we can generate an interpretable output, understandable even without a deep understanding of Mel spectrograms or audio analysis techniques.

4.1.3 Results

The methodology applied in this work unveils crucial aspects of interpretability in synthetic audio detection, illuminating the decision-making processes of the VGG16 model. Through the use of Grad-CAM visualizations, distinct areas contributing to the model's classifications between synthetic and real audio are graphically presented.

In scrutinizing the average saliency maps (Figure 8), the areas within the frequency domain that significantly influence the model's classification decisions are showcased. A visual analysis of these maps reveals a consistent pattern: the model tends to be more active in lower frequency bands across both classes, while a stronger activation is observable in higher frequencies for real audio instances. This pattern might suggest a pivotal role of frequency bands in distinguishing synthetic audio from real ones, constituting a promising area for future exploration.



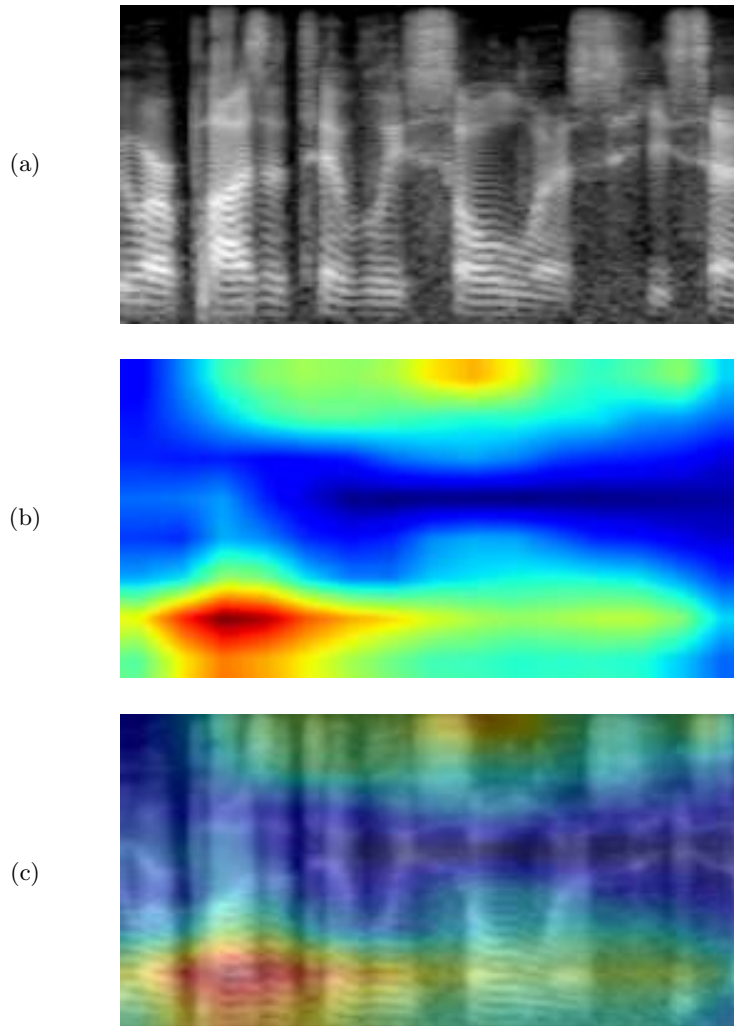


Figure 7. Visualization of the spectrogram, saliency map, and superimposed image of an audio sample. (a) The spectrogram shows the frequency components of the audio sample over time. (b) The saliency map shows the areas of the audio sample that are most salient, or attention-grabbing. (c) The superimposed image shows the audio sample with the saliency map overlaid.

The use of cosine similarity in this context serves as an intuitive metric that provides a window into the model's level of confidence when classifying a new instance. A confidence score, calculated as the absolute difference between the cosine similarity of a new instance with the average saliency map of the synthetic and real audio classes, is introduced. The larger this score, the more confident the model is in its classification.

The histogram in Figure 9 visually portrays the distribution of these confidence scores. The median confidence score is found to be approximately 0.1088, indicating that for a typical sample, the difference in cosine similarities is about this value. This score should not be interpreted as a percentage difference but as an absolute difference, which provides a measure of the model's ability to distinguish between the synthetic and real audio classes. When considered alongside the



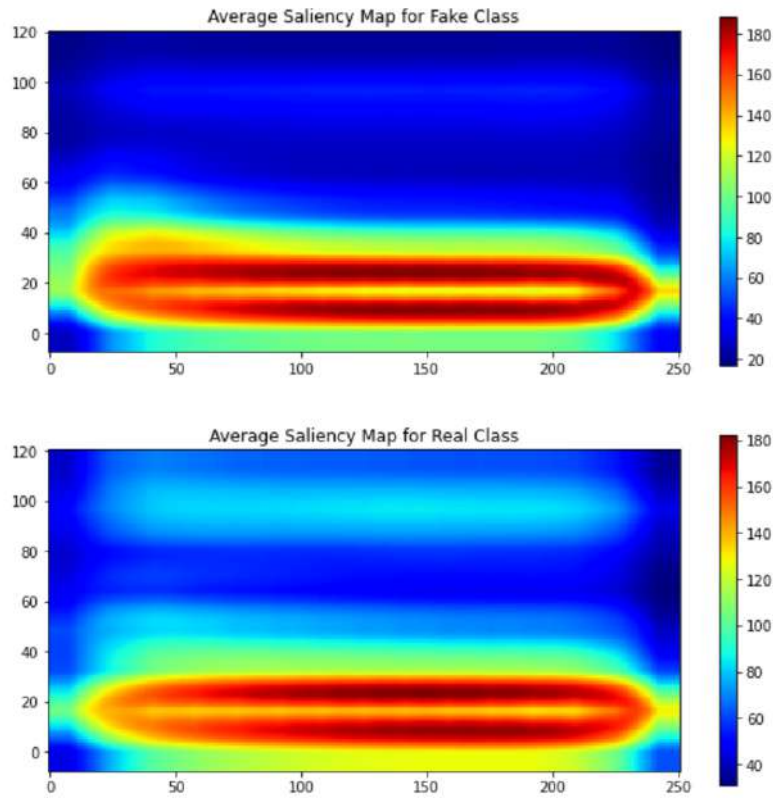


Figure 8. Average saliency maps for the synthetic (top) and real (bottom) audio classes. The x-axis represents the time domain from 0 to 251, and the y-axis shows the Mel frequency bins from 0 to 128. The color scheme, depicted by the legend, illustrates the activation strength from low (blue) to high (red) according to the Jet color map.

observation that most confidence scores lie within a similar range, these findings offer valuable insights into the model’s decision-making process.

However, it is essential to acknowledge the limitations of this approach. For certain instances, the confidence scores, derived from the absolute differences of cosine similarities between real and synthetic audio, might yield values that are quite similar. This scenario may hint at the method’s struggle to make a definitive distinction between classes, leading to potential errors or contradictive results compared to the VGG16 classification result.

Though this work employed the FoR dataset, the flexibility of the methodology allows for the use of other publicly available or custom datasets, paving the way for a broader evaluation of its performance across diverse data sources.

In summary, the methodology developed in this work presents an interpretable AI solution for synthetic audio detection, enabling a deep understanding of the complex decision-making process, accessible to both technical and non-technical audiences.

4.1.4 Relevance to AI4Media use cases and media industry applications

Our methodology is related and intended to be intergrated in UC1’s synthetic audio detection application. Designed for media professionals in fields like journalism, film, and gaming, our system aims to provide more than just a ‘real’ or ‘synthetic’ label for the audio samples tested. The goal is



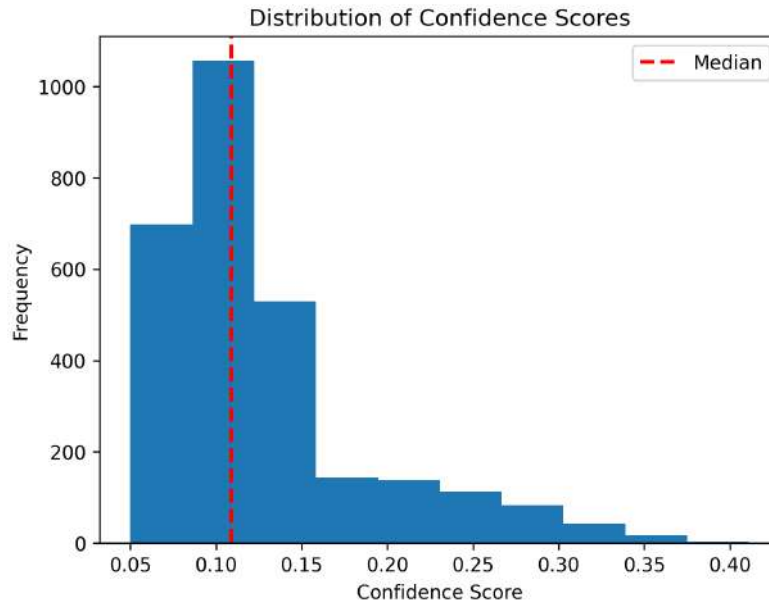


Figure 9. Distribution of confidence scores, calculated as the absolute difference between cosine similarities of new instances with the average saliency maps of synthetic and real audio classes. The median confidence score (depicted by the dashed red line) is approximately 0.1088, indicating that for a typical instance, the difference in cosine similarities is about this value.

to offer a transparent decision-making process through visual cues and confidence scores. A high confidence score will provide assurance in the system’s decision, while a lower score will act as a cue for deeper, human-in-the-loop investigation.

This dual-layered approach is designed with the expectation of significantly enhancing the reliability and credibility of automated synthetic audio detection. We aim to foster greater trust and drive wider adoption among media professionals. Moreover, by making the technology more accessible, we hope to enable a broader range of users in the media industry to effectively utilize these tools, thereby strengthening the industry’s defenses against misinformation.

4.2 Learning Visual Explanations for DCNN-Based Image Classifiers Using an Attention Mechanism

Contributing partners: CERTH

4.2.1 Overview

Gradient-based Class Activation Mapping (CAM) ([83], [77], [84], [85], [86]) and perturbation-based ([87], [88], [89], [90]) approaches have shown promising explanation performance. Given an input image and its inferred class label, these methods generate a CAM, which is re-scaled to the image size providing the so-called Saliency Map (SM); the SM indicates the image regions that the Deep Convolutional Neural Network (DCNN) has focused on in order to infer this class. However, these methods are either based on backpropagating gradients ([77], [84], [85]), producing suboptimal SMs due to the well-known gradient problems [91], or require many forward passes at the inference stage





([87], [88], [89], [90]), thus introducing significant computational overhead. Furthermore, the training dataset is not exploited in the exploration of the internal mechanisms concerning the decision process of the classifier. Driven by the observed limitations, in this section we present two new learning-based CAM methods, called L-CAM-Fm and L-CAM-Img, which utilize an appropriate loss function to train an attention mechanism [92] for generating visual explanations. Both methods can be used to generate explanations for arbitrary DCNN classifiers, are gradient-free and during inference require only one forward pass to derive a CAM and generate the respective SM of an input image.

4.2.2 Methodology

4.2.2.1 Problem formulation Let f be a DCNN model trained to categorize images to one of R different classes. Suppose an input image $\mathbf{X} \in \mathbb{R}^{W \times H \times C}$ that passes through f producing a model-truth label $y \in \{1, \dots, R\}$, i.e. the top-1 class label inferred by f , and K feature maps extracted from f 's last convolutional layer,

$$\mathbf{A} \in \mathbb{R}^{P \times Q \times K}, \quad (6)$$

where, W, H, C and P, Q, K , are the width, height and number of channels of \mathbf{X} and \mathbf{A} , respectively, and $\mathbf{A}_{::,k}$ is the k^{th} feature map. Given the above, the goal of CAM-based methods is to derive an activation map from the K feature maps, the so-called CAM, and based on it generate the respective SM, visualizing the salient image regions that explain f 's decision.

4.2.2.2 Training the attention mechanism Consider a training set of R classes (the same ones used to train f), where each image \mathbf{X} in the dataset is associated with a model-truth label y . This dataset is used to train an attention mechanism $g()$,

$$\mathbf{L}^{(y)} = g(y, \mathbf{A}), \quad (7)$$

where $\mathbf{L}^{(y)} \in \mathbb{R}^{P \times Q}$ is the CAM produced for a specified \mathbf{X} and y . Specifically, the attention mechanism is implemented as follows

$$g(y, \mathbf{A}) = \sum_{k=1}^K w_k^{(y)} \mathbf{A}_{::,k} + b^{(y)} \mathbf{J}, \quad (8)$$

where, the weight matrix $\mathbf{W} = [\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(R)}]^T \in \mathbb{R}^{R \times K}$ and bias vector $\mathbf{b} = [b^{(1)}, \dots, b^{(R)}]^T \in \mathbb{R}^R$ are the parameters of the attention mechanism, the transpose of vector $\mathbf{w}^{(r)} = [w_1^{(r)}, \dots, w_K^{(r)}]^T \in \mathbb{R}^K$ is the r^{th} row of \mathbf{W} , $w_k^{(r)} \in \mathbb{R}$ is the k^{th} element of $\mathbf{w}^{(r)}$, and $\mathbf{J} \in \mathbb{R}^{P \times Q}$ is an all-ones matrix. That is, the model-truth label y at the input of $g()$ is used to select the class-specific weight vector and bias term from the y^{th} row of \mathbf{W} and \mathbf{b} , respectively.

To learn the parameters of the attention mechanism we developed two different approaches, called L-CAM-Fm and L-CAM-Img, with their network architectures depicted in Figs. 10a and 10b, respectively. In both cases, the attention mechanism is placed at the output of the last convolutional layer of the DCNN and the elements of the derived CAM are normalized to $[0, 1]$ using the element-wise sigmoid function $\sigma()$. In L-CAM-Fm, the CAM produced by the attention mechanism is used as a self-attention mask to re-weight the elements of the feature maps. Contrarily, in L-CAM-Img the derived CAM is upsampled and applied to each channel of the input image.

The overall architecture is trained end-to-end using an iterative gradient descent algorithm, where the attention mechanism's weights are updated at every iteration, while the weights of f



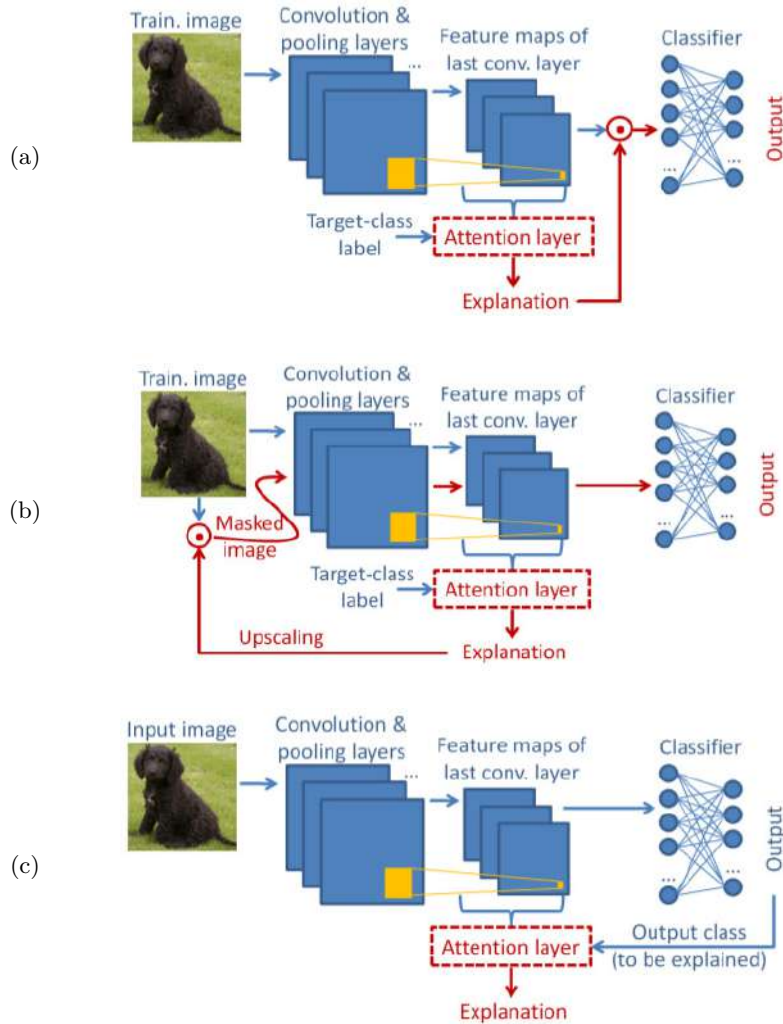


Figure 10. The network architectures of the developed approaches: (a) L-CAM-Fm training, (b) L-CAM-Img training, (c) L-CAM-Fm/-Img inference.

remain fixed to their original values. For both of the developed approaches, we use a loss function that is made of a Cross Entropy (CE) loss term, an average variation term and a total variation term. Intuitively, the latter two terms, in synergy with the CE loss, guide the DCNN to learn more informative, fine-grained SMs, i.e., SMs that contain only a few high-valued elements corresponding to the image regions contributing mostly to the classifier’s decision.

4.2.2.3 Inference of model decision’s explanation At inference stage, the procedure to derive the CAM of a test image is the same for both L-CAM-Fm and L-CAM-Img approaches (see Fig. 10c). That is, the test image is forward-passed through the DCNN to produce the corresponding feature maps and the model-truth label, which are then forwarded to the trained attention mechanism for computing the CAM (Eqs. (7), (8)).



4.2.3 Results

The developed L-CAM-Fm and L-CAM-Img are compared against several top-performing approaches from the literature with publicly-available implementations, namely, the Grad-CAM [77], Grad-CAM++ [84], Score-CAM [87], and RISE [89] approaches. Two sets of experiments are conducted with respect to the employed DCNN classifier, i.e., one using VGG-16 [93] and another using ResNet-50 [94]. In both cases, we use pretrained models from the PyTorch model zoo¹⁵. In terms of data, we utilize the ImageNet dataset [95], which is among the most popular datasets in the visual XAI domain. This dataset contains $R = 1000$ classes, 1.3 million images for training and 50K images for testing. Due to the prohibitively high computational cost of the considered perturbation-based approaches for our comparisons, we use only 2,000 randomly-selected testing images for evaluation, following an evaluation protocol similar to [87]. Finally, in terms of evaluation measures, we use the Average Drop (AD) and Increase in Confidence (IC) [84], calculated by retaining all or the most salient pixels of the SM (i.e., 100%, 50% or 15% of the SM).

The evaluation results are presented in the upper and lower half of Table 12 for VGG-16 and ResNet-50, respectively. As an ablation study, we also report results for the developed methods when trained using only the CE loss (denoted as L-CAM-Fm* and L-CAM-Img*). The number of forward passes, #FW, needed to compute the SM for an input image at the inference stage, is also shown at the last column of this table. The auxiliary masks used by RISE in the VGG-16 experiment are of size 7×7 [88] (which contrasts to the other approaches that use 14×14 feature maps for this experiment). For a fair comparison, we performed an additional experiment with the 7×7 feature maps after the last max pooling layer of VGG-16 using our L-CAM-Img, denoted as L-CAM-Img[†]. The results of this experiment are reported in the last row of the upper half of Table 12, under the L-CAM-Img's results (i.e. the ones obtained using the 14×14 feature maps). In addition, qualitative results are shown in Figure 11, while class-specific SM results for two images containing instances of two different classes are provided in Figure 12. From the obtained results we observe the following:

- i) L-CAM-Img generally outperforms the gradient-based approaches and is comparable in AD, IC scores to the perturbation-based approaches Score-CAM, RISE; though, contrarily to the latter requires only one FW instead of 512-8,000 at the inference stage.
- ii) L-CAM-Img[†] using 7×7 feature maps achieves the best performance in VGG-16; our approach is learning-based and, as the experiments showed, it is easier for it to learn the combination of the feature maps in the lower-dimensional space. This is consistent with the typical behavior of learning methods when working with high-dimensional data that may lay in a low-dimensional manifold (which is often the case with images), i.e. the curse of dimensionality.
- iii) L-CAM-Img outperforms L-CAM-Fm, but the latter still generally outperforms the gradient-based approaches.
- iv) Both of the developed approaches provide smooth SMs focusing on important regions of the image, as illustrated in Figure 11 and can produce class-specific explanations, as depicted in the examples of Fig. 12.
- v) From the ablation study of employing only the CE loss (L-CAM-Fm*, L-CAM-Img*), we see that incorporating the two additional terms in the loss function is very beneficial.

The findings discussed above, point out the advantages of the developed approaches compared to other existing state-of-art approaches from the literature. Additional qualitative analysis results and comparisons, as well as details about the implementation and training of the developed and compared approaches, can be found in the relevant publication.

¹⁵<https://pytorch.org/vision/stable/models.html>



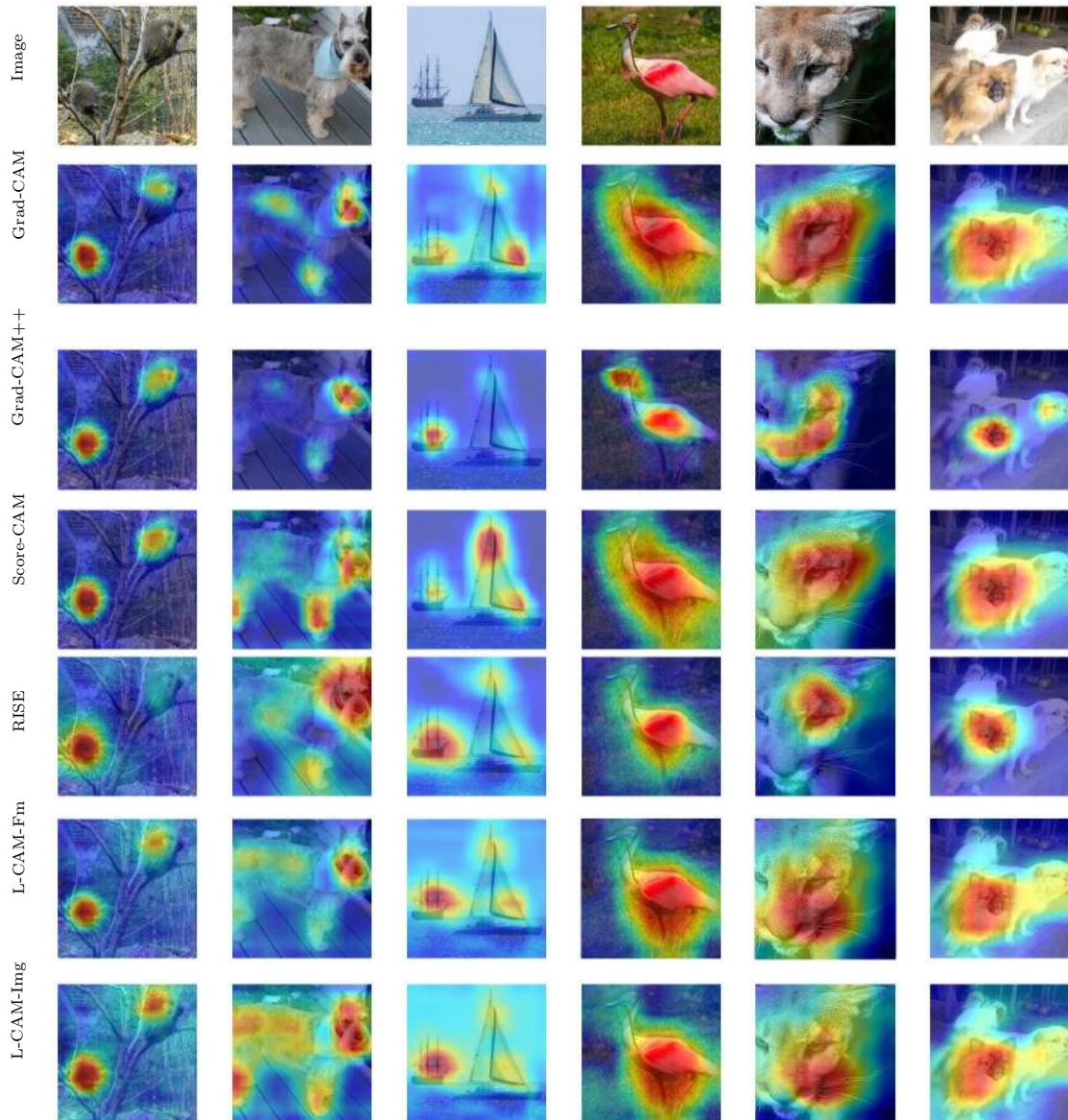


Figure 11. Visualization of SMs from various XAI methods superimposed on the original input image to produce class-specific visual explanations for the VGG-16 (columns 1 to 3) and ResNet-50 (columns 4 to 6) backbones.





	AD(100%)↓	IC(100%)↑	AD(50%)↓	IC(50%)↑	AD(15%)↓	IC(15%)↑	#FW↓
Grad-CAM [77]	32.12	22.1	58.65	9.5	84.15	2.2	1
Grad-CAM++ [84]	30.75	22.05	54.11	11.15	82.72	3.15	1
Score-CAM [87]	27.75	22.8	45.6	14.1	<u>75.7</u>	<u>4.3</u>	512
RISE [88]	8.74	51.3	<u>42.42</u>	<u>17.55</u>	78.7	4.45	4000
L-CAM-Fm*	20.63	31.05	51.34	13.45	82.4	3.05	1
L-CAM-Fm	16.47	35.4	47	14.45	79.39	3.65	1
L-CAM-Img*	18.01	37.2	50.88	12.05	82.1	3	1
L-CAM-Img	12.96	<u>41.25</u>	45.56	14.9	78.14	4.2	1
L-CAM-Img [†]	<u>12.15</u>	40.95	37.37	20.25	74.23	4.45	1
Grad-CAM [77]	13.61	38.1	29.28	23.05	78.61	3.4	1
Grad-CAM++ [84]	13.63	37.95	30.37	23.45	79.58	3.4	1
Score-CAM [87]	11.01	39.55	26.8	24.75	78.72	3.6	2048
RISE [88]	11.12	46.15	36.31	21.55	82.05	3.2	8000
L-CAM-Fm*	14.44	35.45	32.18	20.5	80.66	2.9	1
L-CAM-Fm	12.16	40.2	29.44	23.4	<u>78.64</u>	4.1	1
L-CAM-Img*	15.93	32.8	39.9	14.85	84.67	2.25	1
L-CAM-Img	<u>11.09</u>	<u>43.75</u>	<u>29.12</u>	<u>24.1</u>	79.41	<u>3.9</u>	1

Table 12. Evaluation results for a VGG-16 (upper half) and ResNet-50 (lower half) backbone classifier using 2,000 randomly-selected testing images of ImageNet. The best and 2nd-best performance for a given evaluation measure are shown in bold and underline, respectively.

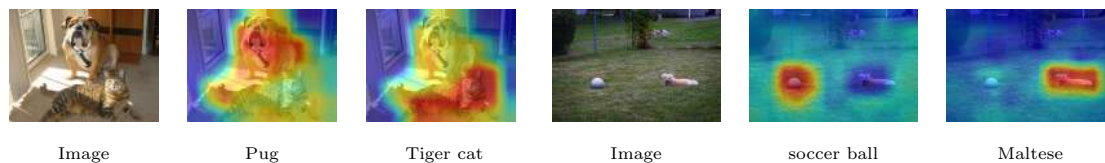


Figure 12. Two examples of using class-specific SMs (superimposed on the input image) produced by L-CAM-Img[†] on VGG-16 for classes “pug” and “tiger cat” (left) and classes “soccer ball” and “Maltese” (right).





4.2.4 Relevant Resources and Publications

Relevant publications:

- I. Gkartzonika, N. Gkalelis, V. Mezaris, “Learning Visual Explanations for DCNN-Based Image Classifiers Using an Attention Mechanism”, Proc. ECCV 2022 Workshop on Vision with Biased or Scarce Data (VBSD), Springer LNCS vol. 13808, pp. 396-411, Oct. 2022. DOI:10.1007/978-3-031-25085-9_23 [96]
Zenodo record: <https://zenodo.org/record/7572371>.

Relevant software and/or external resources:

- The implementation of the reported approaches can be found in <https://github.com/bmezaris/L-CAM>.

4.2.5 Relevance to AI4Media use cases and media industry applications

The developed approaches can facilitate the explanation of CNN image classifier decisions. Given the broad use of these classifiers in several use cases of AI4Media, their output will help to: (i) better assist story development by showing the parts of the image that affected the most the estimates of a CNN-based classifier concerning the image’s relevance with a news story (Use Case 2: AI for News - The Smart News Assistant), and (ii) support the creation of metadata for visual content by providing visual explanations (e.g., in the form of heat-maps) about the detected objects by CNN-based image classifiers (Use Case 3: AI in Vision - High Quality Video Production & Content Automation), and (iii) advance both the re-organization of media collections and the content moderation, by associating the CNN-based classification/categorization labels to images with human-interpretable visual explanations (Use Case 7: AI for (Re-)organisation and Content Moderation).

4.3 Wasserstein loss for Semantic Editing with GANs

Contributing partners: CEA

4.3.1 Overview

4.3.1.1 Context and Limits of the SotA Generative Adversarial Networks (GANs) are known to encode the semantics of the training data in their latent space [97]–[99]. Moving the latent codes in certain directions results in changing specific semantic attributes (*e.g.*, the smile on a face) in the generated images [97]. This ability makes GANs great tools to perform image editing, especially as it can be applied to real images through inversion methods [100]. The process directly links semantic attributes that are understandable by humans to the internal representation of the neural networks. It allows a human user to modify this representation, that is usually a vector with several hundred dimension, with a couple of simple clues she/he can grasp and observe the results. In that sense, it contributes to make the networks more explainable.

The challenge is to identify the manipulations in the latent space that have the desired effect on one attribute without affecting others. To obtain such *disentangled* manipulations, existing supervised methods leverage the semantic knowledge learned by pretrained attribute classifiers operating either in the image domain (*image* classifiers) or directly in the latent domain (*latent* classifiers). The key idea is that manipulated latent codes (or the images they produce) shift the predictions to match the desired outcome [101], [102]. However, classifiers can easily be fooled [103],



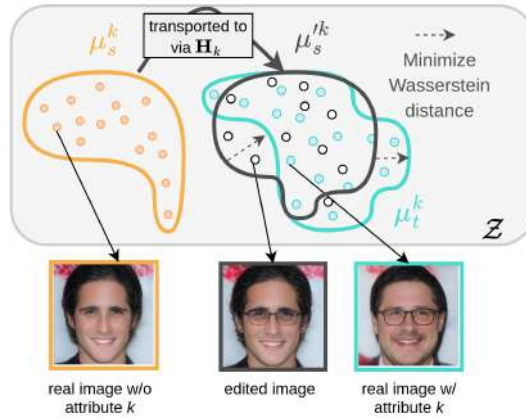


Figure 13. *CEA Method overview.* For each semantic attribute (e.g., “Glasses”) we learn a mapping \mathbf{H}_k that moves the distribution of latent codes lacking the attribute to the distribution of codes having that attribute. We enforce that each latent code is moved near a point that shares similar semantics, thus only changing that attribute. To preserve identity, the resulting distribution does not entirely match the target distribution.

e.g they can classify with high confidence out-of-distribution samples. As illustrated in Fig. 14 (left), the latent classifier of [102] steers latent codes outside the distribution resulting in edited images that are unrealistic. To address this issue we employ an *ad hoc* L_2 -regularization to minimize the norm of the latent editing. While this fixes out-of-distribution edits, Fig. 14 (right) shows that on MultiMNIST [104] this regularization produces adversarial samples [105] instead, i.e., the edited latent codes are correctly classified but the corresponding images remain unchanged. This is not surprising as changing the predicted class while minimizing the L_2 -norm of the edit precisely mimics the search for adversarial examples.

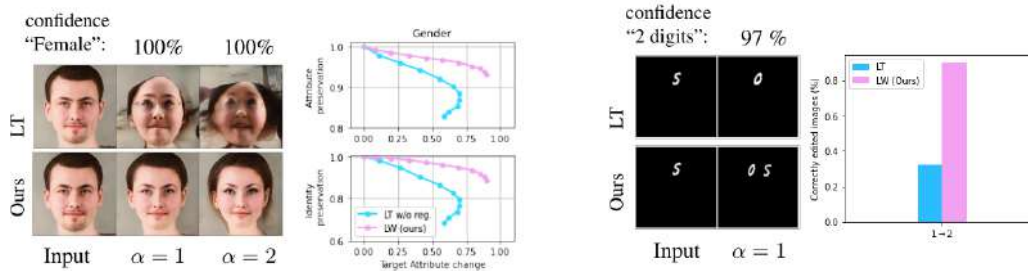


Figure 14. *Failure cases of a classifier-based method.* Latent transformer (LT) [102] learns edits in latent space under the guidance of a latent classifier. (left) On FFHQ [106] for “Male” → “Female”: without L_2 -regularization on the edited codes, the edited images are unrealistic (as shown in the qualitative result on the left) before reaching the desired editing. The classifier leads to out-of-distribution regions as it allocates high confidence to regions larger than that of the training samples [103]. The quantitative analysis on attribute and identity preservation shows highly degraded results. (right) On MultiMNIST [104] “1 digit” → “2 digits”: the edited images remain unchanged (no digit is being added) while the classifier indicates the opposite (predicts 2 digits with high confidence). The classifier leads to regions close to the decision boundaries where there are adversarial samples. The quantitative analysis shows that only 32% of images are correctly edited.



4.3.1.2 Contribution To prevent these issues, we introduce a new formulation for learning semantic edits in the latent space, leading to a core solution that *does not* rely on classifiers that is not subject to the brittle of classifiers. From a global perspective, latent editing can be viewed as an optimal transport problem [107]. Given a distribution of latent codes sharing some semantics, we propose to transport it onto the distribution of latent codes that share the same semantics except for the attribute to be edited. Since the resulting images should not exhibit any other changes than the desired one, the initial points should be transported “close” to points sharing their semantics; that is, the transport should be optimal w.r.t. a cost representing the perceptual similarity (13). To achieve this, we learn transformations in latent space using the guidance of the Wasserstein loss with an Euclidean cost, in latent space, which can be combined with a Wasserstein loss with a cost computed in the attribute space to enforce disentanglement.

We applied our method in the latent space of StyleGAN2 to modify the number of digits and edit facial attributes. We compared quantitatively and qualitatively to the method of Yao *et al.* (LT) [102] that relies exclusively on a latent classifier. Without additional regularization, our method leads to realistic edited images and achieves on-par disentanglement and better identity preservation than a classifier-based method.

4.3.2 Methodology

The direction used to edit an image may not be the same at each point of the full latent space. For this, the latent transformer [102] consists to edit according to $z_{edit} = z + \alpha \mathbf{H}(z)$, where \mathbf{H} is an affine transform in the latent space. Hence, the direction used to edit depends on the starting latent code z . In practice, \mathbf{H} is learned through the guidance of a multilabel classifier in the latent space. To enforce disentanglement, a weighted Mean Square Error (MSE) loss is added to minimize the change of the other attributes, that are not manipulated. They also add a $L2$ regularization on the difference between the original and edited latent code to preserve the global layout of the original image (actually the person’s identity since they work on faces). However, changing the label of a vector to classify while minimizing the $L2$ -norm with its edit is also the basic scheme to create adversarial samples [103], [105]. As a consequence, editing methods guided by classifier can either produce out-of-distribution images that are unrealistic or adversarial samples that have wrong attributes (Figure 14).

To avoid these problems, we introduce a new formulation for learning semantic edits in the latent space of GANs, which does not rely on classifiers and thus avoids the intrinsic shortcomings identified., and rather pose it as an optimal transport problem. Given two probability distributions μ, ν and the set Γ of their couplings (that all the joint probabilities γ such that μ and ν are the marginals of γ), the Wasserstein distance $W(\mu, \nu)$ is the infimum of the expected “cost” between μ and ν , this cost being *e.g.* $\|x - y\|_2$ for $(x, y) \sim \gamma \in \Gamma$. Intuitively, it represents “how much” μ needs to be transformed (transported) into ν . Let us consider an attribute of interest a_k we want to edit without changing the others and let note μ_s^k (resp. μ_t^k) be the distribution of latent codes z_k that are negative (resp. positive) with respect to this binary attribute. To increase the intensity of the attribute \mathbf{a}_k in the generated images, \mathbf{H}_k should transport the distribution of edited latent codes μ_s^k close to μ_t^k , while the information encoding other attributes or properties should remain unchanged (Fig. 13). Hence, we propose to minimize $W(\mu_s^k, \mu_t^k)$, with a cost based on a weighted Euclidean distance that reflects the possible biases in the collections of training latent codes. Moreover, to ensure that the edited latent codes share the same attributes as the initial ones, we propose to minimize a preservation loss $W(\mu_s^k, \mu_s^k)$. In that case, the cost is computed in the *attribute* space, thanks to latent classifiers trained to predict the attributes from the latent codes. It also takes into account the existing correlation between attribute, in order to avoid disentangling naturally correlated attributes (*e.g.* “Smile” and “High Cheekbones”), which would be pointless.





4.3.3 Results

We compared our approach to Latent Transformers on FFHQ and CelebAHQ for face attributes editing and MultiMNIST for a task consisting in editing the number of digits in images having between one and four of them. This last is quite original in the field since the attribute to manipulate is intrinsically discrete while usual evaluations on face consider continuous binary attributes that can vary continuously (young/old, male/female, smiling or not, wearing glasses or not...). We apply the editing in the latent space of StyleGAN2 [106] pretrained on FFHQ or MultiMNIST. The training data are the latent codes of real images previously projected in the $\mathcal{W}+$ latent space using the pSp encoder [100].

We employ respectively the 30k labeled 1024×1024 CelebAHQ images for face editing and 25k 128×128 MultiMNIST images. To learn a transformation, we use the implementation of the Wasserstein loss provided by the GeomLoss [108] library. We set the batch size as the minimum between the number of samples in the source and target distributions and drop the last batch if it causes a strong imbalance between both. We use Adam optimizer with a learning rate of 0.001. To avoid overfitting the target distribution, we perform early stopping on a hold-out validation set. As CelebAHQ contains various biases, we weight the samples and use the disentanglement loss. Optimal value for λ is 1 for all considered attributes except for “Glasses” ($\lambda = 15$). The cost is computed on all 40 attributes of CelebA [109]. Samples are weighted based on the most common attributes. We evaluate the methods with three metrics [102]:

- The *target attribute change* rate indicates the percentage of images for which the target attribute is indeed modified.
- The *attribute preservation* rate corresponds to the average number of attributes, apart from the target attribute, that are preserved.
- The *identity preservation* rate as the average of the cosine similarities between ArcFace [110] features of input and edited images.

All metrics are evaluated on 1,000 images from FFHQ. The attribute and identity preservation rates are reported against the target change for 10 values of $\alpha \in [1 \cdot d, 2 \cdot d]$ where d is chosen such that the target change for a given α is comparable between the different methods.

For facial attributes editing, we consider common attributes (“Glasses”, “Gender”, “Smile”, “Age”) and rarer ones chosen based on their representation and the performances of the image classifiers (“Pale Skin”, “Bangs”, “Blond Hair”, “Wavy Hair”). Qualitative results in Fig. 15 (left) exhibits some advantages of our method. Nose, lips and eyes shape are much better preserved for “Gender” and “Age”. LT also produces “cartoonish” edits for these attributes while ours remain naturalistic. LT ‘Gender’ editing is also heavily entangled with ‘Makeup’ while our approach adds nearly none. These differences have been quantitatively estimated in terms of attribute and identity preservation [111] but both methods are on-par. Our approach has occasionally slightly lower attribute preservation and usually higher identity preservation, that is surprising since we do not explicitly enforce this metric, contrary to LT. We also evaluate the ability of both methods to achieve disentangled and identity preserving editing without any explicit constraint, with an advantage for our approach. As shown in Fig. 15 (right), LT produces highly entangled edits (*e.g.* with the attribute “Smile”) and alters the identity. Without enforcing it explicitly, the Wasserstein-based approach already exhibits a good disentanglement ability and the identity is also well-preserved.

Regarding the task of digit number editing, we tested a change from $n = 1, 2, 3$ to $n + 1 = 2, 3, 4$ in real images from MultiMNIST. Given a change rate of 100% according to a latent classifier, the *actual* change rate measured by an image classifier is from 31% to 64% for LT while our approach reaches a rate of 90% to 99%.



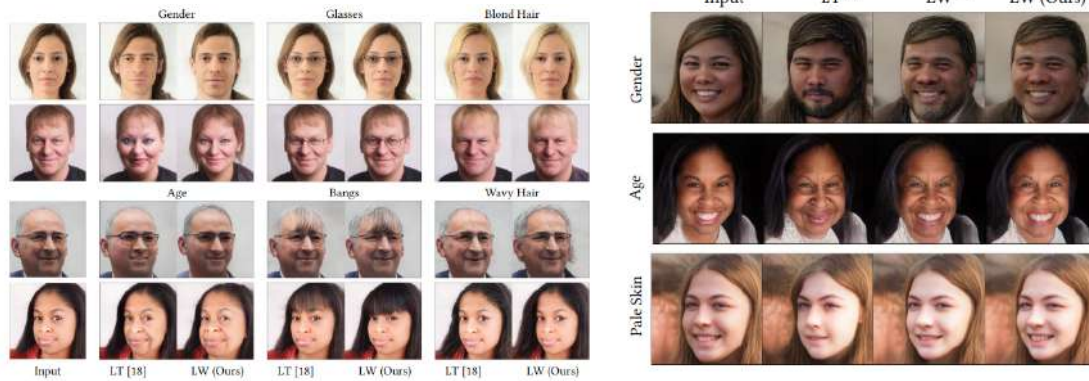


Figure 15. (left) Qualitative results for facial attribute editing. We report the editing results for $\alpha = \pm 2$. We observe that our approach better preserves identity and some facial attributes (e.g expression, absence of makeup) compared to LT. (right) Qualitative comparison between classifier-based edits (2nd col.) and our Wasserstein-based edits w/o any constraint (3rd col.) and w/ disentanglement constraint (4th col.).

4.3.4 Relevant Resources and Publications

Relevant publications:

- P. Doubinsky, N. Audebert, M. Crucianu, and H. Le Borgne, “Wasserstein loss for semantic editing in the latent space of GANs”, in International Conference on Content-Based Multimedia Indexing, Orléans, France, Sep. 2023 [111].
Zenodo record: <https://zenodo.org/record/8112753>.

4.3.5 Relevance to AI4Media use cases and media industry applications

This work deals with the creation of visual content, which aims at being realistic although synthetic. Specifically, we propose an alternative to the usual approaches used to control GANs that lead to adversarial images, thus our work is likely to help in detecting these last. By exploring the inner structure of the generative neural networks, and mapping these vectorial representation to concepts that are understandable by human users it makes these models more explainable. Thus, considering all these points, the work contributes to several use cases of AI4Media and the targeted industry applications:

- UC1.c - Recognising fake AI-generated images and UB3.b deep fake checking
- UC3.c - AI-based data obfuscation (e.g., GAN face anonymisation, logo removal)
- UC6 - AI for human co-creation

4.4 Disentangling Neuron Representations with Concept Vectors

Contributing partners: HES-SO



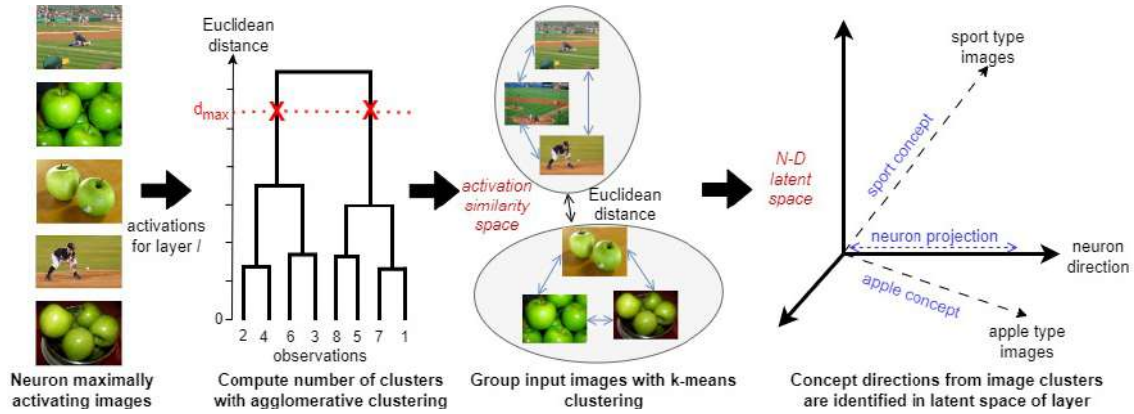


Figure 16. Step 1. A set of images that maximally activate a neuron in a model layer is taken. Step 2. The Euclidean distance between images in activation space is used as the similarity space on which the clustering is performed. This returns the appropriate number of clusters for a given distance threshold. Step 3. K-means clustering computes the cluster membership. Step 4. From the images in each cluster, a concept vector is calculated, which points toward the non-neuron aligned direction in activation space.

4.4.1 Approach

We propose a novel method to find and disentangle monosemantic directions starting from polysemantic neurons. Our method can search for concepts that are fine-grained according to the user’s desired level of concept separation and shows that polysemantic neurons can be disentangled into directions pointing to semantically unique concepts. Polysemanticity happens in neurons that respond to several unrelated features, or concepts [112]–[114], and it is a phenomenon that makes the interpretation of individual neurons challenging, since they cannot be mapped to semantically unique features.

We consider a CNN predicting a classification output (p -dimensional output vector) from an input image. We note that the method can be generalised to other models, but use a CNN for our analysis. We assume the model was already trained, and that we have access to the intermediate representations of an arbitrary layer inside the model. The first step of our method is to calculate the embeddings at a given intermediate layer for the entire dataset. We then apply global average pooling to aggregate the spatial information of the convolutional feature maps. For each neuron n , we then apply the following steps iteratively:

- Take the activations $\{\phi^l(x_i)\}_{i=1}^N$ where $\phi^l(x_i) \in \mathbb{R}^d$ and identify the top N activating images with the highest activation values.
- Measuring the similarity of the pooled maxed activations at the intermediate layer l . We use the Euclidean distance as a distance metric which has been shown by previous work to be highly predictive of perceptual similarity [115].
- Apply agglomerative clustering to identify the optimal number of clusters to be found in the pooled maxed activations.
- Perform k-means clustering on the same measurements of similarity and with the optimal number of clusters identified in the previous step as a hyperparameter.
- Exclude from the analysis non-significant clusters with less than 5 elements.



Figure 17. UMAP of the maximally activating images kept after k -means clusters and outlier removal in latent space: (left) separate clusters for polysemantic neuron 35 (right) a single cluster for monosemantic neuron 16.

4.4.2 Contribution

The contribution is a novel method that disentangles polysemanticity in Convolutional Neural Networks into distinct concept vectors that point at semantically unique features.

4.4.3 Experiments

Inception V3 (IV3) [116] is used in the experiments since it is a de-facto standard convolutional neural network. As this exploratory study only aims at a proof of concept, we focused on an undersampled version of ImageNet, retaining 130 random images for each class. This kept computation accessible to our infrastructure, feasible and light. Our results can easily be scaled to the entire dataset and larger dataset sizes. Where not stated, we consider the concatenation layer *Mixed 7b*, a convolutional layer with 2048 feature maps ($d = 2048$) near the end of the IV3 model. We pick this layer as we expect it to encode complex concepts [112], [117]. A similar analysis can be done on other layers and architectures. We took $N = 100$ top activating dataset examples and set the distance threshold parameter $d_{max} = 15$. Neuron 35 is here used as an example of a polysemantic neuron that activates highly for images of apples, sports, and also three images are dominated by a net-like pattern. Our method identifies 3 clusters, of which the cluster containing the net-like images is removed as it contains < 5 images. The embeddings of the remaining images plotted using UMAP [118] are shown in Figure 17 (on the left). Neuron 16 is here presented as a counter-example of a neuron that does not show polysemanticity at all. The neuron only activates for elliptical shapes as depicted in Figure 17 (on the right). As expected, the same procedure applied to this neuron yields only one cluster, yielding one monosemantic concept vector which has a much higher similarity than the neuron direction to the original images. We note that we found that the majority of the neurons we analysed in layer *Mixed 7b* were found to show some amount of polysemanticity. A possible explanation for this is that the number of features may be very high for the considered later layer in the model as it encodes complex concepts.



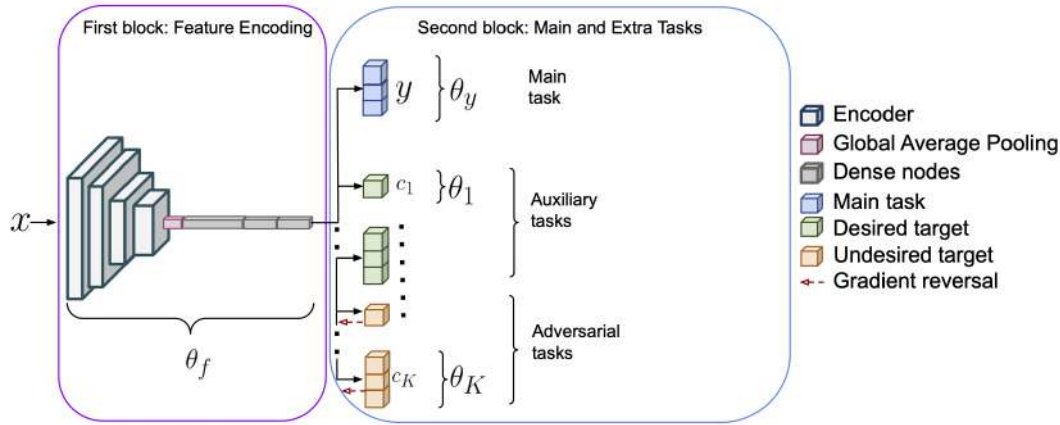


Figure 18. Multi-task adversarial architecture

4.4.4 Relevant Resources and Publications

Relevant publications:

- L. O'Mahony, V. Andrearczyk, H. Müller, and M. Graziani, "Disentangling Neuron Representations with Concept Vectors", In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition [119].
Zenodo record: <https://zenodo.org/record/8146780>.

Relevant software and/or external resources:

- The implementation of this work can be found in <https://github.com/lomahony/sw-interpretability>.

4.4.5 Relevance to AI4Media use cases and media industry applications

This work facilitates the explainability of image classifiers and will serve multiple crucial purposes in various AI4Media applications. Importantly, it enhances fake image recognition in UC1. This approach can discover and identify visual cues that the model has found useful to detect deep fakes, improving our understanding of the method and generating supporting evidence for automated deep fake recognition in the form of visual heatmaps and clustering.

Besides, it advances media collection reorganization and content moderation in UC7, since the images are linked to visual cues that correspond to high level concepts.

4.5 Multitask-Adversarial Learning Architecture

Contributing partners: HES-SO

4.5.1 Approach

We propose a novel convolutional architecture that increases the transparency and control of the learning process. The main technical innovation here is the combination of two successful techniques, namely multi-task learning [120] and adversarial training [121], with the purpose of guiding model training to focus on relevant features, called *desired targets*, and to discard *undesired targets* such



as confounding factors. The joint optimization of main, auxiliary and adversarial task losses is also a novel exploration that presents a main challenge when we combine losses that have different error metrics such as mean squared error and cross entropy. We investigate the impact of a dynamic task re-weighting technique based on the uncertainty estimation of each task during training [122], which is designed on purpose to facilitate the joint optimization of classification and regression objectives. From our analysis, it emerges that this uncertainty-based approach best handles the convergence and stability of the joint optimization. Our results also show a significant increase in the performance and generalization to unseen data.

The architecture is illustrated in Figure 18 and consists of two blocks. The first block is used to extract features from the input images. A state-of-the-art CNN of arbitrary choice without the decision layer is used as a feature encoder generating a set of feature maps. The feature maps are passed through a Global Average Pooling (GAP) operation that is performed to spatially aggregate the responses and connect them to a stack of dense layers. For this specific architecture, we use a stack of three dense layers of 1024, 512, and 256 nodes respectively. The second block comprises one branch per task, taking as input the output of the first block. The main task branch consists of the prediction of the labels \mathbf{y} and has as many dense nodes as there are of unique classes in \mathbf{y} . For binary classification tasks, e.g. discrimination of tumorous against non-tumorous inputs, the main task branch has a single node with a sigmoid activation function. K branches are added to model the extra targets. We refer to *extra* tasks for all the additional targets to the main task whether desired or undesired. *Auxiliary* tasks refer to the modeling of the desired targets, while *adversarial* tasks refer to that of undesired targets. The extra tasks are modeled by linear models as in [123]. For continuous-valued targets, the extra branch consists of a single node with a linear activation function. For categorical targets, the extra branch has multiple nodes followed by a softmax activation function. A gradient reversal operation [121] is performed on the branches of the undesired targets to discourage the learning of these features.

4.5.2 Contribution

Building on top of successfully existing techniques such as multi-task learning, domain adversarial training and concept-based interpretability, we address the challenge of introducing guidance in the training objectives of Convolutional Neural Networks.

4.5.3 Relevant Resources and Publications

Relevant publications:

- M. Graziani, S. Otalora, S. Marchand-Maillet, H. Muller, and V. Andrearczyk, “Learning Interpretable Microscopic Features of Tumor by Multi-task Adversarial CNNs to Improve Generalization”, arXiv preprint arXiv:2008.01478 [124].
Zenodo record: <https://zenodo.org/record/8147031>.

Relevant software and/or external resources:

- The implementation of this work can be found in https://github.com/maragraziani/multitask_adversarial.

4.5.4 Relevance to AI4Media use cases and media industry applications

Guiding a model to learn specific concepts through multi-task learning and adversarial training is a powerful strategy for bolstering the detection of deep fake images (UC1). In multi-task learning, the model simultaneously trains on various related tasks, such as recognizing facial expressions or





lip-syncing accuracy, which are crucial for authentic content. This approach equips the model with a deeper understanding of these specific concepts, making it more adept at discerning anomalies in deep fake creations. Additionally, adversarial training refines the model by pitting it against deep fake generators, forcing it to identify and adapt to evolving deception techniques. Together, these strategies enhance the model's ability to identify subtle artifacts and inconsistencies, contributing to more accurate and robust deep fake detection.

4.6 Explaining Autonomous Driving with Visual Attention and End-to-End Trainable Region Proposals

Contributing partners: UNIFI

4.6.1 Approach

Although autonomous driving vehicles are starting to become a reality, their diffusion worldwide is still slowed down by how such advancements are perceived by society. To ensure the pervasivity of automotive in everyday life, it is fundamental that algorithms and learning models guiding the decisions of autonomous vehicles are trustworthy, transparent and fully understandable. In other words, it is of paramount importance that the technologies that the end user will rely on must be explainable. Explainability in autonomous driving has been largely studied in recent years, especially regarding machine learning and computer vision algorithms that make autonomous navigation possible [125]–[127]. Explanations can be provided in different forms and styles, e.g., presenting factual, contrastive or counterfactual evidence to support cause effect relationships [128] or showing the sensitivity of the decision with reference to parts of the input [125].

In this work, we present a study on how different types of visual attention can be exploited to explain the decisions of a driving agent. We propose a conditional imitation learning approach capable of learning driving policies from RGB frames, trained with an attention block that weighs image regions based on their importance for the task. We design different region proposals, trained end-to-end along with the driving agent. Our full architecture is shown in Figure 19. A preliminary version of this work was described in [127], introducing the first visual attention based driving agent in the literature that learned to assign attention weights to a static grid of regions of interest in the input image. This work differs substantially from [127] in several aspects: (i) we overcome the limitation of having static proposals by developing different dynamic region proposal functions based on either Region Proposal Networks (RPNs) [129] or Spatial Transformer Networks (STNs) [130]; (ii) we provide a comparison with ex-post explainability methods, showing the importance of explicitly modeling visual attention to obtain meaningful interpretations; (iii) we show that the learned attention maps can be used to retrieve hard examples framing the problem as an anomaly detection task.

4.6.2 Results

In Table 13 we show the results of the models varying the number of boxes for STN and RPN. In both cases, when using approximately 100 proposals we obtain the best results.

We have trained two networks with the same architecture, the first fed with RGB frames and the second with attention maps produced by our model. We generate attention maps using the STN model, overlaying each generated box on a reference black image, weighing the RoI with the corresponding attention value.



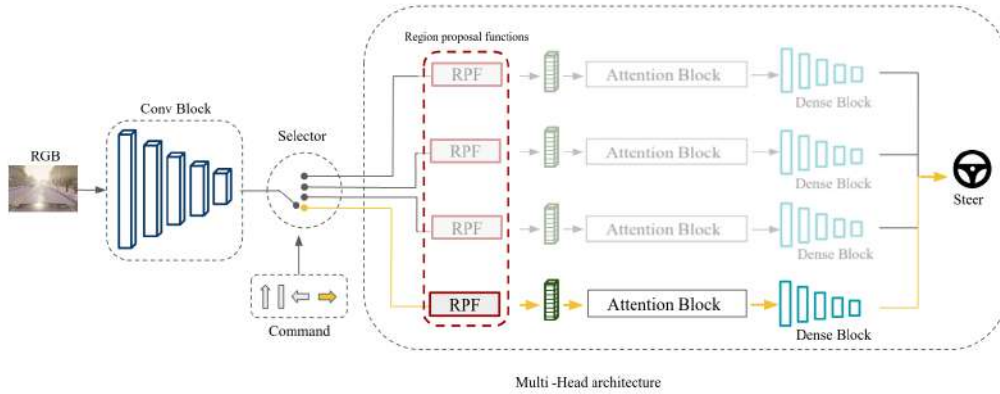


Figure 19. A convolutional backbone generates a feature map. Then, a region proposal function extracts RoIs that are pooled and weighed by an attention layer. Separate region proposal and attention modules are trained for each high level command in order to focus on different regions and output the appropriate steering angle.

	Training conditions						New weather					
	STN			RPN			STN			RPN		
Num boxes	50	100	300	72	108	432	50	100	300	72	108	432
Straight	100	100	100	100	100	100	100	100	100	100	100	100
One turn	100	100	98	80	93	93	94	100	96	84	84	84
Navigation	88	95	91	66	84	84	88	94	86	70	82	72
Navigation dynamic	87	90	90	64	82	84	84	94	86	68	80	76

Table 13. Ablation study. We vary the number of proposals produced by STN and RPN. Both STN and RPN perform better using a number of boxes around 100. In general, STN can obtain higher driving accuracy even with a low number of proposals, compared to RPN.

To test the models we used a test set consisting of 600 episodes extracted from the CARLA benchmark, 300 of which were successfully completed by the model. Failed episodes contain collisions with pedestrians, cars, other objects and/or unusual maneuvers. Our assumption is that failed episodes will contain out of the ordinary events, making the predicted attention anomalous. We thus leverage the reconstruction error of the autoencoders to detect such anomalies. We treat this task as a retrieval task, aiming at automatically identifying failed episodes. To evaluate the task, for each episode we take the maximum reconstruction error and use it to generate precision recall curves, as shown in Fig. 20. The model trained on attention maps reaches an AUC on the precision-recall curve of 71.53, while the model trained on RGB only 50.06. Similarly, computing Average Precision, we obtain 56.24 using attention maps and 37.97 with RGB frames. This experiment demonstrates that modeling attention is also effective in retrieving challenging episodes, which can be used to retrain the model and improve its performance.

4.6.3 Relevant Resources and Publications

Relevant publications:

- L. Cultrera, F. Becattini, L. Seidenari, P. Pala, A. Del Bimbo, “Explaining Autonomous

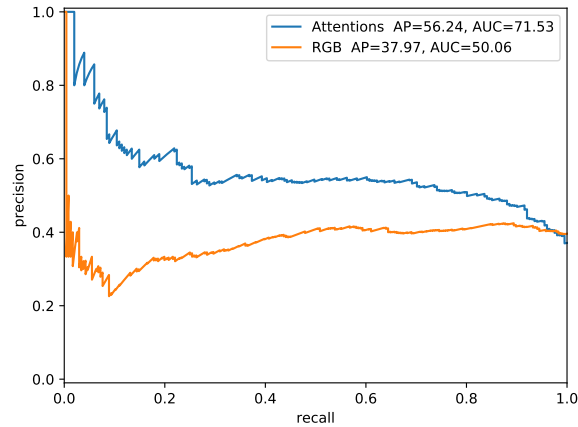


Figure 20. Precision-Recall Curves for detecting failed episodes.

Driving with Visual Attention and End-to-End Trainable Region Proposals” [131].
Zenodo record: <https://zenodo.org/record/8194594>.

4.6.4 Relevance to AI4Media use cases and media industry applications

This method is applied to a self-driving scenario, for which explainability is paramount. In applications like automated cinematography, we are interested in autonomous agents, which include but are not limited to UAVs, that can plan and execute maneuvers independently. The approach presented here can be employed to elucidate the rationale behind these decisions.

In general, while our proposed work primarily focuses on explaining the decisions made by autonomous driving agents, the method can also be adapted to clarify the outcomes of regression problems. In these problems, we assign a continuous score to indicate the likelihood of a particular event occurring, e.g., image tampering. This adaptation is pertinent to UC1, where we aim to explain the results of Deepfake detectors.

4.7 SMACE: Semi-Model-Agnostic Contextual Explainer

Contributing partners: 3IA-UCA

4.7.1 Overview

Interpretability is a pressing issue for decision systems. Many *post hoc* methods have been proposed to explain the predictions of a single machine learning model. However, business processes and decision systems are rarely centered around a unique model. These systems combine multiple models that produce key predictions, and then apply decision rules to generate the final decision. To explain such decisions, we propose the Semi-Model-Agnostic Contextual Explainer (SMACE), a new interpretability method that combines a geometric approach for decision rules with existing interpretability methods for machine learning models to generate an intuitive feature ranking tailored to the end user. We show that established model-agnostic approaches produce poor results on tabular data in this setting, in particular giving the same importance to several features, whereas SMACE can rank them in a meaningful way.

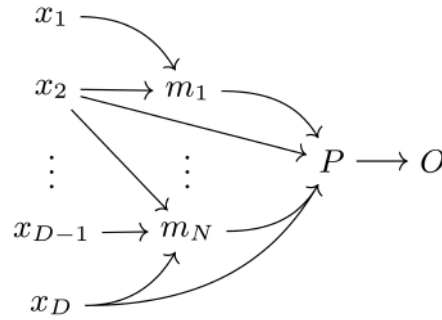


Figure 21. Structure of a composite decision system with D input features x_1, \dots, x_D , and N models m_1, \dots, m_N . A decision policy P (i.e., a set of decision rules) is finally applied to produce an outcome O . Note that in general both the models and the rules take a subset of input features as input, though not necessarily the same.

4.7.2 Approach

4.7.2.1 Setting Let $x \in \mathbb{R}^{Q \times D}$ be the input data, where each row $x^{(i)} = (x_1, \dots, x_D)^\top \in \mathbb{R}^D$ is an instance and D is the cardinality of the *input features set* F . Let the set $M = \{m_1, \dots, m_N\}$ be the set of models. We will refer to their outputs $m_1(x), \dots, m_N(x)$ as the *internal features*, whose values we also denote as $y^{(1)}, \dots, y^{(N)}$ when there is no ambiguity. The union of input and internal features is the set of $D + N$ features to which the decision policy can be applied.

We define $\tilde{x} := (x_1, \dots, x_D, m_1(x), \dots, m_N(x))^\top$ as the completion of x with the outputs of the N models. Likewise, we call $\xi = (\xi_1, \dots, \xi_D)^\top$ the example to be explained and $\tilde{\xi} = (\xi_1, \dots, \xi_D, m_1(\xi), \dots, m_N(\xi))^\top$ its completion. A decision rule R is formally defined by a set of conditions on the features in the form $\tilde{x}_j \geq \tau$, for some cutoff $\tau \in \mathbb{R}$. Figure 21 illustrates the structure of a generic composition of models and decision policies.

4.7.2.2 Assumptions The definition of SMACE is based on three assumptions required to frame the setting. Ideas for solving some of their limitations are discussed in the conclusion.

Assumption one: *Decision rules only refer to numerical values.*

This assumption allows us to take a simple geometric approach for the explainability of the decision tree. Note that this does not imply any restriction on the input of the machine learning models, that can still be categorical.

Assumption two: *Each decision rule is related to a single feature, without taking into account feature interactions.*

For instance, this assumption excludes conditions like **if** $\tilde{x}_1 \geq \tilde{x}_2$. Geometrically, this implies decision trees with splits parallel to the axes, such as CART [132], C4.5 [133], and ID3 [134].

Assumption three: *The machine learning models only use input features to make predictions.*

We disregard the cases in which a machine learning model takes as input the output of other machine learning models. We remark that this is a very reasonable assumption that covers most real-world applications. Note that assumptions *one* and *two* refer to the decision rules, while





Assumption *three* is the only referring to the machine learning models and does not concern their nature.

4.7.2.3 Methodology For each example ξ whose decision we want to explain, we first perform two parallel steps:

- **Explain the results of the models:** for each machine learning model m , we derive the (normalized) contribution $\hat{\phi}_j^{(m)}$ for each input feature j . By default, SMACE relies on KernelSHAP to allocate these importance values;
- **Explain the rule-based decision:** measure the contribution r_j of each feature (that is, each input feature and each internal feature directly involved in the decision policy), through Algorithm 2.

Then, to get the **overall explanations** (see Algorithm 1), we combine these partial explanations. The total contribution of the input feature $j \in F$ to the decision for a given instance is

$$e_j = r_j + \sum_{m \in M} r_m \hat{\phi}_j^{(m)}.$$

That is, we weight the contribution of input features to each model with the contribution of that model in the decision rule, and we add the direct contribution of feature j to the decision rule (if a feature is not directly involved in a decision rule, its contribution is zero).

Finally, once the partial explanations have been obtained, we agglomerate them via the equation above. We thus obtain a measure of the importance of features for a specific decision made by a system combining rules and machine learning models. Our measure of importance highlights the most critical features, those therefore most involved in the decision. In this way, a domain expert can analyse a decision by focusing on these features to make her or his own qualitative assessment.

Algorithm 1 Overview of `smace`.

```

function SMACE_EXPLAIN(rule  $R$  (set of conditions), list of models  $M$ , example to explain
 $\xi \in \mathbb{R}^D$ )
   $\tilde{\xi} \leftarrow \xi$ ,  $\phi \leftarrow \{0\}^N$ ,  $r \leftarrow \{0\}^{D+N}$ ,  $e \leftarrow \{0\}^D$ 
  for  $m \in M$  do
     $\hat{\phi}^{(m)} \leftarrow \text{EXPLAIN\_MODEL}(\xi, m)$  ▷ explain the result of model  $m$ 
     $\tilde{\xi} \leftarrow (\xi_1, \dots, \xi_D, \dots, m(\xi))$ 
  end for
  for  $j = 1, \dots, D + N$  do
     $r_j \leftarrow \text{RULE\_CONTRIBUTION}(R, j, \tilde{\xi})$  ▷ explain the rule-based decision
  end for
  for  $j = 1, \dots, D$  do
     $e_j \leftarrow r_j + \sum_{m \in M} r_m \hat{\phi}_j^{(m)}$  ▷ aggregate
  end for
  return  $e$ 
end function

```





Algorithm 2 Computing RULE_CONTRIBUTION.

```

function RULE_CONTRIBUTION(rule  $R$ , variable  $j$ , example to explain  $\tilde{\xi}$ )
   $S \leftarrow R$  ▷ projection to the decision surface  $S$  generated by  $R$ 
   $\pi_j^{(S)}(\tilde{\xi}) \leftarrow \arg \min_{z \in h_j} \|\tilde{\xi} - z\|_2$ 
  if  $\tilde{\xi}$  satisfies condition on  $j$  then
     $r_j \leftarrow 1 - \left| \tilde{\xi}_j - \pi_j^{(S)}(\tilde{\xi}) \right|$ 
  else
     $r_j \leftarrow \left| \tilde{\xi}_j - \pi_j^{(S)}(\tilde{\xi}) \right| - 1$ 
  end if
  return  $r_j$ 
end function

```

4.7.3 Results

What makes interpretability even more challenging is the lack of adequate metrics to appropriately assess the quality of explanations. In this section we compare the results obtained with SMACE and those obtained by applying the default implementations of SHAP¹⁶ and LIME¹⁷ on the whole decision system. We perform a sanity check on aggregate explanations on three different realistic use cases.

We demonstrate here that SMACE retains an ability to identify the set of features contributing negatively to a decision, regardless of individual attribution. If a feature contributes negatively, it means it must be moved to meet its condition. Correctly identifying negative features is a desirable property: to change the decision, each of them must be moved.

We consider 100 random instances which do not satisfy the rules (described in the supplementary), from three different datasets, and we apply SMACE, SHAP, and LIME. For each method, we extract the set of negative features. Note that to be sure that the rule will be satisfied, each negative feature should be shifted to a specific value: none of the three methods is giving this information. We then generate 1000 samples by shifting negative features with a local perturbation. The average decision made on these perturbed samples is an indicator of the quality of the explanations provided by each of the three methods.

Cancer treatment A machine learning model is trained to predict whether a breast cancer is benign or malignant from information about its size and structure. An automated decision system is then applied to decide on treatment: if the risk of the tumor being malignant is too high, it proceeds in full reliance on the model. If, on the other hand, the probability is low, but the size and composition of the tumor are suspicious, further investigation is carried out. The decision system consists of 30 continuous *input features* and 1 *internal feature* (coming from the model). We use the *Breast Cancer Wisconsin Data Set*.¹⁸

In this example, we want to explain *why* the treatment was not proposed, *i.e.*, which input features are negatively contributing to the decision. Given the large number of parameters to be analyzed, it is useful to order them by importance, in order to speed up the investigation by giving the right priorities. The graph at the top left of the Figure 22 shows the comparison. SMACE curve is always above the others: it is better at detecting negative features.

Fraud Detection A financial authority must track mobile money transactions, promptly

¹⁶<https://github.com/slundberg/shap>

¹⁷<https://github.com/marcotcr/lime>

¹⁸<https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data>

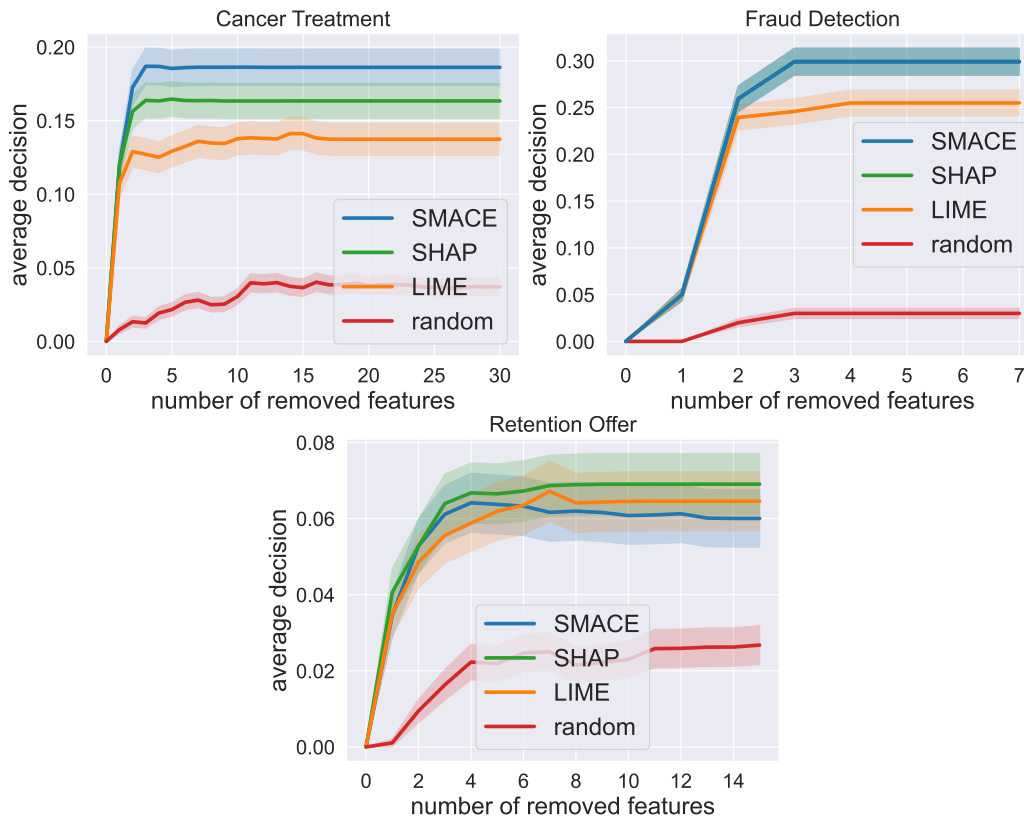


Figure 22. Comparison of SMACE, SHAP, and LIME on the ability to identify the set of features contributing negatively to a decision, regardless of individual attribution. Correctly identifying negative features is a desirable property: to change the decision, each of them must be moved. When the conditions are not met, the three methods are used to extract the negative features, and we generate perturbed samples around the original values. We then compare the average decision made on the samples.

halting anomalous transactions suspected of fraud. The authority uses a decision-making system to approve or block transactions, according to a *fraud score*, computed through a machine learning classifier, and the amount and balanced involved in the transaction. We use the *Synthetic Financial Datasets For Fraud Detection*¹⁹ As before, we extract and perturb the negative features set for each method.

The graph at the top right of Figure 22 shows that SMACE and SHAP are on par. In this decision system, the conditions based on the input features matter significantly more than the one on the model. SMACE and SHAP are able to extract the correct set of negative features. However, we remark that SHAP is likely to assign them the same (negative) contribution: SMACE carries more information.

Retention Offer Let us consider a mobile phone company which wants to predict if a customer is going to leave for a competitor, and to decide if a retention offer should be made, while not spending more on retention than the value of retaining the customer. The decision policy is based on information about the customer and their subscription (input features), and two models (producing internal features) predicting the *churn risk* (*i.e.*, the likelihood that the customer will cancel

¹⁹<https://www.kaggle.com/ealaxi/paysim1>





their subscription) and the *lifetime value* (*i.e.*, the expected revenue generated by the customer if retained). We use the IBM *Telco Churn* dataset.²⁰

In this example, we want to explain *why* a retention offer was not made, in terms of the original input features. In practice, the features that are contributing negatively should be moved to meet the conditions. Note that this use case is characterized by the presence of many categorical input features (see Assumption *one*): this is a stress test for SMACE. Figure 22 shows that SMACE is comparable with the state of the art in extracting the right set of negative features: error bars are overlapping. However, it is only a partial measure of quality, since the ranking of features is ignored.

We compared the ability of SMACE, SHAP, and LIME to extract features that are negatively contributing to a decision and should therefore be moved to change it. SMACE is best when applied to the standard context: one or more models and several continuous features (Cancer Treatment). SHAP tends to extract the same set of negative features as SMACE when the impact of models is absent or insignificant (Fraud Detection). SMACE loses performance when many categorical features are involved in the decision: however, the error bars of the three methods are overlapping (Retention Offer).

Up to the best of our knowledge, it is the first method specifically designed for a decision-making system composed of both machine learning models and decision rules. SMACE approaches the problem with a projection-based solution to explain the rule-based decision and by aggregating it with models explanations. We showed that model-agnostic approaches designed to explain machine learning models are not well-suited for this problem, due to the complications coming with the rules. In contrast, SMACE provides meaningful results by meeting our requirements, *i.e.*, adapting to the needs of the end user.

4.7.4 Relevant Resources and Publications

Relevant publications:

- Lopardo, G., Garreau, D., Precioso, F., and Ottosson, G. (2022, September). SMACE: A New Method for the Interpretability of Composite Decision Systems. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 325-339) [135].

Relevant software and/or external resources:

- The PyTorch implementation of our work “SMACE: Semi-Model-Agnostic Contextual Explainer” can be found in <https://github.com/gianluigilopardo/smace>.

4.7.5 Relevance to AI4Media use cases and media industry applications

Many of the AI services nowadays are based on complex compositions of several AI models: one extracting information from the images in the document, one from the text content, one from the structure of the document, etc. If I am a journalist using such AI service to classify content, retrieve information, “*I want to have human-understandable explanations for AI Services in my journalism tool that fit my context (general journalists or at least AI business experts)*”. This is what this work will provide you with.

²⁰<https://github.com/IBMDDataScience/DSX-DemoCenter/tree/master/DSX-Local-Telco-Churn-master>





4.8 A Sea of Words: An In-Depth Analysis of Anchors for Text Data

Contributing partners: 3IA-UCA

4.8.1 Overview

Anchors [136] is a post-hoc, rule-based interpretability method. For text data, it proposes to explain a decision by highlighting a small set of words (an anchor) such that the model to explain has similar outputs when they are present in a document. In this work, we present the first theoretical analysis of Anchors, considering that the search for the best anchor is exhaustive. After formalizing the algorithm for text classification, we present explicit results on different classes of models when the vectorization step is TF-IDF, and words are replaced by a fixed out-of-dictionary token when removed. Our inquiry covers models such as elementary if-then rules and linear classifiers. We then leverage this analysis to gain insights on the behavior of Anchors for any differentiable classifiers. For neural networks, we empirically show that the words corresponding to the highest partial derivatives of the model with respect to the input, reweighted by the inverse document frequencies, are selected by Anchors.

In this work, we present the first theoretical analysis of Anchors for text data, based on the default implementation available on Github²¹. The main restrictions of our analysis are the simplification of the combinatorial optimization procedure (therefore considering an *exhaustive* version of Anchors), the use of an out of dictionary token when removing words, and the assumption that a TF-IDF vectorization is used as a preprocessing step. Specifically,

- we dissect Anchors' algorithm for text classification, showing that the sampling procedure can be described simply as an i.i.d. Bernoulli's removal of words not belonging to the anchor (Proposition 1);
- we show that the exhaustive version is stable with respect to perturbation of the precision function, justifying our study of the exhaustive Anchors algorithm (Proposition 2 and Proposition 3);
- if the classifier ignores some words, they will not appear in the anchor selected by the exhaustive Anchors (Proposition 4);
- exhaustive Anchors for simple if-then rules provably outputs meaningful explanations, though words can be ignored from the explanation if their multiplicity is too high (Proposition 5);
- exhaustive Anchors picks the words associated to the most positive coefficients reweighted by the inverse document frequency for all linear classifiers (Proposition 6 and Proposition 7);
- we empirically show that exhaustive Anchors picks the words associated to the most positive partial derivatives scaled by the inverse document frequency for neural networks.

All our theoretical claims are supported by mathematical proofs in the full paper of this work [137], and numerical experiments whose code is available at https://github.com/gianluigilopardo/anchors_text_theory. Unless otherwise specified, experiments use the official implementation of Anchors with all default options.

4.8.2 Methodology

The operating procedure of Anchors for text data, as introduced by [136] is based on the key notions of *precision* and *coverage*.

²¹<https://github.com/marcotcr/anchor>





4.8.2.1 Precision and coverage The *precision* of an anchor $A \in \mathcal{A}$ is defined by [136] as the probability for a local perturbation of ξ to be classified as 1. Since we assume $f(\xi) = 1$, the precision can be written as

$$\text{Prec}(A) = \mathbb{E}_A [f(x) = 1] = \mathbb{P}_A (g(\varphi(x)) = 1) , \tag{9}$$

where the expectation is taken with respect to x , a random perturbation of ξ still containing all the words included in the anchor A . For the anchor containing all the words of ξ , the precision is exactly 1, while smaller anchors have, in general, smaller precision.

Of course, large anchors with size comparable to b are not very interesting from the point of view of interpretability (the text in Figure 23 would be completely highlighted). To quantify this idea, one can use the notion of *coverage*, defined in our case as the proportion of documents in the corpus (*i.e.*, the dataset of documents on which the vectorizer is fitted) that contain the anchor. For instance, the coverage of the anchor in Figure 23 is 0.12, meaning that 12% of the reviews contain it. The notions of precision and coverage are paramount to the Anchors algorithm: in a nutshell, **Anchors will look for an anchor of maximal coverage with prescribed precision.**

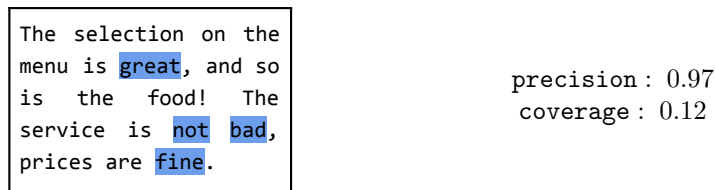


Figure 23. Anchors explaining the positive prediction of a black-box model f on an example ξ from the Restaurant review dataset. The anchor $A = \{\text{great, not, bad, fine}\}$ (in blue), having length $|A| = 4$ is selected. Intuitively, these four words together ensure a positive prediction by f with high probability (*precision* : 0.97), while being not too uncommon (*coverage* : 0.12).

4.8.2.2 The algorithm In practice, the coverage can be costly to compute, and in many cases a corpus is not available when the prediction is explained. Since anchors with smaller length tend to have larger coverage, a natural solution, used in the default implementation, is to minimize the length instead of maximizing the coverage, leading to:

$$\text{Minimize } |A| , \text{ such that } \text{Prec}(A) \geq 1 - \epsilon , \tag{10}$$

where $\epsilon > 0$ is a pre-determined tolerance threshold (set to 0.05 in practice). The lower ϵ is, the harder it is to find an anchor satisfying Eq. (10).

Of course, the exact precision of a specific anchor $A \in \mathcal{A}$ is unknown, since we cannot compute the expectation appearing in Eq. (9) in general. The strategy used by [136] is to approximate $\text{Prec}(A)$ by $\widehat{\text{Prec}}_n(A)$, an empirical approximation. Let us note that the optimization problem in Eq. (10) is generally intractable, whatever the selection function may be. The cardinality of \mathcal{A} is simply too large in all practical scenarios. As a consequence, the default implementation applies the KL-LUCB [138] algorithm to identify a subset of rules with high precision: at the next step, this subset is used as representative of all candidate anchors, finding an approximate solution to Eq. (10). In our work, we do not consider this optimization procedure and consider below an exhaustive version of Anchors.



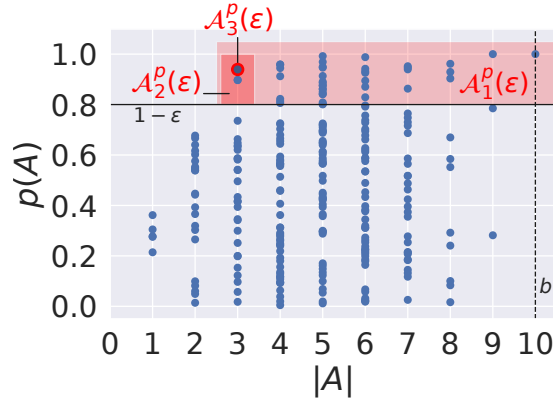


Figure 24. An illustration of Algorithm 3 with evaluation function $p = \text{Prec}$. Each blue dot is an anchor, with x coordinate its length and y coordinate its value for p . Here, $\epsilon = 0.2$ and the maximal length of an anchor is $b = 10$ (the length of ξ). In the end, the anchor A such that $|A| = 3$ and $p(A) = 0.9$ is selected (red circle).

4.8.2.3 Exhaustive p -Anchors In a nutshell, it is a formalized version of the original combinatorial optimization problem of Eq. (10) for any evaluation function $p : \mathcal{A} \rightarrow \mathbb{R}$.

The optimization problem of Eq. (10) can be decomposed in two steps: first, all anchors in \mathcal{A} such that $\text{Prec}(A) \geq 1 - \epsilon$ are selected. We call this first subset of anchors $\mathcal{A}_1(\epsilon)$. Note that $\mathcal{A}_1(\epsilon) \neq \emptyset$ since the full anchor $[b]$ has precision 1. Then, among these anchors, the ones with minimal length are kept, giving raise to $\mathcal{A}_2(\epsilon)$. At this point, it is not clear from Eq. (10) which anchors should be selected, and we settle for the ones with the highest precision. Equality cases can happen at this step (for instance, there can be several anchors with precision 1): we call $\mathcal{A}_3(\epsilon)$ the corresponding set of anchors. If $\mathcal{A}_3(\epsilon)$ is not reduced to a single element, we draw uniformly at random the selected anchor.

Algorithm 3 An overview of exhaustive p -Anchors.

input set of candidate anchors \mathcal{A} , selection function $p : \mathcal{A} \rightarrow \mathbb{R}$, tolerance threshold ϵ
select $\mathcal{A}_1^p(\epsilon) = \{A \in \mathcal{A} \text{ s.t. } p(A) \geq 1 - \epsilon\}$
select $\mathcal{A}_2^p(\epsilon) = \arg \min_{A' \in \mathcal{A}_1^p(\epsilon)} |A'|$
select $\mathcal{A}_3^p(\epsilon) = \arg \max_{A' \in \mathcal{A}_2^p(\epsilon)} p(A')$
select $A^p(\epsilon) \in \mathcal{A}_3^p(\epsilon)$ uniformly at random
return $A^p(\epsilon)$

Algorithm 3 formally describes this procedure for a generic evaluation function $p : \mathcal{A} \rightarrow \mathbb{R}$, which we illustrate in Figure 24. When using p , we write $\mathcal{A}_k^p(\epsilon)$ the sets constructed and $A^p(\epsilon)$ the selected anchor.

The goal here is to have a flexible framework: we can use Algorithm 3 with $p = \widehat{\text{Prec}}_n$ or $p = \text{Prec}$ as a selection function, or any other function which is a good approximation of Prec . When $p = \text{Prec}$, we call this version of the algorithm *exhaustive Anchors*, whereas when $p = \widehat{\text{Prec}}_n$ we call this version *empirical Anchors*. Empirical Anchors is very similar to Anchors; the main difference is that the former is looking at all possible anchors, while the latter uses an efficient approximate procedure, which we do not consider here. A second difference is that empirical Anchors selects anchors with maximal precision in the third step. This is not necessarily the case with the default



implementation, since an approximate procedure is used. We notice, nevertheless, that the chosen anchors tend to have high precision, and the demonstration that empirical Anchors and the default implementation give very similar output in practice can be found in our full article [137].

4.8.3 Results

This work presents the first theoretical analysis of Anchors. Specifically, we formalize the implementation for textual data, in particular giving insights on the sampling procedure. Our analysis shows that Anchors provides meaningful results when applied to these models, which is supported by experiments with the official implementation.

Finally, we exploit our theoretical claim about explainable classifiers to obtain empirical results for neural networks, yielding a surprising result that links the classifier gradient to the importance of words for a prediction. When having access to the model, this result can be used as a faster and more efficient method of obtaining explanations.

This work uncovered some surprising results that emphasize the importance of theoretical analysis in the development of explainability methods. We believe that the insights presented in this work may be valuable for researchers and practitioners in natural language processing who seek to correctly interpret Anchors' explanations. Furthermore, the analysis framework we developed can aid the explainability community in designing new methods based on sound theoretical foundations and in scrutinizing existing ones.

The detailed results of this work can be found in the full version of the article [137].

4.8.4 Relevant Resources and Publications

Relevant publications:

- Lopardo, G., Garreau, D., and Precioso, F. (2022). A Sea of Words: An In-Depth Analysis of Anchors for Text Data. AISTATS 2023, 26th International Conference on Artificial Intelligence and Statistics 2023 [137].

Relevant software and/or external resources:

- The PyTorch implementation of our work “A Sea of Words: An In-Depth Analysis of Anchors for Text Data” can be found in https://github.com/gianluigilopardo/anchors_text_theory.

4.8.5 Relevance to AI4Media use cases and media industry applications

As a journalist, I use a lot textual documents and when I apply an AI service on these documents to classify them, extract some specific information, or retrieve similar content, I would like to understand on the basis of which textual content the AI service has taken its decision. This is the role of this work, being able to highlight which part of the text and with which importance has led to the decision of classification, retrieval, extraction.

4.9 Interpretable Neural-Symbolic Concept Reasoning

Contributing partners: 3IA-UCA



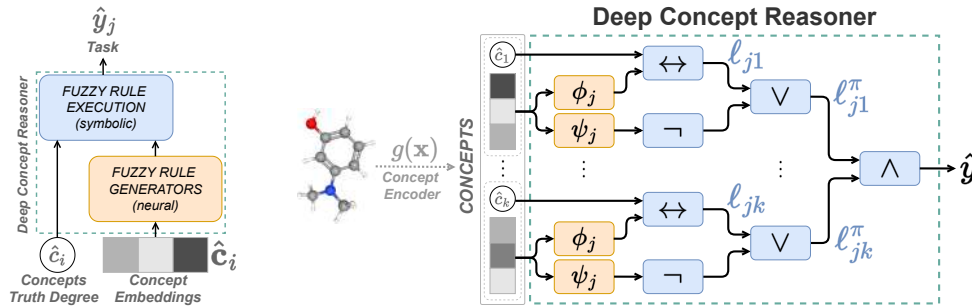


Figure 25. (left) Deep Concept Reasoner (DCR) generates fuzzy logic rules using neural models on concept embeddings. Then DCR executes the rule using the concept truth degrees to evaluate the rule symbolically. (right) Schema of DCR modules: first neural models ϕ and ψ generate the rule, and then the rule is executed symbolically.

4.9.1 Overview

Deep learning methods are highly accurate, yet their opaque decision process prevents them from earning full human trust. Concept-based models aim to address this issue by learning tasks based on a set of human-understandable concepts. However, state-of-the-art concept-based models rely on high-dimensional concept embedding representations which lack a clear semantic meaning, thus questioning the interpretability of their decision process.

To overcome this limitation, we propose the *Deep Concept Reasoner* (DCR), the first interpretable concept-based model that builds upon concept embeddings. In DCR, neural networks do not make task predictions directly, but they build syntactic rule structures using concept embeddings. DCR then executes these rules on meaningful concept truth degrees to provide a final interpretable and semantically-consistent prediction in a differentiable manner. Our experiments show that DCR: (i) improves up to +25% w.r.t. state-of-the-art interpretable concept-based models on challenging benchmarks (ii) discovers meaningful logic rules matching known ground truths even in the absence of concept supervision during training, and (iii), facilitates the generation of counterfactual examples providing the learnt rules as guidance.

4.9.2 Methodology

Let us describe the “Deep Concept Reasoner” (DCR, Figure 25), the first interpretable concept-based model based on concept embeddings.

Similarly to existing models based on concept embeddings, DCR exploits high-dimensional representations of the concepts. However, in DCR, such representations are only used to compute a logic rule. The final prediction is then obtained by evaluating such rules on the concepts’ truth values and not on their embeddings, thus maintaining clear semantics and providing a totally interpretable decision.

Being differentiable, DCR is trainable as an independent module on concept databases, but it can also be trained end-to-end with differentiable concept encoders.

In our work, we describe the different steps of applying our Deep Concept Reasoner: (1) the syntax of the rules we aim to learn, (2) how to (neurally) generate and execute learnt rules to predict task labels, (3) how DCR learns simple rules in specific t-norm semantics, and (4) how we can generate global and counterfactual explanations with DCR.

The main advantage of DCR w.r.t. existing interpretable and black-box methods arises when dealing with challenging tasks where both interpretability and accuracy should be maximized. For simpler tasks, existing interpretable methods, such logistic regression, could be enough. On the other



side, when interpretability is not a hard user requirement, then a simple black-box model would be easier to set up (e.g., it does not require concept labels or concept encoders). However, in all cases where interpretability plays a crucial role for the end user and existing interpretable models fail, then DCR could be preferable. Finally, compared to existing neural-symbolic approaches, DCR has an edge in all settings where the rules are unknown, while other methods (like DeepProbLog [139]) might be more stable when the full set of rules is known in advance. For other limitations/drawbacks, please see our reply to common questions.

One of the main limitations of DCR is that its global behavior may not be directly interpretable, which means that global rules may not perfectly align with the exact reasoning of the model. This could be an issue in cases where a user requires a precise understanding of the global model behavior. Also, the complexity of DCR rules may increase significantly when the difference between two tasks can only be determined by using a very high number of concepts. However, in most real-world cases, and in current benchmark datasets for concept-based models, this issue rarely arises. Finally, DCR requires concept embeddings as inputs, which assumes the existence of concept-based datasets or high-quality concept-discovery methods.

This work presents the *Deep Concept Reasoner* (DCR), the new state-of-the-art of interpretable concept-based models. To achieve this, DCR builds for each sample a weighted logic rule combining neural and symbolic algorithms on concept embeddings in a unified end-to-end differentiable system. In our experiments, we compare DCR with state-of-the-art interpretable concept-based models and black-box models using datasets spanning three of the most common data types used in deep learning: tabular, image, and graph data. Our experiments show that Deep Concept Reasoners: (i) attain better task accuracy w.r.t. state-of-the-art interpretable concept-based models, (ii) discover meaningful logic rules, and (iii) facilitate the generation of counterfactual examples.

While the global behaviour of the model is still not directly interpretable, our results show how aggregating Boolean DCR rules provides an approximation for the global behaviour of the model which matches known ground truth relationships. As a result, our experiments indicate that DCR represents a significant advance over the current state-of-the-art of interpretable concept-based models, and thus makes progress on a key research topic within the field of explainability.

4.9.3 Results

We have conducted an analysis for the following research questions:

- **Generalization** — How does DCR generalize on unseen samples compared to interpretable and neural-symbolic models? How does DCR generalize when concepts are unsupervised?
- **Interpretability** — Can DCR discover meaningful rules? Can DCR re-discover ground-truth rules? How stable are DCR rules under small perturbations of the input compared to interpretable models and local post-hoc explainers? How long does it take to extract a counterfactual explanation from DCR compared to a non-interpretable model?

The full results can be found in our full paper [140]. To sum up, we have shown that:

- DCR outperforms interpretable models (Figure 26)
- DCR matches the accuracy of neural-symbolic systems trained using human rules (Table 14)
- DCR discovers semantically meaningful logic rules (Table 15)
- DCR rules are stable under small perturbations (Figure 27)
- DCR explains prediction mistakes



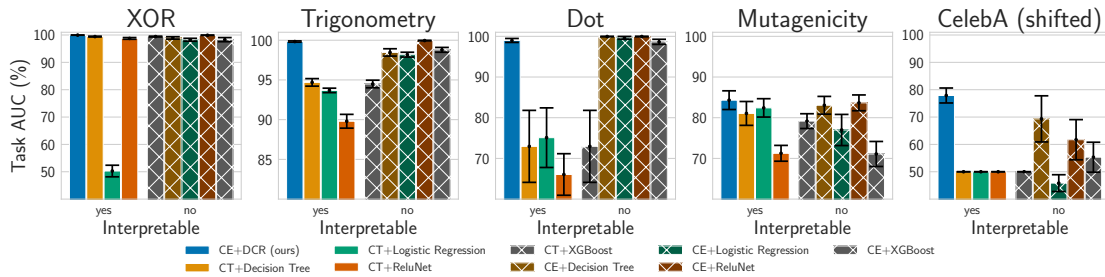


Figure 26. Mean ROC AUC for task predictions for all baselines across all tasks (the higher the better). DCR often outperforms interpretable concept-based models. CE stands for concept embeddings, while CT for concept truth degrees. Models trained on concept embeddings are not interpretable as concept embeddings lack a clear semantic for individual embedding dimensions.

- DCR enables discovering counterfactual examples (Figure 28)

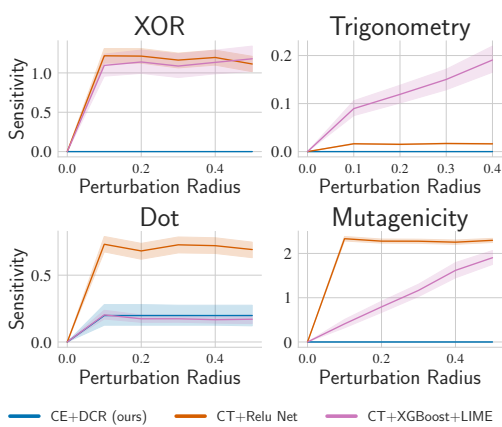


Figure 27. Sensitivity of model explanation when changing the radius of the input perturbation. The lower, the better. DCR explanations engender trust as they are stable under small perturbations of the input. The same does not hold generally for LIME explanations of XGBoost or Relu Net decision rules.

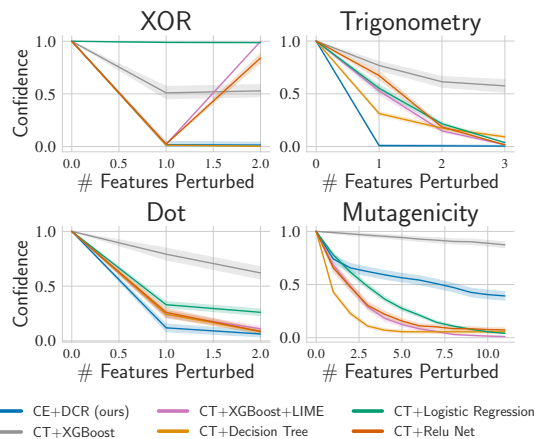


Figure 28. Model confidence as a function of the number of perturbed features on counterfactual examples. The lower, the better. Similarly to interpretable methods, DCR prediction confidence quickly drops after inverting the truth degree of a small set of relevant concepts, facilitating the discovery of counterfactual examples.

4.9.4 Relevant Resources and Publications

Relevant publications:

- Barbiero, P., Ciravegna, G., Giannini, F., Espinosa Zarlenga, M., Magister, L.C., Tonda, A., Lio, P., Precioso, F., Jamnik, M., and Marra, G.. (2023). Interpretable Neural-Symbolic Concept Reasoning. In Proceedings of the 40th International Conference on Machine Learning, 202:1801-1825 [140].

Relevant software and/or external resources:



Table 14. Task accuracy on the MNIST-addition dataset. The neural-symbolic baselines use the knowledge of the symbolic task to distantly supervise the image recognition task. DCR achieves similar performances even though it learns the rules from scratch.

Model	Accuracy (%)
With ground truth rules	
DeepProbLog	97.2 ± 0.5
DeepStochLog	97.9 ± 0.1
Embed2Sym	97.7 ± 0.1
LTN	98.0 ± 0.1
Without ground truth rules	
DCR(ours)	97.4 ± 0.2

Table 15. Error rate of Booleanised DCR rules w.r.t. ground truth rules. Error rate represents how often the label predicted by a Booleanised rule differs from the fuzzy rule generated by our model. The error rate is reported with the mean and standard error of the mean.

Ground-truth Rule	Predicted Rule	Error (%)
XOR		
$y_0 \leftarrow \neg c_0 \wedge \neg c_1$	$y_0 \leftarrow \neg c_0 \wedge \neg c_1$	0.00 ± 0.00
$y_0 \leftarrow c_0 \wedge c_1$	$y_0 \leftarrow c_0 \wedge c_1$	0.00 ± 0.00
$y_1 \leftarrow \neg c_0 \wedge c_1$	$y_1 \leftarrow \neg c_0 \wedge c_1$	0.02 ± 0.02
$y_1 \leftarrow c_0 \wedge \neg c_1$	$y_1 \leftarrow c_0 \wedge \neg c_1$	0.01 ± 0.01
Trigonometry		
$y_0 \leftarrow \neg c_0 \wedge \neg c_1 \wedge \neg c_2$	$y_0 \leftarrow \neg c_0 \wedge \neg c_1 \wedge \neg c_2$	0.00 ± 0.00
$y_1 \leftarrow c_0 \wedge c_1 \wedge c_2$	$y_1 \leftarrow c_0 \wedge c_1 \wedge c_2$	0.00 ± 0.00
MNIST-Addition		
$y_{18} \leftarrow c'_9 \wedge c''_9$	$y_{18} \leftarrow c'_9 \wedge c''_9$	0.00 ± 0.00
$y_{17} \leftarrow c'_9 \wedge c''_8$	$y_{17} \leftarrow c'_9 \wedge c''_8$	0.00 ± 0.00
$y_{17} \leftarrow c'_8 \wedge c''_9$	$y_{17} \leftarrow c'_8 \wedge c''_9$	0.00 ± 0.00

- The PyTorch implementation of our work “Interpretable Neural-Symbolic Concept Reasoning” can be found in https://github.com/pietrobarbiero/pytorch_explain.

4.9.5 Relevance to AI4Media use cases and media industry applications

In most of multimedia databases, the content is multimodal and comes with Metadata. This metadata can be exploited to build an extra knowledge on the content, for instance the location, the date, the author of a picture, sometimes even a description of the content of that picture. This is also true of course for audio and video files. As a journalist, if I want to classify my multimedia content database, or retrieve specific documents in this database, I would like to exploit this extra knowledge. This is exactly what this work is going to be able to provide: a hybrid system taking a decision based jointly on the raw multimedia content and on the associated knowledge.





4.10 Concept Embedding Models: Beyond the Accuracy-Explainability Trade-Off

Contributing partners: 3IA-UCA

4.10.1 Overview

Deploying AI-powered systems requires trustworthy models supporting effective human interactions, going beyond raw prediction accuracy. Concept bottleneck models promote trustworthiness by conditioning classification tasks on an intermediate level of human-like concepts. This enables human interventions which can correct mispredicted concepts to improve the model’s performance. However, existing concept bottleneck models are unable to find optimal compromises between high task accuracy, robust concept-based explanations, and effective interventions on concepts—particularly in real-world conditions where complete and accurate concept supervisions are scarce. To address this, we propose Concept Embedding Models, a novel family of concept bottleneck models which goes beyond the current accuracy-vs-interpretability trade-off by learning interpretable high-dimensional concept representations. Our experiments demonstrate that Concept Embedding Models (1) attain better or competitive task accuracy w.r.t. standard neural models without concepts, (2) provide concept representations capturing meaningful semantics including and beyond their ground truth labels, (3) support test-time concept interventions whose effect in test accuracy surpasses that in standard concept bottleneck models, and (4) scale to real-world conditions where complete concept supervisions are scarce.

4.10.2 Methodology

In real-world settings, where complete concept annotations are costly and rare, vanilla CBMs may need to compromise their task performance in order to preserve their interpretability [141]. While Hybrid CBMs are able to overcome this issue by adding extra capacity in their bottlenecks, this comes at the cost of their interpretability and their responsiveness to concept interventions, thus undermining user trust [142]. To overcome these pitfalls, we propose *Concept Embedding Models* (CEMs), a concept-based architecture which represents each concept as a supervised vector. Intuitively, using high-dimensional embeddings to represent each concept allows for extra *supervised* learning capacity, as opposed to Hybrid models where the information flowing through their *unsupervised* bottleneck activations is concept-agnostic. In the following section, we introduce our architecture and describe how it learns a mixture of two semantic embeddings for each concept (Figure 29). We then discuss how interventions are performed in CEMs and introduce *RandInt*, a train-time regularisation mechanism that incentivises our model to positively react to interventions at test-time.

4.10.2.1 Architecture For each concept, CEM learns a mixture of two embeddings with explicit semantics representing the concept’s activity. Such design allows our model to construct evidence both in favour of and against a concept being active, and supports simple concept interventions as one can switch between the two embedding states at intervention time.

We represent concept c_i with two embeddings $\hat{\mathbf{c}}_i^+, \hat{\mathbf{c}}_i^- \in \mathbb{R}^m$, each with a specific semantics: $\hat{\mathbf{c}}_i^+$ represents its active state (concept is **true**) while $\hat{\mathbf{c}}_i^-$ represents its inactive state (concept is **false**). To this aim, a DNN $\psi(\mathbf{x})$ learns a latent representation $\mathbf{h} \in \mathbb{R}^{n_{\text{hidden}}}$ which is the input to CEM’s embedding generators. CEM then feeds \mathbf{h} into two concept-specific fully connected layers, which learn two concept embeddings in \mathbb{R}^m , namely $\hat{\mathbf{c}}_i^+ = \phi_i^+(\mathbf{h}) = a(W_i^+ \mathbf{h} + \mathbf{b}_i^+)$ and



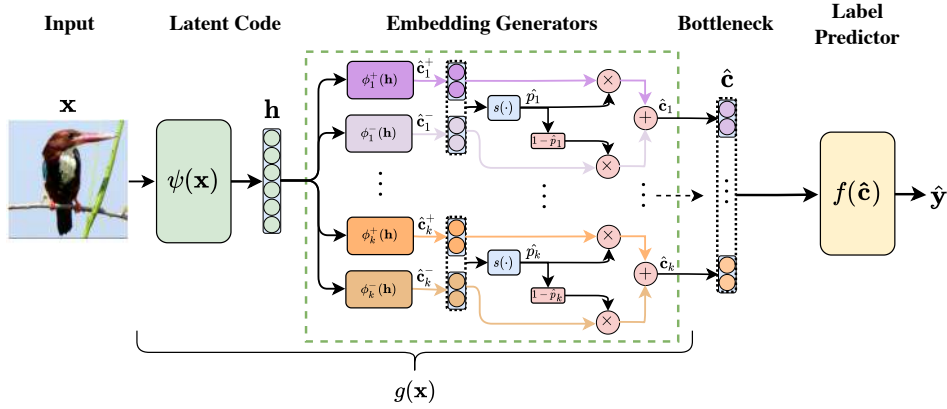


Figure 29. **Concept Embedding Model:** from an intermediate latent code \mathbf{h} , we learn two embeddings per concept, one for when it is active (i.e., $\hat{\mathbf{c}}_i^+$), and another when it is inactive (i.e., $\hat{\mathbf{c}}_i^-$). Each concept embedding (shown in this example as a vector with $m = 2$ activations) is then aligned to its corresponding ground truth concept through the scoring function $s(\cdot)$, which learns to assign activation probabilities \hat{p}_i for each concept. These probabilities are used to output an embedding for each concept via a weighted mixture of each concept’s positive and negative embedding.

$\hat{\mathbf{c}}_i^- = \phi_i^-(\mathbf{h}) = a(W_i^-\mathbf{h} + \mathbf{b}_i^-)$.²² Notice that while more complicated models can be used to parameterise our concept embedding generators $\phi_i^+(\mathbf{h})$ and $\phi_i^-(\mathbf{h})$, we opted for a simple one-layer neural network to constrain parameter growth in models with large bottlenecks. Our architecture encourages embeddings $\hat{\mathbf{c}}_i^+$ and $\hat{\mathbf{c}}_i^-$ to be aligned with ground-truth concept c_i via a learnable and differentiable scoring function $s : \mathbb{R}^{2m} \rightarrow [0, 1]$, trained to predict the probability $\hat{p}_i \triangleq s([\hat{\mathbf{c}}_i^+, \hat{\mathbf{c}}_i^-]^T) = \sigma(W_s[\hat{\mathbf{c}}_i^+, \hat{\mathbf{c}}_i^-]^T + \mathbf{b}_s)$ of concept c_i being active from the embeddings’ joint space. For the sake of parameter efficiency, parameters W_s and \mathbf{b}_s are shared across all concepts. Once both semantic embeddings are computed, we construct the final concept embedding $\hat{\mathbf{c}}_i$ for c_i as a weighted mixture of $\hat{\mathbf{c}}_i^+$ and $\hat{\mathbf{c}}_i^-$:

$$\hat{\mathbf{c}}_i \triangleq (\hat{p}_i \hat{\mathbf{c}}_i^+ + (1 - \hat{p}_i) \hat{\mathbf{c}}_i^-)$$

Intuitively, this serves a two-fold purpose: (i) it forces the model to depend only on $\hat{\mathbf{c}}_i^+$ when the i -th concept is active, that is, $c_i = 1$ (and only on $\hat{\mathbf{c}}_i^-$ when inactive), leading to two different semantically meaningful latent spaces, and (ii) it enables a clear intervention strategy where one switches the embedding states when correcting a mispredicted concept, as discussed below. Finally, all k mixed concept embeddings are concatenated, resulting in a bottleneck $g(\mathbf{x}) = \hat{\mathbf{c}}$ with $k \cdot m$ units (see end of Figure 29). This is passed to the label predictor f to obtain a downstream task label. In practice, following [141], we use an interpretable label predictor f parameterised by a simple linear layer, though more complex functions could be explored too. Notice that as in vanilla CBMs, CEM provides a concept-based explanation for the output of f through its concept probability vector $\hat{\mathbf{p}}(\mathbf{x}) \triangleq [\hat{p}_1, \dots, \hat{p}_k]$, indicating the predicted concept activity. This architecture can be trained in an end-to-end fashion by *jointly* minimising via stochastic gradient descent a weighted sum of the cross entropy loss on both task prediction and concept predictions:

$$\mathcal{L} \triangleq \mathbb{E}_{(\mathbf{x}, y, \mathbf{c})} \left[\mathcal{L}_{task}(y, f(g(\mathbf{x}))) + \alpha \mathcal{L}_{CrossEntr}(\mathbf{c}, \hat{\mathbf{p}}(\mathbf{x})) \right] \quad (11)$$

where hyperparameter $\alpha \in \mathbb{R}^+$ controls the relative importance of concept and task accuracy.

²²In practice, we use a leaky-ReLU for the activation $a(\cdot)$.



4.10.2.2 Intervening with Concept Embeddings As in vanilla CBMs, CEMs support test-time concept interventions. To intervene on concept c_i , one can update \hat{c}_i by swapping the output concept embedding for the one semantically aligned with the concept ground truth label. For instance, if for some sample \mathbf{x} and concept c_i a CEM predicted $\hat{p}_i = 0.1$ while a human expert knows that concept c_i is active ($c_i = 1$), they can perform the intervention $\hat{p}_i := 1$. This operation updates CEM’s bottleneck by setting \hat{c}_i to \hat{c}_i^+ rather than $(0.1\hat{c}_i^+ + 0.9\hat{c}_i^-)$. Such an update allows the downstream label predictor to act on information related to the corrected concept. In addition, we introduce *RandInt*, a regularisation strategy exposing CEMs to concept interventions during training to improve the effectiveness of such actions at test-time. RandInt randomly performs independent concept interventions during training with probability p_{int} (i.e., \hat{p}_i is set to $\hat{p}_i := c_i$ for concept c_i with probability p_{int}). In other words, for all concepts c_i , during training we compute embedding \hat{c}_i as:

$$\hat{c}_i = \begin{cases} (c_i\hat{c}_i^+ + (1 - c_i)\hat{c}_i^-) & \text{with probability } p_{int} \\ (\hat{p}_i\hat{c}_i^+ + (1 - \hat{p}_i)\hat{c}_i^-) & \text{with probability } (1 - p_{int}) \end{cases}$$

while at test-time we always use the predicted probabilities for performing the mixing. During backpropagation, this strategy forces feedback from the downstream task to update only the correct concept embedding (e.g., \hat{c}_i^+ if $c_i = 1$) while feedback from concept predictions updates both \hat{c}_i^+ and \hat{c}_i^- . Under this view, RandInt can be thought of as learning an average over an exponentially large family of CEM models (similarly to dropout [143]) where some of the concept representations are trained using only feedback from their concept label while others receive training feedback from both their concept and task labels.

4.10.3 Results

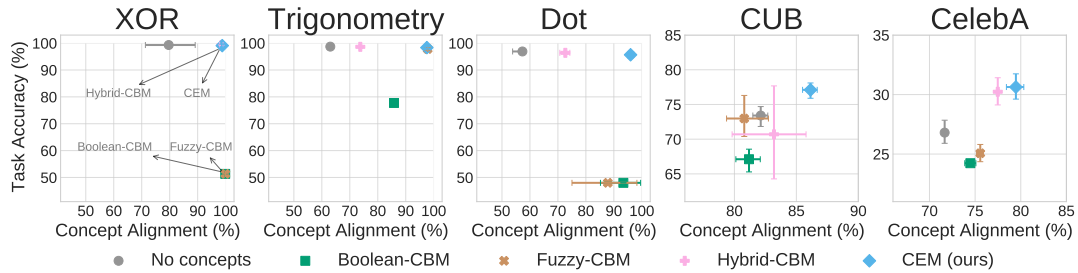


Figure 30. Accuracy-vs-interpretability trade-off in terms of task accuracy and concept alignment score for different concept bottleneck models. In CelebA, our most constrained task, we show the top-1 accuracy for consistency with other datasets.



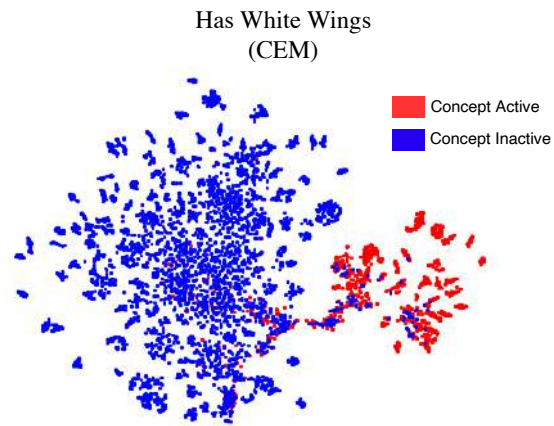


Figure 31. Qualitative results for our CEM with *t*-SNE visualisations of “has white wings” concept embedding learnt in CUB with sample points coloured red if the concept is active in that sample

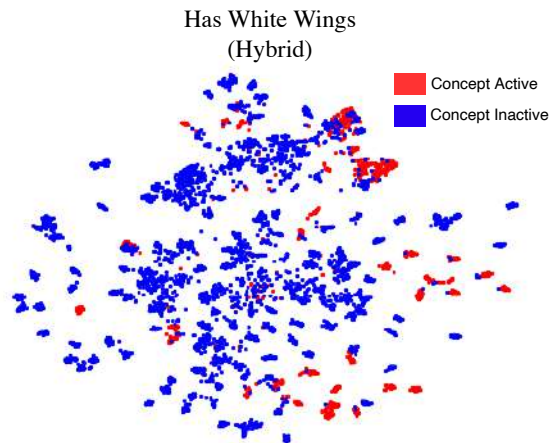


Figure 32. Qualitative results for hybrid with *t*-SNE visualisations of “has white wings” concept embedding learnt in CUB with sample points coloured red if the concept is active in that sample





Figure 33. top-5 test neighbours of CEM’s embedding for the concept “has white wings” across 5 random test samples.

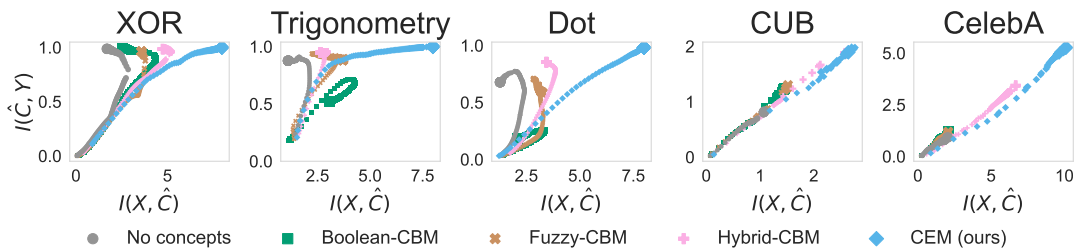


Figure 34. Mutual Information (MI) of concept representations (\hat{C}) w.r.t. input distribution (X) and ground truth labels (Y) during training. The size of the points is proportional to the training epoch.

- CEM improves generalisation accuracy (y-axis of Figure 30)
- CEM overcomes the information bottleneck (Figure 34)
- CEM learns more interpretable concept representations (x-axis of Figure 30)
- CEM captures meaningful concept semantics (Figure 31, Figure 32, and Figure 33)
- CEM supports effective concept interventions and is more robust to incorrect interventions (Figure 35)

Our experiments provide significant evidence in favour of CEM’s accuracy/interpretability and, consequently, in favour of its real-world deployment. In particular, CEMs offer: (i) state-of-the-art task accuracy, (ii) interpretable concept representations aligned with human ground truths, (iii) effective interventions on learnt concepts, and (iv) robustness to incorrect concept interventions. While in practice CBMs require carefully selected concept annotations during training, which can



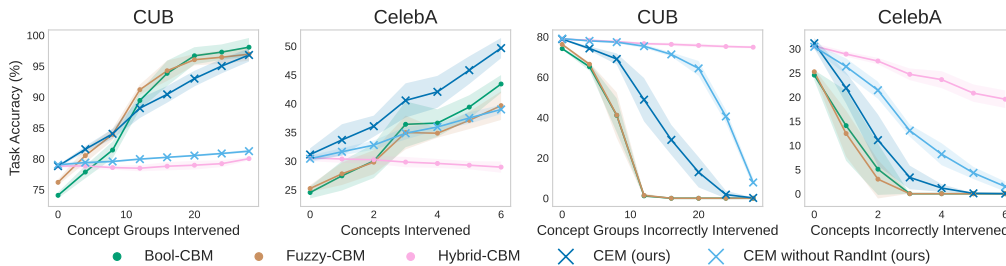


Figure 35. Effects of performing positive random concept interventions (left and center left) and incorrect random interventions (center right and right) for different models in CUB and CelebA. As in [141], when intervening in CUB we jointly set groups of mutually exclusive concepts.

be as expensive as task labels to obtain, our results suggest that CEM is more efficient in concept-incomplete settings, requiring less concept annotations and being more applicable to real-world tasks. While there is room for improvement in both concept alignment and task accuracy in challenging benchmarks such as CUB or CelebA, as well as in resource utilisation during inference/training, our results indicate that CEM advances the state-of-the-art for the accuracy-vs-interpretability trade-off, making progress on a crucial concern in explainable AI.

4.10.4 Relevant Resources and Publications

Relevant publications:

- Espinosa Zarlenga, M., Barbiero, P., Ciravegna, G., Marra, G., Giannini, F., Diligenti, M., Shams, Z., Precioso, F., Melacci, S., Weller, A., Lio, P., and Jamnik, M.. (2022). Concept Embedding Models: Beyond the Accuracy-Explainability Trade-Off. In Advances in Neural Information Processing Systems (NeurIPS), vol.35. [140].

Relevant software and/or external resources:

- The PyTorch implementation of our work “Concept Embedding Models: Beyond the Accuracy-Explainability Trade-Off” can be found in <https://github.com/mateoespinosa/cem>.

4.10.5 Relevance to AI4Media use cases and media industry applications

In most of multimedia databases, the content is multimodal (visual, text, audio, video). If this content is associated with a description, the concepts present in the multimedia content may be described. As a journalist, if I want to classify my multimedia content database, or retrieve specific documents in this database, I would like to condition the content classified or retrieved based on the concepts I expect to find in it or the concepts I do not want to find in it. This is exactly what this work is going to be able to provide: a hybrid system taking a decision based on the raw multimedia content but conditioning its importance by concepts that should be present or not in the expected results.

4.11 First Nice Workshop on Interpretability (NWI)

Contributing partners: 3IA-UCA





4.11.1 Summary

The 1st Nice Workshop on Interpretability (NWI)²³ took place on November 17-18, 2022. It was organized by Damien Garreau and Frédéric Precioso, with the help of their PhD student Gianluigi Lopardo, and brought together around 50 researchers from all across Europe. They discussed the many facets of machine learning models' interpretability. Lasting two days, the workshop was structured around 6 long talks and 11 short talks. Perhaps the main scientific takeaway is the diversity of approaches and the lack of consensus on what a good explanation should be, which led to stimulating discussion. This workshop was also a unique occasion for many researchers affiliated with AI4Media to meet: apart from Damien Garreau and Frédéric Precioso as organizers, Mara Graziani, Gianluigi Lopardo, Vasileios Mezaris, and Gabriele Ciravegna gave a talk. While the main event took place in the Université Côte d'Azur, some speakers and participants attended remotely to the afternoon sessions. NWI received the financial support of LJAD (the maths department of UCA) and AI4Media.

Photographs from the workshop are shown in Figure 36.

4.11.2 List of invited talks

Abstracts for all talks can be found in Appendix A.2.

- **Jenny Benois-Pineau (Université de Bordeaux):** *FEM and MLFEM post-hoc explainers for CNNs and their evaluation with reference-based and no-reference quality metrics*
- **Joao Marques-Silva (IRIT CNRS ANITI):** *Logic-Based Explainability in Machine Learning*
- **Vasileios Mezaris (ITI - CERTH):** *Explaining the decisions of image/video classifiers*
- **Martin Pawelczyk (University of Tübingen):** *On the Trade-Off between Actionable Explanations and the Right to be Forgotten*
- **Tristan Gomez (LS2N):** *Metrics for saliency maps faithfulness evaluation: an application to embryo stage identification*
- **Sebastian Bordt (University of Tübingen):** *From Shapley Values to Generalized Additive Models and back*
- **Hugo Sénétaire (DTU):** *Casting explainability as statistical inference*
- **Gianluigi Lopardo (3IA-UCA):** *A Sea of Words: An In-Depth Analysis of Anchors for Text Data*
- **Gabriele Ciravegna (3IA-UCA):** *Entropy-Based Logic Explanations of Neural Networks*
- **Jean-Michel Loubes (Université Toulouse Paul Sabatier):** *Explainability of a Model under stress*
- **Yann Chevaleyre (Paris Dauphine):** *Learning interpretable scoring rules*
- **Alexandre Benoit (Université Savoie Mont Blanc):** *Explainable AI for Earth Observation*

²³The website of the event can be found at <https://sites.google.com/view/nwi2022/home?authuser=0>





Figure 36. Pictures from NWI: (left) captivated audience; (middle:) discussions during the coffee break; (right:) The speakers and the organization team at the restaurant.

- **Salim Amoukou (Université Paris Saclay):** *Consistent Sufficient Explanations and Minimal Local Rules for explaining regression and classification models*
- **Giorgio Visani (University of Bologna):** *Inspecting Stability and Reliability of Explanations*
- **Hidde Fokkema (Korteweg-de Vries Institute):** *Attribution-based Explanations that Provide Recourse Cannot be Robust*
- **Mara Graziani (IBM Research):** *Reliable AI in healthcare: from model validation to hypothesis generation*
- **Pietro Barbiero (Cambridge University):** *Concept Embedding Models: Beyond the Accuracy-Explainability Trade-Off*





5 Privacy-Enhancing AI (Task 4.4)

Data is the new oil. Never before, so much personal data has been collected and evaluated. Never before, so many technologies have been available to analyze the data and combine this into new insights.

All these advances in Artificial Intelligence (AI) have the important downside that breaching individuals' privacy at scale is also as easy as never before. The European legislation reacted with the General Data Protection Regulation (GDPR) regulating what is allowed and what is not. However, this suggests a trade off between AI performance and privacy. But instead of drawing things black and white, making data privacy a natural enemy of progress, it is important to take a look at technologies that allow the processing of personal data without sacrificing sensitive information held by individuals and organizations. More often than not, cleverly anonymised data is enough.

Within this task (T4.4) we create tools that help protecting private data, while making data analysis required by the AI4Media use cases possible. Our contributions during this reporting period include work on (i) unlearning in the federated learning setting, a new field of work (Section 5.1), (ii) continuing work on `diffprivlib`, a general-purpose library for differential privacy computations in Python (Section 5.2), (iii) a utility-preserving de-identification approach for data publication using relation extraction filtering (Section 5.3), (iv) a tool for combining differential privacy, homomorphic encryption and multiparty computation for secure federated learning (Section 5.4), (v) a graph neural network with differentially private learning guarantees (Section 5.5), and (vi) the use of a reversible transformation to create adversarial examples for training (Section 5.6).

5.1 Federated Unlearning: How to Efficiently Erase a Client in FL?

Contributing partners: IBM

5.1.1 Overview

Recent privacy legislation [144] provide data owners the ability to revoke consent and the right to be forgotten. In the ML context, this requires that the data and any influence of the data on the ML model is removed. This process is also known as machine unlearning. This research focuses on machine unlearning in the context of Federated Learning (FL). We consider the case where a client wants to opt out of federation after the federated learning process, and as a result, wants to remove their contribution from the global model. Existing machine unlearning [145]–[148] approaches can not be directly applied in a federated learning setting due to the differences in the inherent characteristics of ML and FL. The most naive way of implementing federated unlearning is to retrain the model from scratch after removing from the corresponding client(s) the data sample(s) that are requested to be deleted. However, this approach is computationally expensive. Recently, some approximate unlearning approaches for the FL setting have been proposed aiming to speed up this process, but they are either not practical to be used in real-life [149] or require the server to store the history of the parameter updates [150]. Therefore, we introduce a new federated unlearning approach that relies on the client that wants to opt out of federation.

5.1.2 Methodology

After FL training is performed with N clients for the specified T rounds (Figure 37(a)), a client $i \in [N]$ requests to opt out of federation and wants to remove their contribution from the FL model.





We refer to this client as the *target client*. We propose to perform federated unlearning in two phases: (i) local unlearning (Figure 37(b)) and (ii) FL post-training (Figure 37(c)).

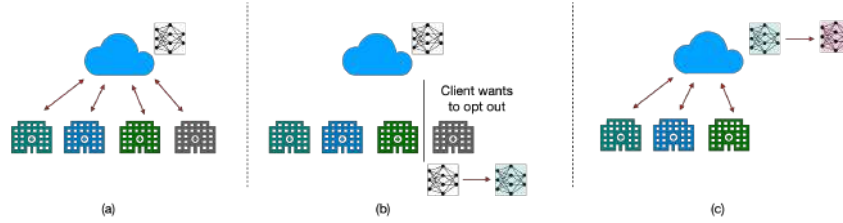


Figure 37. Phases of Federated Unlearning: (a) First, clients and the server participate in a federated learning process to train a global model. (b) One of the clients wants to opt out of federation and wants to unlearn their data. The target client i locally runs Projected Gradient Descent to obtain model \mathbf{w}_i^u . (c) The server and the remaining clients perform a few steps of federated learning with \mathbf{w}_i^u as the initial point to obtain the final ‘unlearned’ model.

Local Unlearning: To motivate our unlearning method, let us consider what happens during a federated training round. In each round, the goal of a client is to learn a local model that *minimizes* the (local) empirical risk, i.e., to solve the following optimization problem:

$$\text{(Train)} \quad \min_{\mathbf{w} \in \mathbb{R}^d} F_i(\mathbf{w}) := \frac{1}{n_i} \sum_{j \in \mathcal{D}_i} L(\mathbf{w}; (\mathbf{x}_j, y_j)), \quad (12)$$

where $L(\mathbf{w}; (\mathbf{x}_j, y_j))$ is the loss of the prediction on example (\mathbf{x}_j, y_j) made with model parameters \mathbf{w} . Each client locally makes several passes of (mini-batch stochastic) *gradient descent* to find a model that has *low* empirical loss.

We argue that a natural idea for unlearning is to *reverse* this learning process. That is, during unlearning, instead of learning model parameters that minimize the empirical loss, the client strives to learn the model parameters to *maximize* the loss. To find a model with *large* empirical loss, the client can simply make several local passes of (mini-batch stochastic) *gradient ascent*. However, simply maximizing the loss with gradient ascent can be problematic, since the loss function can be unbounded. For an unbounded loss, each gradient ascent step moves towards a model that increases the loss, and after several steps, it is likely to produce an arbitrary model similar to a random model.

To tackle this issue, we ensure that the unlearned model is *sufficiently close* to a *reference model* that has effectively learned the other clients’ data distributions. In particular, we propose to use the average of the other clients’ models as a reference model, i.e., $\mathbf{w}_{\text{ref}} = \frac{1}{N-1} \sum_{j \neq i} \mathbf{w}_j^{T-1}$. Note that the target client i can compute this reference model locally as $\mathbf{w}_{\text{ref}} = \frac{1}{N-1} (N \mathbf{w}^T - \mathbf{w}_i^{T-1})$, where \mathbf{w}^T is the global FL model after T rounds and \mathbf{w}_i^{T-1} is the i -th client’s local model update in round $T - 1$. The client i then optimizes over the model parameters that lie in the ℓ_2 -norm ball of radius δ around \mathbf{w}_{ref} . A natural choice for solving this optimization problem is to use *projected gradient descent*. More specifically, let us denote the ℓ_2 -norm ball of radius δ around \mathbf{w}_{ref} as $\Omega = \{\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v} - \mathbf{w}_{\text{ref}}\|_2 \leq \delta\}$. Let $\mathcal{P} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ denote the projection operator onto Ω . Then, for a given step-size η_u , client i uses PGD to iterate the update:

$$\mathbf{w} \leftarrow \mathcal{P}(\mathbf{w} + \eta_u \nabla F_i(\mathbf{w}; b)), \quad (13)$$

where $\nabla F_i(\mathbf{w}; b)$ is the gradient of F_i with respect to \mathbf{w} computed on a batch b . To avoid learning an arbitrary model, we perform early stopping if the ℓ_2 -distance of the target client \mathbf{w}_i^{T-1} to the unlearned model \mathbf{w}_i^u is smaller than a predetermined threshold τ .





FL post-training. To improve the performance of the locally unlearned model on the data of the retained clients, the server and the retained clients perform a few rounds of FL training starting with the unlearned model \mathbf{w}_i^u .

5.1.3 Results

We evaluate the performance of the proposed method on three datasets: MNIST [151], EMNIST (balanced version) [152], and CIFAR-10 [50]. An effective federated unlearning method must remove the contribution of the target client’s data, maintain good performance, and be more efficient than retraining from scratch. To reflect these properties in our evaluation, we use three performance measures: efficacy, fidelity, and efficiency (similar to Warnecke et al. [153]). We use the backdoor triggers [154] as an effective way to evaluate the performance of unlearning methods. We consider two cases: (i) $N = 5$ clients with the target client having 66% of their images backdoored, and (ii) $N = 10$ clients with the target client having 80% of their images backdoored. We compare our proposed unlearning method to retraining from scratch (referred as the baseline approach).

Figure 38 shows the accuracy on a hold-out test set of backdoored images (backdoor accuracy) of each model for each dataset. The high value of backdoor accuracy for the FedAvg model indicates that the FL model has learned the target client’s data consisting of backdoor triggers. We observe that the proposed PGD-based unlearning method substantially reduces the backdoor accuracy, and in fact, achieves similar backdoor accuracy to the baseline approach, demonstrating the high efficacy of our method.

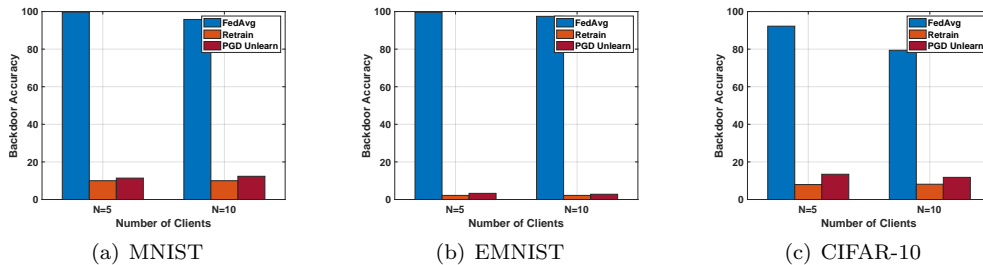


Figure 38. Backdoor accuracy (efficacy) of the fully retrained and the PGD-based unlearned model in each dataset, and their comparison with the FedAvg model before unlearning.

In Figure 39, we show the accuracy on a hold-out test set that consists of clean images (clean accuracy) of the unlearned models obtained by our method and retraining. We observe that our PGD-based unlearning method achieves similar clean accuracy to retraining.

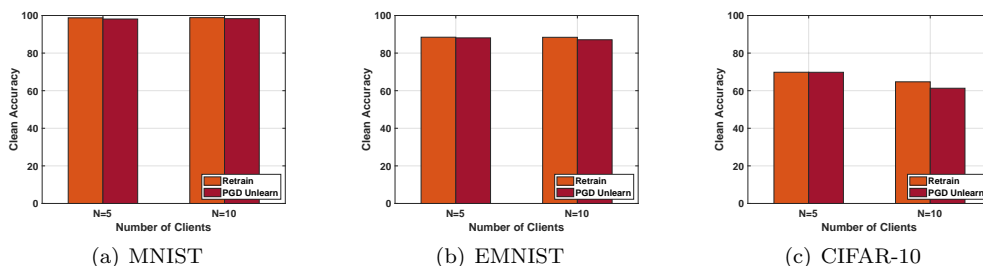


Figure 39. Clean accuracy (fidelity) of the fully retrained and the PGD-based unlearned model in each dataset.



Figure 40 shows the communication cost for various clean accuracy (fidelity) values for $N = 5$ clients. We observe that the proposed unlearning method is more efficient in terms of the communication cost on the retained clients than the baseline of retraining while achieving comparable fidelity and efficacy. We believe that lowering the communication burden on retained clients is appealing in practice since these clients are not incentivized to help the target client in unlearning.

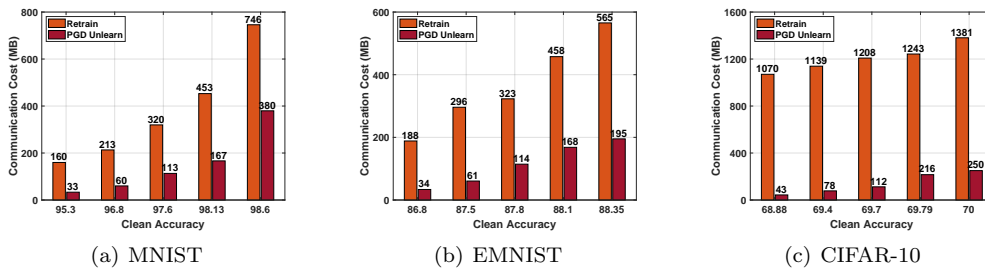


Figure 40. Communication costs (efficiency) of the proposed unlearning method and the baseline approach with respect to the clean accuracy (fidelity) in each dataset for $N = 5$.

5.1.4 Relevant Resources and Publications

Relevant publications:

- A. Halimi, S. Kadhe, A. Rawat, and N. Baracaldo. “Federated Unlearning: How to Efficiently Erase a Client in FL?” International Workshop on Updatable Machine Learning in conjunction with ICML (UpML), 2022 [155].
Zenodo record: <https://zenodo.org/record/8154387>.

Relevant resources:

- The implementation of this work can be found in <https://github.com/IBM/federated-unlearning>.

5.1.5 Relevance to AI4Media use cases and media industry applications

Our approach, with its focus on being compliant with GDPR, is relevant to various media industry use cases. Unlearning empowers users by providing them the ability to request the removal of their data. In content moderation, unlearning data from a federated learning model can be used to improve the moderation models when irrelevant, harmful, or toxic information is found. Unlearning plays an essential role for journalists and researchers by ensuring that the media content is of high quality. Media companies can also use unlearning to reduce bias in recommendation models by removing the impact of biased data points from the model.

5.2 Diffprivlib – A General-Purpose Differential Privacy Library

Contributing partners: IBM

5.2.1 Overview

Owing to its robust mathematical guarantees, generalised applicability and rich body of literature, Differential Privacy (DP) has emerged as the defacto standard in data privacy since its inception in



2006. Diffprivlib [156] was created as a central repository of differential privacy mechanisms, readily available to apply and combine in various application use cases, in conjunction with state of the art standard machine learning practices and tools. The aim is to enable users to (i) experiment with differential privacy, (ii) explore the impact such techniques can have on machine learning accuracy and (iii) build commercial grade applications with differential privacy mechanisms integrated from their inception onwards.

Since D4.1, a number of important additions have been made to diffprivlib to enhance its functionality, safeguard its privacy guarantees and maintain compatibility with dependencies. Highlights of updates include the following:

1. *Seeding*: The ability to seed the Random Number Generator (RNG) for DP noise generation.
2. *Random Forest*: The addition of a differentially-private random forest classifier (including one significant refactoring of the code to enhance performance).
3. *Secure Sampling*: The implementation of secure noise sampling to enhance privacy in floating-point calculations.

5.2.2 Methodology

We outline the methodologies associated with the main additions listed in the previous section.

5.2.2.1 Seeding The ability to seed the random number generator used for generating the noise required to implement DP is important for the reproducibility of results, bug-fixing and testing, and was an important addition to Diffprivlib between D4.1 and the present deliverable [157]. This was achieved by following the pre-existing standards from Numpy and Scikit-Learn, both already heavily integrated with Diffprivlib.

Users can now pass a `random_state` parameter directly to each Diffprivlib function, which can be (i) `None`, (ii) an integer, or (iii) an existing `RandomState` instance (from the Numpy package). This random state is then passed along and used for all sources of randomness with the Diffprivlib function, and any internal subcalls. This allows for scientists to reproduce published results (whenever the seed is provided, as is best practice), and also eases the burden of bug fixing and testing, where the randomness of DP can cause problems.

5.2.2.2 Random Forest The addition of a `RandomForestClassifier` algorithm to Diffprivlib has added another important feather to its cap. The implemented random forest follows state-of-the-art literature in the area, which cleverly apportiones the privacy budget ϵ across each tree in the forest [158].

In summary, the DP version of `RandomForestClassifier` produces a forest of completely random trees without first looking at the data. By “training” the trees by only referencing the metadata (*i.e.*, the range/domain of the data) no privacy budget is expended. Then, the entire dataset is partitioned across each of the trees, and DP counts are taken at each leaf node. Because of the partitioning of the dataset and the nature of binary trees (where each datapoint contributes to precisely one leaf node), the sensitivity of the algorithm can be tightly controlled, thereby adding minimal noise.

5.2.2.3 Secure Sampling One final important addition to Diffprivlib in the reporting period was the addition of secure sampling of noise. It is well known that the sampling of random numbers can present privacy vulnerabilities when looking at the least significant bits of the generated noise.





Existing techniques to overcome these failings were arduous to implement, typically slower than naive sampling, and difficult to generalise to arbitrary distributions [159], [160].

As part of work in this area, we constructed a new technique taking advantage of the (infinite) divisibility of the probability distributions of interest to DP implementations, by generating a single sample from the desired distribution as a sum of variates from its divisors. By way of example, to generate a single sample from the Gaussian (normal) distribution, we propose using a sum of multiple Gaussian variates, thereby preventing the attack presented in [159]. This technique is fast, easy to implement and simple to understand.

The outcome of this research was published at the ESORICS conference in late 2021 [161]. This solution is also used for secure noise generation in the third party Opacus software.²⁴

5.2.3 Results

5.2.3.1 Seeding All Diffprivlib functions can now be seeded to produce repeatable “random” outputs. This is available for all mechanisms, tools and models, and will be especially useful for scientists and engineers looking for reproducible outputs, or for bug fixing and consistent testing of code.

Figure 41 gives an example of the difference of seeding a Diffprivlib function with an integer and a `RandomState` instance.

```
>>> import diffprivlib as dp
>>> dp.mechanisms.Laplace(epsilon=1, sensitivity=1, random_state=42).randomise(0)
-0.8652695764638703
>>> dp.mechanisms.Laplace(epsilon=1, sensitivity=1, random_state=42).randomise(0)
-0.8652695764638703

>>> rng = dp.utils.check_random_state(42)
>>> dp.mechanisms.Laplace(epsilon=1, sensitivity=1, random_state=rng).randomise(0)
-0.8652695764638703
>>> dp.mechanisms.Laplace(epsilon=1, sensitivity=1, random_state=rng).randomise(0)
0.09503204532300952
```

Figure 41. Example code showing the difference of seeding a Diffprivlib function with an integer, and seeding with a `RandomState` instance. In the former, repeated execution returns the same “random” value. In the latter, different values are returned, but repeated execution of the same script will give the same overall output. Typically, the second behaviour is what is desired.

5.2.3.2 Random Forest The new implementation of the `RandomForestClassifier` has added further classification functionality to diffprivlib. Random forests are a commonly-used tool in machine learning and data analysis, and are especially useful in the DP context for their ability to generalise easily to highly dimensional problems.

In the example given in Figure 42, the random forest classifier is trained on a dataset with ten features, but non-private accuracy is nonetheless achieved at $\epsilon = 0.02$, which is good by DP standards.

The refactored code introduced in Diffprivlib v0.6 gave substantial improvements in many areas. The code now follows similar parametrisation requirements to other Diffprivlib models (*i.e.*, necessitating a `bounds` parameter to specify the range of the dataset), and is also much faster by utilising existing functionality within Scikit-Learn (including parallel processing).

²⁴<https://github.com/pytorch/opacus/blob/v1.4.0/opacus/optimizers/optimizer.py#L119-L134>



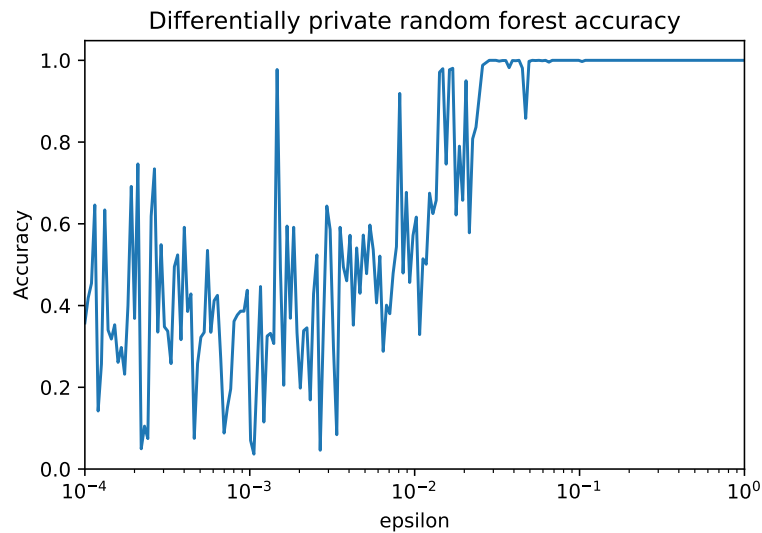


Figure 42. Example performance of the Random Forest classifier with differential privacy across various epsilon (privacy loss) values. The simulations were completed using Scikit-Learn’s `make_blobs` data generator, with 10,000 samples generated over 3 centres, with a 80/20 train/test split. As can be seen from the plot, maximum performance is achieved approximately when $\epsilon = 0.02$.

5.2.3.3 Secure Sampling The implementation of secure sampling has been integrated within Diffprivlib. Although there is a performance penalty with making these changes (as outlined in Figure 43), this comes with the advantage of significant security improvements for outputs. The work also gives scientists and engineers the option of increasing the security of the noise generation even further, by using even more uniform variates in the generation of individual samples. For the purposes of DP however, using 4 uniform variates to generate a single sample is sufficient.

In Figure 43, we demonstrate the performance comparisons across a number of techniques: (i) the naive (insecure) sampling approach (using equation (2) in [161]), (ii) the secure procedure presented in Theorem 1 of [161], (iii) sampling normal Gaussians directly from Python’s `random` module (and summing/scaling accordingly to produce a Laplace sample), and (iv) sampling Gaussians using Numpy. The code for these experiments can be found in Appendix B of [161].

5.2.4 Relevant Resources and Publications

Relevant publications:

- N. Holohan and S. Braghin. “Secure Random Sampling in Differential Privacy”. In Computer Security–ESORICS 2021: 26th European Symposium on Research in Computer Security, Darmstadt, Germany, October 4–8, 2021, Proceedings, Part II 26 (pp. 523-542). Springer International Publishing [161].
Zenodo record: <https://zenodo.org/record/8211750>.
- N. Holohan. “Random Number Generators and Seeding for Differential Privacy”. ArXiv e-prints 2307.03543 [cs.CR] [157].
Zenodo record: <https://zenodo.org/record/8211753>.

Relevant resources:



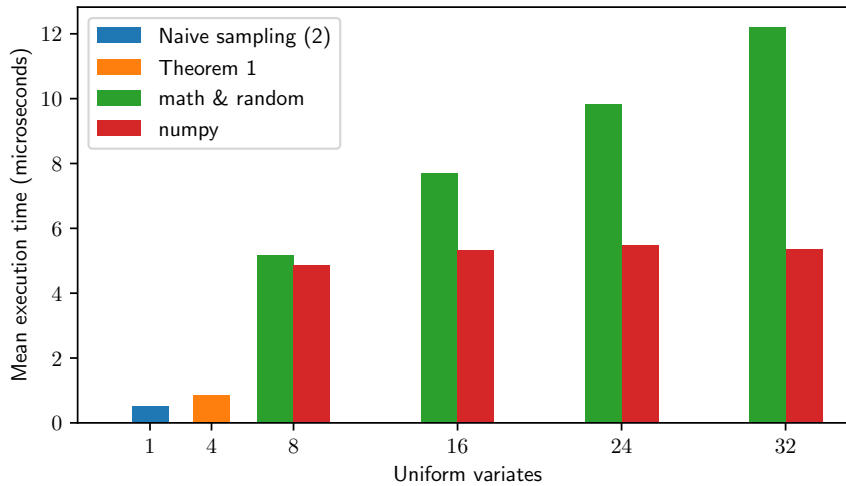


Figure 43. Computation time for secure sampling versus naive (insecure) sampling in blue. The increased sampling time for using more uniform variates is offset by the (exponentially) increased security and attack resistance.

- Diffprivlib Github repository:
<https://github.com/IBM/differential-privacy-library>.

5.2.5 Relevance to AI4Media use cases and media industry applications

Differential privacy has a multitude of potential use cases in the media industry, wherever sensitive data is being collected, stored or ingested, and offers a great opportunity for media companies to meet their data protection obligations while maintaining the usability and functionality of that data. This is particularly important for those in the news industry dealing with sensitive and confidential information, e.g., provided by whistleblowers, where the data itself cannot be released for fear of identifying the whistleblower, or revealing sensitive information about individuals.²⁵ In this example, differential privacy can be used to publish dataset-specific information without compromising the privacy of the individuals involved.

Another relevant application in the media industry is privacy-preserving surveying, where differential privacy principles can also be leveraged through randomised response. In cases where surveys are sought on very sensitive subjects, randomisation can be used to protect participants’ privacy while still collecting accurate statistics [162], [163].

5.3 A Utility-Preserving De-Identification Approach with Relation Extraction Filtering

Contributing partners: IBM

According to a recent report²⁶, about 80% of all data produced will be unstructured by 2025 and much of this ever-increasing amount of data is in the form of free text. Examples of these are reports, contracts, and medical notes, which contain an overwhelming amount of information yet to

²⁵ Cf. a recent data breach at the Police Service of Northern Ireland: <https://www.irishtimes.com/crime-law/2023/08/11/psni-data-leak-qa-reputational-and-financial-damage-trail-security-worries/>

²⁶ <https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf>





be properly leveraged. One of the reasons for this resides in the fact that the de-identification of unstructured documents is a non-trivial task. Existing solutions [164]–[166] attempt to solve the ambiguity issue of free text analysis by leveraging the context, but do not preserve the utility of the de-identified text.

Modern Natural Language Processing (NLP) methods have reached a fundamentally new level²⁷. One of the areas of NLP which is highly useful in the case of de-identification of personal information is Named Entity Recognition (NER). NER methods are able to analyze text and detect a predefined set of concepts (person names, dates, addresses, or more specific ones, like drug names). A detailed explanation of the recent de-identification approaches can be found in the following research works [167]–[170]. The main drawback of these approaches is that they can mark non-sensitive entities as Personal Identifiable Information (PII) producing high false positives. This research introduces a novel framework for the detection of sensitive entities based on Relation Extraction (RE) filtering.

5.3.1 Method

Figure 44 shows the main steps of the proposed method. The pipeline of the framework is organized as follows. We first split the input text into sentences and then pass the obtained sentences through a set of predefined entity detectors, which results in a list of detected entities. These entities are marked as “*potentially sensitive*” and transferred to the relation extraction module. To detect the entities, we adopt a set of rule-based methods proposed in [170] and expand it with state-of-the-art neural-based methods [171]–[173] to improve the detection accuracy.

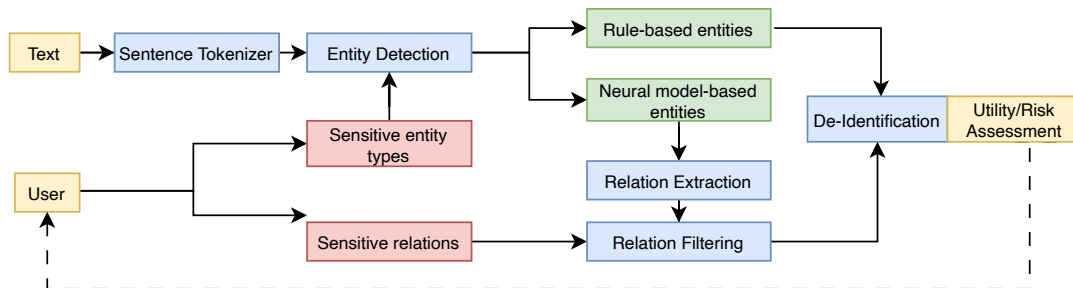


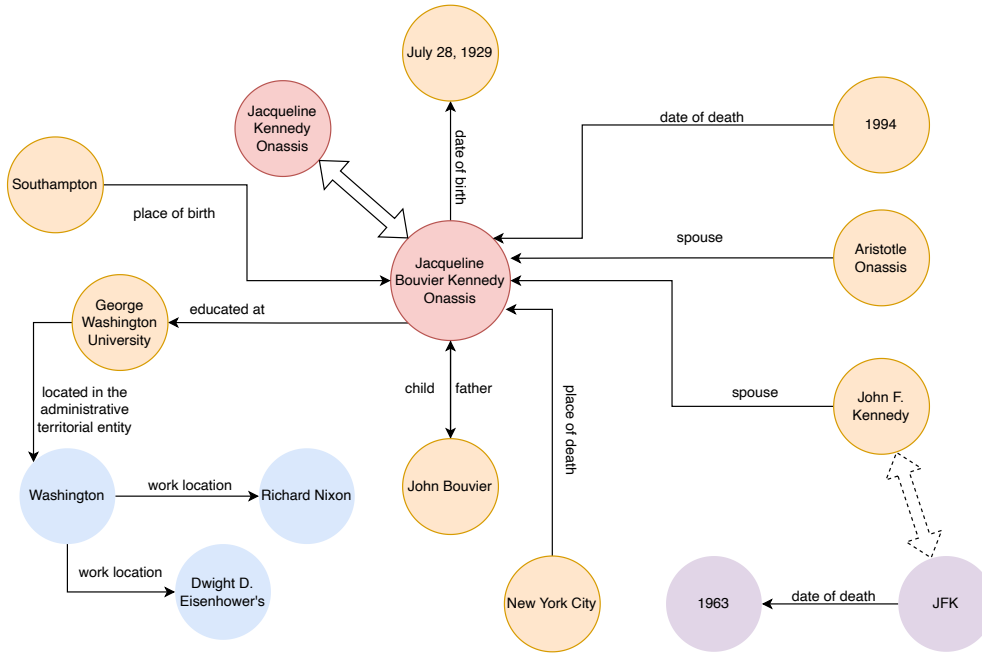
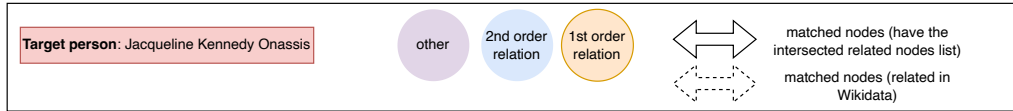
Figure 44. Proposed method diagram.

Via the relation extraction module, we filter the list of previously detected entities and maintain only the sensitive ones. We apply Document-level Relation Extraction with Evidence-guided Attention Mechanism (DREEAM), a RE model proposed in [174], to the list of detected entities to determine the relations between them. The user can provide a list with sensitive relations as input to the relation filtering module. As the output of DREEAM, we obtain a graph-like structure with all the relations between the detected entities. Entities that are connected by a sensitive relation are marked as “*sensitive*” and transferred to the de-identification stage. The remaining entities are marked as “*non-sensitive*” and filtered out.

De-identification is usually performed by redacting sensitive values. However, in order to maintain the utility of the document, more advanced mechanisms can be utilized: tagging (“John” becomes “NAME-1”), masking (“John” becomes “Mark”), generalization (“Paris” becomes “France”), or noise addition (for example, using differential privacy methods), and more. Depending on the requested utility level, the user can parameterize the system with the desired approach.

²⁷<https://openai.com/blog/chatgpt/>





Non-sensitive entities

French Catholic - MISC	Paris - LOC	1951 - TIME	the White House - ORG	Onassis - PER
Irish Catholic - MISC	D.C. - LOC	1953 - TIME	the Washington Times-Herald - ORG	Queen Elizabeth II - PER
French - MISC	B.A. - LOC	1961 - TIME	first - NUM	Janet - PER
Inquiring Camera Girl - MISC	New York - LOC	1968 - TIME		

Jacqueline Bouvier Kennedy Onassis was born on **July 28, 1929**, in **Southampton, New York**. Her father, **John Bouvier**, was a wealthy **New York** stockbroker of **French Catholic** descent, and her mother, **Janet**, was an accomplished equestrienne of **Irish Catholic** heritage.

Upon returning from **Paris**, **Onassis** transferred to **George Washington University** in **Washington, D.C.**, and graduated with a **B.A.** in **French** literature in **1951**. After graduating from college in **1951**, **Onassis** landed a job as the "**Inquiring Camera Girl**" for **the Washington Times-Herald** newspaper. Her job was to photograph and interview various **Washington** residents, and then weave their pictures and responses together in her column. Among her most notable stories were an interview with **Richard Nixon**, coverage of President **Dwight D. Eisenhower's** inauguration and a report on the coronation of **Queen Elizabeth II**.

Jacqueline Kennedy Onassis married **John F. Kennedy** in **1953**. When she became first lady in **1961**, she worked to restore **the White House** to its original elegance and to protect its holdings. After **JFK's** assassination in **1963**, she moved to **New York City** and married **Aristotle Onassis** in **1968**. She died of cancer in **1994**.

Figure 45. Example of entities and relationships

5.3.2 Experimental Evaluation

To evaluate the performance of the proposed framework, we use the Text Anonymization Benchmark (TAB) dataset [175]. The TAB dataset consists of court case records labelled by experts for the purpose of text anonymization. For evaluation, we select 282 documents that satisfy the DREEM input length requirements (1024 tokens).



We compute the percentage of true positives and the number of false positives (TP% and FP, respectively) as well as the utility of the de-identified text in two scenarios: (i) with the RE-based filtering and (ii) without the RE-based filtering. We also compute the *informational content*, which is defined as the negative logarithm of the probability of predicting the masked entity from the text, as proposed in [175]. Thus, the higher the probability, the lower the *informational content*. The informational content represents the weight of inferring the masked entity. We use this value to compute the weighted precision, which we refer to as utility.

Table 16 demonstrates the performance of the proposed method on the TAB dataset. We

Framework	TP%	FP	Utility Increment
Proposed	79%	4,792	+ 560% (0.28)
Proposed w/o RE	99.7%	42,046	baseline (0.05)

Table 16. Performance measured on the TAB dataset.

observe that the proposed method without relation extraction detects 99.7% of the tokens, but at the same time, it suffers from a high number of false positives (over 40,000), which significantly decreases the utility of the approach (0.05). Via the RE-based filtering, the proposed method decreases the number of FPs by up to 9 times, resulting in a significant utility increase (from 0.05 to 0.28) while having a slight drop in the number of true positives. Thus, the proposed method significantly improves the utility of the de-identified documents.

5.3.3 Relevant Resources and Publications

Relevant publications:

- L. Nedoshivina, A. Halimi, J. Bettencourt-Silva, and S. Braghin. “A Utility-Preserving De-Identification Approach with Relation Extraction Filtering”, The 23rd Privacy Enhancing Technologies (PETS) Poster, 2023.
Zenodo record: <https://zenodo.org/record/8279802>.

5.3.4 Relevance to AI4Media use cases and media industry applications

Document de-identification is highly relevant in the media sector. It protects individuals’ privacy rights by identifying and removing sensitive information such as names, addresses, or contact details from documents. At the same time, it helps media organizations comply with the data protection laws. De-identifying documents allow researchers to share datasets or reports without compromising individuals’ privacy. It also promotes responsible and ethical reporting by removing the sensitive information of the individuals involved in news stories, legal cases, or investigations.

5.4 Secure Federated Learning

Contributing partners: FhG-IDMT

Note: T4.4 will be extended to the end of the project (M48) to continue the work described in this section.

Usually, a machine learning model is trained at a central server, with all training data being aggregated from participants / clients prior to the training process. From the perspective of model performance, this approach is probably ideal, but depending on the application, it can come





with serious drawbacks: Providing all training data to a central entity may come with significant cost (think of big media archives and large amounts of data, or small sensor devices with little bandwidth). More importantly, participants may not want to share their data with a central entity at all, due to privacy and confidentiality considerations.

The idea of federated learning is that the training data stays with the participants / clients, and only the model weights are shared: Every client performs a local training, resulting in a local model update. Then, all models are sent to the central server, and the aggregate model is sent back to the clients. While this approach is likely to come with a decrease in performance, ideally, it still performs well, and all clients benefit from each other without having the need to transmit their training data to other actors.

While Federated learning is very helpful when it comes to reduce the amount of shared data, it is however not guaranteed that it avoids privacy and confidentiality concerns, because the trained models can still reveal a lot of information about the clients and their training data. The original federated learning paper mentions this in a footnote [176, p. 2]: “For example, if the update is the total gradient of the loss on all of the local data, and the features are a sparse bag-of-words, then the non-zero gradients reveal exactly which words the user has entered on the device. In contrast, the sum of many gradients for a dense model such as a cnn offers a harder target for attackers seeking information about individual training instances (though attacks are still possible).” Not surprisingly, reconstruction attacks on CNNs turned out to be feasible and effective.

Hence, it is clear that there is a need for securing federated learning, and depending on the given use case and attacker model, there are several candidates technologies regarding privacy enhancement technologies for federated learning:

- Differential Privacy
- Fully Homomorphic Encryption (FHE)
- Secure Multiparty Computation (SMPC)

The secure federated learning tool will consist of a set of modules that allows incorporating selected Privacy Enhancement Technologies (PET) in federated learning frameworks. For AI4Media we decided to select DP and FHE.

5.4.1 Methodology

Depending on the use case, it is important to specify attacker models and define the level of trust put into other participants, and select appropriate security measures based on that. For instance, there might be scenarios where only the (cloud) server is untrusted, while other scenarios will also require protecting against other clients. A basic FHE scheme where all clients share the same key can prevent the server to learn anything on the model data, thereby addressing the former type of scenario, but it will not prevent a malicious client to spy on others, which is necessary for the latter scenario.

In the domain of federated learning, the notion of an *honest, but curious* attacker is common, emphasizing the privacy aspect. Participants in the federated learning systems are suspected to get as much out of the data as they can (or lose them after being hacked), but to do so *passively*. In contrast to an *active* attacker, the aggregator can run model inversion attacks on the individual model updates it receives, but it will not send faulty data to individual clients, which could compromise the overall system performance.

From all possible participant / attack mitigation combinations, we will start with differential and FHE for Federated Averaging: Differential Privacy (DP) can be applied to the training data directly





but DP can also be applied to the resulting model weights of the local training *before* sending it to the aggregator, which is specially useful within the AI4Media context, and will therefore be supported by the tool. This will include investigation of the trade-off between performance degradation and resilience regarding model inversion attacks.

The benefits of FHE come with significant cost in terms of computing time and memory requirements (by orders of magnitudes). This makes direct neural network computations, e.g., encrypted inference, infeasible for many problems relevant for AI4Media. However, a central part of Federated Learning is the computation of the aggregated model, which can be as simple as calculating an average. These operations are a good fit for FHE, and support will be provided by the tool for popular FL frameworks like *flower*²⁸. Practical security issues like key exchange and modifications to the Federated Averaging will be investigated as well.

5.4.2 Results and Updates

We continued working of the integration of FHE into *flower*, selecting CKKS as the default encryption scheme. The initial proof of concept worked well on MNIST classification problems. We are now developing a second version, updated for recent versions of *flower*, that is able to provide more than just encrypted Federated Averaging.

The integration is as transparent as possible, by just adding a few annotations inside the “vanilla” *flower* code, all the messages will be encrypted using CKKS and aggregated in the encrypted domain by the server.

Benchmarks and documentation will be published in the upcoming deliverable D4.7.

5.4.3 Relevance to AI4Media use cases and media industry applications

While privacy is a feature of AI applications that only a few use cases will go without, the proposed approach deals with privacy within Federated Learning systems. Regarding AI4Media, there is no Use Case directly dealing with Federated Learning, yet. Regarding the broader media industry, the outlook of not having to share private data (being it usage, user or content data), and therefore avoiding all the practical hassles of data exchange (usage rights, data exchange contracts, data privacy laws, . . .) is so promising, that there will be real industry applications for Federated Learning. On that premise, applications that improve the privacy (for robustness and fairness, see Section 3.2) of Federated Learning are worth researching and will be relevant in the future as they are already in non-media domains such as medicine or industrial applications.

5.5 Differentially Private Graph Learning

Contributing partners: IDIAP

5.5.1 Overview

Graph Neural Networks (GNNs) have shown superior performance in solving the problems formulated as a machine learning task over graphs, such as node classification, link prediction, and graph classification, in various disciplines from social network analysis and recommendation services to drug discovery and medical diagnosis. However, the graphs used to train such models could be sensitive and contain personal information, and this information can be leaked through the model’s output, when the GNN is released publicly, or when it is offered as a service [177]–[179].

²⁸<https://flower.dev>





For example, a GNN trained over a social network for friendship recommendation may reveal the graph’s linkage information through its predictions. As another example, a GNN trained over the social graph of COVID-19 patients to predict the spread of the disease could be used as a service by government authorities, but an adversary might be able to recover the private graph used for training.

This research aims to prevent the information leakage of the underlying graph in GNNs using DP, which is a widely accepted mathematical framework for measuring the privacy guarantees of algorithms that operate on sensitive data. First, we propose a novel differentially private GNN based on Aggregation Perturbation (GAP), which adds stochastic noise to the GNN’s aggregation function to statistically obfuscate the presence of a single edge (edge-level privacy) or a single node and all its adjacent edges (node-level privacy). To reduce the excessive privacy costs of the aggregation perturbation technique, GAP decouples the neighborhood aggregation steps from the learnable parameters of the model, enabling the aggregations to be pre-computed and perturbed only once, leading to reduced privacy costs.

Next, we investigate the application of progressive learning to privacy-enhancing GNNs, and we show that it can be used to substantially improve the accuracy-privacy trade-off of differentially private GNN models that utilize the aggregation perturbation technique without sacrificing privacy. Our proposed progressive method trains the GNN in a series of steps, with each step building upon the private node embeddings learned by the previous ones. This approach maintains the representational power of GNNs while limiting the incurred privacy costs.

5.5.2 GAP: Differentially Private GNNs with Aggregation Perturbation

We briefly explain GAP, our differentially private GNN model satisfying edge-level privacy, which is also extensible to node-level privacy if combined with standard private learning algorithms such as DP-SGD [180]. As perturbing an edge in the input graph can practically be viewed as changing a sample in a node’s neighborhood aggregation set, GAP enhances edge privacy via *aggregation perturbation*: we add calibrated Gaussian noise to the output of the aggregation function, which can effectively hide the presence of a single edge (edge-level privacy) or a group of edges (node-level privacy). To avoid accumulating privacy costs at every model update, we propose a custom GNN architecture (Figure 46) comprising three individual components: (i) the encoder module, where we pre-train an encoder to extract lower-dimensional node features without relying on the graph structure; (ii) the aggregation module, where we use aggregation perturbation to privately compute multi-hop aggregated node embeddings using the graph edges and the encoded features; and (iii) the classification module, where we train a neural network on the aggregated data for node classification without further querying the graph edges.

Aggregation perturbation allows us to benefit from higher-order, multi-hop aggregations by composing individual noisy aggregations, yet the proposed architecture significantly reduces the privacy costs as the perturbed aggregations are computed once on lower-dimensional embeddings, and reused during training and inference. GAP also provides inference privacy, as the inference of any node relies on the perturbed aggregations, which hide information about neighboring nodes. Due to reusing cached aggregations, the inference step does not incur additional privacy costs beyond that of training.

5.5.2.1 Experimental Results We analyze GAP’s formal privacy guarantees using Rényi Differential Privacy [181], and empirically evaluate its accuracy-privacy performance on three medium to large-scale graph datasets, namely Facebook [182], Reddit [183], and Amazon [184]. The Facebook dataset comprises anonymized Facebook interaction data among UIUC students from September 2005, where users are represented by nodes and friendships by edges. It includes user



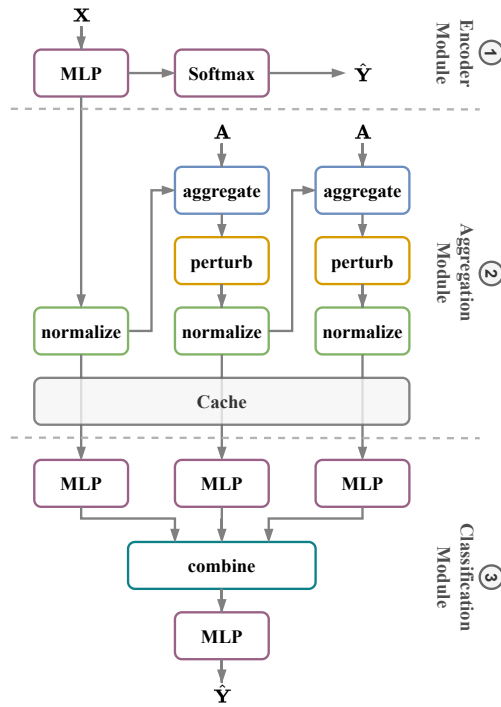


Figure 46. Overview of GAP’s architecture: (1) The encoder is trained using only node features (X) and labels (Y). (2) The encoded features are given to the aggregation module to compute private K -hop aggregations (here, $K = 2$) using the graph’s adjacency matrix (A). (3) The classification module is trained over the private aggregations for label prediction.

attributes like student/faculty status, gender, major, minor, and housing status, and the objective is to predict users’ class year. The Reddit dataset includes Reddit posts, where each node signifies a post and an edge signifies the same user commenting on both posts. Node features are derived from post content embeddings, and the goal is to predict the subreddit (community) to which a post belongs. Finally, the Amazon dataset represents the product co-purchasing network on Amazon, where nodes are products and edges denote products purchased together. The nodes’ features are based on bag-of-words vectors of product descriptions processed by PCA, with the aim of predicting product categories.

We compare GAP with a simple MLP baseline, which does not use graph edges, and also with GraphSAGE, which is a popular GNN architecture. We vary the privacy cost parameter ϵ from 0.1 to 8 for edge-level private methods and from 1 to 16 for node-level private algorithms and report the accuracy of the methods under each privacy budget. The result for both edge-level and node-level privacy settings is depicted in Figure 47. It is evident that GAP’s accuracy surpasses the competing baselines’ at (very) low privacy budgets under both edge-level DP (e.g., $\epsilon \geq 0.1$ on Reddit) and node-level DP (e.g., $\epsilon \geq 1$ on Reddit), and we observe that it always performs on par or better than a naive MLP model which does not utilize the graph’s structural information.

While GAP can work in either edge-level or node-level privacy settings, it must be emphasized that the former setting is suitable only for the use cases where the node-level information (e.g., features or labels) is not sensitive or is publicly available (e.g., the vertically partitioned graph setting described in [185]). Whenever node-level information is private as well (e.g., user profiles in a social network), however, edge-level privacy fails to provide appropriate privacy protection, and

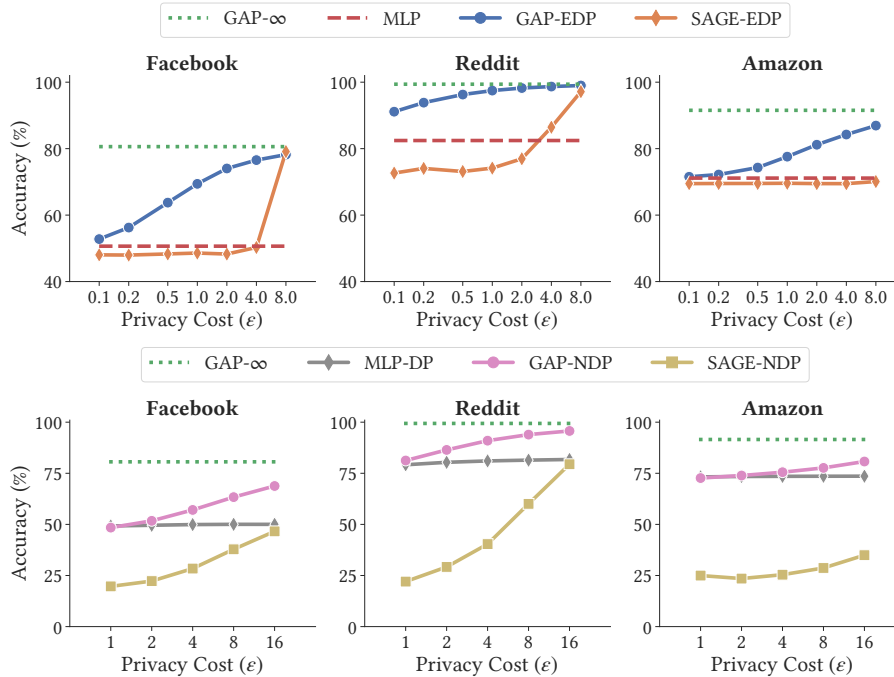


Figure 47. Accuracy vs. privacy cost (ϵ) of edge-level private algorithms (top) and node-level private methods (bottom). $-\infty$, $-EDP$, and $-NDP$ suffixes correspond to non-private, edge-level and node-level DP, respectively.

thus node-level privacy setting has to be enforced.

5.5.3 ProGAP: Progressive GNNs with Aggregation Perturbation

As discussed in the previous section, GAP recursively aggregates node features first, and then trains a classifier over the resulting perturbed aggregations, enabling DP to be maintained without incurring excessive privacy costs. However, despite outperforming relevant baselines, this decoupling approach reduces the representational power of the GNN due to having non-trainable aggregations, leading to suboptimal accuracy-privacy trade-offs.

To address this challenge, we present a novel differentially private GNN, called “**Progressive GNN with Aggregation Perturbation**” (PROGAP), which is depicted in Figure 48. Our new method uses the same AP technique as in GAP to ensure DP. However, instead of decoupling the aggregation steps from the learnable modules, PROGAP adopts a multi-stage, progressive training paradigm to surmount the formidable privacy costs associated with AP. Specifically, PROGAP converts a K -layer GNN model into a sequence of overlapping submodels, where the i -th submodel comprises the first i layers of the model, followed by a lightweight supervision head layer with softmax activation that utilizes node labels to guide the submodel’s training. Starting with the shallowest submodel, PROGAP then proceeds progressively to train deeper submodels, each of which is referred to as a training stage. At every stage, the learned node embeddings from the preceding stage are aggregated, perturbed, and then cached to save privacy budget, allowing PROGAP to learn a new set of private node embeddings. Ultimately, the last stage’s embeddings are used to generate final node-wise predictions.

The proposed progressive training approach overcomes the high privacy costs of AP by allowing

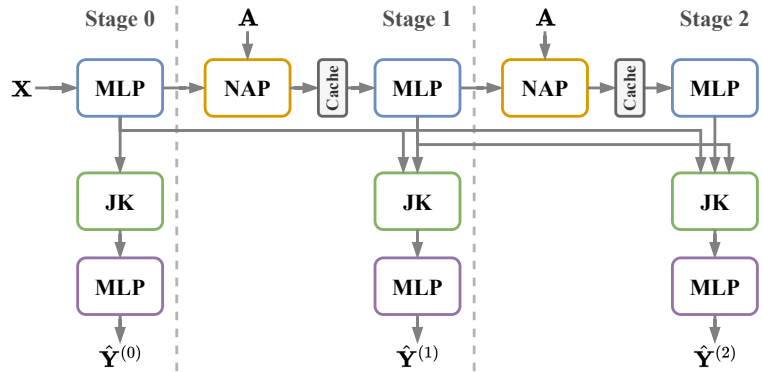


Figure 48. An example PROGAP architecture with three stages. MLP and JK represent multi-layer perceptron and Jumping Knowledge [186] modules, respectively. NAP denotes the normalize-aggregate-perturb module used to ensure the privacy of the adjacency matrix, with its output cached immediately after computation to save privacy budget. Training is done progressively, starting with the first stage and then expanding to the second and third stages, each using its own head MLP. The final prediction is obtained by the head MLP of the last stage.

the perturbations to be applied only once per stage rather than at every training iteration. PROGAP also maintains a higher level of expressive power compared to GAP, as the aggregation steps now operate on the learned embeddings from the preceding stages, which are more expressive than the raw node features. Moreover, we prove that PROGAP retains all the benefits of GAP, such as edge- and node-level privacy guarantees and zero-cost privacy at inference time.

5.5.3.1 Experimental Results We test our proposed method on node-wise classification tasks and evaluate its effectiveness in terms of classification accuracy and privacy guarantees on four datasets: Reddit and Amazon, which were also used to evaluate GAP, and also Facebook-100 [182] and WeNet [187], [188].

We varied the privacy parameter ϵ between 0.25 to 4 for edge-level privacy and 2 to 32 for node-level private algorithms. We then recorded the accuracy of each method for each privacy budget. The outcome for both edge-level and node-level privacy settings is depicted in Figure 49. Notably, we observe that PROGAP achieves higher accuracies than GAP across all ϵ values tested and approaches the non-private accuracy more quickly under both privacy settings. This is because in PROGAP each aggregation step is computed on the node embeddings learned in the previous stage, providing greater expressive power than GAP, which recursively computes the aggregations on the initial node representations.

It is worth noting that the performance discrepancy between ProGAP and GAP is not consistent across all datasets. For instance, this gap in accuracy is less pronounced with the Reddit dataset compared to FB-100. This is due to the specific characteristics and the learning task of each dataset, which require different levels of graph representational power. In Reddit, where the goal is to predict the community of nodes (representing Reddit posts), most of the pertinent information needed is already present in the node features, making the relationships between the posts less crucial for this prediction task. In contrast, the learning task of the FB-100 dataset (predicting students' class year) relies more heavily on the graph structure, necessitating more powerful graph representations. Therefore, the performance difference between PROGAP and GAP is more noticeable in this dataset.

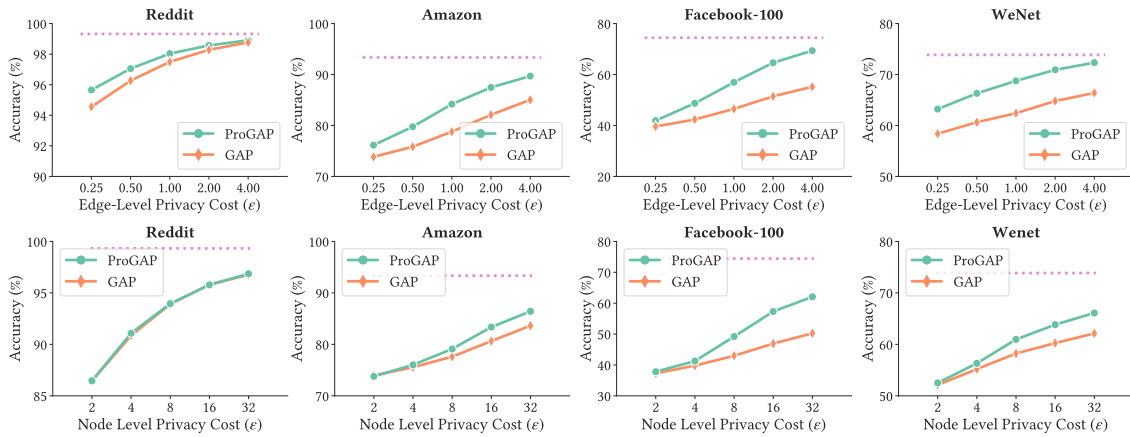


Figure 49. Accuracy-privacy trade-off of edge-level (top) and node-level (bottom) private methods. The dotted line represents the accuracy of the non-private PROGAP.

5.5.4 Relevant Resources and Publications

Relevant publications:

- S. Sajadmanesh, A. Shahin Shamsabadi, A. Bellet, and D. Gatica-Perez. “GAP: Differentially Private Graph Neural Networks with Aggregation Perturbation”. The 32nd USENIX Security Symposium (USENIX Security), Anaheim, CA, USA. Zenodo record: <https://zenodo.org/record/7554788>.
- S. Sajadmanesh and D. Gatica-Perez. “ProGAP: Progressive Graph Neural Networks with Differential Privacy Guarantees”. ArXiv preprint arXiv:2304.08928. <https://arxiv.org/abs/2304.08928>

Relevant resources:

- GAP official implementation: <https://github.com/sisaman/GAP>.
- The official implementation of ProGAP will be released publicly after the publication of the paper.

5.5.5 Relevance to AI4Media use cases and media industry applications

The GAP model, with its focus on differential privacy in Graph Neural Networks, offers a transformative approach to various media industry use-cases. In content personalization, GAP can help media platforms deliver tailored content that respects user privacy, thereby enhancing user engagement and trust. For journalists and researchers, GAP’s capabilities in social media analytics can provide valuable insights into trends and public opinion while adhering to ethical privacy norms. Lastly, in customer segmentation, media companies can use GAP to better understand their audience and deliver targeted services, all while maintaining strict data privacy standards. Overall, GAP presents a scenario that balances the need for advanced analytics with the imperative of user privacy in the media industry.



5.6 Reversible adversarial attacks for privacy protection

Contributing partners: AUTH

5.6.1 Overview

Traditional research on image privacy protection often assumes human adversaries. In other words, privacy risks are usually quantified by how effectively the information contained in images can be picked up by human eyes and brains. As a result, “blurring”, “pixelation”, and “mosaic” are still the most widely used techniques to protect privacy in images, even while their effectiveness against automatic analysis tools is limited [189], [190]. On the other hand, de-identification methods based on universal adversarial attacks [191], almost guarantee that the image data are misclassified by automated analysis systems, while introducing the minimal possible perturbation, maintaining data utility for humans. This is why adversarial attacks are gaining increasing value in privacy protection applications, e.g., they have been employed to disable known automatic face detection/recognition algorithms applied on visual data uploaded by social media users [192], without severely compromising image quality [193], while at the same time, not hiding the person identities to human viewers.

Nevertheless, an important privacy protection aspect is to not only maintain the utility of the de-identified data but to be able to completely restore the original data, upon request. To this end, the most straightforward approach is to maintain a local copy of the original data. An example for this case could be a news video depicting a suspect who is taken into custody. The news company needs to maintain a local copy of the original video, that could be requested by some authority (e.g., the police), while at the same time, create another version of the video that is used for publishing to the general public. When such video duplicates need to be produced for so many cases, this can dramatically increase the storage overhead. Therefore, it would be a lot more useful if the company could just use a single function for calculating the privacy protection transformation, thus only needing a single version of the video file.

Universal adversarial perturbations could be used to this end [194], [195], however, the actual transformation to the images is merely an additive noise, and most importantly, it is the same for any given input image. Thus, the perturbation is easily attainable by a third party with access to a single original and perturbed image pair. Therefore, in privacy protection applications, it is essential that this transformation is also unique for a given input image.

To this end, we propose the Transformation-based Universal Adversarial attacks, where the adversarial perturbation can be obtained by a single transformation function. Using this framework, the adversarial perturbation is unique for any given input, therefore, it is not easily attainable for third parties. To increase its applicability to privacy protection scenarios against automatic classification systems, we formulate two variants where the transformation function is invertible, therefore, we can obtain the original image from its adversarial counterpart. In the first variant, this transformation function is linear, while the second, the transformation function is a reversible GAN. The methodology is detailed in the Subsections below.

5.6.2 Transformation based adversarial attacks

Let $\mathbf{x} \in \mathbb{R}^D$ be a vectorized image sample of dimensions D (D is equal to the image’s height \times width) having a true label index y from a set $\mathcal{Y} = \{y \mid y \in \mathbb{N}, 1 \leq y \leq C\}$. A deep neural network classifier $f(\mathbf{x}; \theta)$, where θ are the model trainable parameters, has learned to classify images by training the operation $\mathcal{X} \mapsto \mathcal{Y}$ in the representative dataset $\mathcal{S} = \{\mathcal{X}, \mathcal{Y}\}$, $|\mathcal{S}| = N$, $\mathcal{X} = \{\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^D\}$.





The universal adversarial perturbation (UAP) [194] is an adversarial attack that generalizes to almost all data samples $\mathbf{x} \in \mathcal{X}$. The optimization problem can be formulated as follows:

$$\begin{aligned} \min_{|\mathbf{n}|} &: f(\mathbf{x} + \mathbf{n}; \boldsymbol{\theta}) \neq y, \quad \forall \mathbf{x} \in \mathcal{X}, \\ \text{s. t. :} & \|\mathbf{n}\|_p < \epsilon, \quad p \in [1, \infty), \end{aligned} \quad (14)$$

where ϵ a parameter for controlling the magnitude of the perturbation. In practice, the perturbation is calculated by accumulating the outputs of DeepFool for all samples $\mathbf{x} \in \mathcal{X}$. As a stopping condition, the function $P(f(\mathbf{x} + \mathbf{n}; \boldsymbol{\theta}) \neq f(\mathbf{x}; \boldsymbol{\theta})) \leq 1 - \delta$ is introduced, where $P(\cdot)$ is a probability function and $0 < \delta < 1$ is a parameter that denotes the target fooling rate to be achieved ($\delta = 0$ denotes a fooling rate of 100%).

The adversarial attack optimization problem can also be viewed as a transformation estimation one, that is expressed as follows:

$$\begin{aligned} \min_{|\boldsymbol{\Phi}|} &: f(\mathbf{g}(\mathbf{x}; \boldsymbol{\Phi}); \boldsymbol{\theta}) \neq y, \\ \text{s. t. :} & \|\mathbf{x} - \mathbf{g}(\mathbf{x}; \boldsymbol{\Phi})\|_p < \epsilon, \quad p \in [1, \infty) \end{aligned} \quad (15)$$

where $\mathbf{g}(\cdot) : \mathbb{R}^D \mapsto \mathbb{R}^D$ is an iterative transformation that maps the data samples of the clean domain \mathcal{X} to an adversarial domain $\tilde{\mathcal{X}}$, while $\boldsymbol{\Phi}$ are the parameters of the transformation. Here, it should be noted that any type of function can be employed in order to solve the proposed optimization problem, i.e., $\mathbf{g}(\cdot)$ could be represent any linear/non-linear transformation, or even a whole neural network. This formulation allows more flexibility in the definition of additional optimization constraints. For instance, the constraint of reversibility, which is very useful in privacy protection settings, could be expressed as an additional optimization constraint, i.e., $\mathbf{g}^{-1}(\tilde{\mathbf{x}}) = \mathbf{x}$.

5.6.3 The linear case

The simplest possible case is that $\mathbf{g}(\cdot)$ denotes a linear transformation that perturbs clean samples from their domain to an adversarial one, such that they are misclassified by the model f . This definition makes more sense in the universal adversarial attack optimization problem. The transformation parameters in this case include a matrix $\mathbf{T} \in \mathbb{R}^{D \times D}$ and a bias term $\mathbf{b} \in \mathbb{R}^D$. Therefore, adversarial samples can be represented as follows:

$$\tilde{\mathbf{x}} = \mathbf{T}\mathbf{x} + \mathbf{b}. \quad (16)$$

Within the scope of AI4Media project, we examined the special case where $\mathbf{b} = \mathbf{0}$, where $\mathbf{0}$ is a vector of zeros. We developed the Multiplicative Universal Adversarial Transformation (MUAT) method, which is a multiplicative noise generator formulated as follows:

$$\begin{aligned} \min_{\|\mathbf{T}\|} &: f(\mathbf{T}\mathbf{x}; \boldsymbol{\theta}) \neq y, \\ \text{s. t. :} & \|\mathbf{x} - \mathbf{T}\mathbf{x}\|_p < \epsilon, \quad p \in [1, \infty), \\ & \mathbf{x} = \mathbf{T}^{-1}\tilde{\mathbf{x}}, \end{aligned} \quad (17)$$

where an additional constraint requiring that the matrix \mathbf{T} is invertible is also imposed. In the standard additive noise-based universal adversarial attacks, the perturbation is attainable by a single adversarial-clean image pair, by a simple subtraction. However, in the multiplicative noise case, the analogous is to reverse engineer the matrix \mathbf{T} from the data, which cannot be obtained, using just a pair of clean-adversarial samples, since the rank of \mathbf{T} is supposed to be larger than 1.





The limitations of the linear case are the following. The transformation matrix is calculated at the input space of the image, therefore, such formulation only allows application is images of specific resolution. When high resolution images are used, this matrix becomes analogously very big, thus it could be very difficult to be optimized and stored. In addition, given a sufficiently large number of images, the transformation matrix can be estimated by a third party.

5.6.4 The non-linear case

In the non-linear case, we examine the case where $g(\cdot)$ represents a whole neural network, and more specifically a Generator. Inspired by Image-to-image translation (I2I), our work considers a clean image domain \mathcal{X} and an adversarial image domain \mathcal{Y} . The adversarial image domain can be obtained implicitly, by training a generator to produce adversarial examples, or explicitly, by using any adversarial attack. Then, our goal is to create an image-to-adversarial image translation model which is approximately invertible by design. The image-to-image translation aims at transferring images from a source to a target domain while retaining content representations [196] [197]. According to [198], the goal is to find the appropriate mapping between two given domains \mathcal{X} and \mathcal{Y} , while minimizing the corresponding loss functions for unpaired training data. To this end, two mappings $F : X \rightarrow Y$ and $F^{-1} : Y \rightarrow X$ are learned, following the cycle-consistency.

In a similar fashion, we create a generator $F : \mathcal{X} \rightarrow \mathcal{Y}$ such that $F(\mathbf{x}_i) = \mathbf{y}_i^{adv}$ in order to generate adversarial examples such that $f(\mathbf{y}_i^{adv}) \neq f(\mathbf{x}_i)$ (untargeted attack). Also, we design an “inverse” generator, $F^{-1} : \mathcal{Y} \rightarrow \mathcal{X}$. Then, F^{-1} is another architecture that produces \mathbf{x}_i^{rec} as approximations of \mathbf{x}_i . Figure 50 depicts the architecture of our model.

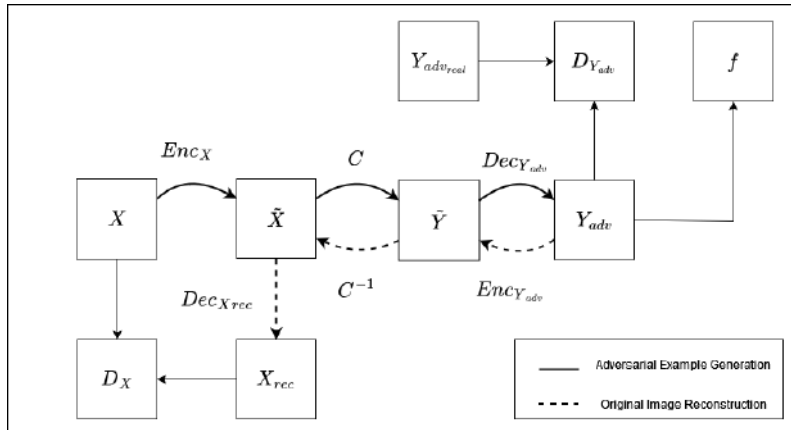


Figure 50. Architecture of the proposed AdvRevGAN model.

The forward mapping of generator F and the backward one of F^{-1} are broken down into three components. X is the original image domain, $Y_{adv,real}$ is the original adversarial image domain while Y_{adv} is the domain of adversarial generated images that are produced by F . We associate a feature space \tilde{X} and \tilde{Y} in higher dimensions for each domain respectively. Mappings between original and adversarial image space are individual and non-invertible. More specifically, for real image space X , we use an encoder $Enc_X : X \rightarrow \tilde{X}$ that extracts the image features of X , lifting the image into a higher dimensionality feature space and a decoder $Dec_{X_{rec}} : \tilde{X} \rightarrow X_{rec}$ that switch the image back to a lower image space in same dimensions as the initial. We follow the same procedure for generated adversarial image domain Y_{adv} using $Enc_{Y_{adv}} : Y_{adv} \rightarrow \tilde{Y}$ and $Dec_{Y_{adv}} : \tilde{Y} \rightarrow Y_{adv}$.





Between feature spaces, we have an invertible core such that $C : \tilde{X} \rightarrow \tilde{Y}$ and $C^{-1} : \tilde{Y} \rightarrow \tilde{X}$. As a result, we demonstrate the full mappings that are:

$$F(X) = Dec_{Y_{adv}} \circ C \circ Enc_X(X) \quad (18)$$

$$F^{-1}(Y_{adv}) = Dec_{X_{rec}} \circ C^{-1} \circ Enc_{Y_{adv}}(Y_{adv}), \quad (19)$$

where \circ denotes the composition of Enc_X , C , $Dec_{Y_{adv}}$ for function F and $Enc_{Y_{adv}}$, C^{-1} , $Dec_{X_{rec}}$ for function F^{-1} . Also for each image space, X and Y_{adv} we use domain-specific discriminators D_X and $D_{Y_{adv}}$ for training with the adversarial loss.

The main advantage of the non-linear case is that this formulation can be potentially used for images of various resolutions, while it remains easy to implement and store. However, the main disadvantage is that the reconstructed image is not exactly equal to the original image, but only an estimated version of it.

5.6.5 Experiments

In order to evaluate the methods, we have conducted experiments in image classification settings. We refer the readers to the relevant publications for more details. In our first set of experiments, we have evaluated the proposed methods on their ability to ensure privacy protection against neural network classifiers, on MNIST dataset. Since both methods are in essence Universal Adversarial attacks, we evaluate them as such. That is, the linear and non-linear transformation methods are evaluated in terms of how much noise they add to the original images in order to achieve a universal adversarial attack constraint, i.e., the mean square error (MSE) between the original images and the adversarial images produced. In addition, the methods are also evaluated in terms of visual similarity between adversarial images the original ones, according to the Structural Similarity (i.e., SSIM [199] metric). For comparison reasons, the methods are evaluated against the SGD-UAP [200], which is the state-of-the-art universal adversarial attack method. As can be observed in Table 17, the proposed methods are able to generate adversarial attacks with less noise when compared to the SGD-UAP.

Table 17. Comparison results on MNIST dataset

	Accuracy (initial dataset)	Accuracy (attacked dataset)	MSE(x, y^{adv})	SSIM(x, y^{adv})
AdvRevGAN	98.4%	0.09%	0.017	0.908
MUAT	98.4%	0.01%	0.056	0.384
SGD-UAP	98.4%	0.07%	0.106	0.300

Example figures of the proposed method can for MNIST dataset be seen in Figure 51. As can be observed, both MUAT and AdvRevGAN can generate adversarial images containing less noise than their legacy Universal Adversarial Perturbation methods. In addition, MUAT is able to fully reconstruct the original image, while AdvRevGan can reconstruct the original image, with a minor noise margin. Here, it should be noted that both methods achieve the reconstruction without requiring access to the original image. Therefore, based on our results and while the research moves towards our path, we could advise the media end-users to only store the de-identified version of the image, as the original image can be (almost) reconstructed upon request.

5.6.6 Conclusion

Two reversible adversarial attack methods have been described, that produce a reversible mapping function that uniquely maps given input images into an adversarial domain, where its inverse


















Original	AdvRevGAN	MUAT	SGD-UAP	$T^{-1}\tilde{x}$	$F^{-1}(Y_{adv})$
	5 	3 	3 		
	3 	5 	3 		
	1 	5 	3 		
	3 	5 	3 		
	3 	5 	3 		

Figure 51. Adversarial examples and reconstructed images on MNIST Dataset. The first column depicts original images x_i , the next three columns are the corresponding adversarial examples y_i^{adv} generated by the proposed method, MUAT and UAP respectively while above them demonstrated the wrong class that predicted by the model. In the last two columns are demonstrated the reconstructed images x_i^{rec} derived by MUAT and our proposed method respectively.

can either fully (linear case) or almost (non-linear case) reconstruct the original input. The developed methods allow the generation of untargeted adversarial examples that are also reversible for different dataset complexities using generative adversarial networks (GANs). The developed methods generate adversarial attacks with less noise when compared to their legacy counterparts. Last but not least, the transformation cannot be obtained by third parties, since it is unique for a given input, and requires access to the transformation matrix or the neural network architecture and parameters.

According to recent research, diffusion models are suggested as a promising alternative to GANs for generating diverse and realistic samples as they use a diffusion process to iteratively transform a noise vector into a sample that matches the data distribution, and they have shown to be more stable and easier to train than GANs. Their ability to capture complex multi-modal distributions makes them a viable alternative for generating synthetic data in scenarios where labeled data is limited or costly to obtain. Future work could include extending the proposed architecture to also accommodate differential privacy constraints in the adversarial attack optimization problem using more complex datasets.

5.6.7 Relevant Resources and Publications

Publications:

- A. Zamichos, V. Mygdalis, and I. Pitas, “Properties of learning Multiplicative Universal Adversarial Perturbations in image data”, In IEEE International Conference on Machine Learning for Signal Processing (MLSP), 2022.
Zenodo record: <https://zenodo.org/record/8276422>.
- S. Altini, V. Mygdalis, and I. Pitas, “AdvRevGan: On Reversible Universal Adversarial Attacks for privacy protection applications”, In IEEE International Conference on Machine





Learning for Signal Processing (MLSP), 2023.
Zenodo record: <https://zenodo.org/record/8276432>.

5.6.8 Relevance to AI4Media use cases and media industry applications

The relevant privacy protection tools can be used in AI4Media UC3: “AI in Vision - High Quality Video Production & Content Automation”, in the following manner. We assume a scenario during news broadcasting, where human faces must be de-identified prior to broadcasting. The media producer may opt to use the proposed technology (accompanied with a face/human detector) in order to select the area where the privacy transformation will be applied. Without this technology, the producer must store two versions of the video, i.e., the broadcasted one, where the privacy transformation has been applied, and the “clean” one, for archiving purposes. Using the proposed technology, the producer may only store a single video file, where he/she can restore the privacy protected broadcasted version and the original “clean” video version at will.





6 Fair AI (Task 4.5)

As machine learning models are fast becoming critical components of every decision making process essential for our society (mortgage lending, prison sentencing etc), it becomes crucial to guarantee that these models do not privilege specific groups or individuals at the disadvantage of others. These models are constructed upon the statistical analysis and properties of training data, which may contain biases due to existing prejudice and/or inaccurate sampling. Hence, if left unchecked unwanted biases can emerge from these models with significant societal consequences.

AI Fairness is typically evaluated either on a group or individual level. When addressing group fairness, a population is divided into groups based on a set of protected attributes (gender, ethnicity, etc.). A fair ML model within this context is a model which seeks some statistical measure to be equal across such groups. On the other hand, when addressing individual fairness, ML models seek to treat individuals similarly regardless of their protected attributes.

Algorithms and metrics designed to address biases in ML models can operate on the training data itself as well as on the trained model. Moreover, they can also occur at various points in the machine learning lifecycle whether at a pre-processing, in-processing, or post-processing phase. T4.5 seeks to apply AI fairness algorithms and metrics at group and individual levels and at various points in the AI lifecycle.

Contributions towards this task include work on (i) datasheets that can be passed along with data and machine learning models to highlight potential risks and recommend appropriate uses (Section 6.1), (ii) fairness of deepfake detection systems (Section 6.2), and (iii) debiasing neural networks using explainable AI (Section 6.3).

6.1 Datasheet for the Dataset on European Press Coverage of Covid-19 Vaccination News

Contributing partners: IDIAP

6.1.1 Goals

The ubiquity of extensive data collection and its use for machine learning (ML) creates a strong demand for fairness, accountability, transparency, and ethics. While fast-paced developments in AI are delivering innovations, the risk of their use is highly dependent on the underlying data they have been trained with [201]. Many examples show that societal biases in the data can be reproduced or amplified by ML models [201], [202]. Therefore, it is crucial that researchers and practitioners are aware of the biases and imbalances in the data they use to train models. A growing body of research, notably initiated by the work by Gebru et al. [201] is investigating the use of context documents that can be passed along with data and ML models to highlight potential risks and recommend appropriate uses. Such documentation methods can increase transparency and accountability within the ML community, mitigate unwanted societal biases in models, facilitate greater reproducibility of results, and help researchers and practitioners to select more appropriate datasets for their chosen tasks [201]. Through the work of an AI4Media Junior Fellow hosted at Idiap, we contributed to these developments by creating a datasheet for a dataset built to investigate the news coverage of Covid-19 vaccinations by European, high-quality press [203]. The datasheet is provided in Appendix A.1.





6.1.2 Method

Datasheets for datasets, first proposed by Gebru et al. [201], are intended to be useful for dataset creators and consumers to encourage ethical reflection about data collection and usage. However, datasheets are also aimed to be valuable to policy makers, journalists, individuals whose data is included in datasets, and people who may be impacted by models trained or evaluated using datasets [201]. Additionally, datasheets facilitate reproducibility by providing researchers and practitioners with detailed useful information. Keeping these goals in mind, the datasheet proposed by [201] contains over 50 questions about the motivation, composition, collection, pre-processing, uses, distribution, and maintenance of a dataset in order to provide the right information to the above-mentioned stakeholders. As discussed by Gebru et al. these questions include [201]:

- Motivation: Articulate the reason for the dataset creation and promote transparency about funding interests.
- Composition: Provide detailed information to dataset consumers so they can make an informed decision whether to use the data.
- Collection: Provide information to help researchers and practitioners to recreate datasets with similar characteristics.
- Pre-processing: Inform dataset consumers about any pre-processing done that might not fit their goals and tasks.
- Uses: Reflect on the tasks that the dataset might or might not be used for.
- Distribution and Maintenance: Provide dataset consumers with infrastructure information and points of contact

It is important to note that a datasheet is a document that evolves over time, as new knowledge regarding the dataset becomes available.

6.1.3 Discussion

Creating a Datasheet for Datasets undoubtedly offers valuable insights into the nature of the underlying data, enabling researchers and practitioners to comprehend the essential aspects of the data they intend to use. However, it is not a task without its own complexities, especially when the datasheet involves information from sensitive or unconventional areas such as news articles serving as social data. In the current context, these complexities are threefold.

Firstly, a datasheet requires a thoughtful consideration of news articles as a form of social data. These articles not only encompass critical insights about the global events or phenomena they represent (like the Covid-19 vaccination debate) but also include intricate details about individuals and entities involved. The process of translating these articles into a structured data form that preserves relevant social nuances is a challenging task. There are multiple dimensions to consider. These range from discerning the role of prominent figures, understanding public sentiment, and interpreting the framing of news stories. Developing a systematic methodology to convert this multi-dimensional data into structured, understandable inputs for a datasheet can be a challenge that future research must address.

Secondly, the creation of a datasheet is compounded by the scarcity of public examples, specifically those related to news data. This presents the challenge of creating an effective datasheet from a relatively limited pool of examples and best practices. While there is existing literature that details the effectiveness of datasheets [204], or domain-specific adaptations of datasheets [205]–[208],





it may not necessarily provide a comprehensive framework for news data. Future work should thus consider developing a comprehensive template or framework for news data, drawing from a broad range of examples and including input from experts in both the data science and journalism fields.

Lastly, there are limitations in the Datasheet for Dataset approach that must be addressed. The datasheet is undeniably a critical starting point for researchers looking to understand their data better and leverage it effectively. It is also a fundamental tool in the journey toward creating trustworthy AI. However, its use should not be limited to a one-time assessment. Instead, it should be continuously updated and maintained to reflect the dynamic nature of data and its underlying context. Furthermore, there is a need to promote its use as a standard practice among researchers and ML practitioners. This can be achieved through training, awareness campaigns, and integrating it as a core part of the data science and ML curriculum.

Overall, while the creation of a Datasheet for Datasets brings numerous benefits, it presents several challenges that future research must strive to address. By deepening the understanding of news articles as social data, developing standardized datasheet frameworks for specific data domains, and promoting the datasheet approach as a common practice, we can move towards more reliable and ethically sound data use in machine learning and AI.

6.1.4 Relevance to AI4Media use cases and media industry applications

Datasheet creation is relevant to all AI4Media use cases, as well as to media industry applications. The work summarized here represents a concrete example of how to do it.

6.2 Exploring Fairness of an AI-based Deepfake Detection Service

Contributing partners: IBM & CERTH

6.2.1 Introduction

This work was completed as part of a virtual Junior Fellow Exchange between IBM and CERTH and is the second of two evaluations of a Deepfake Detection Service created by CERTH - a first evaluation on robustness is detailed in Section 3.1 The MeVer DeepFake Detection (DFD) service [7] was developed by CERTH-ITI for aiding in the detection of manipulated images and videos. The system is comprised of a pre-processing pipeline and model ensemble scheme which is used to obtain a probability score for input images/videos indicating a likelihood of manipulation. The DFD service was initially evaluated using three standard deepfake datasets: FaceForensics++; CelebDF-V2; and WildDeepFake, and was shown to perform competitively with state-of-the-art alternatives.

Section 3.1 details an evaluation of the DFD service with respect to robustness. Whilst this contributes toward one facet of Trustworthy AI²⁹, a comprehensive evaluation of the service, other facets must be considered, such as *fairness*. A limitation of the assessment of the DFD service thus far, which is addressed in this research, is the lack of evaluation with respect to fairness in the output predictions. Bias is an unfortunate and common feature of many machine learning models and presents an ongoing challenge for the developers of AI-services. Bias is a prejudice in favor or against a person, group, or thing that is considered to be unfair. In this context, if a deepfake detector is found to unfairly favour a group and assign (or not assign!) deepfake labels accordingly, certain groups may find their content on social media sites frequently flagged and taken down. The images/videos of these groups may also then rarely appear in publications as journalists using

²⁹<https://research.ibm.com/topics/trustworthy-ai>





biased deepfake detectors for verification purposes might be misled, and so certain groups could be under-represented in media to a greater degree.

In addition, whilst many studies exist in literature evaluating AI models with respect to fairness [209], there is a gap understanding how bias identified in such models changes when subjected to adversarial attacks and when combined with adversarial defences or bias mitigation strategies. Therefore, the focus of this work was to:

- Apply AI Fairness 360 (AIF360) fairness metrics [210] to the DFD service.
- Integrate an AIF360 bias mitigation algorithm with the DFD service to improve fairness.
- Combined use of IBM's Adversarial Robustness Toolbox (ART) [9] and AIF360 to evaluate the fairness of the DFD service under adversarial conditions to ascertain if groups are equally vulnerable to attack.

and subsequently measure:

- any bias present in the DFD service.
- the fairness performance of the DFD service with a bias mitigation algorithm applied.
- the fairness of the DFD service under different adversarial conditions.

6.2.2 Identifying Bias in the DeepFake Detection Service

The driving motivation of this work was to determine if the DFD service showed bias toward a subset of protected groups when scoring images/videos as deepfakes and, if so, attempt to apply a mitigation strategy to improve the fairness of the DFD service.

Xu et al. [211] recently published a comprehensive analysis on the biases prevalent in the most common video data sets used to train state-of-the-art deepfake detection models and conducted an evaluation regarding how such models were subsequently influenced. In their work, five datasets were annotated with 47 additional attributes relating to demographic and non-demographic characteristics of the subjects depicted in real and deepfake videos. The datasets selected for annotation were Celeb-DF, DeepFakeDetection, FaceForensics++, DeeperForensics and the DeepFake Detection Challenge Dataset.

In the context of fairness, some of the attributes annotated can be classified as sensitive or protected characteristics, such as: gender, ethnicity and age. Other annotated attributes are not protected but describe features of the individuals within the videos which are closely associated with protected features, such as: skin, hair, face geometry and accessories (e.g., makeup, eyeglasses).

Xu et al. evaluated three deepfake detection backbone models: Xception, EfficientNet and Capsule-Forensics-v2. In contrast, the MeVer DFD service is an ensemble model which leverages EfficientNet-B4, EfficientNet-V2-m (as opposed to solely EfficientNet-B0 evaluated by Xu et al.) and the Detection Transformer (DETR). Therefore, a fairness evaluation of MeVer's DFD service, discussed in the following sections, provides new insight over how such an ensemble model for deepfake detection compares with the state-of-the-art. Amongst the conclusions drawn by Xu et al. were that many attributes, both demographic and non-demographic, strongly influenced the predictions of the deepfake detectors. The authors did not explore mitigation techniques, a gap which this work addresses.

IBM's AI Fairness 360 (AIF360) toolkit [210] is an open-source library which implements state-of-the-art techniques for detecting and mitigating bias in AI models. The toolkit supports detection and mitigation at multiple points in a machine learning pipeline – pre-processing, training, post-processing - and provides a simple interface for integrating fairness checks and strategies for





reducing the influence of bias in AI models. In this work, the AIF360 toolkit was used to evaluate potential bias inherent in the DFD service and take steps toward evaluating mitigation strategies which could be implemented.

The following four popular fairness metrics were used to assess the predictions made by MeVer’s DFD service on the FF++ and Celeb-DF datasets using the AIF360 toolkit:

- **Disparate Impact (DI):** the ratio of the rate of favourable outcomes for the unprivileged group to the privileged group. A value of 1 indicates the deepfake detector is fair, values contrary to this indicate the deepfake detector is biased.
- **Statistical Parity Difference (SPD):** the difference in the rate of favourable outcomes received between unprivileged and privileged groups. A value of 0 indicates the deepfake detector is fair. Values below 0 indicate bias toward the privileged group. Values greater than 0 indicate bias toward the unprivileged group.
- **Equal Opportunity Difference (EOD):** the difference of the True Positive Rate (TPR) between the unprivileged group and the privileged group. A value of 0 indicates the deepfake detector is fair, whilst other values indicate bias.
- **Average Odds Difference (AOD):** the average difference of False Positive Rate and True Positive Rate between unprivileged and privileged groups. A value of 0 indicates the deepfake detector is fair, whilst other values indicate bias.

Table 18. Fairness assessment of DFD service. Values of 0 (for Statistical Parity difference, Average Odds difference and Equal Opportunity difference) and 1 (for Disparate Impact) indicate a fair deepfake detector.

feature	Statistical Parity		Disparate Impact		Average Odds		Equal Opportunity	
	FF++	Celeb-DF	FF++	Celeb-DF	FF++	Celeb-DF	FF++	Celeb-DF
race	0.07	-	1.13	-	0.05	-	0.10	-
gender	-0.12	-0.19	0.80	0.77	-0.15	0.01	-0.08	-0.03
age	0.15	-	1.33	-	0.33	-	0.03	-
attractive	0.29	0.19	1.72	1.28	0.28	0.12	0.29	0.11
shiny skin	0.55	-	2.82	-	0.55	-	0.51	-
beard	0.28	-0.23	1.56	0.68	0.37	-0.22	0.26	-0.1
face	-0.20	-	0.67	-	-0.20	-	-0.22	-
accessory	0.15	0.2	1.35	1.42	0.11	0.09	0.15	0.12

Table 18 illustrates the results for each fairness metric for 8 attributes of interest, including 3 protected attributes (gender, race, age). The Celeb-DF dataset, as stated by Xu et al., does not contain the same number of annotated attributes found in FF++, hence not all rows could be completed for this dataset. The results indicate that several attributes prevalent in videos of both datasets are influencing the DFD service predictions for deepfake detection. For instance, the SPD for gender is negative indicating that the DFD service is biased toward the privileged group, which in this case is male. This can be interpreted as the DFD service not making positive (deepfake) predictions for videos featuring men and women at an equal rate. This is echoed in the DI result as values are below the ideal value of 1, and the AOD result for gender indicates a difference in deepfake prediction error rate between male and female videos (albeit more so for the FF++ videos). Similarly, the EOD of -0.08 indicates that both genders do not have equal





opportunity for positive deepfake predictions, which also indicates bias is present. In addition to gender, bias was also detected in the remaining two protected features, age and race, as well as non-protected features such as attractiveness and skin.

In their analysis of bias, Xu et al. assessed their deepfake detection models using Relative Performance (RP) - comparing performance of the model when attributes were present versus when they were not. RP values of 0 indicate an attribute does not affect performance of the model when predicting deepfakes - positive values indicate a lower error rate when the attribute is present and negative values indicate lower error rate when the attribute is not present. To address data imbalance of attributes in the test sets, which could cause results to be misconstrued, the authors also opted to analyse the Corrected Relative Performance (CRP), which includes the RP of control groups to correct for the impact of data imbalance.

The RP and CRP were calculated for attributes using the MeVer DFD service predictions with the FF++ and Celeb-DF datasets. Figure 52 illustrates the RP-vs-CRP plots. Attributes closer to the top of the plot (green area) are associated with higher deepfake detection performance, whilst those lower (red area) are associated with higher errors. Points closer to the bisectrix line are less impacted by imbalances within the test data. For example, heavy makeup has a RP value of $\approx 80\%$ to -100% (white stars in plots of Figure 52), which suggests that wearing heavy makeup results in twice as many deepfake detection errors compared to not wearing heavy makeup. Following the analysis of the protected attribute, gender, using the AIF360 toolkit which highlighted bias toward the male class (i.e. that videos with a male subject received a higher proportion of positive deepfake predictions), the blue points in Figure 52 illustrate that when the video had a male subject, the DFD service correctly predicted a deepfake $\approx 20\%$ more often than if the video did not contain a male subject (i.e. contained a female subject).

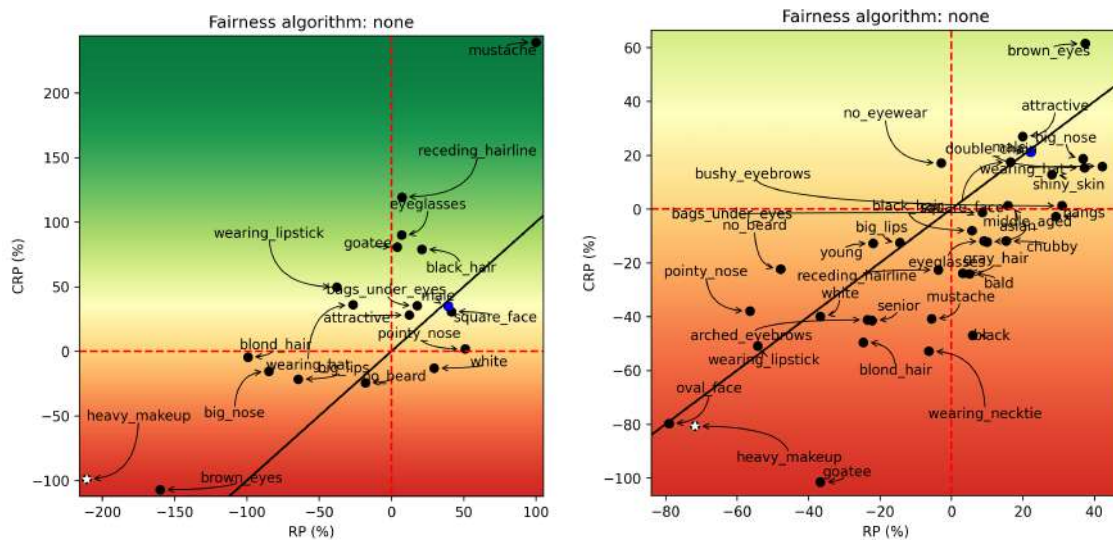


Figure 52. Relative Performance vs Corrected Relative Performance plots for predictions of the DFD service on videos of Celeb-DF (left) and FF++ (right) illustrating bias toward (green) and against (red) attributes.

Following a similar approach, Xu et al also proposed evaluating RP on pristine (real) and fake data alone to further understand the influence of attributes on deepfake detectors for both partitions. Figure 53 illustrates the Pristine Data Relative Performance (PDRP) and the Deepfake Data Relative Performance (DDRP) for attributes and deepfake predictions of the MeVer DFD



service. A negative PDRP for an attribute a , indicates real videos with a are more likely falsely predicted as deepfakes than videos without a . Negative DDRP values indicate that deepfake videos with a are less likely to be predicted as deepfakes. Positive PDRP values indicate real videos with a are more likely to be correctly predicted whilst positive DDRP values indicate fake videos with a are less likely to be predicted as real. For example, the attribute big lips (yellow points in Figure 53), has a positive PDRP ($\approx 50\%$) and negative DDRP (≈ -30 to -200%), which subsequently indicates that in an authentic video, if the subject has big lips, the DFD service makes correct deepfake predictions 50% more often than if big lips were not present. In addition, if a subject in a deepfake video has big lips, the DFD service makes $\approx 30\%+$ more errors than in videos without the attribute. Following the previous analysis for the protected attribute gender, the male attribute has a positive DDRP in both Celeb-DF and FF++ plots. This indicates that when deepfakes with male subjects are generated and passed to the DFD service, correct predictions occur $\approx 25\%$ more often than if the subject was female. The PDRP for the male attribute is shown to differ between the Celeb-DF and FF++ datasets, which could indicate representations of gender in the pristine (real) videos of both datasets differ and the variation was subsequently learned by the DFD service.

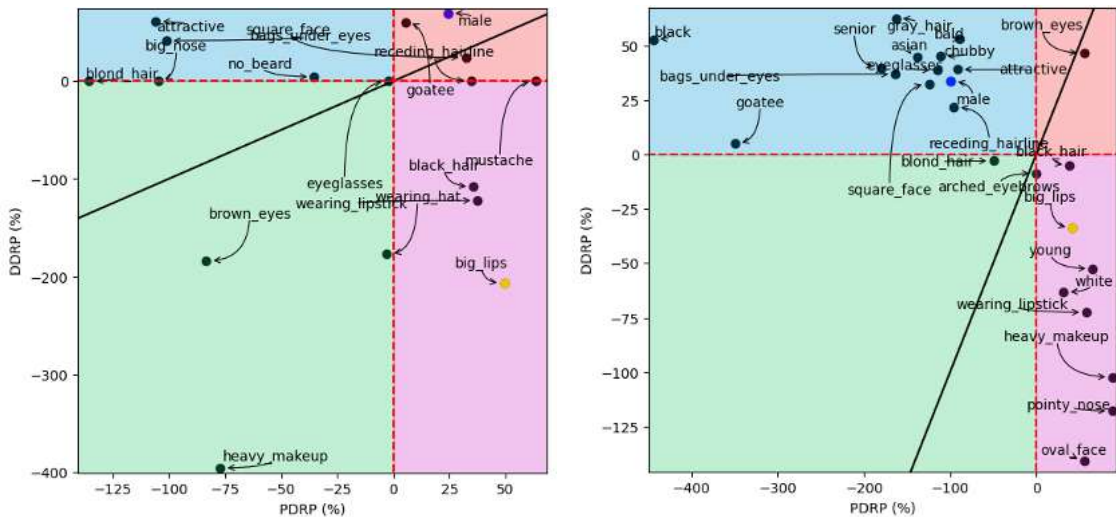


Figure 53. Deepfake Data Relative Performance vs Pristine Data Relative Performance for predictions of the DFD service on videos from Celeb-DF (left) and FF++ (right).

6.2.3 Toward Fair Deepfake Detection

The analysis thus far has sought to evaluate the fairness of deepfake predictions made by the MeVer DFD service across a variety of demographic and non-demographic characteristics present in real and deepfake videos using established metrics from the AIF360 toolkit and those proposed in recent literature. From this analysis it has been shown that both protected and non-protected attributes hold influence over the DFD service and that predictions are not always fair for subjects with certain characteristics – this is consistent with recent studies on fairness in deepfake detection models. To attempt to address this shortcoming, this section details the integration of mitigation strategies with the DFD service and analyses the subsequent results. Three post-processing algorithms were selected from the AIF360 toolkit for analysis:

- **Reject Option Classification (ROC):** this algorithm assigns favourable outcomes to



unprivileged groups and unfavourable outcomes to privileged groups for samples which are uncertain and within a reject-option band (defined by a margin parameter θ) to identify an optimal classification threshold that optimizes a fairness metric, such as Disparate Impact (ROC-DI), Statistical Parity Difference (ROC-SP) or Equal Opportunity Difference (ROC-EOD).

- **Calibrated Equality of Odds (CEO)**: this algorithm optimizes calibrated classifier score outputs to find probabilities with which to change output labels with an equalized odds objective – that is to say, unprivileged groups are assigned new predictions, which aim to minimize disparities in error rates between the privileged and unprivileged groups. This algorithm attempts to minimize one of the following cost-constraints: False Negative Rate (CEO-FNR), False Positive Rate (CEO-FPR) or a weighted combination (CEO-weighted).
- **Equality of Odds (EOO)**: this algorithm attempts to identify probabilities with which to change output labels such that the model performs equally well for different groups and thus reducing quality-of-service harms.

Table 19. Results for fairness metrics using predictions from DFD service and post-processed using Reject-Option Classification, optimized for Statistical Parity Difference.

feature	Statistical Parity		Disparate Impact		Average Odds		Equal Opportunity	
	FF++	Celeb-DF	FF++	Celeb-DF	FF++	Celeb-DF	FF++	Celeb-DF
race	0.03	-	1.03	-	0.01	-	0.04	-
gender	0.02	-0.09	1.02	0.82	-0.04	0.07	0.08	0.07
age	-0.03	-	0.96	-	-0.05	-	-0.02	-
attractive	0.02	0.13	1.02	1.23	-0.06	0.03	0.07	-0.01
shiny skin	0.25	-	18.85	-	0.14	-	0.29	-
beard	0.10	-0.28	1.13	0.52	0.13	-0.24	0.16	-0.27
face	0.19	-	1.34	-	0.17	-	0.18	-
accessories	0.08	0.18	2.00	1.51	0.05	0.07	0.09	0.14

Table 19 depicts the results for four fairness metrics (previously discussed) calculated with predictions made by the MeVer DFD service which were post-processed using the ROC-SP algorithm. Comparing results with Table 18, it is clear the mitigation algorithm is effective at improving the fairness of the DFD service predictions. The protected attributes of race, gender and age are closer to the ideal values of 0 for SPD, AOD and EOD and 1 for DI. Non-protected attributes, such as attractiveness and facial features also show improved fairness. Naturally however, the post-processing of prediction scores also impacts the quality of predictions which are made by the DFD service. Table 20 illustrates the Balanced Accuracy (BA) scores before and after ROC-SP is applied to DFD service predictions and the new associated classification threshold necessary to achieve fair output labels for each attribute. In the case of the FF++ dataset, the new thresholds vary widely across attributes and reduce the BA scores of the detector, albeit none below 50% (random guess). In the case of Celeb-DF, there is some agreement in the selection of threshold values across attributes and whilst BA is reduced, it is not as significant as FF++. A point of contention may be the selection of a classification threshold which maximizes fairness for certain attributes at the expense of accuracy and lower fairness scores for other attributes.



Table 20. Balanced accuracy scores for post-processed (fair) predictions of the DFD service and the associated adjustments made to the classification threshold for each attribute to achieve fairness.

feature	FF++			Celeb-DF		
	BA before	New threshold	BA after	BA before	New threshold	BA after
race	0.695	0.109	0.547	-	-	-
gender	0.695	0.347	0.648	0.833	0.743	0.766
age	0.695	0.079	0.548	-	-	-
attractive	0.695	0.059	0.525	0.833	0.653	0.814
shiny skin	0.695	0.950	0.564	-	-	-
beard	0.695	0.158	0.617	0.833	0.653	0.811
face	0.695	0.376	0.693	-	-	-
accessories	0.695	0.931	0.592	0.833	0.653	0.814

Figure 54 illustrates RP-vs-CRP plots for predictions from the DFD service post-processed using ROC-SP. In the case of the FF++ dataset, the male attribute has an RP value of $\approx 5\%$ (down from $\approx 20\%$ in Figure 1), which indicates the DFD service now correctly predicts a deepfake just $\approx 5\%$ more often when a male subject is present vs a female subject, which is subsequently fairer. Interestingly, in the case of the Celeb-DF dataset, whilst the new RP value for the male attribute is closer to 0 when mitigation is applied ($\approx +40\%$ to $\approx -30\%$), it is also now negative, indicating the new classification threshold causes the DFD to have $\approx 30\%$ more deepfake detection errors when given videos of male subjects rather than female subjects.

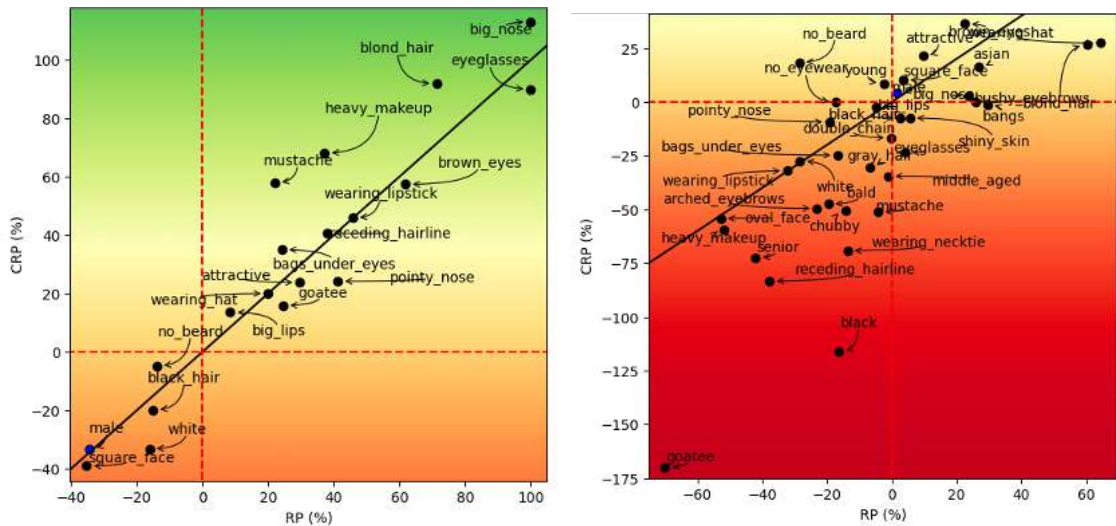


Figure 54. RP-vs-CRP plots for predictions made by the DFD service and post-processed using ROC-SP on videos from Celeb-DF (left) and FF++ (right).



6.2.4 Relevance to AI4Media use cases and media industry applications

The fairness evaluation of the MeVer Deepfake Detection Service is relevant to UC1 as it tackles disinformation detection and, specifically in this evaluation, the performance of deepfake detection across different groups. The results of such evaluations which study how such a deepfake detector can be advantageous or disadvantageous to certain groups of people over others also makes this work relevant to UC4 “AI for Social Sciences and Humanities” as the impact of deploying such an AI system “in-the-wild” without concern for fair treatment of subjects, may introduce biases and scenarios in which certain people are treated unfairly or are discriminated against. As a result, this work is also relevant to UC2 “AI for News”, as a tool which can help journalists discern authentic content from deepfakes, must also ideally be fair and unbiased, which this work strives to achieve.

6.3 Debiasing Neural Models Using Explainable Artificial Intelligence

Contributing partners: 3IA-UCA

6.3.1 Introduction

In recent months, our research has focused on exploring methods to debias existing neural models. The presence of biases in machine learning models poses significant challenges, particularly in sensitive contexts. Our objective was to develop an approach capable to debias an existing model by means of Explainable Artificial Intelligence (XAI) techniques identifying biases within trained models, a subsequent modification of the training samples to reduce the amount of bias in the training set, and a model retraining.

Biases in neural networks can manifest themselves in various ways, leading to unfair and discriminatory outcomes. As shown in [212], a model may learn associations that correlate attractiveness solely with certain characteristics, such as being white or non-chubby. Another bias emerged in the same work, could involve the assumption that old age is negatively correlated with attractiveness. Identifying and rectifying these biases is crucial for ensuring fair and unbiased decision-making.

Standard neural networks lack transparency, making it challenging to comprehend and address biases effectively. It is often unclear which aspects of the training process contribute to the emergence of biases. Consequently, rectifying biases in a trained model becomes an intricate task, hindering often the possibility of making targeted countermeasures.

6.3.2 Debiasing Neural Models Using XAI

To detect biases in the neural models, we employed an eXplainable Artificial Intelligence (XAI) algorithm providing logical explanations of a black-box model behavior. XAI provides interpretability, allowing us to understand how the model reaches its decisions and identify potential biases. By utilizing XAI methods, we aimed to empower users to detect biases in the model’s explanations and facilitate bias mitigation without discarding the entire model.

Indeed, one of the primary causes of biased models is the use of biased training sets. Therefore, our research primarily focuses on addressing biases arising from the training data. The intuition is that if we are capable of keep training our model on a de-biased dataset or eventually retrain it from scratch, we can fundamentally reduce the amount of bias in the model. We devised a method to quantify the satisfaction of bias within the training set and remove data associated with higher biases. This process became feasible with the utilization of XAI algorithms that provide logical explanations of the model’s behavior.





To quantify the satisfaction of the bias (provided by the XAI algorithm by means of a logic rule), we employ the Triangular Norm (T-Norm) operator. By converting the logic rule defining the bias into a numerical constraint, we can compute how much the bias is satisfied (and therefore enforced) by each sample in the training set. This idea follows what has been proposed in [213] within an active learning strategy. It works, however, in the opposite direction: indeed, while [213] was proposing to add all the samples violating a given ground-truth knowledge, in this work we propose to delete all the samples associated to the satisfaction of a found bias. This computation considers both interpretable input features, output predictions of the network, and associated training labels.

Based on the computed satisfaction values, we can identify the samples in the training set associated with the highest levels of bias satisfaction. These highly biased samples are then removed from the training set, and the model is finetuned using the modified dataset. This approach allows us to progressively reduce the bias within the model. In case the finetuning process is not sufficient, the user can decide to retrain the model from scratch to further reduce the bias in the model. In case the removal process would entail many samples, however, the overall performance of the model could be impacted.

In scenarios where new data can be added to the training set, we also define a method to filter and select the most impactful samples for reducing previously identified biases. By incorporating data associated with the highest violations of previously detected biases, we can further refine the training set. A model trained on this latter dataset should not only have lower bias levels, but it should also have comparable (if not higher) predictive performance at inference time.

It may also happen that more than one bias is identified by the user in the provided model explanation. In this case, the debiasing process can be applied to reduce all the detected biases at the same time. To do so, we simply need to consider all the found biases (and the corresponding bias rules), while computing the bias satisfaction level of the training samples.

Our research aims to address the challenging issue of biases in neural models by leveraging the power of XAI e T-Norm operators. By quantifying bias satisfaction using the T-Norm operator, we are able to identify and remove highly biased samples from the training set. Furthermore, our approach accommodates the addition of new data by selecting samples that contribute most significantly to mitigating biases. The proposed methodology offers a promising avenue for debiasing neural models and promoting fairness in their applications.

6.3.3 Relevant Resources and Publications

Relevant publications:

- Ciravegna G., Precioso F., Betti A., Kevin M. and Gori M. “Knowledge-driven Active Learning”. Proceeding of the ECML-PKDD 2023: Joint European Conference on Machine Learning and Knowledge Discovery in Databases [213].
- Ciravegna G., Giannini F., Gori M., Maggini M. and Melacci S., “Human-Driven FOL Explanations of Deep Learning”. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20) [212].

Relevant software and/or external resources:

- The PyTorch implementation of our work “Knowledge-driven Active Learning” can be found in <https://github.com/gabrieleciravegna/Knowledge-driven-Active-Learning>.





6.3.4 Relevance to AI4Media use cases and media industry applications

In most of multimedia databases, the content is multimodal (visual, text, audio, video). If this content is associated with a description, the concepts present in the multimedia content may be described. As a journalist, I can check that an AI service to classify my multimedia content database, or retrieve specific documents in this database, is fair or not. For instance, in the CelebA database, I am going to train a system to classify celeb women faces into attractive vs not-attractive (only asking the system to provide such classification should be questioned, but if we accept this experiment). Then when I check the results on the basis of the knowledge provided for each sample of this dataset, if the class “Attractive” is equivalent to “pale skin” & “not chubby”, I could want to unlearn this class-features relations but just by interacting with the features of the samples. I am then going to ask the system to learn that “Attractive” can be “not pale skin” or “chubby” and let the model retrained. This is what our work intend to do.





7 Contributions to the AI4Media WP8 Use Cases and Media Industry Applications

Work Package 4 holds a unique position in the hierarchy of other technical work packages and their integration with Use Cases (UCs). The tools and components being developed as part of this work package are intended to be used as a trustworthy enhancement to the techniques and modules developed by other technical work packages. These technical work packages comprise WP3 (New Learning Paradigms & Distributed AI), WP5 (Content-centered AI) and WP6 (Human- and Society-centred AI). The technical WP hierarchy is outlined in Figure 55 below, showing how WP4 components feed into all other technical WPs, which then feed into WP8 for deployment in Use Cases.

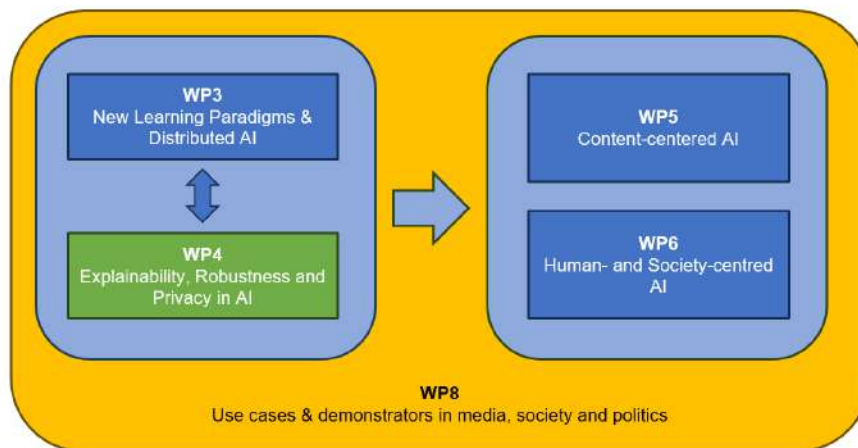


Figure 55. Hierarchy of technical Work Packages, showing the flow of work from top to bottom. WP4 is highlighted in green.

In order to allow for the easy application of WP4 components upon solutions from other technical work packages, a number of concrete steps have been taken. Firstly, all WP4 partners have given an outline of the applicability to AI4Media Use Cases of their work, and/or other potential use cases within the media industry, under the “Relevance to AI4Media use cases and media industry applications” subsection in each section above. These sections were written in non-technical language to allow for their consumption and understanding by professionals in the media industry and other members of the AI4Media consortium for future use of the solutions.

A number of examples were given for the use of AI4Media work by journalists and news organisations, including privacy-preserving surveying of sensitive topics (Section 5.2), and the de-identification of confidential/sensitive documents for publication (Section 5.3). Other strands of the wider media industry are also considered, including the archiving and indexing of image and video material (Section 3.3), the delivery of ads on media platforms that preserve user privacy (Section 5.5), and datasheets to ensure the faithful use and presentation of AI models in published work (Section 6.1). More details can be found in the relevant subsection accompanying each piece of work.

Secondly, WP4 partners compiled a list of all components developed as part of the work package, designed as a first point-of-reference for partners from the other technical WPs looking to improve the trustworthiness of their algorithms. The list, shown in Figure 56, contains a short description of





all components, as well as links to public code repositories. This will be followed up by a catalogue of code available to AI4Media partners for the final phase of the project.

Lastly, the forthcoming plenary meeting of the AI4Media consortium will feature a Speculative Design Workshop to *“produce illustrative scenarios that showcase the potential application and added value of AI4Media technologies that may not all be integrated into the seven use cases.”*



Partner	Name	Description	GitHub repo link	Technology stack (i.e. Language & ML framework)	Relevant publication(s)
IBM	Differential Privacy Library	Library augmenting Scikitlearn ML models with Differentially Private capabilities	https://github.com/IBM/differential-privacy-library	Python, scikitlearn	Holohan, Naoise, Stefano Braghin, Pól Mac Aonghusa, and Kilian Levacher. "Diffprivlib: the IBM differential privacy library." arXiv preprint arXiv:1907.02444 (2019).
IBM	Adversarial Robustness Toolbox	Library providing attacks and defences for deep learning and scikitlearn models	https://github.com/Trusted-AI/adversarial-robustness-toolbox	Python, pytorch, tensorflow or scikitlearn	Nicolae, Maria-Irina, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi et al. "Adversarial Robustness Toolbox v1. 0.0." arXiv preprint arXiv:1807.01069 (2018).
IDIAP	Locally Private Graph Neural Networks	Tool for federated training of Graph Neural Networks (GNNs) with Local Differential Privacy	https://github.com/sisaman/LPGNN	Python, PyTorch	Sina Sajadmanesh and Daniel Gatica-Perez, "Locally Private Graph Neural Networks", in Proceedings of ACM Conference on Computer and Communication Security (CCS), 2021
IDIAP	Differentially Private Graph Neural Networks	Tool for training of and releasing Graph Neural Network models with edge- and node-level differential privacy guarantees	https://github.com/sisaman/GAP	Python, PyTorch	Sajadmanesh, Sina, Shahin Shamsabadi, Ali, Bellet, Aurélien, & Gatica-Perez, Daniel. (2023, August 9). GAP: Differentially Private Graph Neural Networks with Aggregation Perturbation. The 32nd USENIX Security Symposium (USENIX Security), Anaheim, CA, USA.
UNITN	Self-Residual-Calibration Regularization	Tool to address the robust overfitting problem in adversarial training by designing a novel regularization scheme	https://github.com/LynnHqngLiu/AIJ2023-SRC	Python, PyTorch	H. Liu, Z. Zhong, N. Sebe, and S. Satoh, Mitigating Robust Overfitting via Self-Residual-Calibration Regularization, Artificial Intelligence, vol. 137, Article 103877, April 2023.
UNITN	Learning to Attack Real-World Models via Virtual-Guided Meta-Learning	Novel universal attack algorithm (MetaAttack) for person re-ID that can mislead re-ID models on unseen domains by a universal adversarial perturbation	https://github.com/FlyingRoastDuck/MetaAttack_AA_AI21	Python, PyTorch	F. Yang, Z. Zhong, H. Liu, Z. Wang, Z. Luo, S. Li, N. Sebe, and S. Satoh, Learning to Attack Real-World Models for Person Re-identification via Virtual-Guided Meta-Learning. AAAI 2021
FhG-IDMT	FLCrypt	A library to transparently add Fully Homomorphic Encryption to a Federated Learning System	private repository	Python, Pytorch, Tensorflow	TBD
AUTH	Adversarial Robustness by exploiting geometric constraints	Algorithms for training neural networks that increases their robustness to adversarial attacks	private repository	Python, Pytorch	V. Mygdalis and I. Pitas, "Hyperspherical class prototypes for adversarial robustness", Elsevier Pattern Recognition, vol 125, pp 108527, 2022. V. Mygdalis and I. Pitas, "Exploiting One-Class Classification optimization objectives for increasing Adversarial Robustness", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023
AUTH	Adversarial privacy protection algorithms	Algorithms for privacy protection, based on adversarial attacks, against pre-trained NN classifiers	private repository	Python, Pytorch	V. Mygdalis, A. Tefas and I. Pitas, "Introducing K-Anonymity Principles to Adversarial Attacks for Privacy Protection in Image Classification Problems", in IEEE International Workshop on Machine Learning for Signal Processing (MLSP), 2021 A. Zamichos, V. Mygdalis and I. Pitas, "Properties of learning Multiplicative Universal Adversarial Perturbations in image data", in IEEE International Workshop on Machine Learning for Signal Processing (MLSP), 2022
HES-SO	RCVtool	Concept-based interpretability from user queries	https://github.com/maragraziani/rcvtool	Python, Tensorflow, scikit-image	
HES-SO	cdisco	Concept discovery in latent spaces	https://github.com/maragraziani/cdisco	Python, Pytorch, scikit-image	
CEA	Unbiased Control to Generate Synthetic Images	Generative approach to generate synthetic faces, with a method that decrease the biases of the controlled attributes	https://github.com/perladoubinsky/balanced_sampling_gan_controls	Python, Pytorch	P. Doubinsky, N. Audebert, M. Crucianu, H. Le Borgne, "Multi-attribute balanced sampling for disentangled GAN controls", (2022) Pattern Recognition Letters
CERTH	Explainability algorithms for synthetic speech detection models	Algorithms for feature importance and explainability of synthetic speech detection models, trained on spectrograms	private repository	python, pytorch, tensorflow, scikit	

Figure 56. Full listing of WP4 components, compiled by WP4 partners.



8 Ongoing Work and Conclusions

8.1 Ongoing Work

8.1.1 AI Robustness (Task 4.2)

Task 4.2 is due to conclude at the end of the AI4Media project in August 2024 (M48). **AUTH** has concluded its work on this task with the present deliverable.

IBM will continue work on AI Robustness, focusing specifically on evaluating the security and robustness of AI and ML models. As the models and modelling pipelines become more complex, not only in size but also in terms of their use across different domains and applications, new attack vectors emerge. These new threat models require the development of new mitigation approaches. IBM will continue this journey of advancing the threat analysis of emerging AI models and develop tools and techniques to help with their multi-faceted risk analysis.

UNITN will work on improving the robustness to targeted adversarial attacks. Previous works have extensively studied the transferability of adversarial samples in untargeted black-box scenarios. However, it still remains challenging to craft targeted adversarial examples with higher transferability than non-targeted ones. Recent studies reveal that the traditional Cross-Entropy (CE) loss function is insufficient to learn transferable targeted adversarial examples due to the issue of vanishing gradient. In our work, we will address this problem by analysing the logit margin between the targeted and untargeted classes and provide a solution that increases the transferability of adversarial examples.

8.1.2 Explainable AI (Task 4.3)

Task 4.3 is due to conclude at the end of the AI4Media project in August 2024 (M48).

CERTH will continue to refine the existing methodology with an emphasis on enhancing its transparency. This will involve the testing of additional explainable AI techniques aimed at improving the clarity of the synthetic audio detection process. Concurrently, attention will be devoted to enhancing result interpretability by evaluating various metrics. Insights will be provided in textual format to elucidate the methodology's operation and to indicate conditions under which the model's decision-making may be less robust. As advancements in synthetic audio detection models are anticipated from Work Package 6 (WP6), plans are in place to adapt the methodology to these forthcoming models. This adaptation aims to maintain the methodology's applicability and efficacy in the evolving landscape of synthetic audio detection.

Concerning visual analysis, based on our work in WP5 on video summarization, future work will focus on explaining the output of video summarization networks as any gained improvements in this direction would allow a level of understanding about their functionality, increase the users' trust to them, and facilitate the curation of automatically-produced video summaries. Building on our knowledge about the use of attention mechanisms for video summarization, our future work will involve experimentation with: i) various attention-based video summarization architectures and explanation signals (such as the ones using in the NLP domain), and ii) different model-agnostic (e.g. perturbation-based) approaches for spotting the parts of the video that influenced the most the estimates of a video summarization network about the importance of video frames and fragments.

CEA will continue to work on text-to-image (T2I) diffusion models and explore the use of "controlled" synthetic data to enrich the training datasets. Most of works relating to explainable and interpretable AI focus on the statistical models and propose various approaches to provide a useful feedback on their results to the human users. This has also been the case with previous CEA contributions to T4.3 that have been reported in D4.1 and the current report. Our future work for this task for the final year of the project (M37-48) will rather focus on the training data used to





learn the AI models. We expect to allow a human user to augment these datasets with synthetic images that are semantically controlled, such that she has a better understanding of the underlying process. Generating images with T2I images is quite easy nowadays, but a challenge remains in being able to generate actually useful data to improve such models on some specific tasks.

In this vein the CEA will also investigate the quality of the synthetic images generated by T2I diffusion models. It exists some metrics to assess their quality and fidelity [214], [215], to determine whether the generated images contain the specified objects or adhere to certain criteria [216], [217], to evaluate their visual reasoning skills [218] or to delve into the gender depiction disparities enabling the study of potential stereotypes [219]. However, there is still a need to be able to estimate to which extent the generated image content is aligned with the prompt used as input of the T2I model. Studying such an alignment with regard to various possible prompts may provide insights to better explain the behavior of these T2I models.

HES-SO will continue to work on automating the discovery of concepts in imaging models. By focusing on unsupervised approaches that do not require the user input about concepts, or high level features, that are expected to be potentially relevant for the model, we can isolate only the concepts that actually are learned to solve the task. This reduces the risk of user's induced biases in the interpretability analysis. Particularly in the case of deep fake detection, the model may focus on completely unexpected features to identify fake models, and this would foster the development of more robust detection methods. In particular, we plan on focusing on isolating directions in the latent space of a layer (of an imaging model) that point to semantically unique concepts.

3IA-UCA will work on extending SMACE, making it usable in a wider range of applications. A particularly interesting approach to include categorical features in the rules is implemented in CatBoost [220], a gradient boosting toolkit. The idea is to group categories by *target statistics*, which can replace them. SMACE could also be generalized to more complex model configurations, where some models take as input the output of other models.

On the work about Anchor approach, we plan to extend our analysis to other classes of models, such as CART trees, and to more advanced text vectorizers. We also plan to study Anchors' behavior on images and tabular data.

On our works on Concept Embedding Models (CEM), there is room for improvement in both concept alignment and task accuracy in challenging benchmarks such as CUB or CelebA, as well as in resource utilisation during inference/training. Our future effort on this topic will focus on these questions.

UNIFI will continue to work on explainable architectures, extending the work presented in Section 4.6 to transformer based architectures. Our work will be mostly directed towards training agents that work under regimen in which the i.i.d. assumption is violated, such as online driving in which the output of the model influences future inputs. In particular, we will work on approaches able to provide visual explanation of autonomous agent failures, exploiting visual attention.

Moreover, we will work on understanding the reliability of classification results on transformer based architectures addressing uncertainty estimation and out-of-distribution detection. We will study how, visual attention maps obtained from visual transformers can be exploited to train models such as auto-encoders, in order to perform out-of-distribution detection.

8.1.3 Privacy-Enhancing AI (Task 4.4)

T4.4 was due to be completed with the submission of this deliverable (D4.5) in August 2023 (M36). However, it has since been extended to the end of the project (August 2024, M48) to allow **FhG-IDMT** continue working on the task and deliver more research and results in D4.7. The present deliverable marks the conclusion of this task for the other partners, **IBM**, **AUTH**, and **IDIAP**.





FhG-IDMT will focus on the combination of Federated Learning, Differential Privacy and Fully Homomorphic Encryption, for problems with actual relevance in the media sector (e.g., audio classification). The goal is to build a demonstrator that showcases the technologies and the inherent privacy security tradeoff as well as enumerate (attacker) scenarios, where such a combination of privacy enhancing technologies is useful and worth the additional effort.

8.1.4 AI Fairness (Task 4.5)

Like T4.4 above, T4.5 was due to come to a conclusion with the submission of this deliverable (D4.5). However, it too has now been extended until the end of the project (M48) with the agreement of all partners to facilitate more work on the task from **FhG-IDMT** and **3IA-UCA** to be delivered in D4.7. The present deliverable (D4.5) marks the conclusion of work on this task from the other partners, **IBM**, **UCA**, and **IDIAP**.

FhG-IDMT will continue to work on a demonstrator for the detection on mitigation of bias in recommender systems, based on the outcomes of T6.3 (Hybrid, privacy-enhanced recommendation).

3IA-UCA will continue working on debiasing deep networks using XAI rules. Indeed, while our research provides a valuable step forward in debiasing neural models, there are several areas that warrant further investigation. Future studies can explore the combination of multiple XAI algorithms for enhanced bias detection and mitigation. Additionally, evaluating the generalization of debiasing techniques across different datasets and domains will be crucial. It is also important to consider the ethical implications and potential unintended consequences of debiasing methods. Robust frameworks for evaluating the effectiveness and fairness of debiased models should be developed. Furthermore, ongoing research should focus on developing techniques that enable continuous monitoring and mitigation of biases as models evolve over time.

8.2 Conclusions

This deliverable details the considerable volume of work that relevant partners have been conducting in the context of WP4 from M13 to M36. The work has covered all dimensions of trusted AI, covering robustness, explainability, privacy and fairness in AI. The work is backed up by a comprehensive library of scientific publications and open source code. Novel methods have been conceived, authored and delivered, and the existing state-of-the-art has been improved upon across a range of topics. This deliverable represents a valuable checkpoint in the lifetime of the AI4Media project, showcasing the diligent work that has been completed.

The updated version of D4.5 will be provided in M48 (D4.7 – *Final toolset in robust, explainable, fair, and privacy-preserving AI*) and will include the final outcomes of the ongoing work as well as additional investigations regarding the tasks covered in this deliverable.





References

- [1] N. Papernot, P. McDaniel, A. Sinha, and M. Wellman, *Towards the science of security and privacy in machine learning*, 2016. arXiv: 1611.03814 [cs.CR].
- [2] X. Wang, J. Li, X. Kuang, Y.-a. Tan, and J. Li, “The security of machine learning in an adversarial setting: A survey,” *Journal of Parallel and Distributed Computing*, vol. 130, pp. 12–23, 2019, ISSN: 0743-7315. DOI: <https://doi.org/10.1016/j.jpdc.2019.03.003>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0743731518309183>.
- [3] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “Faceforensics++: Learning to detect manipulated facial images,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1–11.
- [4] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-df: A large-scale challenging dataset for deepfake forensics,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3207–3216.
- [5] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, “Wilddeepfake: A challenging real-world dataset for deepfake detection,” in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 2382–2390.
- [6] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” *arXiv preprint arXiv:1706.06083*, 2017.
- [7] S. Baxevasakis, G. Kordopatis-Zilos, P. Galopoulos, L. Apostolidis, K. Levacher, I. Baris Schlicht, D. Teyssou, I. Kompatsiaris, and S. Papadopoulos, “The mever deepfake detection service: Lessons learnt from developing and deploying in the wild,” in *Proceedings of the 1st International Workshop on Multimedia AI against Disinformation*, 2022, pp. 59–68.
- [8] J. Chen, M. I. Jordan, and M. J. Wainwright, “Hopskipjumpattack: A query-efficient decision-based attack,” in *2020 IEEE Symposium on Security and Privacy (SP)*, IEEE, 2020, pp. 1277–1294.
- [9] M.-I. Nicolae, M. Sinn, M. N. Tran, B. Buesser, A. Rawat, M. Wistuba, V. Zantedeschi, N. Baracaldo, B. Chen, H. Ludwig, I. Molloy, and B. Edwards, “Adversarial robustness toolbox v1.2.0,” *CoRR*, vol. 1807.01069, 2018. [Online]. Available: <https://arxiv.org/pdf/1807.01069>.
- [10] G. K. Dziugaite, Z. Ghahramani, and D. M. Roy, “A study of the effect of JPG compression on adversarial images,” *arXiv*, 2016.
- [11] W. Xu, D. Evans, and Y. Qi, “Feature squeezing: Detecting adversarial examples in deep neural networks,” *arXiv preprint arXiv:1704.01155*, 2017.
- [12] H. Li, S. Shan, E. Wenger, J. Zhang, H. Zheng, and B. Y. Zhao, “Blacklight: Scalable defense for neural networks against Query-Based Black-Box attacks,” in *31st USENIX Security Symposium (USENIX Security 22)*, Boston, MA: USENIX Association, Aug. 2022, pp. 2117–2134, ISBN: 978-1-939133-31-1. [Online]. Available: <https://www.usenix.org/conference/usenixsecurity22/presentation/li-huiying>.
- [13] I. Goodfellow, P. McDaniel, and N. Papernot, “Making machine learning robust against adversarial inputs,” *Communications of the ACM*, pp. 56–66, 2018.
- [14] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” *arXiv preprint arXiv:1312.6199*, 2013.





- [15] G. Ortiz-Jiménez, A. Modas, S.-M. Moosavi-Dezfooli, and P. Frossard, “Optimism in the face of adversity: Understanding and improving deep learning through adversarial robustness,” *Proceedings of the IEEE*, pp. 635–659, 2021.
- [16] M. Alzantot, Y. Sharma, A. Elgohary, B.-J. Ho, M. Srivastava, and K.-W. Chang, “Generating natural language adversarial examples,” in *EMNLP*, 2018, pp. 2890–2896.
- [17] X. Dong, A. T. Luu, R. Ji, and H. Liu, “Towards robustness against natural language word substitutions,” in *ICRL*, 2021, pp. 1–14.
- [18] N. Carlini and D. Wagner, “Audio adversarial examples: Targeted attacks on speech-to-text,” in *S&P Workshop*, 2018, pp. 1–7.
- [19] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in *S&P*, 2017, pp. 39–57.
- [20] Y. Dong, F. Liao, T. Pang, H. Su, X. Hu, J. Li, and J. Zhu, “Boosting adversarial attacks with momentum,” in *CVPR*, 2018, pp. 9185–9193.
- [21] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in *ICLR*, 2015, pp. 1–11.
- [22] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” in *ICLR*, 2018, pp. 1–23.
- [23] M. Al-Rubaie and J. M. Chang, “Privacy-preserving machine learning: Threats and solutions,” *IEEE Security & Privacy*, pp. 49–58, 2019.
- [24] D. J. Miller, Z. Xiang, and G. Kesidis, “Adversarial learning targeting deep neural network classification: A comprehensive review of defenses against attacks,” *Proceedings of the IEEE*, pp. 402–433, 2020.
- [25] F. Croce and M. Hein, “Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks,” in *ICML*, 2020, pp. 2206–2216.
- [26] H. Lang, Z. Chao, and Z. Hongyang, “Self-adaptive training - beyond empirical risk minimization,” in *NeurIPS*, 2020, pp. 19 365–19 376.
- [27] J. Cohen, E. Rosenfeld, and Z. Kolter, “Certified adversarial robustness via randomized smoothing,” in *ICML*, 2019, pp. 1310–1320.
- [28] T. Pang, X. Yang, Y. Dong, K. Xu, H. Su, and J. Zhu, “Boosting adversarial training with hypersphere embedding,” in *NeurIPS*, 2020, pp. 7779–7792.
- [29] L. Rice, E. Wong, and J. Z. Kolter, “Overfitting in adversarially robust deep learning,” in *ICML*, 2020, pp. 8093–8104.
- [30] Y. Yang, G. Zhang, D. Katabi, and Z. Xu, “Me-net: Towards effective adversarial robustness with matrix estimation,” in *ICML*, 2019, pp. 7025–7034.
- [31] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, “Theoretically principled trade-off between robustness and accuracy,” in *ICML*, 2019, pp. 7472–7482.
- [32] J. Zhang, X. Xu, B. Han, G. Niu, L. Cui, M. Sugiyama, and M. Kankanhalli, “Attacks which do not kill training make adversarial learning stronger,” in *ICML*, 2020, pp. 11 278–11 287.
- [33] C. Laidlaw, S. Singla, and S. Feizi, “Perceptual adversarial robustness: Defense against unseen threat models,” in *ICLR*, 2021, pp. 1–25.
- [34] N. Das, M. Shanbhogue, S.-T. Chen, F. Hohman, S. Li, L. Chen, M. E. Kounavis, and D. H. Chau, “Shield: Fast, practical defense and vaccination for deep learning using jpeg compression,” in *SIGKDD*, 2018, pp. 196–204.





- [35] C. Guo, M. Rana, M. Cisse, and L. Van Der Maaten, “Countering adversarial images using input transformations,” in *ICLR*, 2017, pp. 1–12.
- [36] H. Salman, M. Sun, G. Yang, A. Kapoor, and J. Z. Kolter, “Denoised smoothing: A provable defense for pretrained classifiers,” in *NeurIPS*, 2020, pp. 21 945–21 957.
- [37] T. Pang, X. Yang, Y. Dong, H. Su, and J. Zhu, “Bag of tricks for adversarial training,” in *ICLR*, 2021, pp. 1–27.
- [38] A. Athalye, N. Carlini, and D. Wagner, “Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples,” in *ICML*, 2018, pp. 274–283.
- [39] A. Bietti, G. Mialon, D. Chen, and J. Mairal, “A kernel perspective for regularizing deep neural networks,” in *ICML*, 2019, pp. 664–674.
- [40] Y. Carmon, A. Raghunathan, L. Schmidt, P. Liang, and J. Duchi, “Unlabeled data improves adversarial robustness,” in *NeurIPS*, 2019, pp. 11 190–11 201.
- [41] M. Cissé, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier, “Parseval networks - improving robustness to adversarial examples,” in *ICML*, 2017, pp. 854–863.
- [42] Y. Wang, D. Zou, J. Yi, B. James, X. Ma, and Q. Gu, “Improving adversarial robustness requires revisiting misclassified examples,” in *ICLR*, 2020, pp. 1–14.
- [43] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, “Robustness may be at odds with accuracy,” in *ICLR*, 2019, pp. 1–23.
- [44] Y.-Y. Yang, C. Rashtchian, H. Zhang, R. Salakhutdinov, and K. Chaudhuri, “A closer look at accuracy vs. robustness,” in *NeurIPS*, 2020, pp. 8588–8601.
- [45] H. Kannan, A. Kurakin, and I. Goodfellow, “Adversarial logit pairing,” *arXiv*, 2018.
- [46] T. Ishida, I. Yamane, T. Sakai, G. Niu, and M. Sugiyama, “Do we need zero training loss after achieving zero training error?” In *ICML*, 2020, pp. 4604–4614.
- [47] T. DeVries and G. W. Taylor, “Improved regularization of convolutional neural networks with cutout,” *arXiv*, 2017.
- [48] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, “Random erasing data augmentation,” in *AAAI*, 2020, pp. 13 001–13 008.
- [49] D. Wu, S.-T. Xia, and Y. Wang, “Adversarial weight perturbation helps robust generalization,” in *NeurIPS*, 2020, pp. 2958–2969.
- [50] A. Krizhevsky, G. Hinton, *et al.*, “Learning multiple layers of features from tiny images,” Citeseer, Tech. Rep., 2009.
- [51] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, “Reading digits in natural images with unsupervised feature learning,” in *NeurIPS Workshop*, 2011, pp. 1–9.
- [52] Y. Le and X. Yang, “Tiny imagenet visual recognition challenge,” *CS 231N*, 2015.
- [53] H. Liu, Z. Zhong, N. Sebe, and S. Satoh, “Mitigating robust overfitting via self-residual-calibration regularization,” *Artificial Intelligence*, vol. 137, Article 103877, 2023.
- [54] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *ICML*, 2017, pp. 1321–1330.
- [55] M. P. Naeini, G. Cooper, and M. Hauskrecht, “Obtaining well calibrated probabilities using bayesian binning,” in *AAAI*, 2015, pp. 2901–2907.
- [56] K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” in *ECCV*, 2016, pp. 630–645.





- [57] S. Zagoruyko and N. Komodakis, “Wide residual networks,” in *BMVC*, 2016, pp. 87.1–87.12.
- [58] E. Wong, L. Rice, and J. Z. Kolter, “Fast is better than free: Revisiting adversarial training,” in *ICLR*, 2020, pp. 1–17.
- [59] A. Shafahi, M. Najibi, A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein, “Adversarial training for free!” In *NeurIPS*, 2019, pp. 1–12.
- [60] J.-B. Alayrac, J. Uesato, P.-S. Huang, A. Fawzi, R. Stanforth, and P. Kohli, “Are labels required for improving adversarial robustness?” In *NeurIPS*, 2019, pp. 12 192–12 202.
- [61] S.-A. Rebuffi, S. Gowal, D. A. Calian, F. Stimberg, O. Wiles, and T. Mann, “Fixing data augmentation to improve adversarial robustness,” *arXiv*, 2021.
- [62] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, “Learning transferable visual models from natural language supervision,” in *ICML*, 2021, pp. 8748–8763.
- [63] S. Fort, J. Ren, and B. Lakshminarayanan, “Exploring the limits of out-of-distribution detection,” in *NeurIPS*, 2021, pp. 7068–7081.
- [64] C. Xie, Y. Wu, L. v. d. Maaten, A. L. Yuille, and K. He, “Feature denoising for improving adversarial robustness,” in *CVPR*, 2019, pp. 501–509.
- [65] A. Goel, A. Agarwal, M. Vatsa, R. Singh, and N. K. Ratha, “Dndnet: Reconfiguring cnn for adversarial robustness,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 22–23.
- [66] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.
- [67] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” in *International Conference on Learning Representations*, 2018.
- [68] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in *Proceedings of the 2017 IEEE Symposium on Security and Privacy*, IEEE, 2017, pp. 39–57.
- [69] A. Mustafa, S. H. Khan, M. Hayat, R. Goecke, J. Shen, and L. Shao, “Deeply supervised discriminative learning for adversarial defense,” *IEEE transactions on pattern analysis and machine intelligence*, 2020.
- [70] V. Mygdalis and I. Pitas, “Hyperspherical class prototypes for adversarial robustness,” *Pattern Recognition*, p. 108 527, 2022.
- [71] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *CVPR*, 2016.
- [72] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, “Reading digits in natural images with unsupervised feature learning,” *In: Neural Information Processing Systems (NIPS) Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.
- [73] J. Welbl, A. Glaese, J. Uesato, S. Dathathri, J. Mellor, L. A. Hendricks, K. Anderson, P. Kohli, B. Coppin, and P.-S. Huang, “Challenges in detoxifying language models,” in *Findings of the Association for Computational Linguistics: EMNLP 2021*, Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2447–2469. DOI: 10.18653/v1/2021.findings-emnlp.210. [Online]. Available: <https://aclanthology.org/2021.findings-emnlp.210>.





- [74] M. Foley, A. Rawat, T. Lee, Y. Hou, G. Picco, and G. Zizzo, “Matching pairs: Attributing fine-tuned models to their pre-trained large language models,” *arXiv preprint arXiv:2306.09308*, 2023.
- [75] Z. C. Lipton, “The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery,” *Queue*, vol. 16, no. 3, pp. 31–57, 2018.
- [76] F. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning,” *arXiv preprint arXiv:1702.08608*, 2017.
- [77] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626. DOI: 10.1109/ICCV.2017.74.
- [78] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.
- [79] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *ICLR*, 2015.
- [80] S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.
- [81] R. Reimao and V. Tzerpos, “For: A dataset for synthetic speech detection,” in *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, IEEE, 2019, pp. 1–10.
- [82] A. Huang, “Similarity measures for text document clustering,” *Proceedings of the sixth New Zealand computer science research student conference (NZCSRSC2008)*, pp. 49–56, 2008.
- [83] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2016, pp. 2921–2929. DOI: 10.1109/CVPR.2016.319. [Online]. Available: <https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.319>.
- [84] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, “Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks,” in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 839–847. DOI: 10.1109/WACV.2018.00097.
- [85] S. Sattarzadeh, M. Sudhakar, K. N. Plataniotis, J. Jang, Y. Jeong, and H. Kim, “Integrated grad-cam: Sensitivity-aware visual explanation of deep convolutional networks via integrated gradient-based scoring,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 1775–1779. DOI: 10.1109/ICASSP39728.2021.9415064.
- [86] S. Desai and H. G. Ramaswamy, “Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization,” in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 972–980. DOI: 10.1109/WACV45572.2020.9093360.





- [87] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, “Score-cam: Score-weighted visual explanations for convolutional neural networks,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2020, pp. 111–119. DOI: 10.1109/CVPRW50498.2020.00020. [Online]. Available: <https://doi.ieeecomputersociety.org/10.1109/CVPRW50498.2020.00020>.
- [88] V. Petsiuk, A. Das, and K. Saenko, “RISE: randomized input sampling for explanation of black-box models,” in *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, BMVA Press, 2018, p. 151. [Online]. Available: <http://bmvc2018.org/contents/papers/1064.pdf>.
- [89] S. Sattarzadeh, M. Sudhakar, A. Lem, S. Mehryar, K. N. Plataniotis, J. Jang, H. Kim, Y. Jeong, S. Lee, and K. Bae, “Explaining convolutional neural networks through attribution-based input sampling and block-wise feature aggregation,” in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, AAAI Press, 2021, pp. 11 639–11 647. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/17384>.
- [90] M. Sudhakar, S. Sattarzadeh, K. N. Plataniotis, J. Jang, Y. Jeong, and H. Kim, “Ada-sise: Adaptive semantic input sampling for efficient explanation of convolutional neural networks,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 1715–1719. DOI: 10.1109/ICASSP39728.2021.9414942.
- [91] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, “Sanity checks for saliency maps,” in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS’18, Montréal, Canada: Curran Associates Inc., 2018, pp. 9525–9536.
- [92] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: <http://arxiv.org/abs/1409.0473>.
- [93] S. Liu and W. Deng, “Very deep convolutional neural network based image classification using small training sample size,” in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, 2015, pp. 730–734. DOI: 10.1109/ACPR.2015.7486599.
- [94] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.
- [95] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.
- [96] I. Gkartzonika, N. Gkalelis, and V. Mezaris, “Learning visual explanations for dcnn-based image classifiers using an attention mechanism,” in *Computer Vision – ECCV 2022 Workshops*, L. Karlinsky, T. Michaeli, and K. Nishino, Eds., Cham: Springer Nature Switzerland, 2023, pp. 396–411, ISBN: 978-3-031-25085-9.
- [97] Y. Shen, C. Yang, X. Tang, and B. Zhou, “InterFaceGAN: Interpreting the disentangled face representation learned by GANs,” *TPAMI*, 2020.





- [98] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, “GANSpace: Discovering interpretable GAN controls,” in *NeurIPS*, 2020.
- [99] Y. Shen and B. Zhou, “Closed-form factorization of latent semantics in GANs,” in *CVPR*, 2021.
- [100] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, “Encoding in style: A StyleGAN encoder for image-to-image translation,” in *CVPR*, 2021.
- [101] X. Hou, X. Zhang, H. Liang, L. Shen, Z. Lai, and J. Wan, “Guidedstyle: Attribute knowledge guided style manipulation for semantic face editing,” *Neural Networks*, vol. 145, pp. 209–220, 2022.
- [102] X. Yao, A. Newson, Y. Gousseau, and P. Hellier, “A latent transformer for disentangled face editing in images and videos,” in *ICCV*, 2021.
- [103] A. Nguyen, J. Yosinski, and J. Clune, “Deep neural networks are easily fooled: High confidence predictions for unrecognizable images,” in *CVPR*, 2015.
- [104] S.-H. Sun, *Multi-digit MNIST for few-shot learning*, GitHub repository, 2019. [Online]. Available: <https://github.com/shaohua0116/MultiDigitMNIST>.
- [105] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” in *ICLR*, 2014, pp. 1–10.
- [106] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of StyleGAN,” in *CVPR*, 2020.
- [107] C. Villani, *Optimal Transport: Old and New*. Springer Berlin Heidelberg, 2008.
- [108] J. Feydy, T. Séjourné, F.-X. Vialard, S.-i. Amari, A. Trounev, and G. Peyré, “Interpolating between optimal transport and MMD using sinkhorn divergences,” in *Intl. Conf. Artificial Intelligence and Statistics (AISTATS)*, 2019, pp. 2681–2690.
- [109] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in *ICCV*, 2015.
- [110] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.
- [111] P. Doubinsky, N. Audebert, M. Crucianu, and H. Le Borgne, “Wasserstein loss for semantic editing in the latent space of gans,” in *International Conference on Content-Based Multimedia Indexing*, Orléans, France, Sep. 2023.
- [112] C. Olah, A. Mordvintsev, and L. Schubert, “Feature visualization,” *Distill*, vol. 2, no. 11, e7, 2017.
- [113] C. Olah, N. Cammarata, L. Schubert, G. Goh, M. Petrov, and S. Carter, “Zoom in: An introduction to circuits,” *Distill*, vol. 5, no. 3, e00024–001, 2020.
- [114] A. Nguyen, J. Yosinski, and J. Clune, “Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks,” *arXiv preprint arXiv:1602.03616*, 2016.
- [115] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.
- [116] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.





- [117] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I 13*, Springer, 2014, pp. 818–833.
- [118] L. McInnes, J. Healy, and J. Melville, “Umap: Uniform manifold approximation and projection for dimension reduction,” *arXiv preprint arXiv:1802.03426*, 2018.
- [119] L. O’Mahony, V. Andrearczyk, H. Müller, and M. Graziani, “Disentangling neuron representations with concept vectors,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3769–3774.
- [120] R. Caruana, “Multitask learning,” *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.
- [121] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, “Domain-adversarial training of neural networks,” *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.
- [122] A. Kendall, Y. Gal, and R. Cipolla, “Multi-task learning using uncertainty to weigh losses for scene geometry and semantics,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7482–7491.
- [123] M. Graziani, V. Andrearczyk, and H. Müller, “Regression concept vectors for bidirectional explanations in histopathology,” in *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, Springer, 2018, pp. 124–132.
- [124] M. Graziani, S. Otalora, S. Marchand-Maillet, H. Muller, and V. Andrearczyk, “Learning interpretable microscopic features of tumor by multi-task adversarial cnns improves generalization,” *Machine Learning for Biomedical Imaging Journal*, 2020.
- [125] D. Omeiza, H. Webb, M. Jirotko, and L. Kunze, “Explanations in autonomous driving: A survey,” *IEEE Trans. on Intelligent Transportation Systems*, 2021.
- [126] É. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord, “Explainability of vision-based autonomous driving systems: Review and challenges,” *arXiv preprint arXiv:2101.05307*, 2021.
- [127] L. Cultrera, L. Seidenari, F. Becattini, P. Pala, and A. Del Bimbo, “Explaining autonomous driving by learning end-to-end visual attention,” in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 340–341.
- [128] B. Y. Lim and A. K. Dey, “Assessing demand for intelligibility in context-aware applications,” in *Proc. of the 11th international conference on Ubiquitous computing*, 2009, pp. 195–204.
- [129] R. Girshick, “Fast r-cnn,” in *Proc. of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.
- [130] M. Jaderberg, K. Simonyan, A. Zisserman, *et al.*, “Spatial transformer networks,” *Advances in neural information processing systems*, vol. 28, 2015.
- [131] L. Cultrera, F. Becattini, L. Seidenari, P. Pala, and A. Del Bimbo, “Explaining autonomous driving with visual attention and end-to-end trainable region proposals,” *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–13, 2023.
- [132] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Routledge, 1984.
- [133] R. J. Quinlan, *C4.5: programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993, ISBN: 1-55860-238-0. [Online]. Available: <http://portal.acm.org/citation.cfm?id=152181>.
- [134] R. J. Quinlan, “Induction of decision trees,” *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.





- [135] G. Lopardo, D. Garreau, F. Precioso, and G. Ottosson, “Smace: A new method for the interpretability of composite decision systems,” in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2022, pp. 325–339.
- [136] M. T. Ribeiro, S. Singh, and C. Guestrin, “Anchors: High-precision model-agnostic explanations,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.
- [137] G. Lopardo, F. Precioso, and D. Garreau, “A sea of words: An in-depth analysis of anchors for text data,” in *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, F. Ruiz, J. Dy, and J.-W. van de Meent, Eds., ser. Proceedings of Machine Learning Research, vol. 206, PMLR, Apr. 2023, pp. 4848–4879.
- [138] E. Kaufmann and S. Kalyanakrishnan, “Information complexity in bandit subset selection,” in *Conference on Learning Theory*, PMLR, 2013, pp. 228–251.
- [139] R. Manhaeve, S. Dumancic, A. Kimmig, T. Demeester, and L. De Raedt, “Deepproblog: Neural probabilistic logic programming,” *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [140] P. Barbiero, G. Ciravegna, F. Giannini, M. Espinosa Zarlenga, L. C. Magister, A. Tonda, P. Lio, F. Precioso, M. Jamnik, and G. Marra, “Interpretable neural-symbolic concept reasoning,” in *Proceedings of the 40th International Conference on Machine Learning*, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., ser. Proceedings of Machine Learning Research, vol. 202, PMLR, Jul. 2023, pp. 1801–1825.
- [141] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, “Concept bottleneck models,” in *International Conference on Machine Learning*, PMLR, 2020, pp. 5338–5348.
- [142] M. W. Shen, “Trust in AI: Interpretability is not necessary or sufficient, while black-box interaction is necessary and sufficient,” *arXiv preprint arXiv:2202.05302*, 2022.
- [143] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [144] P. Voigt and A. Von dem Bussche, “The eu general data protection regulation (gdpr),” *A Practical Guide, 1st Ed.*, Cham: Springer International Publishing, vol. 10, no. 3152676, pp. 10–5555, 2017.
- [145] L. Graves, V. Nagisetty, and V. Ganesh, “Amnesiac machine learning,” *arXiv preprint arXiv:2010.10981*, 2020.
- [146] L. Bourtole, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, and N. Papernot, “Machine unlearning,” in *2021 IEEE Symposium on Security and Privacy (SP)*, IEEE, 2021, pp. 141–159.
- [147] C. Guo, T. Goldstein, A. Hannun, and L. Van Der Maaten, “Certified data removal from machine learning models,” *arXiv preprint arXiv:1911.03030*, 2019.
- [148] A. Sekhari, J. Acharya, G. Kamath, and A. T. Suresh, “Remember what you want to forget: Algorithms for machine unlearning,” *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [149] G. Liu, X. Ma, Y. Yang, C. Wang, and J. Liu, “Federaser: Enabling efficient client-level data removal from federated learning models,” in *2021 IEEE/ACM 29th International Symposium on Quality of Service (IWQOS)*, IEEE, 2021, pp. 1–10.





- [150] C. Wu, S. Zhu, and P. Mitra, “Federated unlearning with knowledge distillation,” *arXiv preprint arXiv:2201.09441*, 2022.
- [151] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. DOI: 10.1109/5.726791.
- [152] G. Cohen, S. Afshar, J. Tapson, and A. Van Schaik, “Emnist: Extending mnist to handwritten letters,” in *2017 international joint conference on neural networks (IJCNN)*, IEEE, 2017, pp. 2921–2926.
- [153] A. Warnecke, L. Pirch, C. Wressnegger, and K. Rieck, “Machine unlearning of features and labels,” *arXiv preprint arXiv:2108.11577*, 2021.
- [154] T. Gu, B. Dolan-Gavitt, and S. Garg, “Badnets: Identifying vulnerabilities in the machine learning model supply chain,” *arXiv preprint arXiv:1708.06733*, 2017.
- [155] A. Halimi, S. Kadhe, A. Rawat, and N. Baracaldo, “Federated unlearning: How to efficiently erase a client in fl?” *arXiv preprint arXiv:2207.05521*, 2022.
- [156] N. Holohan, S. Braghin, P. Mac Aonghusa, and K. Levacher, “Diffprivlib: The IBM differential privacy library,” *ArXiv e-prints*, vol. 1907.02444 [cs.CR], Jul. 2019.
- [157] N. Holohan, *Random number generators and seeding for differential privacy*, Jul. 2023.
- [158] S. Fletcher and M. Z. Islam, “Differentially private random decision forests using smooth sensitivity,” *CoRR*, vol. abs/1606.03572, 2016. arXiv: 1606.03572. [Online]. Available: <http://arxiv.org/abs/1606.03572>.
- [159] I. Mironov, “On significance of the least significant bits for differential privacy,” in *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, ser. CCS ’12, Raleigh, North Carolina, USA: Association for Computing Machinery, 2012, pp. 650–661, ISBN: 9781450316514. DOI: 10.1145/2382196.2382264.
- [160] C. L. Canonne, G. Kamath, and T. Steinke, “The discrete Gaussian for differential privacy,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 15 676–15 688. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/file/b53b3a3d6ab90ce0268229151c9bde11-Paper.pdf>.
- [161] N. Holohan and S. Braghin, “Secure random sampling in differential privacy,” in *Computer Security - ESORICS 2021 - 26th European Symposium on Research in Computer Security, Darmstadt, Germany, October 4-8, 2021, Proceedings, Part II*, E. Bertino, H. Shulman, and M. Waidner, Eds., ser. Lecture Notes in Computer Science, vol. 12973, Springer, 2021, pp. 523–542. DOI: 10.1007/978-3-030-88428-4_26. [Online]. Available: https://doi.org/10.1007/978-3-030-88428-4_26.
- [162] H. Striegel, R. Ulrich, and P. Simon, “Randomized response estimates for doping and illicit drug use in elite athletes,” *Drug and alcohol dependence*, vol. 106, no. 2-3, pp. 230–232, 2010.
- [163] J. J. Donovan, S. A. Dwight, and G. M. Hurtz, “An assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the randomized response technique,” *Human Performance*, vol. 16, no. 1, pp. 81–106, 2003.
- [164] Y. Hu, I. Ameer, X. Zuo, X. Peng, Y. Zhou, Z. Li, Y. Li, J. Li, X. Jiang, and H. Xu, “Zero-shot clinical entity recognition using ChatGPT,” *arXiv preprint arXiv:2303.16416*, 2023.
- [165] N. L. of Medicine, *NLM-Scrubber*, <https://lhncbc.nlm.nih.gov/scrubber/index.html>.





- [166] Microsoft, *Presidio - Data Protection and De-identification SDK*, <https://github.com/microsoft/presidio/>.
- [167] T. Ahmed, M. M. A. Aziz, and N. Mohammed, “De-identification of electronic health record using neural network,” *Scientific reports*, vol. 10, no. 1, pp. 1–11, 2020.
- [168] K. Murugadoss, A. Rajasekharan, B. Malin, V. Agarwal, S. Bade, J. R. Anderson, J. L. Ross, W. A. Faubion, J. D. Haramka, V. Soundararajan, and S. Ardhanari, “Building a Best-in-Class Automated De-identification Tool for Electronic Health Records Through Ensemble Learning,” *medRxiv*, 2021.
- [169] D. Garat and D. Wonsever, “Automatic curation of court documents: Anonymizing personal data,” *Information*, vol. 13, no. 1, 2022.
- [170] S. Braghin, J. H. Bettencourt-Silva, K. Levacher, and S. Antonatos, “An extensible de-identification framework for privacy protection of unstructured health information: Creating sustainable privacy infrastructures,” in *MEDINFO 2019: Health and Wellbeing e-Networks for All*, IOS Press, 2019, pp. 1140–1144.
- [171] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, “spaCy: Industrial-strength Natural Language Processing in Python,” *Zenodo, Honolulu, HI, USA*, 2020. DOI: 10.5281/zenodo.1212303.
- [172] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, “FLAIR: An easy-to-use framework for state-of-the-art NLP,” in *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 2019, pp. 54–59.
- [173] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, “Stanza: A Python natural language processing toolkit for many human languages,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020.
- [174] Y. Ma, A. Wang, and N. Okazaki, “Dreem: Guiding attention with evidence for improving document-level relation extraction,” in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, ser. EACL, Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, (to appear).
- [175] I. Pilán, P. Lison, L. Øvrelid, A. Papadopoulou, D. Sánchez, and M. Batet, “The text anonymization benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization,” *Computational Linguistics*, vol. 48, no. 4, pp. 1053–1101, 2022.
- [176] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-Efficient Learning of Deep Networks from Decentralized Data,” in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, A. Singh and J. Zhu, Eds., ser. Proceedings of Machine Learning Research, vol. 54, Fort Lauderdale, FL, USA: PMLR, Apr. 2017, pp. 1273–1282. [Online]. Available: <http://proceedings.mlr.press/v54/mcmahan17a.html>.
- [177] V. Duddu, A. Boutet, and V. Shejwalkar, “Quantifying privacy leakage in graph embedding,” *arXiv preprint arXiv:2010.00906*, 2020.
- [178] X. He, J. Jia, M. Backes, N. Z. Gong, and Y. Zhang, “Stealing links from graph neural networks,” in *30th {USENIX} Security Symposium ({USENIX} Security 21)*, 2021.
- [179] I. E. Olatunji, W. Nejdil, and M. Khosla, “Membership inference attack on graph neural networks,” *arXiv preprint arXiv:2101.06570*, 2021.





- [180] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, “Deep learning with differential privacy,” in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.
- [181] I. Mironov, “Rényi differential privacy,” in *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, IEEE, 2017, pp. 263–275.
- [182] A. L. Traud, P. J. Mucha, and M. A. Porter, “Social structure of facebook networks,” *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 16, pp. 4165–4180, 2012.
- [183] W. L. Hamilton, R. Ying, and J. Leskovec, “Inductive representation learning on large graphs,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 1025–1035.
- [184] W.-L. Chiang, X. Liu, S. Si, Y. Li, S. Bengio, and C.-J. Hsieh, “Cluster-gcn: An efficient algorithm for training deep and large graph convolutional networks,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 257–266.
- [185] F. Wu, Y. Long, C. Zhang, and B. Li, “Linkteller: Recovering private edges from graph neural networks via influence analysis,” *arXiv preprint arXiv:2108.06504*, 2021.
- [186] K. Xu, C. Li, Y. Tian, T. Sonobe, K.-i. Kawarabayashi, and S. Jegelka, “Representation learning on graphs with jumping knowledge networks,” in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., ser. Proceedings of Machine Learning Research, vol. 80, Stockholmsmässan, Stockholm Sweden: PMLR, Jul. 2018, pp. 5453–5462.
- [187] F. Giunchiglia, I. Bison, M. Busso, R. Chenu-Abente, M. Rodas, M. Zeni, C. Gunel, G. Veltri, A. De Götzen, P. Kun, *et al.*, “A worldwide diversity pilot on daily routines and social practices (2020),” 2021.
- [188] L. Meegahapola, W. Droz, P. Kun, A. de Götzen, C. Nutakki, S. Diwakar, S. R. Correa, D. Song, H. Xu, M. Bidoglia, *et al.*, “Generalization and personalization of mobile sensing-based mood inference models: An analysis of college students in eight countries,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 4, pp. 1–32, 2023.
- [189] P. Chriskos, R. Zhelev, V. Mygdalis, and I. Pitas, “Quality preserving face de-identification against deep cnns,” in *Proceedings of the 2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, IEEE, 2018, pp. 1–6.
- [190] P. Chriskos, J. Munro, V. Mygdalis, and I. Pitas, “Face detection hindering,” in *Proceedings of the 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, IEEE, 2017, pp. 403–407.
- [191] V. Mygdalis, A. Tefas, and I. Pitas, “K-anonymity inspired adversarial attack and multiple one-class classification defense,” *Neural Networks*, vol. 124, pp. 296–307, 2020.
- [192] Y. Liu, W. Zhang, and N. Yu, “Protecting privacy in shared photos via adversarial examples based stealth,” *Security and Communication Networks*, vol. 2017, 2017.
- [193] S. Joon Oh, M. Fritz, and B. Schiele, “Adversarial image perturbation for privacy protection—a game theory perspective,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1482–1491.
- [194] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, “Universal adversarial perturbations,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1765–1773.



- [195] K. T. Co, L. Muñoz-González, L. Kanthan, B. Glocker, and E. C. Lupu, “Universal adversarial robustness of texture and shape-biased models,” in *Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2021, pp. 799–803.
- [196] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.
- [197] Y. Pang, J. Lin, T. Qin, and Z. Chen, “Image-to-image translation: Methods and applications,” *IEEE Transactions on Multimedia*, vol. 24, pp. 3859–3881, 2022. DOI: 10.1109/TMM.2021.3109419.
- [198] T. F. van der Ouderaa and D. E. Worrall, “Reversible gans for memory-efficient image-to-image translation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4720–4728.
- [199] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004. DOI: 10.1109/TIP.2003.819861.
- [200] A. Shafahi, M. Najibi, Z. Xu, J. Dickerson, L. S. Davis, and T. Goldstein, “Universal adversarial training,” *Association for the Advancement of Artificial Intelligence*, vol. 34, no. 04, pp. 5636–5643, Apr. 2020, ISSN: 2374-3468. DOI: 10.1609/aaai.v34i04.6017.
- [201] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford, “Datasheets for datasets,” *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021.
- [202] M. Hanley, A. Khandelwal, H. Averbuch-Elor, N. Snaveley, and H. Nissenbaum, “An ethical highlighter for people-centric dataset creation,” *arXiv preprint arXiv:2011.13583*, 2020.
- [203] D. Alonso del Barrio and D. Gatica-Perez, “How Did Europe’s Press Cover Covid-19 Vaccination News? A Five-Country Analysis,” in *Proc. ACM International Workshop on Multimedia AI against Disinformation*, Jun. 2022.
- [204] K. L. Boyd, “Datasheets for datasets help ml engineers notice and understand ethical issues in training data,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–27, 2021.
- [205] C. Garbin, P. Rajpurkar, J. Irvin, M. P. Lungren, and O. Marques, “Structured dataset documentation: A datasheet for chexpert,” *arXiv preprint arXiv:2105.03020*, 2021.
- [206] M. Miceli, T. Yang, L. Naudts, M. Schuessler, D. Serbanescu, and A. Hanna, “Documenting computer vision datasets: An invitation to reflexive data practices,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 161–172.
- [207] I. Seck, K. Dahmane, P. Duthon, and G. Loosli, “Baselines and a datasheet for the cerema awp dataset,” *arXiv preprint arXiv:1806.04016*, 2018.
- [208] P. Schramowski, C. Tauchmann, and K. Kersting, “Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content?” In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1350–1361.
- [209] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on bias and fairness in machine learning,” *ACM computing surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.



- [210] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, *et al.*, “Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias,” *arXiv preprint arXiv:1810.01943*, 2018.
- [211] Y. Xu, P. Terhörst, K. Raja, and M. Pedersen, “A comprehensive analysis of ai biases in deepfake detection with massively annotated databases,” *arXiv preprint arXiv:2208.05845*, 2022.
- [212] G. Ciravegna, F. Giannini, M. Gori, M. Maggini, and S. Melacci, “Human-Driven FOL Explanations of Deep Learning,” in *IJCAI-PRICAI 2020 - 29th International Joint Conference on Artificial Intelligence and the 17th Pacific Rim International Conference on Artificial Intelligence*, Yokohama, Japan: International Joint Conferences on Artificial Intelligence Organization, Jul. 2020, pp. 2234–2240. DOI: 10.24963/ijcai.2020/309. [Online]. Available: <https://hal.science/hal-03045280>.
- [213] G. Ciravegna, F. Precioso, A. Betti, K. Mottin, and G. Marco, “Knowledge-driven active learning,” in *Proceeding of the ECML-PKDD 2023: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2023*.
- [214] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. Chen, “Improved techniques for training gans,” in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29, Curran Associates, Inc., 2016. [Online]. Available: <https://proceedings.neurips.cc/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Paper.pdf>.
- [215] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium,” in *Advances in Neural Information Processing Systems*, Dec. 2017, pp. 6629–6640.
- [216] S. Ravuri and O. Vinyals, “Classification accuracy score for conditional generative models,” in *Advances in Neural Information Processing Systems*, 2019.
- [217] T. Hinz, S. Heinrich, and S. Wermter, “Semantic object accuracy for generative text-to-image synthesis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1552–1565, Mar. 2022. DOI: 10.1109/tpami.2020.3021209. [Online]. Available: <https://doi.org/10.1109/tpami.2020.3021209>.
- [218] J. Cho, A. Zala, and M. Bansal, “Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers,” *arXiv 2202.04053*, 2022.
- [219] Y. Zhang, L. Jiang, G. Turk, and D. Yang, “Auditing gender presentation differences in text-to-image models,” *arXiv 2302.03675*, 2023. arXiv: 2302.03675 [cs.CV].
- [220] L. O. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “Catboost: Unbiased boosting with categorical features,” in *NeurIPS*, 2018.
- [221] D. Alonso del Barrio and D. Gatica-Perez, “Examining European Press Coverage of the No-Vax Movement: An NLP Framework,” in *in Proc. ACM International Workshop on Multimedia AI against Disinformation*, Jun. 2023.
- [222] D. Alonso del Barrio and D. Gatica-Perez, “Framing the News: From Human Perception to Large Language Model Inferences,” in *in Proc. ACM International Conference on Multimedia Retrieval*, Jun. 2023.





A Appendix

A.1 Datasheet for the dataset on European press coverage of Covid-19 vaccination news

Main contributor: M. Guido (AI4Media Junior Fellow hosted at IDIAP.) Additional contributions and updates by D. Gatica-Perez and D. Alonso del Barrio. Version: July 2023.

In the datasheet, the questions are reproduced verbatim from the original datasheet framework proposed by Gebru et al. [201], in order to maximize compliance with the framework.

A.1.0.1 Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

There is a body of research studying the spread of (dis)information about Covid-19. In this regard, many researchers pointed out the need of understanding the full information ecosystem. This holistic understanding needs to include the examination of existing high-quality media outlets, which played a key role during the pandemic period. The understanding of how European media organizations covered aspects of the pandemic, such as Covid-19 vaccination, has emerged as a research gap so far. This dataset represents a step towards filling this research gap by including over 50,000 articles on Covid-19 vaccination from 19 newspapers, 5 European countries and 4 languages over a period of 22 months. This represents a unique resource to study the media coverage on Covid-19 vaccinations. The dataset also includes a translated version with all the content in English to facilitate comparisons across countries. The dataset was presented and used in [203], and developed as part of the work in WP6, task T6.5 of AI4Media (Perception of Hyper-Local News).

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by the Social Computing group of the Idiap Research Institute.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

The dataset was designed, obtained, and curated in the context of the AI4Media project (European Commission Grant 951911, under the H2020 Programme).

Any other comments? No.

A.1.0.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Each instance represents a newspaper article. All instances are of the same type.

How many instances are there in total (of each type, if appropriate)?

There are 51,320 instances.





Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset contains a sample of newspaper articles about Covid-19 vaccination over a period of 22 months. The sample is considered to be representative of the press coverage in Europe covering six newspapers from France; two from Italy; six from Spain; three from Switzerland; and two from the United Kingdom. The sample shows an imbalanced number of newspapers per country, the details can be found in Table 1 of [203].

What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.

The data can be considered raw data and consists of the title, subheadline as well as text in a newspaper article in original language and translated to English. Additionally, an instance contains the authors, date, a corresponding link, and the newspaper’s name and country. The only feature that was added to the dataset is the number of words in the original article as well as the translated version.

Is there a label or target associated with each instance? If so, please provide a description. Currently, there is no label or target associated with the instances.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

In the subheadline and author columns, some instances do not have any values. For such cases, the word ‘error’ was set to signal the missing value.

Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)? If so, please describe how these relationships are made explicit.

There are no known relations between the instances. However, some of the instances share the same publisher or author. Whenever an article is from the same newspaper or author, it can be made apparent through the corresponding column.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

There are no recommended data splits for the dataset.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

All articles were translated from their original language to English using the DeepL software. It cannot be ruled out that along the translations some minor errors might have occurred. The possible translation errors are unknown.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.





The dataset is self-contained in the sense that all articles in the dataset are included. For completeness purposes, the dataset also contains the links to the websites they were retrieved from. There is no guarantee that these will exist and remain constant over time. The instances themselves provide the information about the content of these links at the time the dataset was created.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

The dataset contains no confidential data as all data stems from public communication of newspapers.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

The dataset contains data about the news coverage of Covid 19 vaccination. This topic can generally be considered as controversial and polarizing, even though newspapers mainly strive for objectivity in this debate. Furthermore, the Covid-19 pandemic and vaccinations might be topics that might be upsetting or sensitive to some readers.

Any other comments? No.

A.1.0.3 Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The data was directly observable. The only exception is the number of words that was inferred from the text column and the English translation of the articles by the software DeepL.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

As a first step, over 30 European newspapers spanning five countries were contacted, requesting authorization to extract and analyze online articles discussing issues related to Covid-19 vaccination. The authorization of 19 of them was obtained. This includes six newspapers from France; two from Italy; six from Spain; three from Switzerland; and two from the United Kingdom. After receiving written email authorization from each newspaper, the articles were extracted using Selenium and BeautifulSoup, which are scraping techniques that allow to extract the text of the articles. For each article the headline, subheadline (if available), main text, authors, date of publication, and link were extracted.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The dataset is a sample of newspaper articles in the European press about Covid-19 vaccination. While the intention of the dataset was to map the European press coverage by a representative sample of geographic areas, the dataset only consists of five countries and 19 newspapers. As stated earlier, a large number of newspapers was originally contacted (based on their reputation), and only a fraction of them responded positively; thus the sampling is a form of convenience sample at the





newspaper level. In other words, only a few European countries are covered, only a few newspapers per country, and some countries have fewer instances than others.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

A researcher from the Social Computing group at Idiap collected and curated the data.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The data was collected for a 22 month period from 01.01.2020 to 31.10.2021. The creation timeframe of the news articles matches the timeframe of the data collection.

Were any ethical review processes conducted (e.g., by an institutional review board)?

If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

The PI at IDIAP made an initial submission to the institute's internal Data and Research Ethics Committee (DREC), for the different IDIAP research activities in the AI4Media project. An update to this submission was then submitted to describe the news dataset collection process in detail.

Any other comments?

No.

A.1.0.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning or labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

As part of preprocessing, expressions that usually appear when web scraping such as `\xa0`, `\xad`, `\u200b`, and any line breaks (`\n`) were removed from the articles.

Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.

The raw data is stored, but in the context of the intended uses of the dataset, it does not add value.

Is the software used to preprocess/clean or label the instances available? If so, please provide a link or other access point.

While a script running a sequence of commands is not available, all software used to process the data is open source and has been specified above.

Any other comments?

No.





A.1.0.5 Uses

Has the dataset been used for any tasks already? If so, please provide a description.

A first analysis was conducted on how the European press treated Covid-19 vaccination-related issues. Experiments with NLP tools such as named entity recognition, topic modeling, and sentiment analysis were performed on the translated dataset. This is reported in [203]. A second analysis involved a subset of the original dataset, involving 1786 headlines of No-Vax movement press articles, to understand how the European press treated the No-Vax movement. This is reported in [221]. Finally, a third analysis involved the same 1786 headline subset, for human and machine labeling of journalistic frames. This is reported in [222].

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

The papers that used the dataset can be found at Idiap publication website: <https://publications.idiap.ch/index.php>

What (other) tasks could the dataset be used for?

The dataset offers the foundation for research on the media coverage of the Covid-19 vaccinations in European newspapers. While the associated papers investigated the application of natural language processing tools to examine how the European press treated the Covid-19 vaccination issue, a mapping into the context of other aspects and public discussions of the pandemic remains interesting to address. For example, one could study how the debate of Covid-19 vaccinations in media coverage occurred in parallel with international, national, or regional regulations and recommendations. Additionally, social media data could offer a different perspective of opinions on the topic, provide an interesting contrast to professional media coverage, and give insights into the evolution of public agenda and opinions alongside the media agenda. Furthermore, the dataset offers a foundation to expand the above mentioned investigations to more geographical areas (e.g., multiple continents).

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned or labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

Overall, there is little risk for harm: the data was public by construction and was published by professional newspapers following their own editorial guidelines and safeguards.

Are there tasks for which the dataset should not be used? If so, please provide a description.

The dataset covers a specific geographical area. Practitioners are not advised to use the dataset to draw conclusions for countries or geographic areas other than Europe.

Any other comments? No.

A.1.0.6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.





The dataset cannot be distributed to third parties. Newspapers were contacted to request the permission to distribute the data, but such authorization was not obtained for the full dataset.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?

See previous point.

When will the dataset be distributed?

See first point in this section.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

See first point in this section.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

See first point in this section.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

See first point in this section.

Any other comments?

No.

A.1.0.7 Maintenance

Who will be supporting/hosting and maintaining the dataset?

David Alonso del Barrio: ddbarrio@idiap.ch

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Daniel Gatica-Perez: gatica@idiap.ch

Is there an erratum? If so, please provide a link or other access point.

There is no erratum.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

The dataset covers the period from January 2020 to October 2021. There are no plans to update the dataset.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

As the dataset is a snapshot of a fixed period of time and will not be updated, the creators of the dataset do not expect obsolescence of the dataset for the few next years.





If others want to extend/augment or build on and contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

There is no mechanism for others to contribute to the dataset.

Any other comments?

No.

A.2 Abstracts from the invited talks at the First Nice Workshop on Interpretability (NWI)

The following are the abstracts for the invited talks from the First Nice Workshop on Interpretability (NWI) outlined in Section 4.11.

Jenny Benois-Pineau (Université de Bordeaux): *FEM and MLFEM post-hoc explainers for CNNs and their evaluation with reference-based and no-reference quality metrics*

In this talk we will present two methods for explanation of trained CNN models we recently developed: FEM and MLFEM. These methods are based on the evaluation of the strength of features in the deepest convolution layer (FEM) or several convolutional layers of a CNN performing image classification task. Without losing generality, the methods can be applied on any data which are being classified with a CNN. Furthermore, we propose evaluation methods for the quality of obtained explanation maps. We consider two general approaches for quality metrics design: reference-based and no-reference. In the case of image classification, as a reference we consider Gaze Fixation Density maps built upon gaze fixations of observers which have participated in psycho-visual experiment with the goal of the recognition of a visual scene. As quality metrics, are used those proposed in vision research for comparison of saliency maps. The no-reference method, by D. Alvarez Melis and T. Jakkola is based on the Lipschitz constant computation. We study the behavior of this metric as a function of strength of degradations induced on regional images. Furthermore, we explore the correlation of reference-based and no-reference metric. Our experimental studies show that FEM and MLFEM methods outperform reference explainers, such as GradCam in the sense of both reference-based and no-reference metrics.

Joao Marques-Silva (IRIT CNRS ANITI): *Logic-Based Explainability in Machine Learning*

The forecast applications of machine learning (ML) in high-risk and safety-critical applications hinge on systems that are deemed robust in their operation, and that can be understood about their decisions, and so trusted. Most ML models are neither robust nor understandable. This talk gives a broad overview of ongoing efforts in applying logic-enabled automated reasoning tools for explaining black-box ML models. The talk details the computation of rigorous explanations for the predictions made by black-box models, and illustrates how these serve to assess the quality of widely used heuristic explanation approaches. Finally, the talk briefly overviews a number of emerging topics of research in logic-enabled explainability.

Vasileios Mezaris (ITI - CERTH): *Explaining the decisions of image/video classifiers*

We will start by discussing the main classes of explainability approaches for image and video classifiers. Then, we will focus on two distinct problems: learning how to derive explanations for the decisions of a legacy (trained) image classifier, and designing a classifier for video event recognition that can also deliver explanations for its decisions. Technical details of our proposed solutions to





these two problems will be presented. Besides quantitative results concerning the goodness of the derived explanations, qualitative examples will also be discussed in order to provide insight on the reasons behind classification errors, including possible dataset biases affecting the trained classifiers.

Martin Pawelczyk (University of Tübingen): *On the Trade-Off between Actionable Explanations and the Right to be Forgotten*

As machine learning (ML) models are increasingly being deployed in high-stakes applications, policymakers have suggested tighter data protection regulations (e.g., GDPR, CCPA). One key principle is the “right to be forgotten” which gives users the right to have their data deleted. Another key principle is the right to an actionable explanation, also known as algorithmic recourse, allowing users to reverse unfavorable decisions. To date, it is unknown whether these two principles can be operationalized simultaneously. Therefore, we introduce and study the problem of recourse invalidation in the context of data deletion requests. More specifically, we theoretically and empirically analyze the behavior of popular state-of-the-art algorithms and demonstrate that the recourses generated by these algorithms are likely to be invalidated if a small number of data deletion requests (e.g., 1 or 2) warrant updates of the predictive model. For the setting of linear models and overparameterized neural networks – studied through the lens of neural tangent kernels (NTKs) – we suggest a framework to identify a minimal subset of critical training points which, when removed, maximize the fraction of invalidated recourses. Using our framework, we empirically show that the removal of as little as 2 data instances from the training set can invalidate up to 95 percent of all recourses output by popular state-of-the-art algorithms. Thus, our work raises fundamental questions about the compatibility of “the right to an actionable explanation” in the context of the “right to be forgotten” while also providing constructive insights on the determining factors of recourse robustness.

Tristan Gomez (LS2N): *Metrics for saliency maps faithfulness evaluation: an application to embryo stage identification*

Due to the black-box nature of deep learning models, there is a recent development of solutions for visual explanations of CNNs. To evaluate the faithfulness of the explanations, various metrics were introduced. First, we critically analyze the Deletion Area Under Curve (DAUC) and Insertion Area Under Curve (IAUC) metrics proposed by Petsiuk et al. (2018). These metrics were designed to evaluate the faithfulness of saliency maps generated by generic methods such as Grad-CAM or RISE. We show that DAUC and IAUC suffer from two issues: (1) they generate out-of-distribution samples and (2) they ignore the saliency scores. To complement DAUC/IAUC, we propose new metrics that quantify the sparsity and the calibration of explanation methods, two previously unstudied properties. Next, we study the behavior of faithfulness metrics applied to the problem of embryo stage identification. We benchmark attention models and post-hoc methods and further show empirically that (1) the metrics produce low overall agreement on the model ranking and (2) depending on the metric approach, either post-hoc methods or attention models are favored. We conclude with general remarks about the difficulty of defining faithfulness and the necessity of understanding its relationship with the type of approach that is favored.

Sebastian Bordt (University of Tübingen): *From Shapley Values to Generalized Additive Models and back*

In explainable machine learning, local post-hoc explanation algorithms and inherently interpretable models are often seen as competing approaches. In this work, offer a novel perspective on Shapley Values, a prominent post-hoc explanation technique, and show that it is strongly connected with Glassbox-GAMs, a popular class of interpretable models. We introduce \mathcal{S} -Shapley Values, a natural extension of Shapley Values that explain individual predictions with interaction terms up to





order k . As k increases, the k -Shapley Values converge towards the Shapley-GAM, a uniquely determined decomposition of the original function. From the Shapley-GAM, we can compute Shapley Values of arbitrary order, which gives precise insights into the limitations of these explanations. We then show that Shapley Values recover generalized additive models of order k , assuming that we allow for interaction terms up to order k in the explanations. This implies that the original Shapley Values recover Glassbox-GAMs. At the technical end, we show that there is a one-to-one correspondence between different ways to choose the value function and different functional decompositions of the original function. This provides a novel perspective on the question of how to choose the value function. We also present an empirical analysis of the degree of variable interaction that is present in various standard classifiers, and discuss the implications of our results for algorithmic explanations. A python package to compute k -Shapley Values and replicate the results in this paper is available [here](#).

Hugo Sénétaire (DTU): *Castig explainability as statistical inference*

A wide variety of model explanation approaches have been proposed recently, all guided by very different rationales and heuristics. We take a new route and cast interpretability as a statistical inference problem. A general deep probabilistic model is designed to produce interpretable predictions. The model's parameters can be learned via maximum likelihood, and the method can be adapted to any predictor network architecture and any type of prediction problem. Our method is a case of amortized interpretability models, where a neural network is used as a selector to allow for fast interpretation at inference time. Several popular interpretability methods are shown to be cases of regularised maximum likelihood for our general model. We propose new datasets with ground truth selection which allow for evaluating the features' importance map. Using these datasets, we show experimentally that using multiple imputations provides a more reasonable interpretation.

Gianluigi Lopardo (3IA-UCA): *A Sea of Words: An In-Depth Analysis of Anchors for Text Data*

Anchors (Ribeiro et al., 2018) is a post-hoc, rule-based interpretability method. For text data, it proposes to explain a decision by highlighting a small set of words (an anchor) such that the model to explain has similar outputs when they are present in a document. We present the first theoretical analysis of Anchors, considering that the search for the best anchor is exhaustive. After formalizing the algorithm for text classification, we present explicit results on different classes of models when the preprocessing step is TF-IDF vectorization, including elementary if-then rules and linear classifiers. We then leverage this analysis to gain insights on the behavior of Anchors for any differentiable classifiers. For neural networks, we empirically show that the words corresponding to the highest partial derivatives of the model with respect to the input, reweighted by the inverse document frequencies, are selected by Anchors.

Gabriele Ciravegna (3IA-UCA): *Entropy-Based Logic Explanations of Neural Networks*

Explainable artificial intelligence has rapidly emerged since lawmakers have started requiring interpretable models for safety-critical domains. Concept-based neural networks have arisen as explainable-by-design methods as they leverage human-understandable symbols (i.e. concepts) to predict class memberships. However, most of these approaches focus on the identification of the most relevant concepts but do not provide concise, formal explanations of how such concepts are leveraged by the classifier to make predictions. In this paper, we propose a novel end-to-end differentiable approach enabling the extraction of logic explanations from neural networks using the formalism of First-Order Logic. The method relies on an entropy-based criterion which automatically identifies the most relevant concepts. We consider four different case studies to demonstrate that: (i) this





entropy-based criterion enables the distillation of concise logic explanations in safety-critical domains from clinical data to computer vision; (ii) the proposed approach outperforms state-of-the-art white-box models in terms of classification accuracy and matches black box performances.

Jean-Michel Loubes (Université Toulouse Paul Sabatier): *Explainability of a Model under stress*

We propose to study another type of explanation : the response of an algorithm when confronted to constraints on the test distribution. In order to avoid outliers we consider distributions that satisfy a stress constraint while being as close as possible to the original distribution. We define entropic projections under constraints that satisfy such conditions and thus provide some theoretical guarantees for such models. The method is analysed here.

Yann Chevaleyre (Paris Dauphine): *Learning interpretable scoring rules*

Interpretability is a quite old topic in machine learning, which recently gained a lot of traction. In this talk, we will discuss about the need for interpretability in machine learning and what is meant by interpretability. Then, we will present the problem of learning interpretable scoring rules, and how this problem can be relaxed into a standard convex optimization problem. Finally, we will show some applications.

Alexandre Benoit (Université Savoie Mont Blanc): *Explainable AI for Earth Observation*

Earth Observation (EO), as for other domains, is subject to impressive advances thanks to the availability of abundant data, modern AI methods and more specifically deep neural networks. However, most of the available EO data is generally unlabelled, generally illustrates very local context with specific orientation, climate and so on such that the generalization behaviours of machine learning models can be limited. In addition, the implication of model inference applied to EO may lead to costly decisions such as infrastructure design or modification or crop yield. Then automatic decisions should be justified or explained. However, in the era of deep learning-based models, opening those black boxes is a challenge in itself. In this talk, we will present a variety of activities related to EO and explainable AI at LISTIC Lab. A focus on contributions related to explainable AI relying on 3 complementary directions : black box explanation, explanation by model design and redescription mining. These contributions highlight the interest of explanation methods combinations in order to present more concise and focused explanation to the human experts.

Salim Amoukou (Université Paris Saclay): *Consistent Sufficient Explanations and Minimal Local Rules for explaining regression and classification models*

To explain the decision of any model, we extend the notion of probabilistic Sufficient Explanations (P-SE). For each instance, this approach selects the minimal subset of features that is sufficient to yield the same prediction with high probability, while removing other features. The crux of P-SE is to compute the conditional probability of maintaining the same prediction. Therefore, we introduce an accurate and fast estimator of this probability via random Forests for any data (X, Y) and show its efficiency through a theoretical analysis of its consistency. As a consequence, we extend the PSE to regression problems. In addition, we deal with non-binary features, without learning the distribution of X nor having the model for making predictions. Finally, we introduce local rule-based explanations for regression/classification based on the P-SE and compare our approaches w.r.t other explainable AI methods. These methods are publicly available as a Python package.

Giorgio Visani (University of Bologna): *Inspecting Stability and Reliability of Explanations*

Explanations of automated decision systems are extremely important in highly regulated domains.





One of the most well-known solutions to obtain model explanations is the LIME technique. In the talk, we will discuss the technique in general, with a special focus on its reliability. Ad-hoc Stability Indices are going to be presented as a tool to discern whether the explanations can be trusted. Building on the Stability Indices, the OptiLIME policy focuses on obtaining stable and reliable LIME explanations. Stability Indices and the OptiLIME policy represent an important step toward LIME compliance, from a regulatory point of view.

Hidde Fokkema (Korteweg-de Vries Institute): *Attribution-based Explanations that Provide Recourse Cannot be Robust*

When automated machine learning decisions lead to undesirable outcomes for users, recourse methods from explainable machine learning can inform users how to change the decisions. It is often argued that such explanations should be robust to small measurement errors in the users' features. We show that, unfortunately, this type of robustness is impossible to achieve for any method that also gives useful explanations whenever possible. We further discuss possible ways to work around our impossibility result, for instance by allowing the output to consist of sets with multiple attributions. Finally, we strengthen our impossibility result for the restricted case where users are only able to change a single attribute of x , by providing an exact characterization of the functions f to which impossibility applies.

Mara Graziani (IBM Research): *Reliable AI in healthcare: from model validation to hypothesis generation*

Deep learning models in healthcare are yielding exceptional results for the characterization of cancer biomarkers in imaging and molecular data, at times even exceeding human performance. However, assessing the reliability of the predicted outcomes is still a challenge, with predictions lacking robustness to covariate shifts. Moreover, it is still unclear what informative patterns lead to the high performance gains given by the deep models. In this talk, I discuss how model interpretability and reliable AI development can address the tasks of model validation. After briefly introducing the terminology related to reliable AI, I will provide examples on semi-transparent model designs that can be used to introduce desired inductive biases during model training. Finally, I will look at the future potential of interpretability developments for accelerating scientific discovery. In particular, I will discuss the potential of attention mechanisms for scientific hypotheses generation in histopathology.

Pietro Barbiero (Cambridge University): *Concept Embedding Models: Beyond the Accuracy-Explainability Trade-Off*

Human trust in deep neural networks is currently an open problem as their decision process is opaque. Current methods such as Concept Bottleneck Models make the models more interpretable at the cost of decreasing accuracy (or vice versa). To address this issue, we propose Concept Embedding Models, a novel family of concept bottleneck models which goes beyond the current accuracy-vs-interpretability trade-off by learning interpretable high-dimensional concept representations.

