Universitas Mercatorum

Università telematica delle Camere di Commercio Italiane

PhD Programme in

Big Data and Artificial Intelligence

Curriculum "Big data management for the digital transition"

38th Cycle

PhD Dissertation in

AI-generated Deepfakes: Detection and Bias Analysis

Dr. Vittorio Stile

Std N° DT00100006

Programme Coordinator

Prof. Barbara Martini

Supervisor

Prof. Roberto Caldelli

Co-Supervisor

Dr. Elena Santi

Academic Year 2024 / 2025

Finanziato dall'Unione europea NextGenerationEU | Ministero dell'Università e della Ricerca | Italiadomani Piano Nazionale di Ripresa e Resilienza | Universitas Mercatorum Università telematica delle Camere di Commercio Italiane

A chi amo e a chi mi ama.
Alle persone di grande cultura,
che sono fonte di ispirazione per il mondo.
Alle persone dal limpido intelletto,
che non hanno avuto le condizioni per coltivarlo.
Alle persone dalla profonda empatia,
che portano il peso della leggerezza altrui.
A chi ogni giorno lotta per il proprio posto,
ma viene messo in ombra da chi ha percorsi agevolati.
A chi si è smarrito lungo la strada della vita,
perché la dignità non si misura con la forza.
A chi mi ha sostenuto in questo lungo cammino.
Questa pagina è dedicata a tutti voi.

Candidate's declaration

I hereby declare that this thesis submitted to obtain the academic degree of Philosophiæ Doctor (PhD) in *Big Data and Artificial Intelligence* is my own unaided work, that I have not used other than the sources indicated, and that all direct and indirect sources are acknowledged as references.

I hereby declare no financial conflicts of interest related to the concepts, technologies, or entities discussed. Affiliations with organizations in enterprise automation and AI did not influence the study design, analysis, or interpretation. All data were collected and analysed independently, following academic ethical standards. Institutional disclosures were made, and confidentiality was ensured by anonymizing collaborators, datasets, and proprietary tools in compliance with GDPR.

During the preparation of this work, were used GPT-5.1, GPT-4o, o3, DeepL Translator and Mate Translate in order to perform translation, grammar and spelling check. After using these tools and services, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

Parts of this dissertation have been published in international journals and/or conference articles (see list of the author's publications in the Appendix "*Scientific outputs arising from the PhD*" at the end of the thesis). All figures and tables are original or used with proper permission.

Napoli, April 28, 2026

Vittorio Stile

Abstract

DeepFakes, synthetic manipulations of faces produced with generative artificial intelligence, threaten the authenticity of content and expose detectors to the tough task of dealing with a multiplicity of content and great variability, compression levels and deepfake generation pipelines. Against this backdrop, this doctoral thesis investigates how misclassifications in DeepFake detection relate to high-level facial attributes, and how this knowledge can guide more robust and interpretable detectors. The work proceeds in two stages. In a first analysis, a frame-level classifier distinguishes manipulated from authentic content and its errors are examined post hoc. Videos from the dataset are preprocessed by detecting and cropping faces with a cascade classifier. The dataset is enriched through a facial-attribute labeling pipeline that starts from a small manually annotated seed and expands on the whole dataset with per-attribute semi-supervised classifier to derive labels such as gender, hair color, hair length, ear visibility, and ethnicity. Subsequently, was created a DeepFake classifier that delivers achieves good results on the primary subject in each video. Attribute-wise error analysis (including label-level metrics and statistical dependence measures) reveals systematic patterns: in particular, ear visibility and hair length emerge as influential contextual factors that can affect decisions. In an extension of the analysis, insights are stress-tested via controlled exclusion experiments that remove one or more values of a given attribute during training, and the related models are evaluated on the complete test set. The results show that some characteristics impact model performance and decision behavior; for example, removing training exposure to certain visibility conditions degrades the detector's ability at test time. These findings motivate data curation that balances key attribute conditions, applies targeted augmentations, and assesses the influence of attributes on the final outcome. Overall, the thesis contributes a scalable semi-supervised pipeline for attribute labeling and practical guidelines for bias-aware training. The study advances interpretability and tackles the field's central generalization problem by showing that explicit attribute information can guide data curation and training so that models become more reliable to real-world variability.

Keywords: DeepFake Detection, Facial Attributes, Semi-Supervised Labeling, Attribute-Aware Training, Bias Analysis, Generalization.

Abstract in lingua italiana

I DeepFakes, manipolazioni di volti prodotti con intelligenza artificiale generativa, minacciano l'autenticità dei contenuti ed espongono i *detector* al difficile compito di gestire una molteplicità di contenuti e una grande variabilità, livelli di compressione e *pipeline* di generazione di deepfake. In questo contesto, questa tesi di dottorato indaga come gli attributi facciali influenzino le classificazioni errate nel rilevamento DeepFake e come queste informazioni possano aiutare a migliorare i modelli di *detection*. Il lavoro procede in due fasi. In una prima analisi, un classificatore distingue contenuti manipolati da contenuti autentici e i relativi errori vengono esaminati a posteriori. I video del dataset sono preelaborati estraendo i singoli frame, rilevando e ritagliando i volti tramite uno specifico classificatore. Il test set è arricchito mediante l'etichettatura di attributi facciali quali genere, colore dei capelli, lunghezza dei capelli, visibilità delle orecchie ed etnia, si parte da un piccolo nucleo annotato manualmente e poi tramite un classificatore semi-supervisionato si etichetta il resto. Successivamente viene fatto il setup di un classificatore di DeepFake e viene fatta un'analisi degli errori per attributo, con metriche per etichetta e misure di dipendenza statistica fra queste, rilevando pattern sistematici, in particolare la visibilità delle orecchie e la lunghezza dei capelli emergono come fattori influenti sulla qualità del modello. In un'estensione dell'analisi, gli *insight* emersi vengono verificati con esperimenti di esclusione controllata, che rimuovono in addestramento uno o più attributi, e i modelli così prodotti valutati su un test set senza esclusioni. I risultati mostrano che la rimozione in training di specifiche caratteristiche incide in modo sostanziale sulle prestazioni del modello, degradando la capacità del *detector* in fase di test. Queste evidenze motivano una gestione dei dati che bilanci alcuni attributi chiave, *augmentations* mirate e stime del peso degli attributi sul risultato finale. Nel complesso, la tesi fornisce una pipeline semi-supervisionata scalabile per l'etichettatura degli attributi e linee guida pratiche per un'addestamento *bias*-consapevole. Lo studio migliora l'interpretabilità e affronta il problema centrale del settore, ovvero la generalizzazione, dimostrando che avere informazioni sugli attributi facciali, può aiutare a creare modelli più affidabili e rispondenti alla variabilità del mondo reale.

Keywords: DeepFake Detection, Facial Attributes, Semi-Supervised Labeling, Attribute-Aware Training, Bias Analysis, Generalization.

Acknowledgements

This dissertation was carried out within the PhD Programme in *Big Data and Artificial Intelligence* at Universitas Mercatorum.

The author, Vittorio Stile, gratefully acknowledges support from a three–year PhD fellowship awarded under **D.M. n. 352/2022** and co-funded with resources from the European Union – NextGeneration EU within Italy's Piano Nazionale di Ripresa e Resilienza (PNRR), Mission 4 "*Education and Research*", Component 2 "*From Research to Enterprise*", Investment 3.3 "*Introduction of innovative PhD programmes that meet the needs of enterprises and promote the recruitment of researchers by enterprises*" – CUP D83C22001880003.

The fellowship was co–funded by PricewaterhouseCoopers Business Services S.r.l. and formally assigned by Rectoral Decree n. 112/2022 of 9 December 2022.

The author wishes to thank **Prof. Roberto Caldelli** for his supervision and guidance throughout the doctoral studies. Sincere thanks are also due to **PricewaterhouseCoopers Business Services S.r.l.** for the collaboration, hosting, and resources that enabled parts of this work. The author is grateful to **Prof. Inmaculada Medina-Bulo**, from the **University of Cádiz (UCA)**, for her supervision during the period abroad and the **UCASE research group** for hosting the research stay abroad and for the scientific discussions that contributed to the development of the thesis. Unless otherwise stated, all opinions and any errors remain the author's responsibility.

Executive Summary

DeepFakes, synthetic manipulations of faces produced with generative artificial intelligence, threaten the authenticity of content and expose detectors to the tough task of dealing with a multiplicity of content and great variability, compression levels and deepfake generation pipelines. Against this backdrop, this doctoral thesis investigates how errors made by DeepFake detectors relate to high-level facial attributes, and how this knowledge can be used to design more robust and interpretable systems. The work is positioned in computer vision for face forensics and addresses a central challenge in the field, namely the lack of generalization across manipulation families, compression levels, and capture conditions. The thesis contributes a scalable semi-supervised pipeline for attribute labeling, an empirical analysis that links false positives and false negatives to specific appearance factors, and a set of attribute-aware training strategies that improve reliability without sacrificing transparency. While the manuscript surveys the state of the art and reports several educational and industrial activities, the scientific core is concentrated in Chapter 4, *Analysis of DF Detection through Attribute Labeling*, and Chapter 5, *Attribute-Aware Training Strategies*. This summary emphasizes those two chapters while outlining the end-to-end approach and the resulting guidelines.

Introduction: Problem framing and research questions. DeepFake generation has progressed quickly, which reduces the saliency of traditional visual artifacts and stresses detectors that were tuned for early datasets. As new pipelines and editing chains proliferate, model failures concentrate on specific slices of the data, for example under certain occlusions or appearance conditions. The thesis therefore asks three questions. **RQ1** concerns interpretability, namely whether facial attribute labeling can make model behavior more transparent. **RQ2** concerns reliability, namely which attributes correlate the most with false positives and false negatives at video level. **RQ3** concerns design, namely whether detector training and evaluation can be reorganized around attribute information to improve robustness.
Datasets, preprocessing, and detection baselines. Experiments are carried out on the FaceForensics++ corpus, focusing on pristine YouTube videos and their DeepFakes

counterparts at the `c40` compression tier. Videos are decoded, frames are uniformly sampled with a fixed skip factor to reduce redundancy, and faces are detected with a classical cascade. Tightly cropped face patches at `(224 x 224)` feed an image-level classifier. The reference detector is intentionally simple in order to isolate the effect of attributes. A VGG16 backbone with ImageNet initialization is used as a frozen feature extractor, followed by a compact classification head. Training, validation, and test splits are balanced by class and kept disjoint at video level. Video decisions are obtained by aggregating frame scores and by selecting the principal face track when multiple faces are present. With this configuration the system reaches high accuracy at video level, with precision and recall balanced across classes, and with stable behavior under the adopted compression and aggregation protocol. These baselines establish a transparent testbed where attribute-conditioned effects can be measured.

State of the Art. The thesis is grounded in a structured literature review that systematizes methods, datasets, and evaluation protocols in face deepfake detection. The review follows a documented search-and-screen protocol across Scopus, Institute of Electrical and Electronics Engineers (IEEE) Xplore, ScienceDirect and targeted snowballing, with inclusion criteria privileging reproducible evaluations on public dataset such as FaceForensics++, Celeb–DF, and DeepFake Detection Challenge Dataset (DFDC). The surveyed methods cover frame-level artifact detectors, face-centric physiological and geometric cues, and spatio–temporal pipelines, complemented by multimodal audio–visual approaches and fairness analyses. Surveys and benchmarks are used to map strengths and failure modes under compression, editing chains, and cross–dataset shift, which motivates the thesis focus on interpretability and robustness. This evidence base exposes two persistent gaps that directly shape Chapters 4 and 5: limited understanding of how high–level attributes relate to misclassifications, and a lack of training workflows that explicitly control attribute exposure. The review therefore provides both the methodological landscape for the baseline detector and the conceptual rationale for a semi–supervised attribute labeling pipeline and attribute–aware training strategies.

Industry activities in PwC The industrial period at PricewaterhouseCoopers Business Services Italia S.r.l. translated the research agenda into production–minded prototypes and evaluation practices. A staged sequence of projects consolidated the tooling and informed design choices later adopted in the thesis: a transfer–learning binary image classifier established a lightweight baseline for data hygiene and augmentation; a compact MNIST workflow standardized end–to–end training and reporting; a classical Olivetti face identification exercise validated reproducible splits and error inspection; finally, a video–level deepfake detector combined face cropping, pre–trained convolutional backbones, and simple temporal aggregation. These projects were developed with Python and common DL frameworks, presented in internal seminars, and used to stress operational constraints such as storage, compute budgets, compression tolerance, and threshold calibration. The lessons learned informed Chapter 4 by fixing a robust preprocessing and evaluation protocol for the attribute–enriched analysis, and guided Chapter 5 by prioritizing bias–aware curation, targeted augmentations, and attribute–conditioned training and reporting that can be adopted in enterprise settings without excessive complexity.

Analysis of DF Detection through Attribute Labeling. Chapter 4 introduces a semi-supervised pipeline that scales facial attribute annotation with limited human effort. A seed set of fifty real videos is manually labeled for gender, hair color, hair length, ear visibility, and ethnicity, together with a set of boolean context flags. The manipulated companions inherit the same labels since identity swap keeps stable appearance traits. A per-attribute classifier, based on a compact convolutional backbone, is then trained on the seed, used to pseudo-label the remaining videos, and iteratively refined by including high-confidence predictions. This procedure yields a labeled testbed for downstream analysis without incurring prohibitive annotation costs. The chapter documents safeguards adopted to control error propagation, such as confidence thresholds for acceptance, manual spot checks on the tail of the confidence distribution, and the exclusion of attributes that collapse to a single value in the seed. The labeled corpus enables a systematic link between detector outputs and appearance factors. The analysis proceeds in three layers. First, descriptive statistics

quantify error rates per attribute value, for example the share of false positives for short hair and for long hair. Second, dependence measures are computed, including group-wise differences in precision and recall, effect sizes on confusion-matrix entries, and non-parametric association coefficients for categorical variables. Third, simple interpretable models are fitted on the test predictions, for example logistic regressions that use attribute dummies to explain the probability of a misclassification, together with partial-dependence visualizations. Across these views, two factors emerge consistently. **Ear visibility** correlates with detector behavior, with elevated error rates when ears are occluded by hair or accessories. **Hair length** also influences outcomes, particularly in profiles and semi-profiles where the face contour and side hair interact with compression. These patterns persist after controlling for class balance, video aggregation, and the presence of multiple faces, which indicates that they are not artifacts of sampling alone. The chapter further reports calibration slices by attribute, showing that score distributions shift across groups. In particular, the same confidence threshold yields different precision–recall trade-offs when ears are not visible, which suggests group-aware thresholding or improved calibration as practical remedies. The multi-face setting is addressed explicitly. Since the dataset includes crowd scenes and interviews with more than one person in frame, the analysis compares three aggregation rules, namely majority voting across faces, selection of the largest face, and selection of the most persistent track over time. The last option, which approximates the main subject, reduces spurious errors introduced by bystanders, and clarifies that the attribute effects highlighted above are not driven by incidental faces. Taken together, the chapter provides an empirical map of where the baseline detector breaks, and it does so in terms that a human analyst can verify on representative clips.

Attribute-Aware Training Strategies. Building on the evidence gathered in Chapter 4, Chapter 5 investigates how facial attributes can be used not only for post hoc analysis but also to probe the behaviour of a DeepFake detector under controlled distribution shifts. The chapter introduces an attribute-aware pipeline in which a `VGG16`-based classifier is trained multiple times on *FaceForensics++*, each time

excluding one attribute value from the training data and then evaluating all models on a common, complete test set. This controlled exclusion design, combined with subgroup metrics and minimum support thresholds, provides an exploratory view of how the absence of specific attribute configurations affects accuracy, AUC and the balance between TPR and TNR. The results suggest a moderate but recurrent effect for `hair_length` and a more pronounced sensitivity to `is_ears_visible`, with the removal of visible-ear cases in training associated with a higher rate of FAKE → REAL errors. The chapter concludes with cautious operational recommendations: curating training data to ensure explicit coverage of critical attributes, monitoring performance by subgroup, and considering attribute-aware calibration or thresholding in deployment. These findings do not provide definitive bias guarantees, but they indicate that simple attribute-aware training and evaluation strategies may help to organise robustness checks and to support more transparent assessment of DeepFake detectors.

Conclusions and Future Work The concluding chapter revisits the three research questions through the lens of the empirical results and frames the contribution of the thesis as primarily methodological and exploratory. The work combines a standard `VGG16`-based DeepFake detector on *FaceForensics++* with a semi-supervised `ResNet18` attribute labeller, and uses the resulting annotations to study misclassification patterns and simple attribute-aware training variants. The analysis suggests that certain appearance factors, in particular hair length and ear visibility, are associated with variations in error rates and may therefore offer useful context for post hoc monitoring and bias-aware inspection, although no causal conclusions are drawn. The chapter also highlights the main limitations of the study, including the focus on a single dataset and compression level, a limited and imbalanced attribute set, reliance on univariate statistics, and the use of a deliberately simple architecture. Within these constraints, the thesis contributions can be seen as: *(i)* a reproducible semi-supervised pipeline for facial attribute labeling on DeepFake benchmarks; *(ii)* an empirical indication that attribute-conditioned error analysis may help organise the evaluation of DeepFake detectors; and *(iii)* a set of preliminary ideas

for attribute-aware training and calibration. Possible extensions include applying the same protocol to more diverse datasets and richer attribute taxonomies, adopting multivariate analytical tools, and exploring the integration of attribute information into data curation, sampling and threshold selection to support more robust and fairness-aware detection in future work.

Keywords: DeepFake Detection, Facial Attributes, Semi-Supervised Labeling, Attribute-Aware Training, Bias Analysis, Generalization.

Contents

Acronyms

AI Artificial Intelligence. 63, 127–129, 131, 132, 136, 141, 142, 144, 145, 149

AI Act European Union Artificial Intelligence Act. 131

AIDEA Accademia Italiana di Economia Aziendale. 133

API Application Programming Interface. 129

AUC *Area Under the ROC Curve*: aggregate measure of a classifier's performance across all thresholds, equal to the probability that a positive sample receives a higher score than a negative one, ranging in [0,1]. 73, 109, 110, 112, 114, 117, 119

AUROC Area Under the Receiver-Operating-Characteristic curve. 144

BI Business Intelligence. 131

BIM Building Information Modeling. 127–129

C2PA Coalition for Content Provenance and Authenticity. 10

CAMEL Capital Adequacy, Asset quality, Management, Earnings, Liquidity. 144

CCII Italian Code of Business Crisis and Insolvency. 145

CEH Cognitive Enterprise Hub. 132

cGAN conditional Generative Adversarial Network. 129

CIDE Conference on Creativity and Innovation in Digital Economy. 128, 138, 139, 148

CNN Convolutional Neural Network. 66, 81, 82, 90

CRF Constant Rate Factor (H.264 quality scale). 85

CV Computer Vision. 129

DAI Distributed Artificial Intelligence. 141

DFDC DeepFake Detection Challenge Dataset. xii

DGA Data Governance Act. 145

DL Subfield of machine learning that uses multi layer neural networks to learn hierarchical representations from data. Typical architectures include convolutional, recurrent, and transformer models.. 86, 129

E Expected Frequence. 104

ECAI European Conference on Artificial Intelligence. 135

edu4AI Workshop on Education for Artificial Intelligence. 135

EDUCON IEEE Global Engineering Education Conference. 136

ETL Extract, Transform, Load. 129

EU European Union. 131, 145

FL Federated Learning. 142

FN False Negative. 96

FOMM Self-supervised video reenactment framework that transfers pose and expressions from a driving frame or video to a source image while preserving identity. Motion is represented by a small set of learned keypoints and their local first-order transformations (Jacobians). A dense motion network fuses

these into a dense motion field and an occlusion map. A generator, commonly a U-Net, warps and refines the source using these signals to synthesize the target frame. 14

FP False Positive. 96

GDPR General Data Protection Regulation. 131, 142, 145

HR Human Resources. 145

IDS Information Delivery Specification. 129

IEEE Institute of Electrical and Electronics Engineers. xii

IFC Industry Foundation Classes. 127, 129

IHSI International Conference on Human Intelligent Systems Integration. 141

IoT Internet of Things. 127, 129, 136

ISF Ingegneria Specialistica Fontanella. 128

ISM International Conference on Industry of the Future and Smart Manufacturing. 145

IT Information Technology. 132

itAIS Italian Chapter of AIS. 131, 132, 142

KPI Key Performance Indicator. 129

LLM Large Language Model. 129

LSTM Long Short-Term Memory. xxx, 72, 74, 76

MI Mutual Information. 103–105, 107

ML Machine Learning. 135

MNIST Modified National Institute of Standards and Technology. 63, 64, 66

NLP Natural Language Processing. 144

O Observed Frequence. 104

OIN Ordine degli Ingegneri della Provincia di Napoli. 128

PhD Philosophiæ Doctor. iii, ix, 3, 19, 78, 127, 145, 148

PI Physical Internet. 136, 138, 139, 141

PNRR Piano Nazionale di Ripresa e Resilienza. ix

PWC PricewaterhouseCoopers Business Services Italia S.r.l.. 63, 64, 66, 67, 69–72

QTO Quantity Take-Off. 129

RESER Rethinking Services for Society 5.0. 143

RL Branch of machine learning in which an agent learns a policy by interacting with an environment to maximize expected cumulative reward, often modeled as a Markov decision process and solved with dynamic programming or function approximation. 136

ROI Return on Investment. 145

SHAP SHapley Additive exPlanations. 145

SIEM Security platform that centralizes log collection and normalization, correlates events from multiple sources, detects threats, raises alerts and reports, and supports forensic investigations and compliance, often integrated with Security Orchestration, Automation, and Response (SOAR) and Unified Endpoint Management (UEM) for automation and response. xxvi, xxvii, 136

SME Medium-Sized Enterprise. 132, 144, 145

SOAR Security platform that orchestrates tools and workflows, automates incident-response playbooks, and manages cases to reduce mean time to detect and respond, typically integrated with Security Information and Event Management (SIEM) for alert ingestion and enrichment. xxvi, xxvii

TAM Technology Acceptance Model. 129

TN True Negative. 96

TNR *True Negative Rate* (specificity): proportion of actual negatives correctly identified. Formula: TNR = $\frac{TN}{TN+FP}$. 112, 117

TP True Positive. 96

TPR *True Positive Rate* (sensitivity, *recall*): proportion of actual positives correctly identified. Formula: TPR = $\frac{TP}{TP+FN}$. 112, 117

UCASE Software Engineering Research Group at the University of Cádiz. 78

UEM Centralized management of endpoints such as laptops, mobiles, IoT and servers, covering configuration, compliance, patching, app distribution and remote actions, often connected to SIEM and SOAR to enforce policies and isolate compromised devices. xxvi

UTAUT Unified Theory of Acceptance and Use of Technology. 129

XFL Explainable Federated Learning. 146, 148, 149

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Context and threat

The rapid progress of generative models for human faces has lowered the cost of producing visually convincing manipulations, which complicates the reliability of video communication, news ecosystems, and evidence collection. Face swapping tools replace the target face with a source identity across all frames while preserving the target video's pose, illumination, and motion; in effect, it transfers identity. Facial reenactment tools drive a target face's expressions and lip articulation from an external driver (video, audio, or motion-capture); in effect, it transfers expression and motion, not identity. These tools spread through academic prototypes and community software, this democratization accelerates the definition of various real-world scenarios in which detectors must operate, including diverse capture conditions, camera viewpoints, motion patterns, codecs, and compression settings. This thesis focuses on the forensic side, that is the automatic recognition of manipulated face content in images and videos, with particular attention to the robustness of detectors when data, manipulations, and protocols change.

1.1.1 Targets and uses of deepfakes

To explain why detection matters, it is helpful to consider who is targeted and how manipulated content is used. Across the public internet and in illicit markets,

deepfake videos circulate on social media and video-sharing platforms, within digital advertising and blogs, and at times even through official news broadcasts, as well as on the dark web. For analytic clarity, they can be grouped into four, partly overlapping, categories:

1. *Satirical*. Parody or humour, usually signalled to the audience, often circulated for entertainment or commentary. These cases may tolerate visible artifacts, yet rapid sharing can still mislead out of context.

2. *Deceptive*. Content crafted to misinform or to impersonate individuals in news, politics, finance, or private communications. Here the intent is to pass as genuine, which raises the bar for realism and undermines trust.

3. *Pornographic*. Non-consensual sexual content that targets private individuals or public figures. This class poses severe harms for privacy, dignity and safety, and often motivates takedown and forensic investigation.

4. *Technology demonstration*. Research or community showcases that illustrate synthesis capabilities. These clips usually declare their artificial nature, but they still inform adversarial evolution and detector stress tests.

These classes differ in intent, dissemination channels and legal exposure, yet they share technical traits relevant for forensics, such as compression patterns, blending traces, and temporal inconsistencies. The mix of targets also introduces potential bias, since demographic and contextual imbalances in publicly shared material can shape both training data and evaluation. This taxonomy motivates robust, interpretable and dataset-aware detectors, and it connects directly to the research aims and questions that follow.

1.2 Research aims

The research aims of the PhD programme are threefold. First, to map families of deepfake–detection methods that are most promising for generalization and interpretability, including spectral fingerprints at frame level, physiological or

geometric indicators at face level, and spatio–temporal models at video level. Second, to study how performance and failure modes vary across datasets, manipulation techniques, and evaluation pipelines, with reference to widely used dataset. Third, to analyze the role of high–level facial attributes in the detector's errors, connecting the findings to bias–aware training and to the design of a ready-to-use detection pipeline.

1.3 Research Questions and Intended Contribution

Modern deepfake generation techniques are rapidly closing the gap with human perception, which makes the distinction between authentic and manipulated faces increasingly difficult. This situation raises practical risks for communication, trust, and evidence, and it introduces potential biases linked to visual characteristics that may systematically affect detector outputs.

1.3.1 Initial framing at project start.

At the beginning of the doctoral work the inquiry was intentionally broad and guided by two questions aimed at positioning the problem within human–AI collaboration and computer vision:

> *How can artificial intelligence assist human intelligence in discerning true from false?*
>
> *Within computer vision, how does artificial intelligence recognise deepfakes of people that could be mistaken for real individuals?*

1.3.2 Refined research questions.

As the project progressed and the methodological direction became clearer, the questions were refined to align with the activities carried out and the empirical findings:

RQ1. *How can facial attribute labeling improve the interpretability of deepfake detection models?*

RQ2. *Which attributes are most correlated with false positives and false negatives at the video level?*

RQ3. *Based on the attribute analysis, can we design a workflow that improves the performance of current deepfake detection models?*

1.3.3 Intended contribution.

The thesis contributes an attribute–aware perspective on deepfake detection by: (i) introducing a semi–supervised pipeline for facial attribute labeling that can be coupled with standard detectors, (ii) quantifying the relationships between selected attributes and common failure modes at the video level, and (iii) proposing a bias–aware workflow that leverages these insights to improve robustness and interpretability. The outcome is a reasoned mapping of methods together with a ready–to–use, attribute–informed detection pipeline that connects analysis to practical design choices.

1.4 Overview of DeepFake detection

Detection approaches can be grouped by the type of evidence they exploit. Frame-level methods model spectral or demosaicing artifacts and other subtle traces left by synthesis and compression. Face-centric methods measure physiological or geometric plausibility, for example eye blinking statistics or head-pose consistency. Video-level architectures integrate temporal dynamics, for example motion fields and sequence encoders, in order to capture inter-frame cues that single frames cannot express.

1.4.1 Deepfake generation

Deepfake face manipulation is commonly organised into two families, *identity swap* and *facial reenactment*. In identity swap, the target's facial identity is replaced

with that of a source while preserving the target video's head pose, illumination, and motion; typical implementations use shared-encoder dual-decoder autoencoders or graphics pipelines. In facial reenactment, expressions and lip articulation from a driver stream are transferred onto a target while keeping target identity; implementations range from graphics-based blendshape fitting to neural motion-transfer schemes such as First-Order Motion Model (keypoint detector, dense motion network, and generator).

Across practitioner tools and research prototypes the ecosystem has evolved along three broadly recognised "generations" of content quality. **First generation** material exhibits visible seams at mask boundaries, colour mismatches, and inter-frame flicker (Figure 1.1); representative examples include *FakeApp* and *Faceswap* (autoencoder pipelines for identity swap) and *Face2Face* for facial reenactment based on 3D model fitting. **Second generation** pipelines improve face parsing and warping and add adversarial refinements, which increases robustness to scene variability, lighting, camera distance, and pose (Figure 1.2); examples include *DeepFaceLab* (multi-AE with parsing/warping), *Faceswap-GAN v2.2* (AE+GAN with self-attention for gaze and occlusion), *Avatarify* (FOMM-style real-time reenactment), *Zao* (cloud-based swap), and identity-preserving GANs such as *FaceShifter* and *SimSwap*. **Third generation** systems further suppress visible cues and narrow the gap to human perception (Figure 1.3); examples include diffusion-based identity transfer such as *IP-Adapter* (image-prompt adapter for diffusion) and *InstantID* (ID encoder with ControlNet-guided diffusion), as well as proprietary multi-modal stacks such as *HeyGen*. In this chapter the term "generation" is a descriptive grouping used in the community to track the progression in realism and robustness of tools and datasets, not a formal taxonomy.

Table 1.1 summarises representative tools across these families and generations, spanning both community software and research code, together with their indicative architectures (`AE`, `AE+GAN`, `3D blendshape` fitting, diffusion with adapters, `FOMM`-style motion transfer) and typical use. These trajectories explain the steady increase in realism and the corresponding rise in detection difficulty. For a more exhaustive list, see the extended Table 1.2 at the end of this chapter.

Tool	Gen	Features
FakeApp	Gen1	Early desktop autoencoder-based face swap (TensorFlow).
Faceswap	Gen1	Denoising autoencoder with face parsing/warping; identity swap for video.
Face2Face	Gen1	Real-time facial reenactment via 3D blendshape fitting.
DeepFake–tf	Gen2	TensorFlow AE; MTCNN/dlib extraction; DSSIM reconstruction loss.
DeepFaceLab	Gen2	Shared-encoder, dual-decoder AE; face parsing/warping; pairwise swap.
Faceswap–GAN	Gen2	AE+GAN with VGGFace; v2.2 adds self-attention for realistic eye/gaze movement.
FaceShifter	Gen2	Two-stage GAN with identity-preservation module; high-fidelity swap.
HeyGen	Gen3	Proprietary cloud reenactment/avatars; multi-modal pipeline.
InstantID	Gen3	ID encoder + ControlNet-guided diffusion; photorealistic ID/style transfer.

Table 1.1: Representative deepfake creation tools, their generation, and key features.

1.4.2 Detection methods

Two preeminent approaches are commonly emphasized. **Detection through visual artifacts** focuses on spatial inconsistencies that arise from synthesis and compositing, such as unnatural blending boundaries, color or illumination discontinuities between the pasted face and the surrounding context, and local rendering defects that persist even when global quality is high. **Detection through temporal features** exploits motion and dynamics, where optical–flow cues and sequence models capture frame–to–frame inconsistencies in facial geometry, expression trajectories, and background motion coherence. The latter connects to research that leverages optical flow based convolutional models to distinguish real from manipulated videos by modeling temporal regularities that forgers often fail to reproduce. Together, these lines of work reflect a complementary strategy, with spatial cues providing localized evidence and temporal cues providing sequence–level validation.

1.4.3 Datasets

The field relies on benchmark datasets that reflect the generations outlined above. *FaceForensics++* extends an earlier expression–manipulation corpus with

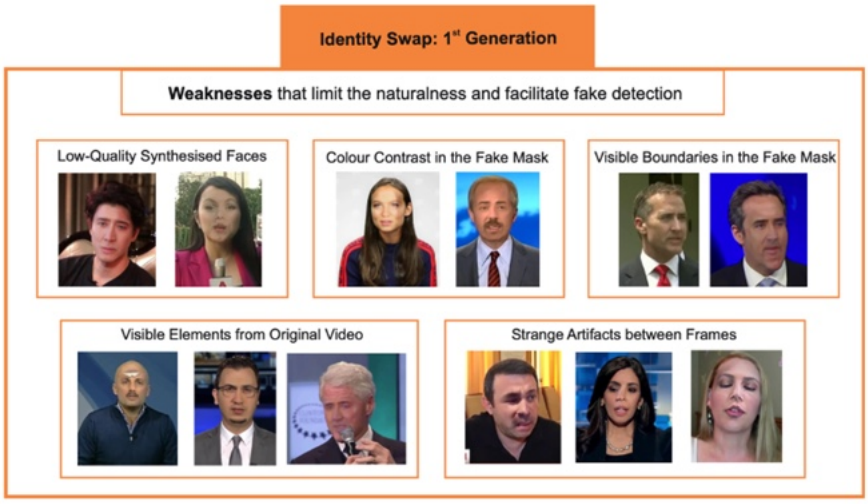


Figure 1.1: Identity swap, first generation. Typical weaknesses that limit naturalness and facilitate detection: (i) low-quality synthesised faces, (ii) colour contrast within the fake mask, (iii) visible mask boundaries, (iv) residual elements from the original video, (v) inter-frame artifacts. The panels show representative examples of these cues.

four automated methods, built from YouTube videos selected to contain trackable, mostly frontal, occlusion–free faces, which enables systematic and reproducible manipulation at scale. *DFDC* broadened scope and scale for a large–scale challenge, stimulating advances in both modeling and evaluation. These resources are routinely used to frame the problem as binary classification between authentic and fake content and to compare spatial and temporal detection pipelines under controlled conditions. The same materials clarify why detection difficulty increases over time, since second-generation and third–generation data intentionally vary scenes, lighting, camera distance, and pose to inhibit overfitting to obvious artifacts.

1.5 Research gap

1.5.1 Challenges and limitations

First–generation forgeries exposed strong weaknesses that simplified detection, for example visible mask borders, color contrast within the pasted region, and artifacts between frames. As data and tools matured, improvements in scene diversity, illumination, camera distance, and pose variability reduced those cues and raised

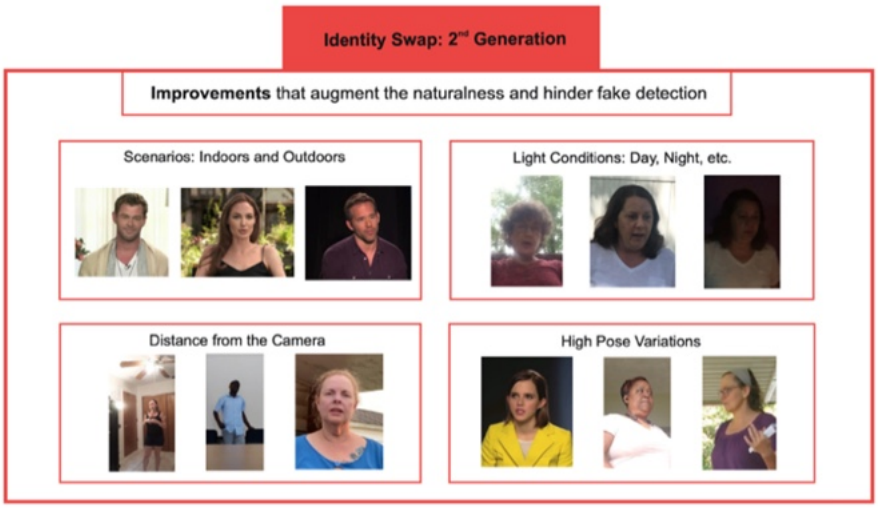


Figure 1.2: Identity swap, second generation. Improvements that increase naturalness and hinder detection: (i) diverse scenarios indoors and outdoors, (ii) varied illumination across day and night, (iii) changes in camera distance and scale, (iv) wide head-pose variations. The panels illustrate these factors.

the bar for generalization. This creates several open challenges: maintaining robustness under distribution shift across datasets and "generations," balancing spatial local evidence with temporal sequence reasoning, preventing over–reliance on spurious artifacts that can vanish as generators improve, and addressing real–world heterogeneity in capture conditions. At the same time, understanding use cases and targets is essential for risk assessment, since deepfakes appear as satire, deception, pornography, and technology demonstrations, each with different harm profiles and operational constraints. These factors jointly motivate detection strategies that integrate artifact–centric and temporal analysis while remaining resilient to evolving manipulation pipelines.

1.5.2 Open issue

Nowadays the main problem for the detector is to have the ability to maintain performance on content created by novel diffusion pipelines, including different samplers, schedulers, and fine-tuned weights, standardized evaluations find sizable drops when detectors face novel domains. An open issue that remains unresolved is represented by strong H.264/HEVC compression, and complex editing chains. In this scenario content provenance and watermarking are gaining traction as complementary

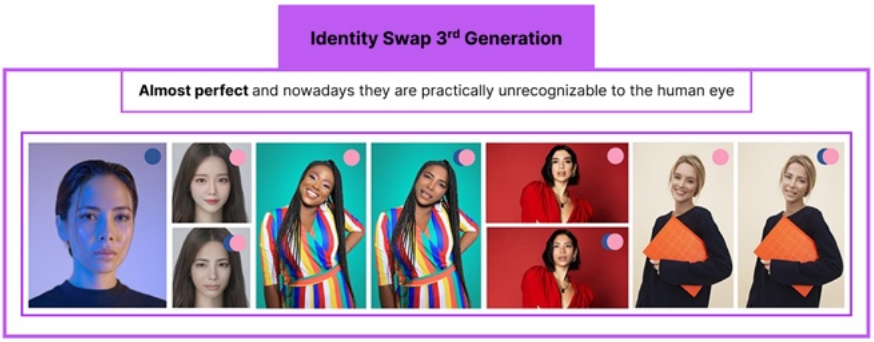


Figure 1.3: Identity swap, third generation. This is a SelfSwapper examples, in all panels the far–left portrait provides the *target* identity that is inserted into other *source* images. Blue dots mark source images, pink dots mark untouched target identities, and pink,+,blue dots mark the compositions produced by SelfSwapper, where the target identity is inserted into the source context.

safeguards, with Coalition for Content Provenance and Authenticity (C2PA) "Content Credentials" and Google *DeepMind's SynthID* (now spanning image, video, audio, and text) helping to verify the origin of media. However, these watermarks are neither universally adopted nor guaranteed to survive common post–processing operations, so they cannot be treated as a standalone solution and must instead evolve together with forensic detection methods as part of a combined defence strategy [1].

1.5.3 Author's Perspective and Contributions

During the three years of this PhD, which coincided with a historic inflection point for synthetic media, the research agenda changed in a concrete way. Early work often treated DeepFake detection as a static pattern-recognition task, optimised for near-perfect accuracy on constrained, curated benchmarks. Today the goal is different: to design dependable systems that *generalise* across datasets and manipulations, explain their decisions to humans, remain well-calibrated under distribution shift, and operate safely within real workflows. In short, the paradigm has shifted from maximising accuracy in fixed settings to delivering trustworthy, auditable, human-centred detection.

As a result, evaluation must consider additional dimensions that were

[1] Remark by Edward J. Delp during his lecture at the IEEE Summer School on Signal Processing (S3P-2025), San Vincenzo (LI), Italy, 24 September 2025, personal communication.

previously secondary: interpretability, calibration and uncertainty; robustness across domains, codecs, compression and editing chains; signals of content provenance and watermarking; governance, audit logging and policy compliance; and human-in-the-loop review to manage operational risk.

This thesis adopts that perspective and contributes along four lines. First, it introduces a scalable semi-supervised pipeline to label high-level facial attributes, enabling attribute-aware analyses at dataset scale. Second, it quantifies attribute-wise errors and bias through label-level metrics and statistical dependence measures, highlighting conditions that systematically affect decisions. Third, it performs controlled exclusion experiments to stress-test *generalisation*, showing how missing attribute coverage at training time influence performance at test time. Fourth, it derives practical guidance for data curation and attribute-aware training that prioritises robustness and interpretability over raw in-dataset accuracy, the entire study was produced with scientific rigor, with transparent technical protocols and fixed seeds to support reproducibility.

1.6 Thesis structure

This thesis is organised into six chapters and two appendices, which together cover the scientific, industrial, and multidisciplinary dimensions of the doctoral work.
Chapter 1, *Introduction*, outlines the overall research context, focusing on DeepFake technologies, their associated risks, and the motivation for studying attribute aware DeepFake detection. It introduces the main research questions and summarises the methodological approach adopted in the thesis.
Chapter 2, *State of the Art*, reviews the relevant literature on DeepFake generation and detection, with particular attention to convolutional and spatio temporal models, fairness and bias analysis, and the use of facial or semantic attributes. This chapter positions the thesis within existing work and highlights the specific gaps that the subsequent chapters aim to address.
Chapter 3, *Industry activities in PwC*, describes the industrial activities carried out during the PhD, with a focus on data driven and AI related tasks performed

in a consultancy setting. The chapter documents how methods and tools related to machine learning and computer vision were applied in practice, and how these experiences informed the design and constraints of the academic study.

Chapter 4, *Analysis of DF Detection through Attribute Labeling*, presents the core methodological contribution of the thesis. It introduces the dataset construction, the semi supervised facial attribute labeling pipeline, the DeepFake detection model, and the first set of analyses on how prediction errors relate to selected visual attributes. The chapter also details the statistical measures used to explore possible associations between attributes and misclassifications.

Chapter 5, *Attribute-Aware Training Strategies*, extends the previous analysis by studying controlled exclusion experiments on attribute subgroups. It investigates how removing specific attribute values from the training data affects performance, both globally and within subgroups, and it derives preliminary operational recommendations for bias aware data curation and evaluation in DeepFake detection.

Chapter 6, *Conclusions and Future Work*, summarises the main findings, with emphasis on the exploratory nature of the results, and discusses their potential implications for attribute informed analysis of DeepFake detectors. It also outlines possible extensions, including more diverse datasets, richer attribute taxonomies, multivariate analyses, and training strategies that may incorporate attribute information in a more direct way.

Appendix 1, *Multidisciplinary Application of Artificial Intelligence*, briefly reports on selected AI projects carried out in domains different from DeepFake detection, such as engineering, education, or organisational settings. These examples illustrate the broader multidisciplinary background of the author and provide context on how AI methods were adapted to heterogeneous application areas.

Appendix 2, *Scientific outputs arising from the PhD*, lists the scientific publications, conference contributions, and other research outputs produced by the author during the doctoral programme.

Table 1.2: Summary of face-manipulation tools by generation, architecture, output and usage

Tool	**Generation**	**Year**	**Architecture**	**Output quality**	**Primary use / Modality**	**Expertise Required**
FaceSwap *(Kowalski)*	Pre-DF	2019	Landmark warping + blending	Low	Still-image swap based on classical Computer Vision techniques, Landmark warping + blending; acceptable only for small, near-frontal faces under uniform lighting; frequent boundary seams, color mismatch and ghosting; weak identity preservation under expression changes; no temporal consistency since image-only	Medium
FakeApp	Gen1	2018	Autoencoder pipeline	Medium–Low	Desktop application built with Google open source library TensorFlow, early AE outputs with soft detail and blur; temporal flicker and blending artifacts are common; best with frontal, well-lit faces and matched skin tone; quality degrades with occlusions, motion and profile views; typically low-resolution face crops	Medium

Continues on next page

Table 1.2: Summary of face-manipulation tools by generation, architecture, output and usage

Tool	**Generation**	**Year**	**Architecture**	**Output quality**	**Primary use / Modality**	**Expertise Required**
Faceswap *(Tora et al.)*	Gen1	2018	Denoising Autoencoder (AE)	Medium	Graphics based approach to swap the identity of subjects. It can be applied to an unlimited number of subjects, though it is more prone to severe artifacts if internal modules fail	Medium
Face2Face	Gen1	2016	3D model-based, blendshape fitting	Medium	Graphics reenactment method published by computer vision, a method that transfers facial expressions from one person to a realistic digital avatar in real time	High
Avatarify	Gen2	2019	FOMM-style keypoints + generator	Medium	Real time facial reenactment and motion transfer with First-Order Motion Model (FOMM), keypoint detector + dense motion network + generator, architecture; able to drive a target portrait from webcam or video stream	Low–Medium
DeepFake–tf	Gen2	2019	Autoencoder (TensorFlow); DSSIM loss; MTCNN, dlib	Medium–High	Identity swap that supports multiple face extraction models (MTCNN, dlib) and uses DSSIM loss to reconstruct the face, implemented in TensorFlow	Medium

Continues on next page

Table 1.2: Summary of face-manipulation tools by generation, architecture, output and usage

Tool	**Generation**	**Year**	**Architecture**	**Output quality**	**Primary use / Modality**	**Expertise Required**
DeepFaceLab	Gen2	2019	Multi-AE + face parsing/warping	High	Identity swapping method trained on two subjects; a single autoencoder is trained for the pair, sharing one encoder and using two decoders that reconstruct each subject; at test time the decoders are inverted to obtain the final face identity manipulation	Medium–High
Zao	Gen2	2019	Proprietary cloud AE/GAN	High	Identity swap with a proprietary client-server architecture and CNN pipeline, likely AE/GAN hybrid; feature one-shot consumer app, fast results, closed source code	Low
Faceswap-GAN *(v2.2)*	Gen2	2019	AE + GAN, self-attention	High	Identity swap with GAN-based pipeline; adds VGGFace to the original Faceswap model, advanced version 2.2 adds self-attention for realistic eye/gaze movement, better occlusion and pose robustness, improved color/texture consistency	High

Continues on next page

Table 1.2: Summary of face-manipulation tools by generation, architecture, output and usage

Tool	**Generation**	**Year**	**Architecture**	**Output quality**	**Primary use / Modality**	**Expertise Required**
FaceShifter	Gen2	2019	Two-stage GAN with identity module	High	Strong identity preservation with robust pose and expression transfer, minor artifacts under occlusion or extreme yaw	High
SimSwap	Gen2	2020	AEI-Net / GAN	High	Reliable one-shot swaps with good color and geometry consistency, quality drops under severe pose or illumination mismatch	Medium
IP-Adapter	Gen3	2023	Diffusion with image-prompt adapter	High (*stylised*)	Excels at artistic identity transfer in images, not intended for strict photorealism or video consistency	Medium
HeyGen	Gen3	2022	Proprietary cloud multi-modal stack	Photorealistic	Production-grade avatars and video within template constraints, stable lip-sync and motion when inputs are well covered	Low

Continues on next page

Table 1.2: Summary of face-manipulation tools by generation, architecture, output and usage

Tool	**Generation**	**Year**	**Architecture**	**Output quality**	**Primary use / Modality**	**Expertise Required**
InstantID	Gen3	2024	ID encoder + ControlNet-guided diffusion	Photorealistic	High-fidelity identity preservation from a single reference with controllable style; robust to moderate pose and lighting changes, slight softness under extreme angles or busy backgrounds; image-only without temporal consistency	Medium

Chapter 2

State of the Art

Understanding the state of the art in deepfake detection requires a structured literature review. The field evolves rapidly due to new generation techniques, shifting benchmarks, and frequent updates to evaluation protocols, therefore any assessment risks becoming incomplete if not methodically scoped. This thesis adopts a literature review to synthesize the most relevant methods, datasets, and findings, clarifying strengths, limitations, and open problems that directly inform the research questions and experimental design. The review should be read as a time-bounded snapshot that reflects the best available evidence at the moment of publication, while acknowledging the constant evolution of the domain. It has been developed progressively during the three years of the author's PhD in *Big Data and Artificial Intelligence*, ensuring continuity between the surveyed evidence and the empirical contributions of the thesis.

2.1 Aim and Scope of a Literature Review

This literature review reconstructs in a systematic and transparent manner the analytical pathway adopted in the thesis on *deepfake detection* with a focus on human faces in images and videos. The review targets: reference datasets (*FaceForensics++*, *Celeb-DF*, *DFDC*), frame-level and video-level methods, spectral and spatial cues, physiological and geometric indicators, spatio-temporal modeling, and survey or overview papers used to contextualize the field. It also frames the empirical study of

generalization across datasets and manipulation types, and the observed relationship between classifier errors and high-level facial attributes.

2.2 Selection Protocol

Primary sources. Proceedings and journals in computer vision and multimedia forensics, including IEEE/CVF venues (e.g., CVPR, ICCV, WIFS, ICASSP, AVSS, CVPRW), IJCAI, relevant IEEE journals, MDPI outlets in imaging, and Springer CCIS volumes, together with institutional dataset pages.

Time window. 2014 to 2025 to cover early work on facial attributes, first detection baselines, and the main benchmark datasets.

Language. English, with selected Italian contributions when directly connected to the thesis or research group outputs.

Document types included. Dataset papers, detection methods, survey and overview papers, works on facial attributes and spatio-temporal cues, proceedings or book chapters that document techniques reused in the thesis.

Exclusions. Position papers without empirical evidence, reports lacking methodological detail, generation-only works that are not used as detection baselines or benchmarks.

2.3 Inclusion and Exclusion Criteria

Inclusion. (i) Reproducible description of method or dataset, (ii) evaluation on public or well-described data, (iii) direct relevance to face manipulation forensics in images or videos, or to spectral and temporal cues, (iv) explicit citation or use within the thesis results.

Exclusion. (i) Absence of quantitative results, (ii) scope unrelated to face-focused media forensics, (iii) redundant contributions superseded by more complete survey or benchmark papers.

2.4 Search Strategy and Traceability

The search protocol is based on Boolean combinations anchored to *deepfake detection*, enriched with families of cues and benchmark datasets, and filtered by venue quality and time window. Queries were issued across Scopus, ScienceDirect, IEEE Xplore, and Google Scholar, with terms applied to title, abstract, and keywords when supported. Results were screened in two stages (title–abstract, then full text) against predefined inclusion/exclusion criteria, with duplicates removed. Each included work is recorded with database of origin, query string, screening decision, and rationale for inclusion, enabling full traceability to the research questions in Section 1.3.

Databases and indexing services. Searches were conducted on *Scopus*, *ScienceDirect*, *IEEE Xplore*, and *Google Scholar*. These sources jointly cover major computer vision and multimedia forensics venues, together with publisher platforms relevant to the topic.

Query design. Searches used a fixed anchor term, `"deepfake detection"`, combined with an **OR**-group of five thematic keywords as defined in Section 2.4.1. For database engines that support fielded queries, the terms were applied to title, abstract, and keywords. For Google Scholar, the same strings were issued as all-fields queries and screened on the first relevant pages.

Filtering and screening. Results were filtered by year and subject area when supported by the platform. A two-stage screening followed: title–abstract screening to remove off-topic items, then full-text screening against the inclusion and exclusion criteria in Section 2.3. Duplicates across platforms were removed.

Seed papers and snowballing. Dataset and benchmark papers central to this thesis, namely *FaceForensics++*, *Celeb-DF*, and the *DFDC* preview, acted as seeds. Backward and forward snowballing from these seeds complemented the database queries and helped to refine this review and select the final set of papers.

Traceability. Each included work is mapped to the research questions in Section 1.3 and to one of the keyword sets in Section 2.4.1. The mapping records the database of origin, query string, screening decision, and rationale for inclusion.

2.4.1 Keywords and Search Strings

All strings assume the fixed block **"deepfake detection"** combined with an **OR**-group of five thematic keywords. Below are three alternative sets aligned with the methodological axes of the thesis. Each line is a self-contained query. To structure the search, four query sets were defined. Set 0 `"deepfake detection" AND ("face" OR "faces")` provides a broad introduction to the domain, while Sets A–C focus on specific families of cues that are central to this thesis.

- **Set A (physiological and geometric cues):** `"deepfake detection" AND ("head pose" OR "eye blinking" OR "lip-sync" OR "facial landmarks" OR "physiological signals")`
- **Set B (spectral domain and artifacts):** `"deepfake detection" AND ("frequency domain" OR "spectral artifacts" OR "compression artifacts" OR "color filter array" OR "noise patterns")`
- **Set C (video dynamics and graphs):** `"deepfake detection" AND ("optical flow" OR "temporal consistency" OR "LSTM" OR "graph neural networks" OR "spatio-temporal")`

2.4.2 Domain scoping via a broad search

A preliminary, broad search was issued on Google Scholar trough title, keywords and all the related metadata to quantify the overall size of the literature relevant to *deepfake detection* and faces. These counts are used only to frame the problem space, they do not feed into the eligibility screening.

Engine	**Set 0**	**Set A**	**Set B**	**Set C**
Google Scholar	11,600	3,640	3,340	5,290

Table 2.1: Broad domain scoping on Google Scholar. Counts indicate raw result volumes and serve only to contextualize the search space.

2.4.3 Targeted retrieval in article keywords and metadata

The actual screening corpus was collected with fielded queries that constrain matches to article keywords, across *Scopus*, *ScienceDirect*, *IEEE Xplore*, and *Google Scholar*. Set 0 still serves as a reference for the magnitude of the "faces + deepfake detection" literature, while Sets A, B and C drive the focused retrieval aligned with the method families investigated in this thesis.

Database	**Set 0**	**Set A**	**Set B**	**Set C**
Scopus	379	19	71	105
ScienceDirect	75	30	30	52
IEEE Xplore	395	41	161	256
Google Scholar	6,670	2,180	1,900	3,410

Table 2.2: Targeted retrieval using keyword-field constraints for Sets 0, A, B, and C across four databases. Values are raw results prior to deduplication and screening.

The Table 2.2 show that the generic faces query (Set 0) yields large volumes, while the method-focused Sets A, B and C substantially narrow the corpus. *IEEE Xplore* and *Scopus* contribute the largest share of technical records for Sets B and C, which is consistent with their coverage of computer vision, signal processing, and multimedia forensics venues. The resulting records were then deduplicated and screened according to Section 2.3, with the selection process summarized in

Figure 2.1 and the final inclusions detailed in the evidence map (Table 2.5).

2.4.4 Focused query volumes (Sets A–C) and database policy

The focused retrieval based on Sets A–C produced the following raw volumes prior to deduplication and screening:

Database	**Total (A–C)**
Scopus	195
ScienceDirect	111
IEEE Xplore	445
Google Scholar	7490
Total	8241
Total without Google Scholar	751

Table 2.3: Result volumes for Sets A–C across the four sources, before deduplication and screening.

Following this step, records from all sources were merged and deduplicated. From the subsequent screening stage onward, Google Scholar was not used as a primary corpus. The rationale is quality and metadata stability: a large fraction of Scholar results are non–peer reviewed, preliminary project notes, technical repositories, or partial reports. High–value scholarly items surfaced by Scholar are, in most cases, indexed as final versions in Scopus, ScienceDirect, or IEEE Xplore. Scholar remains useful for discovery, seeding, and snowballing, but inclusion decisions for this scientific review rely on publisher and index platforms with consolidated peer–review and metadata.

Operational policy.

- Primary screening corpus: *Scopus*, *ScienceDirect*, *IEEE Xplore*.
- *Google Scholar*: used for discovery and snowballing only, not for quantitative counting or inclusion.

- Preference for peer–reviewed venues and publisher platforms with stable identifiers and metadata.
- Grey literature and preprints are excluded unless a peer–reviewed version is identified and used, or the item defines a dataset or benchmark that is essential for reproducibility.

This policy is consistent with the inclusion and exclusion criteria in Section 2.3, the search protocol in Section 2.4, the selection flow summarized in Figure 2.1, and the final set of included studies reported in the evidence map (Table 2.5).

2.4.5 Other sources

In addition to the keyword–driven retrieval on *Scopus*, *ScienceDirect*, and *IEEE Xplore*, a limited number of studies reached the author through professional activities and ancillary channels. These include conference attendance and seminars, reviewer or editorial suggestions, collaborations, dataset and challenge websites, and tool repositories that pointed to citable papers. All items gathered through these routes are counted under the node *Other* in the PRISMA diagram (Figure 2.1), for a total of **9** records.

Admission criteria and safeguards. To preserve scientific rigor, "Other" records were considered only if they met the same inclusion and exclusion criteria defined in Section 2.3. The following operating rules were applied:

- **Peer review preference.** Priority was given to peer–reviewed venues. Preprints were retained only when they are the authoritative reference for a dataset or benchmark, or when a peer–reviewed version was not yet available at screening time. When a final version appeared, it replaced the preprint.
- **Unique contribution.** The record must provide a contribution not already covered by the corpus, such as a new dataset or manipulation type, a distinct method family, or a robust analysis relevant to the thesis research questions.

- **Accessibility and reproducibility.** Full text had to be accessible, with sufficient methodological detail to support reproducibility or well–defined evaluation protocols.

- **De–duplication and version control.** Duplicates were removed and only the highest–quality version was kept, giving preference to publisher versions with stable identifiers.

- **Traceability.** Each inclusion is tagged with provenance "Other" in the evidence map (Table 2.5), together with the reason for inclusion and the research question mapping.

This policy allows the review to capture valuable signals that often emerge first through professional networks and challenge ecosystems, while maintaining comparability with the database–driven corpus and limiting selection bias. The final counts and their role in study selection are summarized in Figure 2.1.

2.4.6 Duplicates Screening

To ensure a clean and traceable corpus, deduplication was performed at two levels: (i) *within each query set* (A, B, C) across the three primary databases (*Scopus*, *ScienceDirect*, *IEEE Xplore*), and (ii) *across query sets* by merging the three within-set lists and removing cross-list duplicates. In line with the database policy in Section 2.4.4, *Google Scholar* was used for discovery only and excluded from quantitative counts and deduplication.

Matching keys and precedence. The primary identity key is a normalized DOI (lowercased, stripped of the `doi.org` prefix and trailing punctuation). When a DOI is unavailable, a normalized title is used (Unicode NFKD, lowercased, punctuation removed, single-space collapsed). When duplicate records exist for the same work, the retained entry follows a fixed precedence that favors metadata stability and publisher indexing: **Scopus > IEEE Xplore > ScienceDirect**. All other fields are carried over from the retained record.

Results. Table 2.4 reports the volumes before and after deduplication within each set, followed by the final cross-set merge. The cross-set step also reveals limited overlap between method families, with intersections A∩B=21, A∩C=3, B∩C=2, and no triple intersection, consistent with the targeted design of Sets A–C.

Scope	Input	Duplicates removed	Unique
Query A (physio/geometry)	90	6	84
Query B (spectral/artifacts)	262	25	237
Query C (spatio-temporal)	413	76	337
Cross-set merge (A+B+C)	765	107	658
Cross-set merge (A+B+C) *without overlap*			632

Table 2.4: Deduplication results by query set and final cross-set merge. Inputs aggregate Scopus, ScienceDirect, and IEEE Xplore records; Google Scholar is excluded from these counts.

The cleaned, unified list (**632** unique records) is used for screening and synthesis in the subsequent sections (see Figure 2.1 and Table 2.5). The complete search logs, per–database exports, deduplicated corpora for Sets A–C, the cross–set merged list, and the scripts used for parsing and reconciliation are publicly available in a GitHub repository maintained by the author, which accompanies this thesis.[1]

2.4.7 Title and Abstract Screening

Screening was conducted on the deduplicated corpus obtained from Scopus, ScienceDirect, and IEEE Xplore (Section 2.4.6). The database merge produced **632** unique records across Sets A–C. In addition, **9** items of high relevance were identified through professional activities and ancillary channels ("Other", Section 2.4.5), yielding **641** records for title and abstract screening.

Two reviewers screened titles and abstracts against the inclusion criteria defined by the research questions (RQ1–RQ3), with conflicts resolved by discussion. Inclusion required a primary contribution to face deepfake *detection* on images or videos or a field-synthesizing survey directly informing protocols, datasets, or metrics used in

[1]Public GitHub Repository created by the author: https://github.com/vstile/deepfake-detection-review (visited on 17 December 2025).

face deepfake detection. Exclusion reasons were logged for transparency and later aggregated in the PRISMA-style diagram (Figure 2.1).

After this phase, **579** records were excluded for the following primary reasons: (i) not focused on face manipulation detection or out of scope (n=148), (ii) generation-only or attack papers without detector evaluation (n=139), (iii) non-video or audio-only deepfake topics (n=98), (iv) non–peer-reviewed or incomplete records such as short notes without methods (n=78), (v) duplicates or obsolete versions (n=54), (vi) not available in English and no equivalent English source (n=26), (vii) insufficient methodological detail for appraisal (n=36). A total of **62** studies proceeded to full-text assessment.

2.4.8 Full Text Assessment for Eligibility

Full texts were retrieved for **62** studies and appraised against the eligibility criteria: (i) explicit treatment of face deepfake detection or surveys that systematize the face deepfake detection landscape; (ii) reproducible or comparable evaluation, preferably on public datasets (e.g., FF++, Celeb-DF, DFDC) or with clearly stated metrics; (iii) alignment with at least one research question (RQ1: detection methods, RQ2: datasets, generalization and evaluation protocols, RQ3: attribute-supervision, interpretability and bias-aware analysis).

Following full-text assessment, **38** studies were excluded with reasons: no reproducible evaluation on public datasets or incompatible metrics (n=13), not face-focused or mixed-modality without face-specific analysis (n=9), redundant variants of the same approach without added contribution (n=6), dataset-only or challenge reports that did not inform the RQs (n=4), outside the methodological scope or time window (n=2), overlapping editorial content already covered by selected papers (n=4). The remaining **24** studies were included for qualitative synthesis and evidence mapping (Table 2.5).

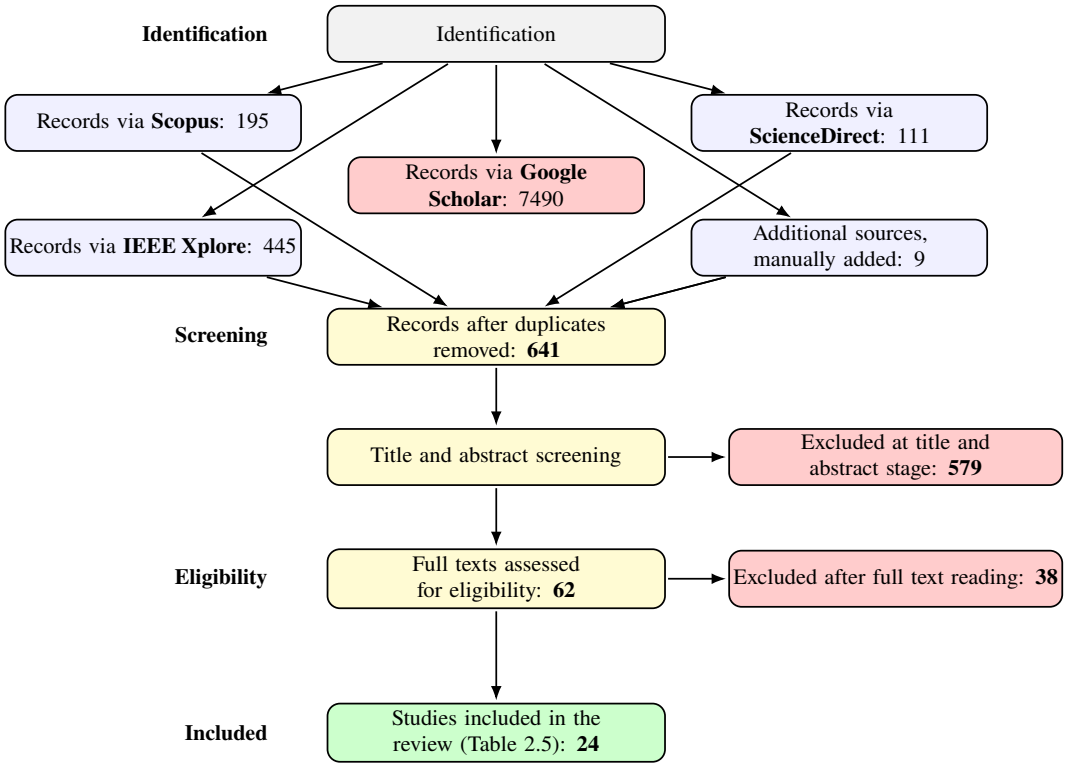


Figure 2.1: PRISMA-style flow summary of study selection. Sources include Scopus, ScienceDirect, IEEE Xplore, and Google Scholar. The final box references the evidence map in Table 2.5.

2.5 Studies Included in the Review

The final set comprises **24** studies. These cover foundational datasets and surveys that structure the field (RQ2), method families at frame level and video level that define the technical landscape (RQ1), and works that directly support the thesis focus on attribute supervision, interpretability, and bias-aware analysis (RQ3). Four items were admitted from the "Other" channel due to their centrality for this thesis (explicitly flagged in the Table 2.5) and are also reflected in the PRISMA block for *Other* sources as shown in Figure 2.1.

2.5.1 Review characteristics

The final evidence base consists of **24** primary studies that span more than a decade of research, from early works on facial attribute modelling to the most recent contributions on multimodal deepfake detection, fairness and self–learning systems. In line with the PRISMA–style flow and the evidence map reported in Table 2.1,

this section briefly characterizes the corpus in terms of publication year, venue and disciplinary area, geographical origin, and provides a structured publication list.

Year of publication. The selected studies are distributed over the period 2014–2025, reflecting the rapid evolution of deepfake generation and detection techniques. Two early works on facial attributes and pose modelling[2][3] predate the emergence of deepfakes and serve as methodological background for the attribute–based analysis carried out in this thesis.

The bulk of the primary studies is concentrated between 2018 and 2020, which corresponds to the first wave of deepfake research driven by the public release of tools and datasets. In this interval the corpus includes frame–level detectors, temporal pipelines and early dataset papers such as *FaceForensics++* and *VidTIMIT*–based

[2]N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev, *PANDA: Pose aligned networks for deep attribute modeling*, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2014, pp. 1637–1644.

[3]G. Levi and T. Hassner, *Age and gender classification using convolutional neural networks*, in: *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Boston, MA, USA June 2015, pp. 34–42, ISBN: 9781467367592, DOI: 10.1109/CVPRW.2015.7301352, URL: http://ieeexplore.ieee.org/document/7301352/ (visited on 12/17/2025).

deepfakes[4567891011121314 15].[16] This period accounts for more than half of the corpus and captures the transition from ad hoc cues to data–driven CNN baselines.

From 2021 onwards, the literature becomes more diversified. Recent years introduce adversarial threat analyses, large–scale fairness and demographic bias

[4]D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, *MesoNet: a Compact Facial Video Forgery Detection Network*, in: *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, IEEE, Hong Kong Dec. 2018, pp. 1–7, ISBN: 9781538665367, DOI: 10.1109/WIFS.2018.8630761, URL: https://ieeexplore.ieee.org/document/8630761/ (visited on 10/27/2025).

[5]D. Guera and E. J. Delp, *Deepfake Video Detection Using Recurrent Neural Networks*, in: *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, IEEE, Auckland, New Zealand Nov. 2018, pp. 1–6, ISBN: 978-1-5386-9294-3, DOI: 10.1109/AVSS.2018.8639163, URL: https://ieeexplore.ieee.org/document/8639163/ (visited on 06/10/2025).

[6]P. Korshunov and S. Marcel, *DeepFakes: a New Threat to Face Recognition? Assessment and Detection*, arXiv:1812.08685, Dec. 2018, DOI: 10.48550/arXiv.1812.08685, URL: http://arxiv.org/abs/1812.08685 (visited on 10/27/2025).

[7]Y. Li, M.-C. Chang, and S. Lyu, *In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking*, in: *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, IEEE, Hong Kong Dec. 2018, pp. 1–7, ISBN: 9781538665367, DOI: 10.1109/WIFS.2018.8630787, URL: https://ieeexplore.ieee.org/document/8630787/ (visited on 10/27/2025).

[8]I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, *Deepfake Video Detection through Optical Flow Based CNN*, in: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, IEEE, Seoul, Korea (South) Oct. 2019, pp. 1205–1207, ISBN: 9781728150239, DOI: 10.1109/ICCVW.2019.00152, URL: https://ieeexplore.ieee.org/document/9022558/ (visited on 10/27/2025).

[9]A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, *FaceForensics++: Learning to Detect Manipulated Facial Images*, in: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Seoul, Korea (South) Oct. 2019, pp. 1–11, ISBN: 9781728148038, DOI: 10.1109/ICCV.2019.00009, URL: https://ieeexplore.ieee.org/document/9010912/ (visited on 10/27/2025).

[10]E. Sabir, W. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, *Recurrent convolutional strategies for face manipulation detection in videos*, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* 2019, pp. 1–9.

[11]X. Yang, Y. Li, and S. Lyu, *Exposing Deep Fakes Using Inconsistent Head Poses*, in: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Brighton, United Kingdom May 2019, pp. 8261–8265, ISBN: 9781479981311, DOI: 10.1109/ICASSP.2019.8683164, URL: https://ieeexplore.ieee.org/document/8683164/ (visited on 10/27/2025).

[12]Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, *Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics*, in: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Seattle, WA, USA June 2020, pp. 3204–3213, ISBN: 9781728171685, DOI: 10.1109/CVPR42600.2020.00327, URL: https://ieeexplore.ieee.org/document/9156368/ (visited on 10/27/2025).

[13]B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, *The DeepFake Detection Challenge (DFDC) Dataset*, 2020, DOI: 10.48550/ARXIV.2006.07397, URL: https://arxiv.org/abs/2006.07397 (visited on 12/18/2025).

[14]R. Durall, M. Keuper, and J. Keuper, *Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions*, in: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 2020, pp. 7887–7896, DOI: 10.1109/CVPR42600.2020.00791.

[15]R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, *Deepfakes and beyond: A survey of face manipulation and fake detection*, in: *Information Fusion* 64 (2020), pp. 131–148.

[16]L. Verdoliva, *Media Forensics and DeepFakes: An Overview*, in: *IEEE Journal of Selected*

studies, challenges and benchmarks, as well as works on multimodal audio–visual detection and self–learning systems[1,7,18,19,20,21,22,23,24,25].[26] Overall, the temporal distribution shows that the review captures both the foundational phase of deepfake detection and the most recent methodological and ethical developments, which is consistent with the time frame of the PhD programme.

Topics in Signal Processing 14.5 (Aug. 2020), pp. 910–932, ISSN: 1932-4553, 1941-0484, DOI: 10.1109/JSTSP.2020.3002101, URL: https://ieeexplore.ieee.org/document/9115874/ (visited on 10/27/2025).

[17] L. Guarnera et al., *The Face Deepfake Detection Challenge*, en, in: *Journal of Imaging* 8.10 (Sept. 2022), p. 263, ISSN: 2313-433X, DOI: 10.3390/jimaging8100263, URL: https://www.mdpi.com/2313-433X/8/10/263 (visited on 10/27/2025).

[18] P. Neekhara, B. Dolhansky, J. Bitton, and C. C. Ferrer, *Adversarial Threats to DeepFake Detection: A Practical Perspective*, in: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Nashville, TN, USA June 2021, pp. 923–932, ISBN: 9781665448994, DOI: 10.1109/CVPRW53098.2021.00103, URL: https://ieeexplore.ieee.org/document/9522903/ (visited on 12/17/2025).

[19] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, *Deepfake Detection: A Systematic Literature Review*, in: *IEEE Access* 10 (2022), pp. 25494–25513, ISSN: 2169-3536, DOI: 10.1109/ACCESS.2022.3154404, URL: https://ieeexplore.ieee.org/document/9721302/ (visited on 10/27/2025).

[20] C. Hazirbas, J. Bitton, B. Dolhansky, J. Pan, A. Gordo, and C. C. Ferrer, *Towards Measuring Fairness in AI: The Casual Conversations Dataset*, in: *IEEE Transactions on Biometrics, Behavior, and Identity Science* 4.3 (July 2022), pp. 324–332, ISSN: 2637-6407, DOI: 10.1109/TBIOM.2021.3132237, URL: https://ieeexplore.ieee.org/document/9634168/ (visited on 12/17/2025).

[21] A. Beckmann, A. Hilsmann, and P. Eisert, *Fooling State-of-the-Art Deepfake Detection with High-Quality Deepfakes*, arXiv:2305.05282, May 2023, DOI: 10.48550/arXiv.2305.05282, URL: http://arxiv.org/abs/2305.05282 (visited on 10/27/2025).

[22] V. S. Katamneni and A. Rattani, *Contextual Cross-Modal Attention for Audio-Visual Deepfake Detection and Localization*, arXiv:2408.01532, Aug. 2024, DOI: 10.48550/arXiv.2408.01532, URL: http://arxiv.org/abs/2408.01532 (visited on 12/17/2025).

[23] A. Anshul, S. Gopal, D. Rajan, and E. S. Chng, *Intra-modal and Cross-modal Synchronization for Audio-visual Deepfake Detection and Temporal Localization*, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Honolulu, Hawaii, USA Oct. 2025, pp. 13826–13836.

[24] G. Guerrero-Contreras, S. Balderas-Díaz, A. García-Pascual, and A. Munõz, *Self-Learning Systems for Enhanced Traffic Management in Urban Settings*, enc, in: (June 2024), DOI: 10.5281/ZENODO.11917270, URL: https://zenodo.org/doi/10.5281/zenodo.11917270 (visited on 11/28/2025).

[25] G. Guerrero-Contreras, S. Balderas-Díaz, A. García-Pascual, and A. Muñoz, *Adaptive Vehicle Detection in Urban Environments: A Self-learning Approach*, en, in: *Ambient Intelligence – Software and Applications – 15th International Symposium on Ambient Intelligence*, ed. by P. Novais, P. B. D., I. Satoh, V. J. Inglada, S. R. González, E. Jove Pérez, J. Parra Domínguez, P. Chamoso, and R. S. Alonso, vol. 1279, Springer Nature Switzerland, Cham, Switzerland 2025, pp. 25–34, ISBN: 9783031831164 9783031831171, DOI: 10.1007/978-3-031-83117-1_3, URL: https://link.springer.com/10.1007/978-3-031-83117-1_3 (visited on 10/27/2025).

[26] V. Stile, R. Caldelli, G. Guerrero-Contreras, S. Balderas-Díaz, and I. Medina-Bulo, *Analysis of DeepFake Detection through Semi-Supervised Facial Attribute Labeling*, ENG, in: *Proceedings of the 11th Spanish-German Symposium on Applied Computer Science (SGSOACS 2025)*, vol. 2831, Communications in Computer and Information Science (CCIS), Springer Cham, Wien, Austria July 2025, pp. XX, 138, ISBN: 978-3-032-14815-5, URL: https://link.springer.com/book/9783032148155.

Type of publications, disciplines and journal ranking. The corpus is dominated by computer vision and multimedia forensics research, complemented by contributions from responsible AI, fairness analysis and self–learning systems. A first group of studies proposes concrete detection pipelines, mostly published in leading computer vision and signal processing venues. This group includes compact frame–level CNNs and mesoscopic networks[27],[28] temporal models based on RNNs and optical flow[29][30],[31] methods exploiting geometric or physiological cues such as head pose and eye blinking[32],[33] and multimodal audio–visual approaches with attention and synchronization modules[34][35].[36] These works are mainly associated with flagship conferences and workshops (CVPR, ICCV, ICASSP, WIFS and related events), and can be considered state of the art in terms of methodological innovation.

A second group is formed by surveys and systematization papers that frame the domain and the available tools. Verdoliva[37] and Tolosana et al.[38] provide broad overviews of media forensics and face manipulation, while Rana et al.[39] conduct a systematic literature review of deepfake detection techniques. These studies are published in high–impact journals in signal processing and information fusion, typically ranked in the top quartiles of their categories, and they support the conceptual organization of this review.

A third cluster consists of dataset, benchmark and challenge papers, which define

[27] Afchar, Nozick, Yamagishi, and Echizen, "MesoNet", op. cit.

[28] Durall, Keuper, and Keuper, "Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions", op. cit.

[29] Guera and Delp, "Deepfake Video Detection Using Recurrent Neural Networks", op. cit.

[30] Amerini, Galteri, Caldelli, and Del Bimbo, "Deepfake Video Detection through Optical Flow Based CNN", op. cit.

[31] Sabir, Cheng, Jaiswal, AbdAlmageed, Masi, and Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos", op. cit.

[32] Yang, Li, and Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses", op. cit.

[33] Li, Chang, and Lyu, "In Ictu Oculi", op. cit.

[34] Neekhara, Dolhansky, Bitton, and Ferrer, "Adversarial Threats to DeepFake Detection", op. cit.

[35] Katamneni and Rattani, *Contextual Cross-Modal Attention for Audio-Visual Deepfake Detection and Localization*, op. cit.

[36] Anshul, Gopal, Rajan, and Chng, "Intra-modal and Cross-modal Synchronization for Audio-visual Deepfake Detection and Temporal Localization", op. cit.

[37] Verdoliva, "Media Forensics and DeepFakes", op. cit.

[38] Tolosana, Vera-Rodriguez, Fierrez, Morales, and Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection", op. cit.

[39] Rana, Nobi, Murali, and Sung, "Deepfake Detection", op. cit.

the evaluation landscape. *FaceForensics++*,[40] *DFDC*,[41] *Celeb–DF*[42] and the Face Deepfake Detection Challenge[43] introduce large–scale corpora and competitions that are now standard references. Hazirbas et al.[44] contribute the Casual Conversations dataset for fairness analysis, which is used to study demographic robustness of winning DFDC models. These works sit at the intersection of computer vision, multimedia forensics and responsible AI.

Finally, several studies focus explicitly on attributes, bias and self–learning. Early facial attribute and pose–aligned modelling[45][46] provides the conceptual basis for attribute supervision, while the thesis paper itself[47] and the self–learning traffic management works illustrate how semi–supervised labelling and iterative self–training can be used to analyse errors and adapt models over time[48].[49] Together with bias and fairness oriented works, these contributions connect deepfake detection to broader discussions on trustworthy and bias–aware AI[50].[51]

Country of origin. The primary studies are geographically diverse, although there is a clear prevalence of European research groups. Several key contributions originate from Germany and Italy, for example *FaceForensics++* and high–quality deepfake generation[52],[53] as well as optical–flow based detection and challenge organization in

[40]Rössler, Cozzolino, Verdoliva, Riess, Thies, and Nießner, "FaceForensics++", op. cit.

[41]Dolhansky, Bitton, Pflaum, Lu, Howes, Wang, and Ferrer, *The DeepFake Detection Challenge (DFDC) Dataset*, op. cit.

[42]Li, Yang, Sun, Qi, and Lyu, "Celeb-DF", op. cit.

[43]Guarnera et al., "The Face Deepfake Detection Challenge", op. cit.

[44]Hazirbas, Bitton, Dolhansky, Pan, Gordo, and Ferrer, "Towards Measuring Fairness in AI", op. cit.

[45]Zhang, Paluri, Ranzato, Darrell, and Bourdev, "PANDA: Pose aligned networks for deep attribute modeling", op. cit.

[46]Levi and Hassner, "Age and gender classification using convolutional neural networks", op. cit.

[47]Stile, Caldelli, Guerrero-Contreras, Balderas-Díaz, and Medina-Bulo, "Analysis of DF Detection through Semi-Supervised Labeling", op. cit.

[48]Guerrero-Contreras, Balderas-Díaz, García-Pascual, and Munõz, "Self-Learning Systems for Enhanced Traffic Management in Urban Settings", op. cit.

[49]Guerrero-Contreras, Balderas-Díaz, García-Pascual, and Muñoz, "Adaptive Vehicle Detection in Urban Environments", op. cit.

[50]Hazirbas, Bitton, Dolhansky, Pan, Gordo, and Ferrer, "Towards Measuring Fairness in AI", op. cit.

[51]Beckmann, Hilsmann, and Eisert, *Fooling State-of-the-Art Deepfake Detection with High-Quality Deepfakes*, op. cit.

[52]Rössler, Cozzolino, Verdoliva, Riess, Thies, and Nießner, "FaceForensics++", op. cit.

[53]Beckmann, Hilsmann, and Eisert, *Fooling State-of-the-Art Deepfake Detection with High-Quality Deepfakes*, op. cit.

the Italian multimedia forensics community[54].[55] Spain and other European countries contribute both survey work and self–learning systems[56][57].[58]

North American and industry–led teams play a central role in dataset and fairness oriented contributions, including *DFDC* and Casual Conversations[59],[60] as well as adversarial threat analyses.[61] Asian institutions contribute to early physiological and geometric cue methods and to later multimodal detection frameworks[62][63][64].[65] This geographical spread confirms that the reviewed evidence reflects both academic and industrial perspectives on deepfake detection, across different regions and regulatory contexts.

Publication list. For completeness, the 25 primary studies included in the final evidence base are reported in the following subsubsections, each presented with title, authors and abstract, and cited using the keys adopted throughout the thesis. This structured list can be used as a reference catalogue for the subsequent methodological analysis and for future replication or extension of the review.

[54]Amerini, Galteri, Caldelli, and Del Bimbo, "Deepfake Video Detection through Optical Flow Based CNN", op. cit.

[55]Guarnera et al., "The Face Deepfake Detection Challenge", op. cit.

[56]Tolosana, Vera-Rodriguez, Fierrez, Morales, and Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection", op. cit.

[57]Guerrero-Contreras, Balderas-Díaz, García-Pascual, and Munõz, "Self-Learning Systems for Enhanced Traffic Management in Urban Settings", op. cit.

[58]Guerrero-Contreras, Balderas-Díaz, García-Pascual, and Muñoz, "Adaptive Vehicle Detection in Urban Environments", op. cit.

[59]Dolhansky, Bitton, Pflaum, Lu, Howes, Wang, and Ferrer, *The DeepFake Detection Challenge (DFDC) Dataset*, op. cit.

[60]Hazirbas, Bitton, Dolhansky, Pan, Gordo, and Ferrer, "Towards Measuring Fairness in AI", op. cit.

[61]Neekhara, Dolhansky, Bitton, and Ferrer, "Adversarial Threats to DeepFake Detection", op. cit.

[62]Li, Chang, and Lyu, "In Ictu Oculi", op. cit.

[63]Yang, Li, and Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses", op. cit.

[64]Li, Yang, Sun, Qi, and Lyu, "Celeb-DF", op. cit.

[65]Anshul, Gopal, Rajan, and Chng, "Intra-modal and Cross-modal Synchronization for Audio-visual Deepfake Detection and Temporal Localization", op. cit.

2.6 Literature reference papers

Afchar et al. (2018) – MesoNet: a Compact Facial Video Forgery Detection Network[66]

Authors: Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen

Abstract. *This paper presents a method to automatically and efficiently detect face tampering in videos, and particularly focuses on two recent techniques used to generate hyper-realistic forged videos: Deepfake and Face2Face. Traditional image forensics techniques are usually not well suited to videos due to the compression that strongly degrades the data. Thus, this paper follows a deep learning approach and presents two networks, both with a low number of layers to focus on the mesoscopic properties of images. We evaluate those fast networks on both an existing dataset and a dataset we have constituted from online videos. The tests demonstrate a very successful detection rate with more than 98% for Deepfake and 95% for Face2Face.*

Comment. `MesoNet` advanced the state of the art by deliberately targeting mesoscopic cues that survive compression, showing that shallow `CNNs` can be competitive in realistic video settings without heavy backbones. For this thesis it serves as a robustness aware baseline that prioritises signals likely to persist after codec transformations. Link to *RQ1* the compact architecture is well suited to feature attribution and error analysis, which supports the interpretability goals of the thesis and failure analysis more tractable.

Amerini et al. (2019) - Deepfake Video Detection through Optical Flow based CNN[67]

Authors: Irene Amerini, Leonardo Galteri, Roberto Caldelli, Alberto Del Bimbo

Abstract. *Recent advances in visual media technology have led to new tools for*

[66] Afchar, Nozick, Yamagishi, and Echizen, "MesoNet", op. cit.

[67] Amerini, Galteri, Caldelli, and Del Bimbo, "Deepfake Video Detection through Optical Flow Based CNN", op. cit.

processing and, above all, generating multi-media contents. In particular, modern AI-based technologies have provided easy-to-use tools to create extremely realistic manipulated videos. Such synthetic videos, named Deep Fakes, may constitute a serious threat to attack the reputation of public subjects or to address the general opinion on a certain event. According to this, being able to individuate this kind of fake information becomes fundamental. In this work, a new forensic technique able to discern between fake and original video sequences is given; unlike other state-of-the-art methods which resorts at single video frames, we propose the adoption of optical flow fields to exploit possible inter-frame dissimilarities. Such a clue is then used as feature to be learned by CNN classifiers. Preliminary results obtained on FaceForensics++ dataset highlight very promising performances.

Comment. By shifting focus from spatial artifacts to inter frame motion, this work helped broaden the state of the art toward temporal reasoning. The optical flow representation operationalises the intuition that synthesis pipelines struggle with dynamics. Link to RQs: *RQ1* the explicit motion cue provides an interpretable physical signal, which complements the thesis objective of explaining why a detector succeeds or fails beyond per frame texture artefacts; *RQ2* suggests that some false positives and negatives may correlate with attributes that affect motion visibility, for example hair length or occlusions, which motivates our attribute conditioned analysis.

Anshul (2025) - Intra-modal and Cross-modal Synchronization for Audio-visual Deepfake Detection and Temporal Localization[68]

Authors: Ashutosh Anshul, Shreyas Gopal, Deepu Rajan, and Eng Siong Chng

Abstract. *Recent deepfake detection algorithms focus solely on uni-modal or cross-modal inconsistencies. While the former disregards audio-visual correspondence entirely rendering them less effective against multimodal attacks, the latter overlooks inconsistencies in a particular modality. Moreover, many*

[68] Anshul, Gopal, Rajan, and Chng, "Intra-modal and Cross-modal Synchronization for Audio-visual Deepfake Detection and Temporal Localization", op. cit.

models are single-stage supervised frameworks, effective on specific training data but less generalizable to new manipulations. To address these gaps, we propose a two-stage multimodal framework that first learns intra-modal and cross-modal temporal synchronization on real videos, capturing audio-visual correspondences crucial for deepfake detection and localization. We introduce a Gaussian-targeted loss in our pretraining model to focus on learning relative synchronization patterns across multimodal pairs. Using pretrained features, our approach not only enables classification on fully manipulated videos but also supports a localization module for partial deepfakes with only specific segments spoofed. Moreover, the pretraining stage does not require fine-tuning, thus reducing complexity. Our model, tested on various benchmark datasets, demonstrates strong generalization and precise temporal localization.

Comment. This paper advances the state of the art in multimodal deepfake detection by learning intra modal and cross modal synchronisation on real videos and by supporting temporal localisation of partially manipulated segments. It shows that correspondence learning on genuine content improves generalisation to new attacks. Link to *RQ1* the focus on synchrony provides a clear, time resolved explanation of decisions that is complementary to the thesis emphasis on interpretable visual attributes.

Beckmann et al. (2023) - Fooling State-of-the-Art Deepfake Detection with High-Quality Deepfakes[69]

Authors: Arian Beckmann, Anna Hilsmann, and Peter Eisert

Abstract. *Due to the rising threat of deepfakes to security and privacy, it is most important to develop robust and reliable detectors. In this paper, we examine the need for high-quality samples in the training datasets of such detectors. Accordingly, we show that deepfake detectors proven to generalize well on multiple research datasets still struggle in real-world scenarios with well-crafted fakes. First, we propose a*

[69]Beckmann, Hilsmann, and Eisert, *Fooling State-of-the-Art Deepfake Detection with High-Quality Deepfakes*, op. cit.

novel autoencoder for face swapping alongside an advanced face blending technique, which we utilize to generate 90 high-quality deepfakes. Second, we feed those fakes to a state-of-the-art detector, causing its performance to decrease drastically. Moreover, we fine-tune the detector on our fakes and demonstrate that they contain useful clues for the detection of manipulations. Overall, our results provide insights into the generalization of deepfake detectors and suggest that their training datasets should be complemented by high-quality fakes since training on mere research data is insufficient.

Comment. This study strengthened the state of the art by demonstrating that detectors trained on standard research corpora degrade severely on carefully crafted high quality forgeries, and by showing that fine tuning on such material can partially recover performance. It reframes generalisation as a coverage problem, where training data must include strong, realistic attacks. Link to *RQ2* the systematic failures on particular visual conditions motivate the thesis focus on correlating false positives and false negatives with explicit facial attributes, in order to disentangle quality gaps from attribute driven biases.

Dolhansky et al. (2020) - The DeepFake Detection Challenge (DFDC) Dataset[70]

Authors: Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer

Abstract. *Deepfakes are a recent off-the-shelf manipulation technique that allows anyone to swap two identities in a single video. In addition to Deepfakes, a variety of GAN-based face swapping methods have also been published with accompanying code. To counter this emerging threat, we have constructed an extremely large face swap video dataset to enable the training of detection models, and organized the accompanying DeepFake Detection Challenge (DFDC) Kaggle competition. Importantly, all recorded subjects agreed to participate in and have their likenesses*

[70]Dolhansky, Bitton, Pflaum, Lu, Howes, Wang, and Ferrer, *The DeepFake Detection Challenge (DFDC) Dataset*, op. cit.

modified during the construction of the face-swapped dataset. The DFDC dataset is by far the largest currently and publicly available face swap video dataset, with over 100,000 total clips sourced from 3,426 paid actors, produced with several Deepfake, GAN-based, and non-learned methods. In addition to describing the methods used to construct the dataset, we provide a detailed analysis of the top submissions from the Kaggle contest. We show although Deepfake detection is extremely difficult and still an unsolved problem, a Deepfake detection model trained only on the DFDC can generalize to real "in-the-wild" Deepfake videos, and such a model can be a valuable analysis tool when analyzing potentially Deepfaked videos. Training, validation and testing corpuses can be downloaded from https://ai.facebook.com/datasets/dfdc.

Comment. DFDC reshaped the state of the art by providing a very large scale, method diverse benchmark and a challenge with a hidden test set, enabling more realistic evaluation. The analysis of top submissions shows that, although some models transfer to in the wild material, generalisation remains difficult. Link to *RQ2* the dataset highlights that error rates vary substantially across content types and subjects, which motivates the thesis decision to introduce attribute labels and to quantify how specific groups contribute to false positives and false negatives.

Durall et al. (2020) - Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions[71]

Authors: Ricard Durall, Margret Keuper, and Janis Keuper **Abstract.** *Generative convolutional deep neural networks, e.g. popular GAN architectures, are relying on convolution based up-sampling methods to produce non-scalar outputs like images or video sequences. In this paper, we show that common up-sampling methods, i.e. known as up-convolution or transposed convolution, are causing the inability of such*

[71] Durall, Keuper, and Keuper, "Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions", op. cit.

models to reproduce spectral distributions of natural training data correctly. This effect is independent of the underlying architecture and we show that it can be used to easily detect generated data like deepfakes with up to 100% accuracy on public benchmarks. To overcome this drawback of current generative models, we propose to add a novel spectral regularization term to the training optimization objective. We show that this approach not only allows to train spectral consistent GANs that are avoiding high frequency errors. Also, we show that a correct approximation of the frequency spectrum has positive effects on the training stability and output quality of generative networks.

Comment. By exposing frequency spectrum mismatches introduced by common up sampling operators, this work clarified a mechanistic origin of some forensic cues and proposed spectral regularisation to correct them. It refines the state of the art from pure black box detection to generator aware understanding of artefacts. Link to *RQ1* the explicit link between spectral behaviour and detection performance supports the thesis aim of interpretability, and it reminds that analyses of attribute related errors must control for spectral and compression effects that could otherwise be confounded with semantic attributes.

Guarnera et al. (2021) - The Face Deepfake Detection Challenge[72]

Authors: Luca Guarnera, Oliver Giudice, Francesco Guarnera, Alessandro Ortis, Giovanni Puglisi, Antonino Paratore, Linh M. Q. Bui, Marco Fontani, Davide Alessandro Coccomini, Roberto Caldelli, Fabrizio Falchi, Claudio Gennaro, Nicola Messina, Giuseppe Amato, Gianpaolo Perelli, Sara Concas, Carlo Cuccu, Giulia Orrù, Gian Luca Marcialis, and Sebastiano Battiato

Abstract. *Multimedia data manipulation and forgery has never been easier than today, thanks to the power of Artificial Intelligence (AI). AI-generated fake content, commonly called Deepfakes, have been raising new issues and concerns, but also*

[72]Guarnera et al., "The Face Deepfake Detection Challenge", op. cit.

new challenges for the research community. The Deepfake detection task has become widely addressed, but unfortunately, approaches in the literature suffer from generalization issues. In this paper, the Face Deepfake Detection and Reconstruction Challenge is described. Two different tasks were proposed to the participants: (i) creating a Deepfake detector capable of working in an "in the wild" scenario; (ii) creating a method capable of reconstructing original images from Deepfakes. Real images from CelebA and FFHQ and Deepfake images created by `StarGAN`*,* `StarGAN-v2`*,* `StyleGAN`*,* `StyleGAN2`*,* `AttGAN` *and* `GDWCT` *were collected for the competition. The winning teams were chosen with respect to the highest classification accuracy value (Task I) and "minimum average distance to Manhatta" (Task II). Deep Learning algorithms, particularly those based on the EfficientNet architecture, achieved the best results in Task I. No winners were proclaimed for Task II. A detailed discussion of teams' proposed methods with corresponding ranking is presented in this paper.*

Comment. This challenge advanced the state of the art through community evaluation on diverse GAN families and through an additional reconstruction task that proved considerably harder than detection. EfficientNet based architectures achieved the best performance, yet generalisation across methods and conditions remained problematic. Link to RQs: *RQ1* the competition highlights the need for detectors whose behaviour can be interpreted and compared beyond raw accuracy; *RQ2* the reported variations across generators and subjects align with the thesis approach of analysing subgroup disparities via facial attributes and of using attribute wise metrics to characterise failure modes.

Güera and Delp (2018) - Deepfake Video Detection Using Recurrent Neural Networks[73]

Authors: David Güera and Edward J. Delp

Abstract. *In recent months a machine-learning-based free software tool has made it*

[73] Guera and Delp, "Deepfake Video Detection Using Recurrent Neural Networks", op. cit.

easy to create believable face swaps in videos that leave few traces of manipulation, in what are known as "deepfake" videos. Scenarios where these realistic fake videos are used to create political distress, blackmail someone, or fake terrorism events are easily envisioned. This paper proposes a temporal-aware pipeline to automatically detect deepfake videos. Our system uses a convolutional neural network (CNN) to extract frame-level features. These features are then used to train a recurrent neural network (RNN) that learns to classify whether a video has been subject to manipulation or not. We evaluate our method against a large set of deepfake videos collected from multiple video websites. We show how our system can achieve competitive results in this task while using a simple architecture.

Comment. As an early temporal baseline, this paper helped normalise sequence models in the state of the art, showing that recurrent aggregation over frame level CNN features improves video level classification over frame independent baselines. Link to RQs: *RQ1* the clear separation between spatial feature extraction and temporal aggregation simplifies the interpretation of temporal errors; *RQ2* the results suggest that attributes that affect motion visibility, such as hair coverage or occlusions, may influence temporal models differently, which motivates the thesis choice to correlate misclassifications with such attributes at the video level.

Guerrero-Contreras et al. (2024) - Self-Learning Systems for Enhanced Traffic Management in Urban Settings[74]

Authors: Gabriel Guerrero-Contreras, Sara Balderas-Díaz, Abel García-Pascual, and Andrés Muñoz

Abstract. *As urban populations grow, the challenges associated with managing city traffic and ensuring the smooth flow of vehicles become increasingly complex. Traditional vehicle detection systems, often reliant on extensive human supervision and manual data labeling, struggle to keep pace with the dynamic conditions of urban environments. The need for smarter, adaptive technologies that can autonomously*

[74]Guerrero-Contreras, Balderas-Díaz, García-Pascual, and Munõz, "Self-Learning Systems for Enhanced Traffic Management in Urban Settings", op. cit.

evolve and respond to changing urban landscapes is more pressing than ever. This research introduces a self-learning approach designed to enhance vehicle detection capabilities using urban camera networks. The objective is to minimize human involvement in the updating and accuracy maintenance of vehicle detection models and to allow these models to adapt autonomously to diverse urban scenarios. The methodology centers on a self-learning algorithm that utilizes initial manual data labeling followed by ongoing automatic data labeling. This approach iteratively improves vehicle detection accuracy. The algorithm was implemented and tested in a high-traffic urban setting in Madrid, using a dataset that encompasses a wide range of vehicle types. The implementation of the self-learning algorithm demonstrated a significant improvement in the accuracy of vehicle detection. The results show that the algorithm was able to effectively adapt to new vehicle types and varying traffic conditions without additional human input. The system's ability to continuously learn and adapt has potential implications for broader applications in intelligent transportation systems. Future work will explore the integration of additional data types and the extension of the algorithm to other urban applications.

Comment. Methodologically, this paper strengthens the state of the art in deployment oriented learning by reducing labelling costs through iterative self training under distribution shift. Although the domain is traffic monitoring, the underlying ideas of self learning and mixed labelled–unlabelled pipelines are transferable. Link to *RQ3* the proposed self learning framework informs the design of the thesis workflow for semi supervised facial attribute labelling and for progressively improving deepfake detectors using automatically labelled samples.

Guerrero-Contreras et al. (2025) - Adaptive Vehicle Detection in Urban Environments: A Self-learning Approach[75]

Authors: Gabriel Guerrero-Contreras, Sara Balderas-Díaz, Abel García-Pascual, and Andrés Muñoz

[75]Guerrero-Contreras, Balderas-Díaz, García-Pascual, and Muñoz, "Adaptive Vehicle Detection in Urban Environments", op. cit.

Abstract. *With increasing urbanization, efficient urban traffic management is a critical challenge that requires smarter and more adaptable systems. This paper introduces a self-learning algorithm designed to enhance the adaptability and effectiveness of vehicle detection models using urban camera infrastructures. By leveraging these ubiquitous devices, the study aims to capture and analyze real-time traffic data, a task traditionally limited by the need for extensive manual data labeling and the limitations of pre-trained models under varying urban conditions. Our self-learning algorithm addresses these challenges by reducing reliance on manual labeling and enabling continuous model adaptation to new conditions without direct human intervention. Implemented in the dynamic urban environment of the city of Madrid, Spain, this study evaluates the algorithm's capacity to enhance vehicle detection, considering a diverse range of vehicle types. The core of the algorithm comprises an iterative self-training process that refines model performance using both labeled and unlabeled data, thus progressively enhancing detection accuracy. Our findings reveal significant improvements in the ability of the model to accurately identify and classify vehicles, highlighting the potential of self-learning algorithms in urban traffic management.*

Comment. This follow up consolidates self learning for continuous adaptation in non stationary environments, showing how models can be updated with new unlabelled data while retaining performance. Link to *RQ3* the adaptive loop provides a cross domain reference for designing an operational workflow in which attribute statistics guide targeted retraining and data selection for deepfake detection, with the goal of improving robustness over time.

Hazirbas et al. (2022) - Towards Measuring Fairness in AI: The Casual Conversations Dataset[76]

Authors: Caner Hazirbas, Joanna Bitton, Brian Dolhansky, Jacqueline Pan, Albert Gordo, and Cristian Canton Ferrer

[76]Hazirbas, Bitton, Dolhansky, Pan, Gordo, and Ferrer, "Towards Measuring Fairness in AI", op. cit.

Abstract. *This paper introduces a novel dataset to help researchers evaluate their computer vision and audio models for accuracy across a diverse set of ages, genders, apparent skin tones and ambient lighting conditions. Our dataset is composed of 3,011 subjects and contains over 45,000 videos, with an average of 15 videos per person. The videos were recorded in multiple U.S. states with a diverse set of adults in various age, gender and apparent skin tone groups. A key feature is that each subject agreed to participate for their likenesses to be used. Additionally, our age and gender annotations are provided by the subjects themselves. A group of trained annotators labeled the subjects' apparent skin tone using the Fitzpatrick skin type scale. Moreover, annotations for videos recorded in low ambient lighting are also provided. As an application to measure robustness of predictions across certain attributes, we provide a comprehensive study on the top five winners of the DeepFake Detection Challenge (DFDC). Experimental evaluation shows that the winning models are less performant on some specific groups of people, such as subjects with darker skin tones, and thus may not generalize to all people. In addition, we also evaluate state-of-the-art apparent age and gender classification methods. Our experiments provide a thorough analysis on these models in terms of fair treatment of people from various backgrounds.*

Comment. By enabling fairness analysis across ages, genders, apparent skin tones, and lighting, this dataset moved the state of the art from aggregate accuracy to disaggregated robustness. The study documents performance drops on darker skin tones for top DFDC models and provides concrete metrics for demographic bias. Link to *RQ3* the work directly motivates the thesis proposal of attribute aware evaluation protocols and balanced sampling strategies, showing how structured attribute labels can be used to design workflows that monitor and mitigate disparities in deepfake detection.

Katamneni and Rattani (2024) - Contextual Cross-Modal Attention for Audio-Visual Deepfake Detection and Localization[77]

Authors: Vinaya Sree Katamneni and Ajita Rattani

Abstract. *In the digital age, the emergence of deepfakes and synthetic media presents a significant threat to societal and political integrity. Deepfakes based on multi-modal manipulation, such as audio-visual, are more realistic and pose a greater threat. Current multi-modal deepfake detectors are often based on the attention-based fusion of heterogeneous data streams from multiple modalities. However, the heterogeneous nature of the data (such as audio and visual signals) creates a distributional modality gap and poses a significant challenge in effective fusion and hence multi-modal deepfake detection. In this paper, we propose a novel multi-modal attention framework based on recurrent neural networks (RNNs) that leverages contextual information for audio-visual deepfake detection. The proposed approach applies attention to multi-modal multi-sequence representations and learns the contributing features among them for deepfake detection and localization. Thorough experimental validations on audio-visual deepfake datasets, namely FakeAVCeleb, AV-Deepfake1M, TVIL, and LAV-DF datasets, demonstrate the efficacy of our approach. Cross-comparison with the published studies demonstrates superior performance of our approach with an improved accuracy and precision by 3.47% and 2.05% in deepfake detection and localization, respectively. Thus, obtaining state-of-the-art performance. To facilitate reproducibility, the code and the datasets information is available at https://github.com/vcbsl/audiovisual-deepfake/.*

Comment. The use of contextual attention over audio visual sequences moves the state of the art beyond naive fusion and improves both detection and localisation on several multimodal benchmarks. The attention mechanism highlights which temporal segments and modalities contribute most to the decision. Link to *RQ1* this aligns with the thesis emphasis on interpretability, since attention weights can be inspected alongside visual attribute information to obtain richer, more transparent explanations

[77] Katamneni and Rattani, *Contextual Cross-Modal Attention for Audio-Visual Deepfake Detection and Localization*, op. cit.

of deepfake decisions.

Korshunov and Marcel (2018) - DeepFakes: a New Threat to Face Recognition? Assessment and Detection[78]

Authors: Pavel Korshunov and Sébastien Marcel

Abstract. *It is becoming increasingly easy to automatically replace a face of one person in a video with the face of another person by using a pre-trained generative adversarial network (GAN). Recent public scandals, e.g., the faces of celebrities being swapped onto pornographic videos, call for automated ways to detect these Deepfake videos. To help developing such methods, in this paper, we present the first publicly available set of Deepfake videos generated from videos of VidTIMIT database. We used open source software based on GANs to create the Deepfakes, and we emphasize that training and blending parameters can significantly impact the quality of the resulted videos. To demonstrate this impact, we generated videos with low and high visual quality (320 videos each) using differently tuned parameter sets. We showed that the state of the art face recognition systems based on VGG and Facenet neural networks are vulnerable to Deepfake videos, with 85.62% and 95.00% false acceptance rates respectively, which means methods for detecting Deepfake videos are necessary. By considering several baseline approaches, we found that audio-visual approach based on lip-sync inconsistency detection was not able to distinguish Deepfake videos. The best performing method, which is based on visual quality metrics and is often used in presentation attack detection domain, resulted in 8.97% equal error rate on high quality Deepfakes. Our experiments demonstrate that GAN-generated Deepfake videos are challenging for both face recognition systems and existing detection methods, and the further development of face swapping technology will make it even more so.*

Comment. This early risk assessment shaped the state of the art by quantifying how GAN based deepfakes can fool face recognition systems and by testing several

[78] Korshunov and Marcel, *DeepFakes*, op. cit.

baseline detectors, including an audio visual lip sync approach that proved ineffective on high quality forgeries. Link to RQs: *RQ1* it underscores the need for detectors that offer interpretable signals about failure modes rather than only scores; *RQ2* the sensitivity to blending quality and synthesis parameters motivates the thesis decision to examine whether specific facial attributes, such as occlusions or pose, systematically correlate with misclassifications.

Levi and Hassner (2015) - Age and gender classification using convolutional neural networks[79]

Authors: Gil Levi and Tal Hassner

Abstract. *Automatic age and gender classification has become relevant to an increasing amount of applications, particularly since the rise of social platforms and social media. Nevertheless, performance of existing methods on real-world images is still significantly lacking, especially when compared to the tremendous leaps in performance recently reported for the related task of face recognition. In this paper we show that by learning representations through the use of deep-convolutional neural networks (CNN), a significant increase in performance can be obtained on these tasks. To this end, we propose a simple convolutional net architecture that can be used even when the amount of learning data is limited. We evaluate our method on the recent Adience benchmark for age and gender estimation and show it to dramatically outperform current state-of-the-art methods.*

Comment. This paper advanced attribute prediction with compact CNNs under limited training data, providing a practical state of the art for age and gender estimation in unconstrained images. Link to *RQ3* the proposed architecture and training strategy offer concrete guidance for building reliable attribute classifiers, which in the thesis are integrated into a workflow where attribute labels support bias analysis and the design of attribute aware training and evaluation procedures for deepfake detection.

[79] Levi and Hassner, "Age and gender classification using convolutional neural networks", op. cit.

Li et al. (2018) - In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking[80]

Authors: Yuezun Li, Ming-Ching Chang, and Siwei Lyu

Abstract. *The new developments in deep generative networks have significantly improved the quality and efficiency in generating realistically looking fake face videos. In this work, we describe a new method to expose fake face videos generated with deep neural network models. Our method is based on detection of eye blinking in the videos, which is a physiological signal that is not well presented in the synthesized fake videos. Our method is evaluated over benchmarks of eye-blinking detection datasets and shows promising performance on detecting videos generated with DNN-based software DeepFake.*

Comment. Detecting missing or unrealistic eye blinks provided a simple physiological cue for early deepfake detection and influenced several first generation pipelines. As generators improved, the brittleness of this cue became apparent, illustrating the lifecycle of targeted forensic artefacts. Link to *RQ1* the work exemplifies human understandable explanations for decisions, and it motivates the thesis choice to move beyond single, fragile cues towards more structured, attribute based interpretations of model behaviour.

Li et al. (2020) - Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics[81]

Authors: Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu

Abstract. *AI-synthesized face-swapping videos, commonly known as DeepFakes, is an emerging problem threatening the trustworthiness of online information. The need to develop and evaluate DeepFake detection algorithms calls for datasets of DeepFake videos. However, current DeepFake datasets suffer from low visual quality and do not resemble DeepFake videos circulated on the Internet. We present a new*

[80] Li, Chang, and Lyu, "In Ictu Oculi", op. cit.
[81] Li, Yang, Sun, Qi, and Lyu, "Celeb-DF", op. cit.

large-scale challenging DeepFake video dataset, Celeb-DF, which contains 5,639 high-quality DeepFake videos of celebrities generated using improved synthesis process. We conduct a comprehensive evaluation of DeepFake detection methods and datasets to demonstrate the escalated level of challenges posed by Celeb-DF.

Comment. Celeb DF raised the state of the art in benchmarking by curating higher quality, internet like swaps that expose generalisation gaps in detectors trained on earlier corpora. Models that perform well on legacy datasets often degrade substantially on Celeb DF. Link to *RQ2* the dataset provides a challenging testbed for analysing how attribute related error patterns change with visual quality, which is essential for assessing whether attribute aware methods learnt on one corpus transfer to more realistic content.

Neekhara et al. (2021) - Adversarial Threats to DeepFake Detection: A Practical Perspective[82]

Authors: Paarth Neekhara, Brian Dolhansky, Joanna Bitton, and Cristian Canton Ferrer

Abstract. *Facially manipulated images and videos or DeepFakes can be used maliciously to fuel misinformation or defame individuals. Therefore, detecting DeepFakes is crucial to increase the credibility of social media platforms and other media sharing web sites. State-of-the art DeepFake detection techniques rely on neural network based classification models which are known to be vulnerable to adversarial examples. In this work, we study the vulnerabilities of state-of-the-art DeepFake detection methods from a practical stand point. We perform adversarial attacks on DeepFake detectors in a black box setting where the adversary does not have complete knowledge of the classification models. We study the extent to which adversarial perturbations transfer across different models and propose techniques to improve the transferability of adversarial examples. We also create more accessible attacks using Universal Adversarial Perturbations which pose a*

[82]Neekhara, Dolhansky, Bitton, and Ferrer, "Adversarial Threats to DeepFake Detection", op. cit.

very feasible attack scenario since they can be easily shared amongst attackers. We perform our evaluations on the winning entries of the DeepFake Detection Challenge (DFDC) and demonstrate that they can be easily bypassed in a practical attack scenario by designing transferable and accessible adversarial attacks.

Comment. By demonstrating black box and universal adversarial perturbations that transfer across DFDC winning models, this paper added a security perspective to the state of the art in deepfake detection. Link to RQs: *RQ1* it highlights the need for calibrated, uncertainty aware detectors whose explanations can reveal when inputs are atypical; *RQ2* the analysis suggests that adversarially induced errors may interact with particular visual conditions, which motivates the thesis plan to track false positives and false negatives by attribute group, including under synthetic perturbations.

Rana et al. (2022) - Deepfake Detection: A Systematic Literature Review[83]

Authors: Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, and Andrew H. Sung

Abstract. *Over the last few decades, rapid progress in AI, machine learning, and deep learning has resulted in new techniques and various tools for manipulating multimedia. Though the technology has been mostly used in legitimate applications such as for entertainment and education, etc., malicious users have also exploited them for unlawful or nefarious purposes. For example, high-quality and realistic fake videos, images, or audios have been created to spread misinformation and propaganda, foment political discord and hate, or even harass and blackmail people. The manipulated, high-quality and realistic videos have become known recently as Deepfake. Various approaches have since been described in the literature to deal with the problems raised by Deepfake. To provide an updated overview of the research works in Deepfake detection, we conduct a systematic literature review (SLR) in*

[83] Rana, Nobi, Murali, and Sung, "Deepfake Detection", op. cit.

this paper, summarizing 112 relevant articles from 2018 to 2020 that presented a variety of methodologies. We analyze them by grouping them into four different categories: deep learning-based techniques, classical machine learning-based methods, statistical techniques, and blockchain-based techniques. We also evaluate the performance of the detection capability of the various methods with respect to different datasets and conclude that the deep learning-based methods outperform other methods in Deepfake detection.

Comment. This survey consolidates the state of the art up to 2020, organising methods into deep learning, classical machine learning, statistical, and blockchain based approaches, and reviewing datasets and metrics. It also notes fragmentation in evaluation protocols. Link to RQs: *RQ1* it supports the thesis focus on interpretable analyses, since current categories often mix heterogeneous cues; *RQ2* it motivates an attribute centred perspective that cuts across method families and allows a more granular view of error patterns.

Rössler et al. (2019) - FaceForensics++: Learning to Detect Manipulated Facial Images[84]

Authors: Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner

Abstract. *The rapid progress in synthetic image generation and manipulation has now come to a point where it raises significant concerns for the implications towards society. At best, this leads to a loss of trust in digital content, but could potentially cause further harm by spreading false information or fake news. This paper examines the realism of state-of-the-art image manipulations, and how difficult it is to detect them, either automatically or by humans. To standardize the evaluation of detection methods, wepropose an automated benchmark for facial manipulation detection. In particular, the benchmark is based on DeepFakes, Face2Face [59], FaceSwap and NeuralTextures as prominent representatives for facial manipulations at random*

[84]Rössler, Cozzolino, Verdoliva, Riess, Thies, and Nießner, "FaceForensics++", op. cit.

compression level and size. The benchmark is publicly available and contains a hidden test set as well as a database of over 1.8 million manipulated images. This dataset is over an order of magnitude larger than comparable, publicly available, forgery datasets. Based on this data, we performed a thorough analysis of data-driven forgery detectors. We show that the use of additional domain-specific knowledge improves forgery detection to unprecedented accuracy, even in the presence of strong compression, and clearly outperforms human observers.

Comment. FaceForensics++ shaped the state of the art by establishing a widely used benchmark with multiple manipulation families, controlled compression levels, and a large number of samples. It enabled more systematic comparison of detectors and protocols. Link to *RQ2* FF++ is the core dataset used in this thesis for attribute labelling and bias analysis, and it provides the basis for measuring which facial attributes are most strongly associated with false positives and false negatives under different training conditions.

Sabir et al. (2019) - Recurrent convolutional strategies for face manipulation detection in videos[85]

Authors: Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan

Abstract. *The spread of misinformation through synthetically generated yet realistic images and videos has become a significant problem, calling for robust manipulation detection methods. Despite the predominant effort of detecting face manipulation in still images, less attention has been paid to the identification of tampered faces in videos by taking advantage of the temporal information present in the stream. Recurrent convolutional models are a class of deep learning models which have proven effective at exploiting the temporal information from image streams across domains. We thereby distill the best strategy for combining variations in these models along with domain specific face preprocessing techniques through extensive experimentation*

[85] Sabir, Cheng, Jaiswal, AbdAlmageed, Masi, and Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos", op. cit.

to obtain state-of-the-art performance on publicly available video-based facial manipulation benchmarks. Specifically, we attempt to detect Deepfake, Face2Face and FaceSwap tampered faces in video streams. Evaluation is performed on the recently introduced FaceForensics++ dataset, improving the previous state-of-the-art by up to 4.55% in accuracy.

Comment. This paper refined the state of the art by systematically combining face centric preprocessing with CNN plus RNN variants for video based detection, documenting gains over purely image based baselines on FF++. Link to RQs: *RQ1* the explicit spatio temporal architecture facilitates analysis of how temporal cues contribute to decisions; *RQ2* the results motivate the thesis choice to compare attribute related error patterns across models that use only frame level information and models that incorporate temporal aggregation.

Tolosana et al. (2020) - Deepfakes and beyond: A survey of face manipulation and fake detection[86]

Authors: Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia

Abstract. *The free access to large-scale public databases, together with the fast progress of deep learning techniques, in particular Generative Adversarial Networks, have led to the generation of very realistic fake content with its corresponding implications towards society in this era of fake news. This survey provides a thorough review of techniques for manipulating face images including DeepFake methods, and methods to detect such manipulations. In particular, four types of facial manipulation are reviewed: i) entire face synthesis, ii) identity swap (DeepFakes), iii) attribute manipulation, and iv) expression swap. For each manipulation group, we provide details regarding manipulation techniques, existing public databases, and key benchmarks for technology evaluation of fake detection methods, including a summary of results from those evaluations. Among all the aspects discussed in the*

[86]Tolosana, Vera-Rodriguez, Fierrez, Morales, and Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection", op. cit.

survey, we pay special attention to the latest generation of DeepFakes, highlighting its improvements and challenges for fake detection. In addition to the survey information, we also discuss open issues and future trends that should be considered to advance in the field.

Comment. This survey updated the state of the art taxonomy across manipulation families, datasets, and detectors, and articulated key open challenges such as generalisation to unseen conditions. Link to RQs: *RQ1* it supports an attribute aware interpretability agenda that spans different types of face manipulation; *RQ2* it informs the choice of facial attributes that are likely to matter for detector behaviour, helping to position the thesis contributions within the broader landscape of open problems.

Verdoliva (2020) - Forensics and DeepFakes: An Overview[87]

Authors: Luisa Verdoliva

Abstract. *With the rapid progress in recent years, techniques that generate and manipulate multimedia content can now provide a very advanced level of realism. The boundary between real and synthetic media has become very thin. On the one hand, this opens the door to a series of exciting applications in different fields such as creative arts, advertising, film production, and video games. On the other hand, it poses enormous security threats. Software packages freely available on the web allow any individual, without special skills, to create very realistic fake images and videos. These can be used to manipulate public opinion during elections, commit fraud, discredit or blackmail people. Therefore, there is an urgent need for automated tools capable of detecting false multimedia content and avoiding the spread of dangerous false information. This review paper aims to present an analysis of the methods for visual media integrity verification, that is, the detection of manipulated images and videos. Special emphasis will be placed on the emerging phenomenon of deepfakes, fake media created through deep learning tools, and on modern data-driven forensic methods to fight them. The analysis will help highlight the limits of current forensic*

[87] Verdoliva, "Media Forensics and DeepFakes", op. cit.

tools, the most relevant issues, the upcoming challenges, and suggest future directions for research.

Comment. This overview positioned deepfake detection within the broader field of multimedia forensics and highlighted both technical advances and societal risks. It synthesised methods for media integrity verification and emphasised the limits of current tools. Link to RQs: *RQ1* it aligns with the thesis push toward more transparent and accountable detectors; *RQ2* it motivates the analysis of generalisation gaps and bias, which in this work are approached via facial attribute labelling and attribute conditioned evaluation.

Yang et al. (2019) - Exposing Deep Fakes Using Inconsistent Head Poses[88]

Authors: Xin Yang, Yuezun Li, and Siwei Lyu

Abstract. *In this paper, we propose a new method to expose AI-generated fake face images or videos (commonly known as the Deep Fakes). Our method is based on the observations that Deep Fakes are created by splicing synthesized face region into the original image, and in doing so, introducing errors that can be revealed when 3D head poses are estimated from the face images. We perform experiments to demonstrate this phenomenon and further develop a classification method based on this cue. Using features based on this cue, an SVM classifier is evaluated using a set of real face images and Deep Fakes.*

Comment. Head pose inconsistency provided a clean geometric cue and influenced several early detection tools, especially for lower quality forgeries generated by splicing. Link to *RQ1* the method offers an interpretable mechanism that can be explained to users and analysts, and it exemplifies how physically meaningful features can be linked to model decisions, a principle that the thesis extends to semantic facial attributes.

[88] Yang, Li, and Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses", op. cit.

Zhang et al. (2014) - PANDA: Pose aligned networks for deep attribute modeling[89]

Authors: Ning Zhang, Manohar Paluri, Marc'Aurelio Ranzato, Trevor Darrell, and Lubomir Bourdev

Abstract. *We propose a method for inferring human attributes (such as gender, hair style, clothes style, expression, action) from images of people under large variation of viewpoint, pose, appearance, articulation and occlusion. Convolutional Neural Nets (CNN) have been shown to perform very well on large scale object recognition problems. In the context of attribute classification, however, the signal is often subtle and it may cover only a small part of the image, while the image is dominated by the effects of pose and viewpoint. Discounting for pose variation would require training on very large labeled datasets which are not presently available. Part-based models, such as poselets and DPM have been shown to perform well for this problem but they are limited by shallow low-level features. We propose a new method which combines part-based models and deep learning by training pose-normalized CNNs. We show substantial improvement vs. state-of-the-art methods on challenging attribute classification tasks in unconstrained settings. Experiments confirm that our method outperforms both the best part-based methods on this problem and conventional CNNs trained on the full bounding box of the person.*

Comment. PANDA advanced the state of the art in attribute prediction by combining pose alignment with CNNs under strong viewpoint and pose variability, improving robustness for attribute classification in the wild. Link to *RQ3* the pose aligned attribute learning strategy provides a methodological basis for building reliable attribute estimators, which in the thesis are integrated in an attribute aware workflow aimed at improving the performance and robustness of deepfake detectors through better supervision and bias analysis.

[89] Zhang, Paluri, Ranzato, Darrell, and Bourdev, "PANDA: Pose aligned networks for deep attribute modeling", op. cit.

Table 2.5: Evidence map linking included works to research questions and thesis focus.

Citation	Level	Dataset(s)	RQ	Notes / Rationale
Zhang et al. (2014)	Frame	Multiple	RQ3	Pose-aligned attribute learning; supports attribute supervision block.
Levi and Hassner (2015)	Frame	Adience	RQ3	Foundational facial modeling for pipeline design based on gender attribute.
Afchar et al. (2018)	Frame	FF++ (edited)	RQ1	Compact CNN baseline for face forgery.
Güera and Delp (2018)	Video	FF++ (edited)	RQ1, RQ2	Temporal modeling from frame sequences.
Li et al. (2018)	Frame	VidTIMIT	RQ1	Eye-blinking inconsistency detection (physiological cue).
Korshunov and Marcel (2018)	Video	VidTIMIT	RQ1, RQ2	Early baselines; AV lip-sync limits.
Rössler et al. (2019)	n/a	FaceForensics++	RQ2	Core benchmark used in thesis; seed for snowballing.
Sabir et al. (2019)	Video	FF++	RQ1, RQ2	Spatio–temporal pipeline for videos.
Yang et al. (2019)	Frame	UADFV	RQ1	Inconsistent head pose features (geometric cue).
Amerini et al. (2019)	Video	FF++	RQ1, RQ2	Motion/optical-flow based artifacts.
Li et al. (2020)	n/a	Celeb-DF	RQ2	High-quality fakes increase difficulty; cross-dataset discussion.
Verdoliva (2020)	n/a	Multiple	RQ1, RQ2	Field overview framing method families and protocols.

Continued on next page

Table 2.5 — continued from previous page

Paper	Level	Dataset(s)	RQ(s)	Notes / Rationale
Tolosana et al. (2020)	n/a	Multiple	RQ1, RQ2	Taxonomy of manipulations and detectors.
Dolhansky et al. (2020)	n/a	DFDC	RQ2	Diversity and baseline metrics; seed paper.
Durall et al. (2020)	Frame	FF++	RQ1	Frequency-aware detector focused on spectral artifacts.
Neekhara et al. (2021)	Video	DFDC (edited)	RQ1, RQ2	Cross-modal inconsistency detection (audio–visual).
Rana et al. (2022)	n/a	Multiple	RQ1, RQ2	Methods, datasets, and metrics synthesis.
Guarnera et al. (2022)	Video	Multiple	RQ1, RQ2	Benchmark tasks, winning approaches, and insights on generalization; useful for framing methods and metrics.
Hazirbas et al. (2022)	Video	DFDC	RQ3	Fairness analysis for face-forgery research, quantifies demographic bias and recommends balanced sampling and evaluation protocols.
Beckmann et al. (2023)	Frame	Custom	RQ2	Detectors fail on High Quality fakes; fine-tuning effects.
Guerrero et al. (2024)	Frame	Custom	RQ3	Useful as methodological background on self-learning pipelines.
Katamneni and Rattani (2024)	Video	Multiple	RQ1	Cross-modal detection via audio-visual consistency, reports gains in generalization and stability.

Continued on next page

Table 2.5 — continued from previous page

Paper	Level	Dataset(s)	RQ(s)	Notes / Rationale
Anshul et al. (2025)	Video	Multiple	RQ1	Intra-modal deepfake detector with spatio-temporal cues, shows better cross-dataset robustness.
Guerrero et al. (2025)	Frame	Custom	RQ3	Included as cross-domain reference for self-learning and automatic labeling ideas.

Chapter 3

Industry activities in PricewaterhouseCoopers Business Services Italia S.r.l.

This chapter reports the industry activities carried out within the Data Science & Artificial Intelligence (AI) team at PricewaterhouseCoopers Business Services Italia S.r.l. (PWC), Rome site, in the framework of the industrial Ph.D. on DeepFake recognition. The placement started on 07/06/2023 and progressed through a set of incremental machine–learning projects planned to consolidate tooling, data handling, and evaluation practices, then to converge toward the target application domain of video manipulation detection. The formal scope and timeline of the first period, as well as the hosting entity and project title, are documented in the internal report and project sheets, which also summarize the tasks executed and the software stack adopted, primarily Python with deep–learning frameworks (`TensorFlow`, `PyTorch`).

Operational planning relied on periodic alignment meetings with the company Co–Supervisor Dr. Elena Santi (Director, Data Science & AI Team Lead), with Dr. Matteo Ciacagliani (Senior Data Scientist) for coordination and onboarding, with Dr. Giovanni Sansaro (Data Scientist) for technical feedback on prototypes, and with Dr. Alberto Grassi (Senior Data Scientist) for the orchestration of knowledge–sharing activities. Intermediate results were disseminated in the internal *Tech Wednesday* seminars, where the Modified National Institute of Standards and

Technology (MNIST) digit classifier was presented on 18/10/2023 and the Olivetti face–classification experiment on 20/03/2024, each followed by discussion and Q&A. These sessions served both as milestones and as internal validation of reproducibility and clarity of the adopted pipelines. These projects are described in detail in Section 3.2 and Section 3.3 and the entire commented code is available on the author's GitHub profile [1].

The work plan followed a progressive structure. First, binary image classification on a custom two–class dataset was implemented and presented, then a supervised benchmark on MNIST was developed to exercise end–to–end data ingestion, model training and evaluation. A face–classification task on the Olivetti dataset was subsequently completed to explore higher–dimensional patterns and subject–level evaluation. Finally, DeepFake detection pipelines were explored using face–cropped videos and temporal modeling. All project details, including data collection, preprocessing, and the CNN + LSTM sequence-classification pipeline, are documented extensively and in full compliance with privacy and corporate policies. This documentation is maintained in private project notes and Git repositories, and has been disseminated via restricted channels and the periodic reports required by the university. Early alignment materials from the 06/09/2023 meeting with PWC confirm the project breakdown and record the use of `TensorFlow` utilities for image pipelines and augmentation in the binary classifier prototype, which were later reused across experiments to mitigate overfitting and to standardize. The overarching research aim, namely the study of state–of–the–art computer–vision methods for DeepFake recognition with a view to practical evaluation in enterprise contexts, is specified in the project brief and underpins the transition from didactic prototypes to application–driven investigations.

3.1 Binary classification: Jaguar vs Capybara

Before addressing DeepFake detection directly, the doctoral work included a set of preparatory experiments on simpler visual tasks. These activities were designed to

[1] The author's GitHub profile: https://github.com/vstile (visited on 26 December 2025).

build practical familiarity with deep learning frameworks, transfer learning workflows, and basic good practices for data preprocessing, augmentation, and evaluation, in a controlled setting with limited scope. The classifier presented here is one such exercise, used to validate an end to end pipeline that could later be adapted to more complex forensic problems. This activity explored a lightweight image classifier for two visually distinct classes, *jaguar* vs *capybara*, using transfer learning with `MobileNetV2` in `TensorFlow`. The objective was to validate a pragmatic pipeline that combines a compact pre-trained backbone with minimal task-specific fine-tuning and standard image augmentation, suitable for small custom datasets and rapid iteration in enterprise settings. The public project repository documents the task definition, the transfer-learning setup, and the augmentation-driven training strategy for binary classification.

3.1.1 Dataset and preprocessing

A custom two–class image corpus, distributed by the author on GitHub as `dataset.zip` and organized in class-labelled directories, was employed.[2]. Basic normalization was applied, and standard data-augmentation operations were enabled during training to increase effective diversity without altering class semantics. To clarify dataset characteristics, training dynamics, and validation behavior, exemplar images are included as shown in Figure 3.1 and an augmentation mosaic is shown in Figure 3.2.

3.1.2 Project development

The classifier adopts `MobileNetV2` pre-trained on `ImageNet` as fixed or partially unfrozen feature extractor, with a task-specific dense head for two-way classification. Fine-tuning proceeds after a brief warm-up with the backbone frozen, then selectively unfreezes upper blocks to align high-level representations to domain statistics. The loss is binary cross-entropy optimized with Adam; early stopping monitors validation

[2]Package: `dataset.zip` (size = 85.7 MB), MD5: `70b2dd7d3d65099176ed51eaa8e2391a`; public repository created by the author: https://github.com/vstile/01-binary-classification (visited on 17 December 2025). These metadata support reproducibility and pre-training integrity checks. Before training, images were decoded and resized to the input resolution required by the chosen backbone.

(a) *Capybara* exemplar **(b)** *Jaguar* exemplar

Figure 3.1: Class exemplars from the custom corpus used for transfer learning. Images are representative of the typical appearance, background clutter, and scale variance encountered in the dataset.

performance to prevent overfitting. The training workflow is implemented in a Jupyter notebook included in the repository, together with configuration and augmentation routines for training and validation streams.

Evaluation follows a standard hold-out split, reporting accuracy on the validation set and monitoring loss and accuracy curves for signs of divergence. As documented in the project description, the trained model achieves high accuracy on the task, confirming that transfer learning with a lightweight backbone is sufficient for cleanly separable visual categories under moderate intra-class variability. Error inspection indicates that failure cases typically involve low-contrast crops or atypical poses, which can be mitigated by targeted augmentation or by modestly unfreezing deeper layers during fine-tuning.

3.2 Digit classification on the MNIST dataset

This section presents an educational baseline for digit classification on the *MNIST* dataset, developed at PWC to demonstrate an end–to–end deep learning workflow with minimal complexity. The goal is to train a compact Convolutional Neural Network (CNN) in `Keras/TensorFlow`, starting from canonical 28×28 grayscale images of handwritten digits, and to document a clean protocol for preprocessing, model selection, and reporting on the standard train and test partitions. *MNIST* is a widely adopted benchmark for teaching and testing pattern recognition pipelines; we

Figure 3.2: Augmented training samples for the Jaguar vs Capybara binary classifier. The grid shows examples generated on the fly by Keras `ImageDataGenerator`: rotations up to ±10°, horizontal and vertical translations up to 15%, shear up to 5°, and zoom in the [0.7, 1.3] range, followed by MobileNetV2 preprocessing. Augmentation is applied only to the training split to improve invariance and reduce overfitting, while validation uses only normalization.

therefore use it to illustrate reproducible training with fixed seeds, saved weights, and simple visualization of predictions. The focus is didactic rather than competitive performance, and the resulting template provides a lightweight reference for later applications on more challenging data.

3.2.1 Dataset and preprocessing

The Modified National Institute of Standards and Technology (MNIST) database contains 70,000 grayscale images of handwritten digits, with 60,000 samples for training and 10,000 for testing. Each image is a 28×28 single–channel raster and is labeled with an integer class in $\{0, \ldots, 9\}$. In our implementation the dataset was loaded from `keras.datasets.mnist`, cast to `float32`, and scaled to $[0, 1]$ by division by 255. A validation split of 10% was carved out from the training portion to monitor generalization during optimization. These settings follow the instructional material used internally at PWC to introduce supervised image classification and benchmark procedures on MNIST.

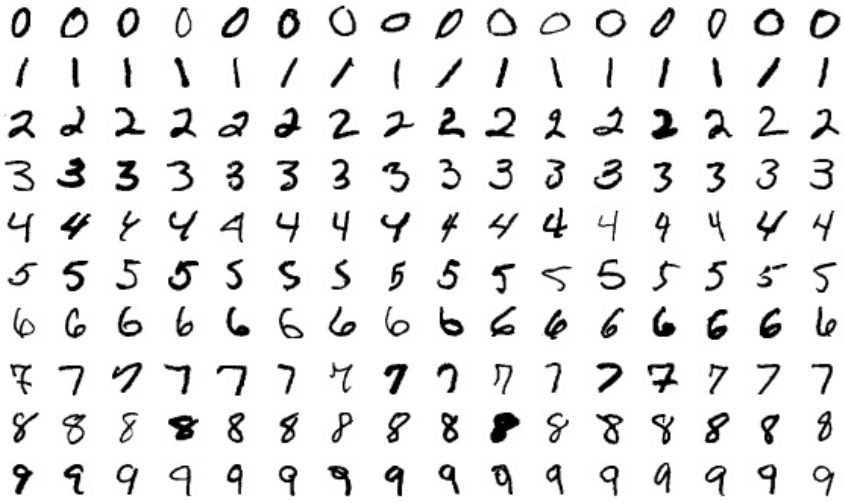
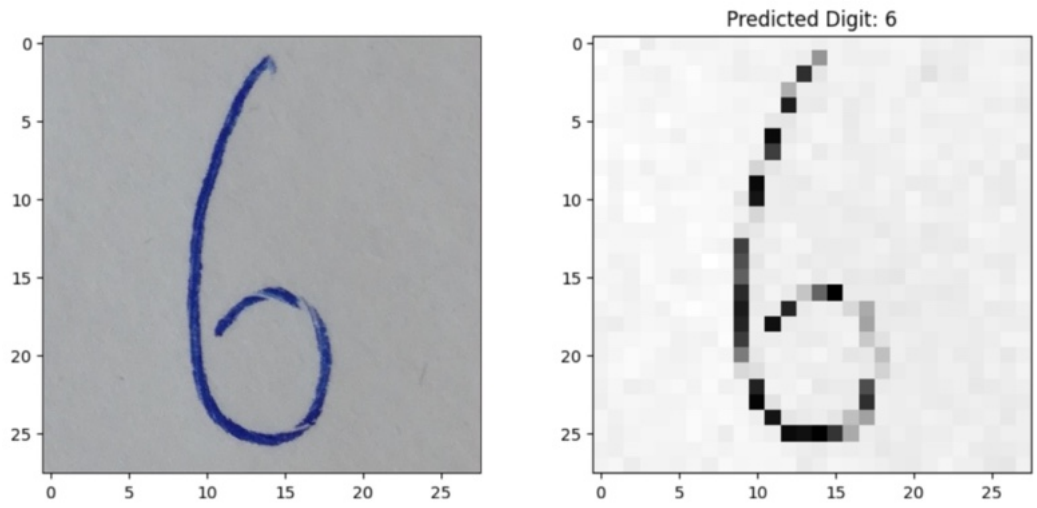


Figure 3.3: Examples from the MNIST handwritten digit dataset (28×28 grayscale). Rows depict instances of digit classes 0–9, illustrating intra-class variability in stroke thickness, slant, and shape after size normalization and centering.

3.2.2 Project development

A compact convolutional network was built with `keras.Sequential` to keep the focus on the end-to-end pipeline. The model accepts $28 \times 28 \times 1$ inputs and stacks two Conv2D blocks with 3×3 kernels (32 and 64 filters, ReLU activations), each followed by 2×2 max-pooling, then a `Flatten` layer and two dense layers (128 units with ReLU, and a 10-way softmax head). The network was compiled with the Adam optimizer, `sparse_categorical_crossentropy` loss, and the `accuracy` metric. Training ran for 10 epochs with mini-batches of 128 and a validation split of 0.1; after convergence the model achieved about 99% test accuracy with low cross-entropy loss, and a simple inference demo on a user-drawn digit confirmed correct classification after the same preprocessing used at training time (grayscale conversion and resize to 28×28). These choices mirror the didactic example presented in the internal slide deck and code excerpts shared with attendees. Our findings are consistent with prior reports on MNIST: a compact CNN with two convolutional layers (kernel 5×5), two 2×2 pooling layers, one fully connected layer, and a softmax output—trained end-to-end on MNIST—can reach 100% training accuracy and $\sim$ 99.25% test accuracy, confirming that even shallow architectures

saturate this benchmark.[3]



(a) *Original image* exemplar **(b)** *Model prediction* exemplar

Figure 3.4: Example of handwritten digit recognition. The figure shows an original handwritten sample of the digit "6" together with the corresponding model output whose predicted label is also "6". The classifier correctly recognizes the digit, illustrating the end-to-end preprocessing and inference on 28×28 grayscale input.

The MNIST exercise was presented at *Tech Wednesday* on 18 October 2023 at PWC as an introductory, hands-on illustration of a modern vision workflow: loading a standard benchmark, defining a minimal CNN, training with proper validation, evaluating with accuracy/loss, and testing on custom inputs. The session emphasized both the advantages of MNIST for rapid prototyping and pedagogy, and its known limitations—restricted visual diversity, relatively neat handwriting samples, and the risk of overfitting that makes high accuracy on MNIST an insufficient indicator of performance on more complex, real-world problems. Our results align with established baselines on MNIST, where shallow CNNs already saturate performance (up to 100% train and $\sim$ 99.25% test accuracy).[4] The live demonstration and discussion attracted significant interest and questions from colleagues regarding architecture choices, evaluation practice, and pathways from this baseline toward harder datasets and tasks. As a companion to this work, the author released the full

[3] Y. Gong and P. Zhang, *Research on Mnist Handwritten Numbers Recognition based on CNN*, in: *Journal of Physics: Conference Series* 2138.1 (Dec. 2021), p. 012002, ISSN: 1742-6588, 1742-6596, DOI: 10.1088/1742-6596/2138/1/012002, URL: https://iopscience.iop.org/article/10.1088/1742-6596/2138/1/012002 (visited on 12/16/2025).

[4] Ibid.

code and notebook in a public GitHub repository [5].

3.3 Face classification on the Olivetti dataset

This section reports an internal exploratory study on face identification using the Olivetti dataset, carried out within PWC to exercise a complete classical vision pipeline on a small, well controlled benchmark. The objective is to establish a transparent baseline that covers data loading, preprocessing, model training, and evaluation with reproducible seeds and a clear split between training, validation, and test. The Olivetti collection provides low–resolution grayscale faces with limited intra–class variability, which makes it suitable for rapid iteration and for assessing the effect of simple feature extraction and shallow classifiers before moving to more complex settings. The emphasis is educational and methodological, not on state–of–the–art accuracy: the study serves to consolidate best practices for data hygiene, cross–validation, and error inspection that are reused in subsequent projects.

3.3.1 Dataset and preprocessing

The Olivetti Faces collection comprises 400 grayscale face images from 40 distinct subjects (ten images per subject) as shown in the Figure 3.5, acquired between April 1992 and April 1994 with variations in illumination, facial expression, and detail [6]. Each sample is a 64×64 pixel array on a black background. In this project, pixel intensities were scaled to the $[0, 1]$ interval and each image was flattened to a 4096-dimensional vector. The corpus was then partitioned into training and test sets with a 75/25 split using a fixed random seed (`random_state=42`) to guarantee reproducibility and to keep the subject distribution uniform across splits.

[5] Public GitHub Repository created by the author: https://github.com/vstile/02-digit-classification-mnist (visited on 17 December 2025). The repository includes the Jupyter notebook, reproducibility seeds, saved weights, and simple inference/visualization utilities.

[6] Dataset: *The AT&T Database of Faces* (formerly the ORL, Olivetti Research Laboratory, Face Database), created by AT&T Laboratories Cambridge and hosted by the University of Cambridge Computer Laboratory. Available at: https://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html (visited on 17 December 2025)

Figure 3.5: Olivetti Faces overview: one exemplar per subject (IDs 0–39). The dataset contains 400 grayscale faces (40 identities, 10 images each, 64×64 px) with variations in illumination and expression, used for the multi-class face classification baseline.

3.3.2 Project development

A classical machine-learning baseline was implemented using a `RandomForestClassifier` trained on the raw, flattened pixel vectors of the training set and evaluated on the held-out test set. Without any face alignment, data augmentation, or dimensionality reduction, the model achieved a test accuracy of about 88%. The accompanying classification report indicated macro-average precision = 0.88, recall = 0.90, and F1 = 0.87. To support qualitative assessment, a tiled grid of test images annotated with the corresponding predicted identities was produced, enabling quick visual inspection of correct recognitions and typical misclassifications, as shown in Figure 3.6. The overall pipeline thus serves as a compact, fully reproducible reference for multi-class face recognition with conventional features. This exercise demonstrates that a simple, well-controlled baseline on the Olivetti data can deliver solid performance with minimal preprocessing, while providing a clear tutorial example for benchmarking future improvements (e.g., feature extraction or deep models). The work and its intermediate findings were presented during the PWC "Tech Wednesday" on 20 March 2024, where the session prompted interest and questions from numerous colleagues and helped align subsequent experimentation

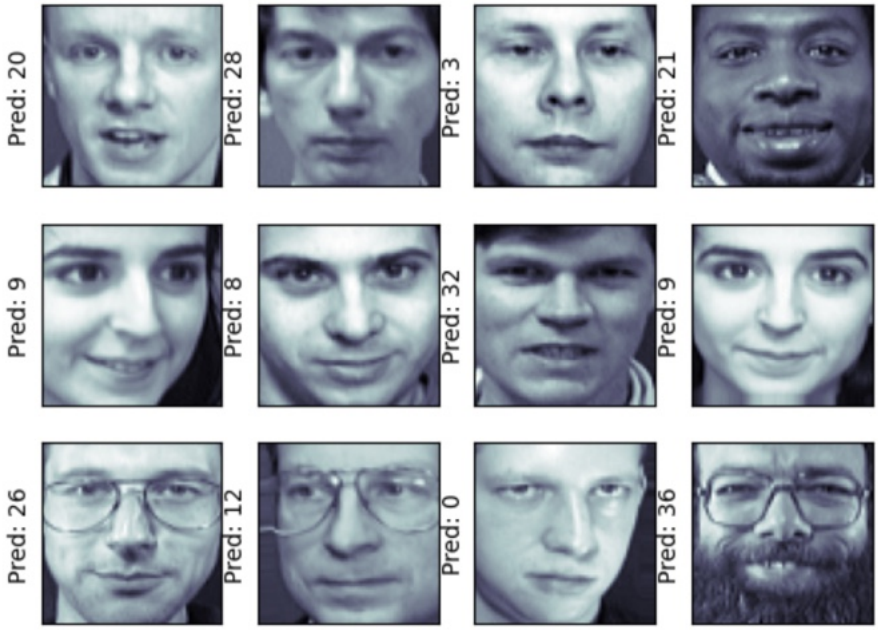


Figure 3.6: Qualitative predictions on the Faces test split. Each tile shows a 64×64 grayscale face with the predicted subject index indicated as `Pred: k`. The examples illustrate typical correct identifications across variations of pose, expression, and illumination.

directions.

3.4 DeepFake detection with ResNet–50 features and temporal aggregation

This section presents an internal, educational study on video-level DeepFake *detection* conducted at PWC. The aim is not to chase state-of-the-art scores, rather to build a clear baseline that uses two ingredients only, frame features extracted with a pre-trained `ResNet-50` and a light temporal aggregation stage that maps per-frame descriptors to a video-level decision. Temporal aggregation is implemented with a compact LSTM for study purposes, while simple pooling or majority voting are considered as ablations, since the focus is on understanding data requirements, protocol choices, and limits to generalisation across sources. The overall workflow follows, with minor adaptations, a public reference pipeline that inspired this study,

see Fig. 3.8.[7]

Within this framework we adopt a face-centric preprocessing stage to mitigate background leakage, we use fixed random seeds and an explicit train, validation, and test protocol, and we report accuracy, area sotto la curva ROC (AUC), and confusion matrices with a discussion of threshold effects at frame and video level. The resulting implementation serves as a transparent, reproducible starting point for subsequent investigations that prioritise robustness and out-of-distribution generalisation over raw performance.

3.4.1 Dataset and preprocessing

We employed a custom video corpus assembled for this study by merging three public sources: 2,115 clips from FaceForensics++, 2,347 clips from the DeepFake Detection Challenge, and 1,989 clips from the Celeb dataset, for a total of 6,452 videos (decompressed size =1.68 GB). Each video is labeled as `REAL` or `FAKE` in the accompanying `Global_metadata.csv`. The dataset was split at the *video level* into *training* (60%), *validation* (20%), and *test* (20%) partitions, preserving the REAL/FAKE proportions of the merged set. For preprocessing, every video was decoded into frames, faces were detected and tightly cropped, and crops were resized to the target spatial resolution before feature extraction. Frame embeddings were computed with a `ResNet-50` backbone (`ImageNet` pre-training), optionally fine–tuned on the face crops, yielding a sequence of per–frame descriptors subsequently aggregated into a video–level decision.

3.4.2 Project development

The main objective of this study was not to push for peak accuracy, but to review existing temporal aggregation strategies and assess their behavior with respect to generalization. We implemented a simple frame–level baseline that averages per–frame scores and a study variant that feeds the `ResNet-50` embeddings to an

[7] A. Jadhav, *Deepfake Video Detection Using Long Short-Term Memory*. Available at: https://abhijithjadhav.medium.com/deepfake-video-detection-using-long-short-term-memory-df3674f83ecc (visited on 17 December 2025).

LSTM head in order to capture local temporal dynamics. The LSTM option was included *for learning purposes* and to replicate a didactic pipeline popularized in public tutorials[8]. Training used early stopping on validation ROC–AUC, light spatial augmentation on face crops, and simple temporal jitter. Video decisions were obtained either by mean–pooling frame scores with a calibrated threshold or by applying the LSTM head followed by time–pooling. As a study baseline for this topic, we reproduced the workflow shown in Fig. 3.8, which was inspired by the public tutorial *"Deepfake Video Detection Using Long Short-Term Memory"*[9].

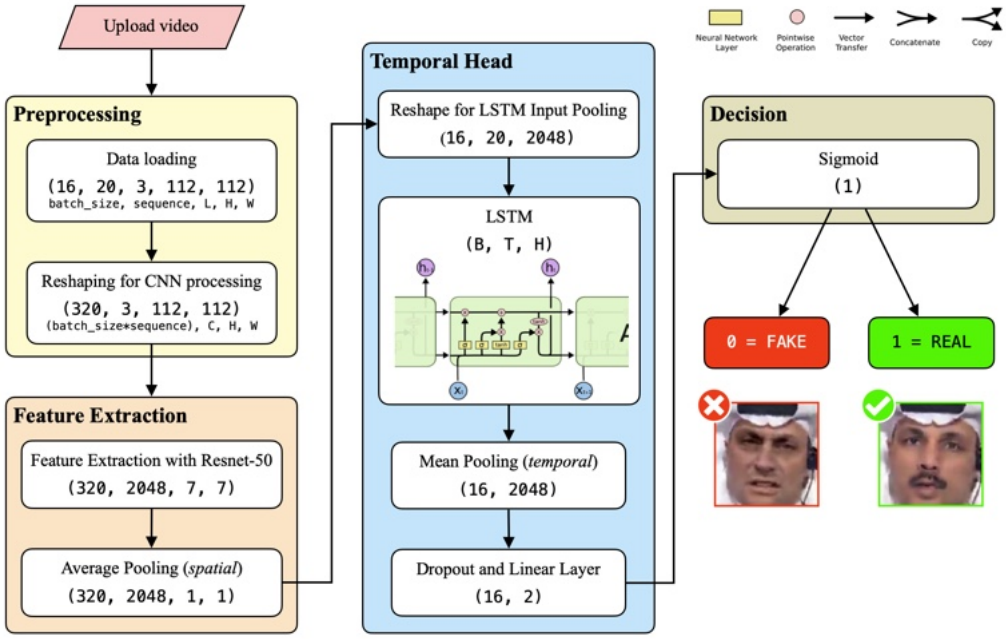


Figure 3.7: Proposed workflow reproduced in our study for DeepFake detection. Videos are decoded and faces cropped; frame tensors are reshaped and passed through a **ResNet–50** backbone to extract 2048-D descriptors. Spatial average pooling yields per-frame features that are reshaped into sequences and temporally aggregated with a lightweight LSTM. Mean pooling followed by a dropout+linear head produces the video-level logit; the binary output follows the convention 0 = FAKE, 1 = REAL.

This study surveyed frame–to–video aggregation options using a consistent preprocessing and a `ResNet-50` feature backbone. Results indicated that the LSTM

[8]Deepfake Video Detection Using Long Short–Term Memory, tutorial article. Available at https://abhijithjadhav.medium.com/deepfake-video-detection-using-long-short-term-memory-df3674f83ecc (visited on 17 December 2025). We reproduced the high–level idea as a study exercise and adapted it by replacing the feature extractor with `ResNet-50` and by standardizing the face–crop preprocessing.

[9]A. Jadhav, https://abhijithjadhav.medium.com/deepfake-video-detection-using-long-short-term-memory-df3674f83ecc (visited on 17 December 2025)

head can match the mean–pooling baseline in–dataset, while still being sensitive to protocol choices and cross–dataset shifts. Accordingly, subsequent work focuses on attribute–aware training and evaluation rather than on designing ever more complex temporal heads, Figure 3.7 illustrates one of the study pipelines we reproduced for analysis: frame-level feature extraction with **ResNet–50** and a small LSTM for temporal aggregation, used here primarily to investigate generalization rather than to optimize raw accuracy.

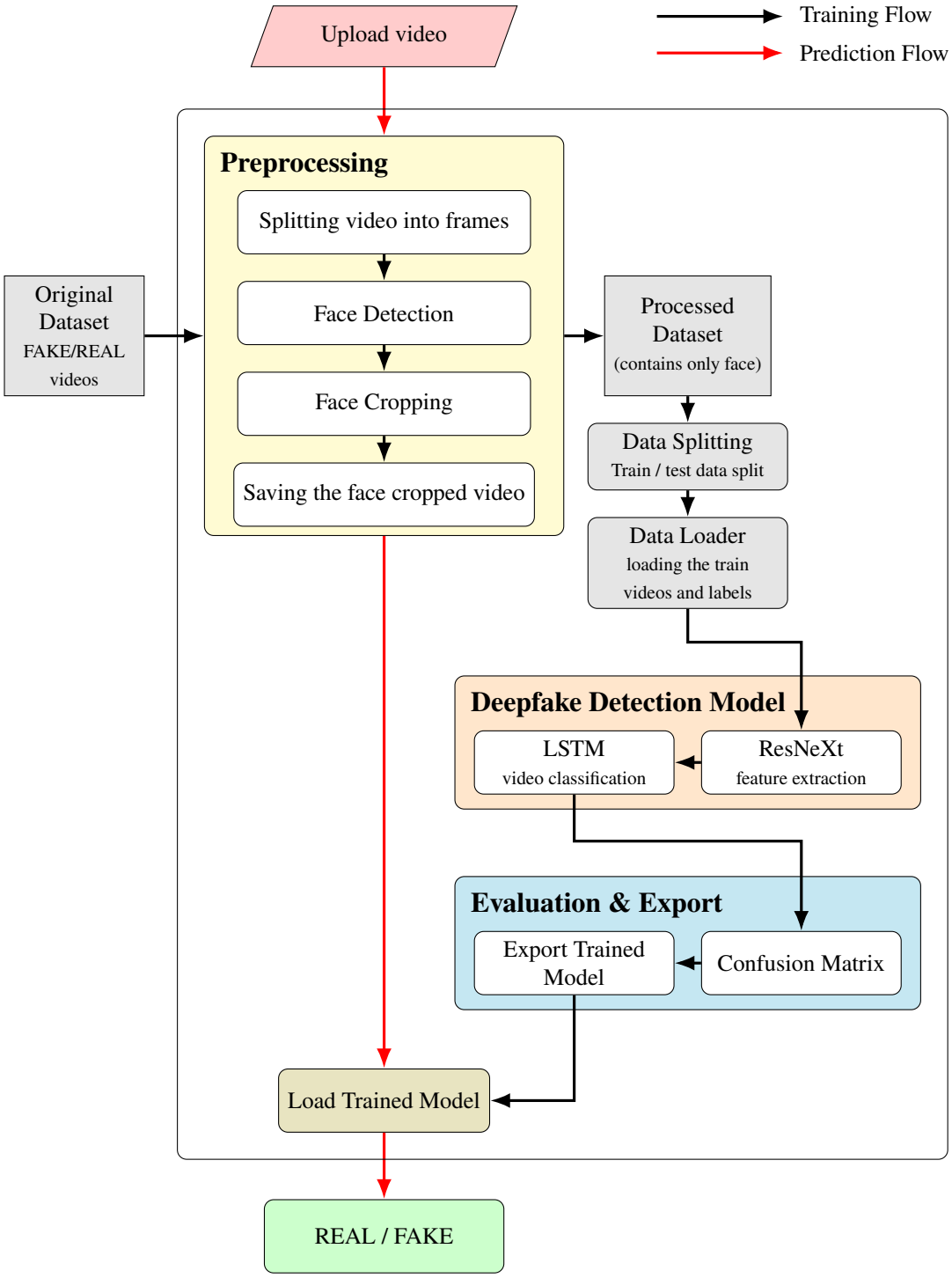


Figure 3.8: Study prototype for video–level DeepFake detection. Frames are face–cropped, encoded with `ResNet-50`, then aggregated either by score averaging or, for analysis purposes, by an LSTM head. The LSTM branch was included to inspect temporal modelling effects, not as a core contribution.

Chapter 4

Analysis of DeepFake Detection through Semi-Supervised Facial Attribute Labeling

4.1 Educational context (Cádiz)

This work was carried out during an international mobility period at the , hosted from 24/06/2024 to 23/03/2025 under the supervision of Profs. Inmaculada Medina-Bulo and Fernando Pérez Peña, with the PhD Tutor Prof. Roberto Caldelli at Universitas Mercatorum. The mobility objective, formally recorded in the final report, was to develop an advanced system for classifying and detecting DeepFake content using semi-supervised learning and convolutional neural networks. The project's twofold goal was: (i) to automatically label the *FaceForensics++* dataset with high-level visual and contextual attributes, and (ii) to use these labels to study whether specific attributes facilitate or hinder detection, ultimately supporting a jointly authored scientific publication.

The collaboration was structured around a shared pipeline discussed in the laboratory of the Software Engineering Research Group at the University of Cádiz (UCASE) in the Puerto Real campus. The workflow combined semi-supervised attribute labeling with conventional face-centric preprocessing and a `ResNet-50` baseline for frame-level classification, followed by an already tested pipeline for the

detection. The labeling protocol adopted manual labeling a small part of samples and automatic propagation for the remaining part, while the dataset preprocessing relied on `Haar-cascade` face detection and cropping.

Training practices were standardised from the outset. Images were normalised, and the dataset was split into 72% training, 14% validation, and 14% test sets, using stratification across REAL and FAKE classes to preserve class balance during model selection and final reporting. This protocol provided a stable baseline for subsequent analyses and comparisons. Building on this foundation, the doctoral work proceeded along four main directions: completing semi supervised labeling on *FaceForensics++*, refining the detector by integrating categorical facial attributes to improve robustness, performing a comparative bias analysis across attribute groups, and consolidating the resulting evidence into a manuscript summarising the findings and their implications for DeepFake detection.

The educational value of the Cádiz period was therefore twofold. First, it provided methodological alignment on a reproducible pipeline for attribute-aware DeepFake detection that the team could share across institutions. Second, it created the conditions for systematic experimentation on generalization and bias with a fixed evaluation protocol, paving the way for the analyses presented in the following sections. These aims, tools and milestones were consistently tracked in progress materials and meeting notes dedicated to the collaboration.

4.2 Introduction

DeepFake technologies have rapidly evolved in recent years, posing increasing threats to digital security, privacy, and public trust. These synthetic media techniques are capable of generating highly realistic face-swapped or manipulated videos, making them difficult to distinguish from authentic content[1].[2] Despite substantial progress in DeepFake detection, current systems often struggle to generalize across different manipulation types and compression levels, especially in uncontrolled

[1]Tolosana, Vera-Rodriguez, Fierrez, Morales, and Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection", op. cit.

[2]Verdoliva, "Media Forensics and DeepFakes", op. cit.

real-world conditions,[3],[4].[5] Traditional detection approaches, such as those based on frequency artifacts[6] or inconsistencies in facial dynamics, typically rely on fixed patterns that may not transfer well across datasets or manipulation methods, limiting their robustness,[7].[8] Recent research has explored the use of interpretable features and semantic attributes to enhance model explainability and resilience,[9].[10] However, these studies often depend on manual annotation or are limited in scope. Automatic facial attribute labeling techniques, particularly semi-supervised approaches, remain underexplored in the context of DeepFake detection. This motivates our work, in which we investigate whether there exists a correlation between prediction errors and high-level visual characteristics derived through semi-supervised attribute labeling. By identifying such patterns, our aim is to support the development of more bias-aware, interpretable, and generalizable DeepFake detection systems. This study aims to analyze if DeepFake detection algorithms depend on visual attributes and, consequently, if detection could be enhanced by the knowledge of these attributes estimated through an automatic attribute labeling. To address this, we apply a semi-supervised labeling approach to the *FaceForensics++* dataset[11] and integrate it into an image-level DeepFake detection pipeline. We aim to evaluate which visual facial attributes most influence classifier decisions on manipulated versus authentic content. Facial attribute information may be incorporated as auxiliary inputs or conditioning variables during training. For example, one could design a bias-aware classification pipeline, where a preliminary model estimates facial attributes and feeds this information into a secondary DeepFake classifier. Such strategies would allow the system to dynamically adjust its decision threshold based on the attribute-informed context. This chapter is structured as follows:

[3]Li, Yang, Sun, Qi, and Lyu, "Celeb-DF", op. cit.

[4]Y. Li, M.-C. Chang, and S. Lyu, *Celeb-DF (v2): A new dataset for deepfake forensics*, 2020, URL: https://cse.buffalo.edu/~siweilyu/celeb-deepfakeforensics.html (visited on 10/27/2025).

[5]Rössler, Cozzolino, Verdoliva, Riess, Thies, and Nießner, "FaceForensics++", op. cit.

[6]Durall, Keuper, and Keuper, "Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions", op. cit.

[7]Li, Chang, and Lyu, "In Ictu Oculi", op. cit.

[8]Yang, Li, and Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses", op. cit.

[9]Zhang, Paluri, Ranzato, Darrell, and Bourdev, "PANDA: Pose aligned networks for deep attribute modeling", op. cit.

[10]Guera and Delp, "Deepfake Video Detection Using Recurrent Neural Networks", op. cit.

[11]Rössler, Cozzolino, Verdoliva, Riess, Thies, and Nießner, "FaceForensics++", op. cit.

- This Section *Introduction* introduces the challenges of DeepFake detection, proposes the integration of facial attributes.
- Section *Related Work* it's about CNN-based detection models and interpretable feature extraction pipelines.
- Section *The proposed methodology* details the methodology, including data preparation for semi-supervised labeling, model architecture, training procedure, and includes an extensive description of the dataset in the Subsection *Dataset used in this study*, detailing its organization, the REAL/FALSE balanced subsets used in this study, and the semi-supervised labeling pipeline adopted to annotate facial attributes.
- Section *Analysis of the relations between labels and wrong predictions* presents the experimental results, including model accuracy with respect to facial attributes. This section also covers the impact of outlier filtering and threshold selection on model behavior.
- Finally, Section *Conclusions and future works* outlines conclusions and future research directions.

4.3 Related Work

The effectiveness of CNNs in DeepFake detection was first demonstrated by `MesoNet`[12] and the *FaceForensics++* benchmark using `XceptionNet`.[13] Subsequent research has introduced enhancements such as spatio-temporal models, frequency-based features, and attention mechanisms[14][15].[16] Other studies incorporate multimodal signals like lip sync and compression artifacts.[17] Recent efforts focus on

[12] Afchar, Nozick, Yamagishi, and Echizen, "MesoNet", op. cit.

[13] Rössler, Cozzolino, Verdoliva, Riess, Thies, and Nießner, "FaceForensics++", op. cit.

[14] Sabir, Cheng, Jaiswal, AbdAlmageed, Masi, and Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos", op. cit.

[15] Durall, Keuper, and Keuper, "Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions", op. cit.

[16] H. Dang, F. Liu, H. Stehouwer, X. Liu, and A. K. Jain, *Detection of deepfake videos using multi-attentional convolutional neural networks*, in: *European Conference on Computer Vision*, Springer 2020, pp. 660–676.

[17] Neekhara, Dolhansky, Bitton, and Ferrer, "Adversarial Threats to DeepFake Detection", op. cit.

interpretable semantic clues e.g., eye blinking, head pose, and facial expressions[18].[19] Despite these advances, generalization remains a challenge. Several studies have recently explored interpretable and bias-aware approaches to DeepFake detection. For instance, Tolosana offers a comprehensive overview of face manipulation methods and detection techniques, including considerations of how certain manipulation strategies challenge generalization across datasets or focusing on interpretable semantic clues e.g., eye blinking[20].[21] More targeted efforts include,[22] which introduces attention-based temporal features using recurrent neural networks, and,[23] which proposes a method based on inconsistencies in head poses to detect forgeries. These works aim to improve detection accuracy through architectural or handcrafted cues directly embedded into the classifier. Similarly, Verdoliva emphasizes the importance of robust forensics pipelines, highlighting limitations in detection systems under real-world variability.[24] By contrast, our contribution is positioned differently. Rather than enhancing the detection model itself, we propose a systematic and empirical evaluation of its misclassification patterns through semi-supervised labeling of high-level visual attributes (e.g., gender, hair color, head visibility).

Our work extends this direction by applying structured facial attributes through semi-supervised labeling. The labeling procedure is based on a self-learning method, who combine lightweight CNNs with inferred labels to adapt detection systems to evolving contexts.[25] This work adopt semi-supervised facial attribute labeling as a foundation for a structured performance analysis of DeepFake detection models and validate it using both descriptive and statistical correlation metrics. While previous literature has explored the use of demographic or semantic features in detection or interpretability, no prior work has combined manual annotation and automated labeling in this fashion to investigate performance disparities. Our

[18] Li, Chang, and Lyu, "In Ictu Oculi", op. cit.

[19] Yang, Li, and Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses", op. cit.

[20] Tolosana, Vera-Rodriguez, Fierrez, Morales, and Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection", op. cit.

[21] Li, Chang, and Lyu, "In Ictu Oculi", op. cit.

[22] Guera and Delp, "Deepfake Video Detection Using Recurrent Neural Networks", op. cit.

[23] Yang, Li, and Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses", op. cit.

[24] Verdoliva, "Media Forensics and DeepFakes", op. cit.

[25] Guerrero-Contreras, Balderas-Díaz, García-Pascual, and Muñoz, "Adaptive Vehicle Detection in Urban Environments", op. cit.

approach aims to uncover potential characteristics or appearance-related biases in post hoc prediction analysis, providing insights that can inform future fairness-aware model design. This novel strategy enables large-scale, interpretable evaluations and highlights potential biases in a replicable and extensible way. Unlike much of the prior work in DeepFake detection, which primarily seeks performance gains through architecture-level innovations or audio–visual fusion, this thesis introduces an attribute-informed diagnostic pipeline that focuses on error analysis and model behaviour rather than on accuracy alone.[26] We seek to understand the relationship between facial attributes and misclassification patterns. A central aim of this study is to use semi-supervised attribute labeling to conduct a structured analysis of DeepFake detection performance.

4.4 The proposed methodology

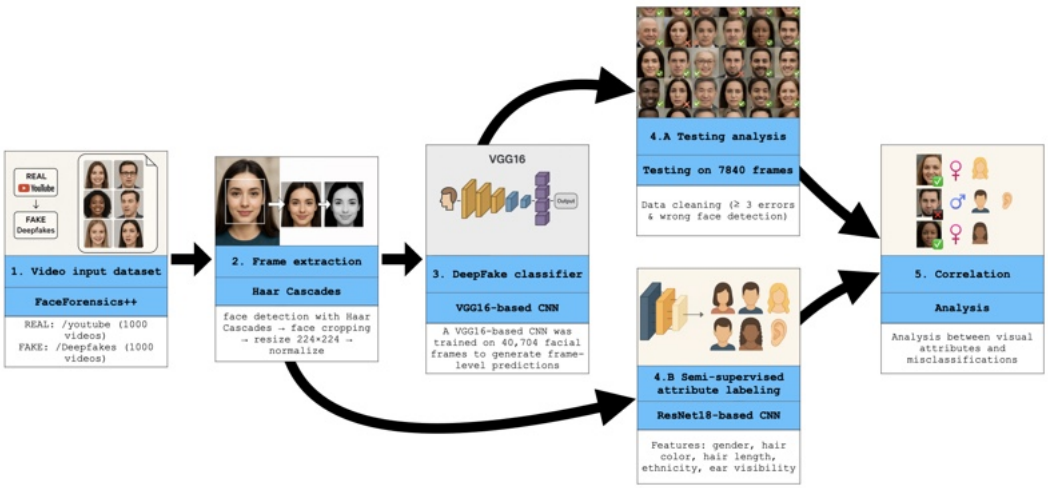


Figure 4.1: Overview of the proposed methodology. The pipeline begins with video-level input from the *FaceForensics++* dataset, proceeds through frame extraction and face cropping, followed by DeepFake classification at the frame level. A semi-supervised labeling process is then used to annotate facial attributes, and finally, correlations between those attributes and misclassification patterns are analyzed.

The methodology that has been followed primarily foresees to define a specific dataset containing real and DeepFake videos and then frames are extracted; these

[26] Verdoliva, "Media Forensics and DeepFakes", op. cit.; Neekhara, Dolhansky, Bitton, and Ferrer, "Adversarial Threats to DeepFake Detection", op. cit.

frames are successively processed and split into training, validation and test set in order to trained a neural network that is able to carry out an image-level classification task to distinguish between pristine and fake contents (see Figure 4.1). In parallel, the test set which is used for the evaluation of the model has been labeled according to some facial attributes by resorting to a semi-supervised procedure. Based on the output of the labeling process, we investigate whether the predictions produced by the trained neural network exhibit systematic errors correlated with specific high-level facial attributes. Such possible relations could be exploited to improve DeepFake detection. Hereafter all the different phases of the adopted pipeline are described in detail.

4.4.1 FaceForensics++ Dataset

We utilize *FaceForensics++*, a benchmark dataset widely adopted in DeepFake research.[27] The original corpus consists of 1,000 pristine YouTube videos, manually screened to ensure near frontal, occlusion free faces at ≥480p, which serve as targets for automated manipulations. The manipulated portion of the dataset is organized into six families: *Deepfakes* (identity swap with autoencoders), *FaceSwap* (graphics based identity swap, commonly known as *FaceSwap Kowalsky* [28]), *Face2Face* (expression reenactment), *NeuralTextures* (reenactment with learned textures), *FaceShifter* (identity swap), and *DeepFakeDetection* (identity swap). Each family comprises 1,000 videos generated from the same 1,000 sources, for a total of 6,000 manipulated and 1,000 real videos; per frame ground truth masks of the edited regions are available for localization research.[29] To mimic real distribution channels, videos are provided as RAW and in two H.264 compressed variants, commonly denoted `c23` and `c40`, corresponding to higher and lower visual fidelity respectively.

Directory layout of the DataSet. We worked with the dataset as downloaded from the project's scripts and verified the following folders and counts in our copy:

[27] Rössler, Cozzolino, Verdoliva, Riess, Thies, and Nießner, "FaceForensics++", op. cit.

[28] The MarekKowalski/FaceSwap project, created and maintained by Marek Kowalski, GitHub repository https://github.com/MarekKowalski/FaceSwap/ (visited on 2025-12-26).

[29] Rössler, Cozzolino, Verdoliva, Riess, Thies, and Nießner, "FaceForensics++", op. cit.

- `original`: 1,000 REAL videos.
- `Deepfakes`: 1,000 FAKE videos.
- `Face2Face`: 1,000 FAKE videos.
- `FaceSwap`: 1,000 FAKE videos.
- `NeuralTextures`: 1,000 FAKE videos.
- `FaceShifter`: 1,000 FAKE videos.
- `DeepFakeDetection`: 1,000 FAKE videos.
- `csv`: 10 CSV files with video metadata.

In total, our working copy contains 7,010 files, of which 7,000 are `.mp4` videos (6,000 FAKE + 1,000 REAL) and 10 are CSV metadata files.

Compression tiers. All subsets are available as RAW uncompressed format and in two compressed variants. A moderate compressed `c23` version, which corresponds to H.264 with Constant Rate Factor (H.264 quality scale) (CRF) 23 that recreate an high quality, nearly lossless visual fidelity and a more compressed`c40` version, that results in a low quality visual fidelity useful for stress testing detectors and comparative analysis.

Table 4.1: FaceForensics++ video encodings and compression settings.

Label	Codec	Parameter	Description
RAW	Uncompressed	n/a	Source-quality frames
`c23`	H.264	CRF=23	Moderate compression, high fidelity
`c40`	H.264	CRF=40	Strong compression for stress testing

CRF = Constant Rate Factor (H.264 quality scale).

Metadata. For each video we record from the CSV files: *Index*, *File path*, *Label* (REAL/FAKE), *Frame count*, *Width*, *Height*, *Codec*, and *File size* (MB). These fields are used to drive preprocessing in our pipeline.

Table 4.2: FaceForensics++ contents with manipulation families. The six FAKE subsets correspond to canonical forgery types widely used in the literature.

Folder	Type	# Videos	Manipulation family
original	REAL	1,000	N/A
Deepfakes	FAKE	1,000	Identity swap *Deep Learning (DL)*
FaceSwap	FAKE	1,000	Identity swap *graphics*
Face2Face	FAKE	1,000	Reenactment *graphics*
NeuralTextures	FAKE	1,000	Reenactment *Neural Rendering*
FaceShifter	FAKE	1,000	High-fidelity identity swap *DL*
DeepFakeDetection	FAKE	1,000	Identity swap *DL*
	Total	**7,000**	
		Breakdown:	6,000 FAKE + 1,000 REAL

Notes on splits and benchmarking. The reference work reports fixed splits used for detection experiments (training/validation/test at 720/140/140 videos) and also provides a public benchmark with a hidden test set for standardized comparison in realistic post–processing conditions.[30]

4.4.2 Data collection phase

In this study, we selected the 1,000 REAL videos from the `/youtube` folder and their corresponding 1,000 FAKE videos from the `/Deepfakes` folder, created with faceswap software [31], focusing our analysis on this specific forgery method. All videos used in this study refer to the **C40 compression level**[32], a setting that applies moderate, anyway higher than `c23`, compression and has shown to produce stable and consistent detection performance across experiments. Accordingly, all reported results are based on this compression configuration. To enable frame-level classification, we converted the original video dataset into an image-based dataset. This was done by applying face detection techniques, specifically using Haar Cascades, on all videos in both `REAL` and `FAKE` categories. Detected face regions were cropped and resized, resulting in a large corpus of `56,000` facial images.

[30] Ibid.

[31] The deepfake/faceswap project, created and maintained by Matt Tora (known by his pseudonym @torzdf), Andrey Ivanov (@andenixa), and Bryan Lyon (@bryanlyon), GitHub repository https://github.com/deepfakes/faceswap (visited on 2025-12-26).

[32] In all the experiments is used the compressed `c40` version, which corresponds to H.264 with constant rate factor 40.

4.4.3 Rationale for high–level attribute labeling

The CSV metadata distributed with *FaceForensics++* provide only technical descriptors, namely *Index*, *File path*, *Label* (`REAL` or `FAKE`), *Frame count*, *Width*, *Height*, *Codec*, and *File size* (`MB`). These fields are operationally useful and we use them to drive preprocessing choices in our pipeline, for example decoding, frame sampling, and compression stratification. They do not, however, encode qualitative or semantic information about the visual content. As a consequence, they do not support analyses that are central to interpretable DeepFake research, such as attribute–conditioned performance reporting, bias diagnosis, or failure–mode inspection tied to concrete visual traits. Given that each manipulated video in *FaceForensics++* is derived from a specific pristine source, and that many human–centric properties remain stable across the manipulation, enriching the corpus with human–interpretable attributes enables controlled comparisons between original and counterfeit pairs, isolates the effect of the forgery from confounding content variation, and facilitates downstream explainability. To our knowledge, neither the official dataset release nor commonly available online resources provide such semantic annotations, which motivates our targeted attribute labeling of video content and, in particular, the physical characteristics of depicted subjects.

4.4.4 Labeling phase

To annotate the dataset, we adopt a semi–supervised labeling approach.[33] As a first step, we establish a ground truth by manually inspecting the first 50 REAL videos from the `/youtube` folder of *FaceForensics++*, specifically `000.mp4` through `049.mp4`. Each video is reviewed frame by frame to assign a comprehensive set of high–level visual and semantic labels.

The annotation covers categorical features such as `gender` (FEMALE, MALE),

[33] G. Guerrero-Contreras, S. Balderas-Díaz, A. García-Pascual, and A. Muñoz, *Adaptive Vehicle Detection in Urban Environments: A Self-learning Approach*, en, in: *Ambient Intelligence – Software and Applications – 15th International Symposium on Ambient Intelligence*, ed. by P. Novais, P. B. D., I. Satoh, V. J. Inglada, S. R. González, E. Jove Pérez, J. Parra Domínguez, P. Chamoso, and R. S. Alonso, vol. 1279, Springer Nature Switzerland, Cham, Switzerland 2025, pp. 25–34, ISBN: 9783031831164 9783031831171, DOI: 10.1007/978-3-031-83117-1_3, URL: https://link.springer.com/10.1007/978-3-031-83117-1_3 (visited on 10/27/2025).

ethnicity (AFRICAN, ASIAN, MIXED, WHITE, OTHER), expression (ANGRY, DISGUSTED, HAPPY, LAUGHING, SERIOUS, SURPRISED), hair_color (BLONDE, BLACK, BROWN, GREY, LIGHT BROWN, RED, OTHER), hair_length (BALD, LONG, PONYTAIL, SHORT), action (HUGGING, STILL, WALKING), and position (SITTING, STANDING, UNKNOWN). In parallel, we collect boolean attributes, including whether the subject is alone, whether ears or forehead are visible, whether glasses are worn, and whether the video exhibits camera movement, pan effects, shakiness, or active talking, each recorded as TRUE or FALSE. All labels are assigned based on traits that are visible and consistent throughout the clip.

For this study, we assume that each manipulated video generated from a given real video via the *DeepFakes* method, and stored in /Deepfakes, inherits the same facial-attribute labels as its source. This assumption is plausible because *DeepFakes* primarily performs an identity swap while preserving most coarse facial characteristics of the target subject, including stable traits such as apparent gender, and in practice the manipulation does not systematically alter these attributes across the dataset. In particular gender, the characteristic that could give the most uncertainty, proved to be highly stable in our setting. All the 50 real videos manually annotated in the seed subset preserved the same gender presentation in their corresponding DeepFake counterparts, and additional spot checks on randomly selected video pairs across the dataset did not reveal any counterexamples. Based on this consistent evidence, we generalize the assumption and treat gender as an invariant attribute for the full 1,000-video of the /Deepfakes subset, transferring the gender label from each real source video to its manipulated version. For example if 005.mp4, in the real /youtube folder, is labeled gender=FEMALE and hair_length=LONG, the corresponding manipulated sample 005_010.mp4, in the */Deepfakes* folder, is assumed to retain these properties.

These enriched labels are added alongside the original binary ground truth is_real, which is the only supervisory signal provided in the dataset. The resulting annotations aim to support the development of more interpretable detection models and attribute–aware evaluations. The initial set of 50 manually labeled REAL videos

is then used to train attribute classifiers within a semi–supervised pipeline.

Using 5% of the `/youtube` videos (50 clips) as human–labeled seeds, a self–labeling procedure produces pseudolabels for the remaining 95% (950 clips). The auto–labeling models use a *ResNet18* backbone and are trained separately for each attribute, as binary or multiclass classifiers depending on the label type.[34] Training follows an iterative loop. First, the 50 seeds are split 60–20–20 into train, validation, and test sets (30, 10, and 10 videos). A pretrained *ResNet18* is fine–tuned on the training portion, with early stopping on validation. The resulting model is evaluated on the test split to establish a baseline, then applied to pseudolabel the 950 unlabeled clips. Only samples with confidence above 95% are admitted into the augmented training set, which is then used for further fine–tuning. The augmentation–retraining loop repeats until convergence on the held–out test set or until no additional high–confidence samples remain.

From the initial pool of attributes, features such as `gender`, `hair_color`, and `is_ears_visible` showed higher predictive utility and were retained. Attributes with a single observed class in the seed set (for example `is_real=True` or `is_glasses=False`) were excluded to avoid severe class imbalance. Table 4.3 lists the attributes selected for analyzing how specific visual features influence DeepFake detection performance.

Table 4.3: Selected facial attributes used for analysis.

Attribute	**Labels**
`hair_color`	BLONDE, BLACK, BROWN, GREY, LIGHT BROWN, RED, OTHER
`hair_length`	BALD, LONG, PONYTAIL, SHORT
`gender`	FEMALE, MALE
`ethnicity`	AFRICAN, ASIAN, MIXED, WHITE, OTHER
`is_ears_visible`	TRUE, FALSE

[34] Guerrero-Contreras, Balderas-Díaz, García-Pascual, and Munõz, "Self-Learning Systems for Enhanced Traffic Management in Urban Settings", op. cit.; Guerrero-Contreras, Balderas-Díaz, García-Pascual, and Muñoz, "Adaptive Vehicle Detection in Urban Environments", op. cit.

4.4.5 Face Extraction and Preprocessing

The objective of this preprocessing step was to extract face-only image frames from the *FaceForensics++* dataset. Face detection was performed using the classical *Haar Cascades* method, the *OpenCV* `detectMultiScale` function was applied on grayscale versions of each frame. The `scaleFactor` parameter, set to 1.3, controls the image pyramid reduction ratio during face detection, enabling the model to detect faces at different scales. The `minNeighbors` parameter, set to 5, defines the minimum number of adjacent rectangles required to retain a detection, acting as a threshold for eliminating false positives. These values were selected empirically, balancing detection accuracy and noise reduction. Detected faces were cropped, resized to 224×224 pixels (standard size for CNNs) and saved as individual `.jpeg` files, this pipeline in represented in Figure 4.2. While effective, this pipeline presents notable challenges in terms of data volume introducing complexity in terms of both storage and computational overhead during training. To mitigate this, in this phase we adopted a frame skipping strategy with a fixed sampling rate of `skip = 20`, meaning one frame was retained for every twenty. This choice was guided by two considerations: it preserves a sufficient level of temporal representation of the source videos and it reduces redundancy, helping to prevent overfitting and model bias from near-duplicate frames.

4.4.6 Dataset used in this study

The final dataset is organized into a structured directory tree designed to support both training and evaluation. It includes the following main components:

- `_info/`: Contains metadata files and auxiliary information about dataset construction.
- `attribute/`: Stores both ground-truth and automatically generated attribute labels for each video.
- `youtube/` (REAL videos) and forgery folders such as `DeepFakes/`: Each organized into two compression levels, `c23/` and `c40/`, containing:

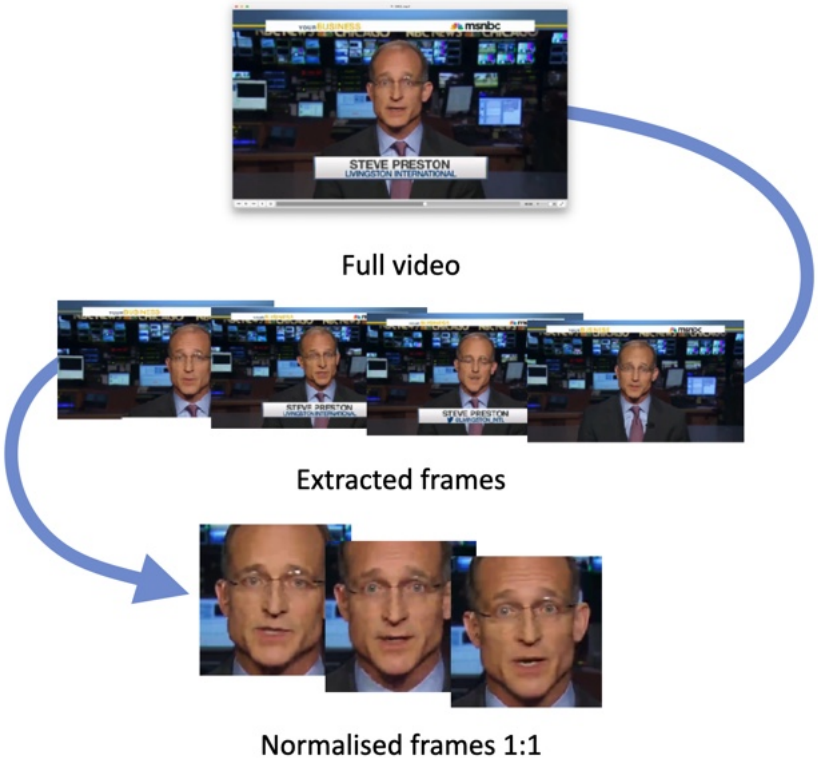


Figure 4.2: Frame extraction pipeline. From each input video we uniformly sample frames, detect the face, crop and align it, then resize to a square 1:1 format (e.g., 224×224) and normalize intensities. The resulting face patches feed the downstream CNN for training and evaluation.

- `faces/`: Cropped facial images.
- `frames/`: Raw video frames extracted from the original source.
- `info...skip=20.txt`: Summary files indicating the number of frames extracted per video using a frame-skipping strategy.
- `videos/`: Original video files from the dataset.

The dataset consists of 28,483 REAL facial frames extracted from the 1,000 videos in the `youtube` folder, and 28,051 FAKE facial frames extracted from the 1,000 videos in the `Deepfakes` forgery folder[35].

To ensure full reproducibility of the experiment, the project repository includes the complete training and analysis code under the `python_code/` directory, as well

[35]The difference in the total number of frames between the two classes is due to the face detection and cropping process using *Haar Cascades*. For instance, in the REAL video `002.mp4`, 40 face frames were extracted, while in the corresponding manipulated video `002_006.mp4`, only 38 frames were successfully processed.

as a `results/` folder containing Python logs and tabular outputs, organized by date. Additionally, the `_info/` folder includes scripts and documentation for downloading the full *FaceForensics++* dataset.[36]

Dataset characteristics

The final dataset used for this study is balanced between `REAL` and `FAKE` frames and consists of 28,000 real facial frames extracted from the 1,000 videos in the `/Youtube` folder and 28,000 fake facial frames extracted from the 1,000 videos in the `/Deepfakes` folder. From this collection, we constructed three subsets for training, validation, and testing. The **training set** contains a total of 40,320 facial images, equally divided between `FAKE (20,160)` and `REAL (20,160)` instances. The **validation set** consists of 7,840 images, again balanced between classes. Similarly, the **test set** includes 7,840 images with an equal number of `REAL` and `FAKE` faces.

Although various convolutional neural networks were evaluated during the frame-level DeepFake detection phase, including architectures such as `ResNet50`, we ultimately report results only for the `VGG16`-based model. This decision stems from the observation that performance across different architectures was highly comparable in terms of accuracy, precision, and F1-score. The `VGG16` model, initialized with ImageNet pre-trained weights and with all layers frozen during training, demonstrated reliable and consistent behavior, particularly on `C40`-compressed facial images and due to its ease of deployment on standard hardware[37], it was selected as the reference architecture for this study.

4.4.7 Labeling phase

To annotate the dataset, a semi-supervised labeling approach was adopted.[38] As a first step, we established a ground truth by manually inspecting the first 50 REAL videos from the `/youtube` folder of the *FaceForensics++* dataset, specifically

[36] Rössler, Cozzolino, Verdoliva, Riess, Thies, and Nießner, "FaceForensics++", op. cit.

[37] For a standard hardware we refer to a consumer-grade computer: a *MacBook Air* equipped with an *Apple M2* processor, 8-core CPU, 8-core GPU, 8GB of RAM.

[38] Guerrero-Contreras, Balderas-Díaz, García-Pascual, and Muñoz, "Adaptive Vehicle Detection in Urban Environments", op. cit.

those named from `000.mp4` to `049.mp4`. Each video was carefully reviewed, frame by frame, to assign a comprehensive set of high-level visual and semantic labels. The annotation included a range of categorical features such as `gender` (with values FEMALE or MALE), `ethnicity` (AFRICAN, ASIAN, MIXED, WHITE, OTHER), `expression` (ANGRY, DISGUSTED, HAPPY, LAUGHING, SERIOUS, SURPRISED), `hair_color` (BLONDE, BLACK, BROWN, GREY, LIGHT BROWN, RED, OTHER), `hair_length` (BALD, LONG, PONYTAIL, SHORT), `action` (HUGGING, STILL, WALKING) and `position` (SITTING, STANDING, UNKNOWN). In parallel, a set of boolean attributes was also collected, including indicators such as whether the subject is alone, whether the ears or forehead are visible, whether glasses are worn, and whether the video exhibits camera movement, pan effects, shakiness, or active talking, each having a TRUE or FALSE value. All labels were assigned based on visible and consistent traits throughout the selected video clips. For the purposes of this study, we assume that each manipulated video, generated from a corresponding real video via the DeepFakes method and located in the `/Deepfakes` folder, inherits the same attribute labels as its original. For example, if the REAL video `005.mp4` is labeled as having gender: FEMALE and hair_length: LONG, its corresponding DeepFake version `005_010.mp4` is assumed to retain these same properties.

These enriched labels were added to the original binary classification label `is_real`, which was the only ground truth available in the original *FaceForensics++* dataset. The resulting annotated set aims to be a foundation for the development of more interpretable detection models. This initial annotated subset of 50 REAL videos serves as the foundation for training the attribute labeling models in the subsequent semi-supervised pipeline. The output of this phase was the creation of a csv file called `attribute_youtube.csv` as represented by way of example in Table 4.4 and publicly accessible in the GitHub repository prepared by the author [39].

Then using 5% (50 videos) of manually labeled instances from the youtube dataset, the self-labeling approach enables the generation of pseudolabels for the

[39]File available on the public GitHub Repository created by the author: https://github.com/vstile/deepfake-attribute-detection/blob/main/dataset/attribute/ground-truth/attribute_youtube.csv (visited on 26 December 2025).

video_name	gender	ethnicity	hair_color	hair_length	is_ears_visible
000.mp4	MALE	OTHER	OTHER	UNKNOW	TRUE
001.mp4	FEMALE	ASIAN	BLACK	LONG	TRUE
002.mp4	MALE	WHITE	GREY	SHORT	TRUE
003.mp4	MALE	WHITE	OTHER	BALD	TRUE
004.mp4	MALE	WHITE	BROWN	SHORT	TRUE
005.mp4	FEMALE	WHITE	BLACK	LONG	FALSE
006.mp4	MALE	WHITE	GREY	SHORT	TRUE
007.mp4	FEMALE	WHITE	BROWN	PONYTAIL	TRUE
008.mp4	FEMALE	WHITE	BLACK	LONG	TRUE
009.mp4	MALE	WHITE	LIGHT BROWN	SHORT	TRUE
010.mp4	FEMALE	ASIAN	BLACK	PONYTAIL	TRUE
...	...	...	...	...	...
049.mp4	FEMALE	WHITE	BLACK	SHORT	TRUE

Table 4.4: Sample of ground-truth attribute annotations from `attribute_youtube.csv`.

remaining 95% (950 videos). The auto-labeling process is based on a *ResNet18* backbone, trained separately for each visual attribute (visibility of ears, ethnicity, forehead coverage, gender, hair color and hair length) using a dedicated binary or multiclass classifier depending on the label type.

Table 4.5: Summary of manual attribute labeling by the author, on the FaceForensics++ `/youtube` subset (REAL videos), and automatic semi-supervised labeling with confidence ≥ 0.95.

Label source Notes	# videos	Share
Manual seed annotation	50	5%
Automatically labeled	950	95%
Total labeled videos	1000	100%

The training follows an iterative semi-supervised loop designed to progressively expand the labeled dataset. Initially, the manually annotated subset (50 videos) is partitioned into training, validation, and test splits following a 60-20-20 scheme (30-10-10 videos). In the first iteration, the pretrained *ResNet18* model is fine-tuned on the training set, using the validation set to prevent overfitting through an early stopping strategy. This fine-tuned model is then evaluated on the test split to establish the baseline performance. Additionally, it is employed to pseudolabel the remaining

unlabeled data (950 videos), thereby generating an augmented training dataset comprising the 30 manually labeled videos and up to 950 pseudolabeled instances. Only pseudolabeled instances with a confidence score above 95% are added to the augmented dataset, which is then used to further fine-tune the pretrained *ResNet18*. This process of augmentation and retraining repeats until convergence on the test split or no additional high-confidence samples remain.

From an initial set of visual attributes, features such as gender, hair color, and ear visibility demonstrated high predictive power and were retained. Table 4.3 provides an overview of the attributes selected to analyze the influence of specific visual features on DeepFake detection performance.

4.4.8 Detection phase

Evaluation metrics

To rigorously assess the performance of the DeepFake detection model, we employed a comprehensive set of evaluation metrics combined with frame-level analysis. These metrics and diagnostic tools offer insights not only into the global performance of the model, but also into its behavior on specific video frames, helping identify potential biases or limitations in generalization.

During model training, we tracked accuracy and loss metrics over each epoch for both the training and validation sets. This helped detect signs of overfitting or underfitting. The training curves were visualized using Matplotlib:

- `Accuracy Curves`: Indicate how well the model is learning to classify REAL and FAKE frames correctly over time.
- `Loss Curves`: Show how the model's error evolves, revealing stability or divergence between training and validation.

Predictions on the test set were generated using a sigmoid output activation with a classification threshold of 0.5. We computed the following binary classification metrics:

- `Accuracy`: Overall percentage of correctly classified frames.

- `Precision`: $\frac{TP}{TP+FP}$ portion of predicted FAKE frames that were truly FAKE.
- `Recall` (Sensitivity): $\frac{TP}{TP+FN}$ portion of real FAKE frames the model was able to detect.
- `F1-Score`: Harmonic mean of precision and recall, offering a balanced measure especially useful in the presence of class imbalance.
- `Confusion Matrix`: Highlights the counts of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN), offering a breakdown of prediction correctness.

To gain a deeper understanding of model behavior we continue with frame-level statistics:

- We computed and compared the number of frames predicted as FAKE or REAL.
- We calculated how many of those predictions were correct and identified FP (REAL predicted as FAKE) and FN (FAKE predicted as REAL).
- We also tracked the proportion of these outcomes relative to the ground truth distribution, allowing us to identify trends or systemic biases.

We end the pipeline with a comparative analysis on misclassified frames. Each misclassified frame (false positive or false negative) was cross-referenced with its associated metadata and automatically labeled attributes (e.g., gender, hair color, ear visibility). This allowed us to:

- Investigate whether certain **visual characteristics** are overrepresented in erroneous predictions.
- Identify **biases** in the model's behavior, such as systematically misclassifying individuals with specific attributes.
- Use this insight to guide **future iterations** of the model design or data preprocessing steps (e.g., balancing datasets, feature augmentation).

This diagnostic phase not only strengthens the interpretability of our model but also supports the overarching research question: whether and how specific facial features influence DeepFake detection accuracy.

Testing our DeepFake detection pipeline

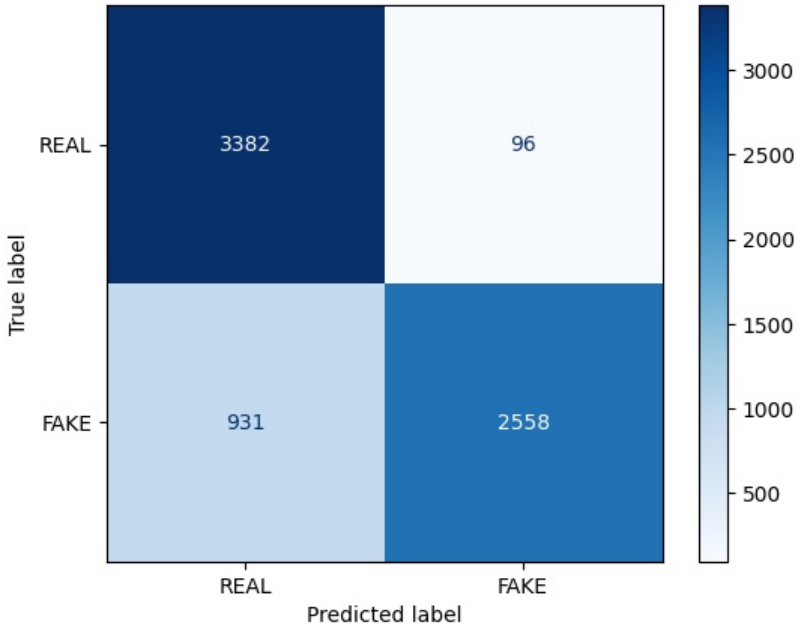


Figure 4.3: Frame-level Confusion Matrix

The experimental evaluation aimed to assess the effectiveness of our DeepFake detection pipeline based on visual features extracted from facial frames. The model was trained on a balanced dataset of 40,320 facial images (20,160 REAL and 20,160 FAKE), with a validation set of 7,840 images (equally split between classes), and a test set of 7,840 frames also balanced across the two classes, as shown in Table 4.6.

The classifier was based on the VGG16 architecture, with pre-trained ImageNet weights and all convolutional layers frozen. A simple custom head was added, consisting of a `Flatten` layer followed by `Dense(512, relu)` and a final `Dense(1, sigmoid)` output. Training was performed with the Adam optimizer (learning rate = 0.0001), binary cross-entropy loss, batch size of 16, and 5 epochs. The complete configuration is reported in Table 4.7. After training, the model reached a training accuracy of 94.58% and a validation accuracy of 86.96%. The test set was then

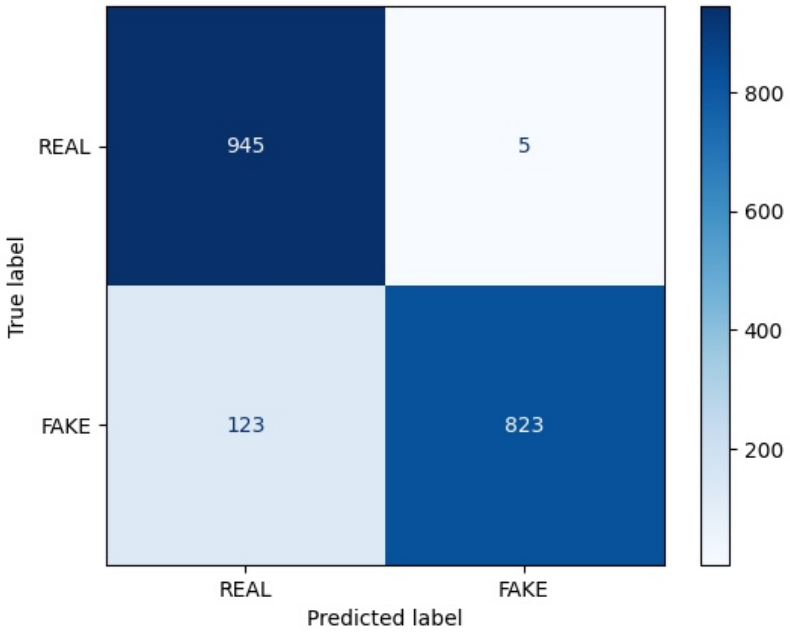


Figure 4.4: Video-level Confusion Matrix

Table 4.6: Dataset split by set and class.

Set	REAL	FAKE	Total
Training	20,160	20,160	40,320
Validation	3,920	3,920	7,840
Test	3,920	3,920	7,840

used to evaluate the final model. Before evaluation, a face-filtering step was applied so that, for each video, only one face track was retained, corresponding to the main subject. In the implementation, this primary track is denoted as `face0`, and it is treated as the candidate manipulated (or authentic) face for that clip.[40]

The model's frame-level classification performance is summarized in Table 4.8, and the confusion matrix is illustrated in Figure 4.3. These results highlight the model's precision in detecting `FAKE` content (96.38%) and its strong recall for `REAL` frames (97.24%), with overall balanced performance across both classes. The performance of the model on the frame-level test set (`face0`-only) resulted in an accuracy of 85.26% across 6,967 frames. This confirms the model's strong

[40]When multiple faces are detected in a frame, the `Haar Cascades` detector stores them with incremental indices such as `face0`, `face1`, `face2`, and so on. In the FaceForensics++ YouTube videos considered here, there is usually a single central speaker; consequently, `face0` almost always corresponds to the most prominent face in the scene and is used as the reference face track for evaluation.

Table 4.7: Model architecture and training configuration.

Component	Details
Base architecture	VGG16 (`include_top=False`, pre-trained on *ImageNet*)
Frozen layers	All pre-trained layers are frozen
Custom head	`Flatten` → `Dense(512, relu)` → `Dense(1, sigmoid)`
Optimizer	Adam (learning rate = 0.0001)
Loss function	Binary Crossentropy
Batch size	16
Epochs	5

generalization capability, particularly its high recall on `REAL` frames and precision on `FAKE` ones, as detailed in Table 4.8 and visualized in the confusion matrix in Figure 4.3. Although classification is performed at the frame-level, system reliability in DeepFake

Table 4.8: Frame-level classification report on the face0-only test set.

Class	Precision	Recall	F1-score	Support
REAL	0.7841	0.9724	0.8682	3478
FAKE	0.9638	0.7332	0.8328	3489
Total frames				6967

detection is typically evaluated at the video-level. According to established literature, it is not methodologically sound to mark an entire video as misclassified based on a single erroneous frame. Rössler et al.[41] and Sabir et al.[42] suggest aggregation strategies such as mean score, majority voting, or top-K frame voting to derive a more robust video-level decision.

A widely accepted criterion in the literature considers a video to be misclassified only if it contains *at least three wrongly predicted frames*, or if a *significant proportion* of its frames (typically ≥10%) are incorrect. This threshold is particularly appropriate in our case, as each video in the test set contains an average of 28 extracted frames. Adopting this strategy helps mitigate the impact of isolated frame-level anomalies and enables a more robust and representative evaluation. As a direct consequence, we avoid marking as erroneous a large portion of the dataset composed of 284 videos with

[41] Rössler, Cozzolino, Verdoliva, Riess, Thies, and Nießner, "FaceForensics++", op. cit.

[42] Sabir, Cheng, Jaiswal, AbdAlmageed, Masi, and Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos", op. cit.

only one misclassified frame and 115 videos with exactly two misclassified frames. These videos will henceforth be considered correctly classified in all subsequent video-level analyses. In addition to this literature-based thresholding, we further examined the distribution of wrongly predicted frames per video. Some videos exhibited disproportionately high error counts, as illustrated in Figure 4.9. To better understand these outliers, we isolated the subset of 13 videos with more than five misclassified frames and manually analyzed each one through visual inspection. Among these, several were found to exhibit specific singularities. In video `548_632`, the frame labeled as `face0` corresponds to a background face depicted within a picture frame as can be seen in the Figure 4.8. Although the video is labeled as FAKE, this face is not the actual DeepFake target; the model correctly classifies it as REAL, thus introducing misleading noise. Similarly, in video `305_513` as can be seen in the Figure 4.6, two faces are present, and the one identified as `face0` is not the manipulated subject. Lastly, in video `554_572` as can be seen in the Figure 4.7, the main detected face corresponds to a shadow (like reflection on the wall) again, not the DeepFake target.

To ensure dataset integrity, we excluded these three videos from the subsequent analysis. We also discarded all videos with fewer than three wrongly predicted frames, following the threshold rationale described earlier. The video `186_170.mp4` shows a television anchorwoman with additional faces on a background screen (Figure 4.5), but in this case the target face of the DeepFake has been correctly identified and therefore this does not lead to spurious detections unrelated to the manipulated subject.

To better reflect real-world deployment conditions, where decisions are typically made at the video level rather than per frame, we aggregated predictions by assigning each video the label associated with its corresponding `face0` frame. After defining criteria for classifying a video as mispredicted and analyzing several edge cases, a total of 128 videos were identified as misclassified at the video level. These misclassified instances are characterized by containing between 3 and 8 wrongly predicted frames, indicating a consistent frame-level ambiguity within those videos. Within the evaluation protocol adopted in this thesis, the detector misclassified

Figure 4.5: Video `186_170.mp4`, many faces but the main face is recognized correctly.

Figure 4.6: Video `305_513.mp4`, the main face is not the target of the DeepFake.

Figure 4.7: Video `554_572.mp4`, a face in a photograph in the background is recognized as main face.

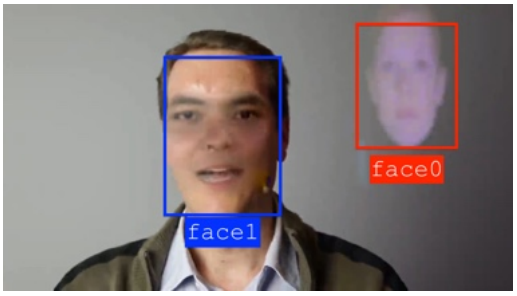


Figure 4.8: Video `548_632.mp4`, a shadow face in the background is recognized as the main face.

123 genuinely `FAKE` videos as `REAL` and 5 genuinely `REAL` videos as `FAKE`. This asymmetry is largely driven by the evaluation protocol, where predictions are first produced at the frame-level and only later aggregated to the video-level. Following established practice in the literature, a video is considered misclassified only if it contains at least three frames whose predicted label does not match the ground truth. The distribution of videos based on the number of frame-level mispredictions is illustrated in Figure 4.10.

When tested at this video granularity, the system achieved a video-level test accuracy of 93.25%. This substantial improvement over the frame-level result confirms the robustness of `face0` as a representative frame for classification. The resulting confusion matrix is shown in Figure 4.4, while a detailed classification report for video-level predictions is presented in Table 4.9. This analysis offers a more application-driven perspective on the effectiveness of the system in practice, where predictions are typically needed per video unit.

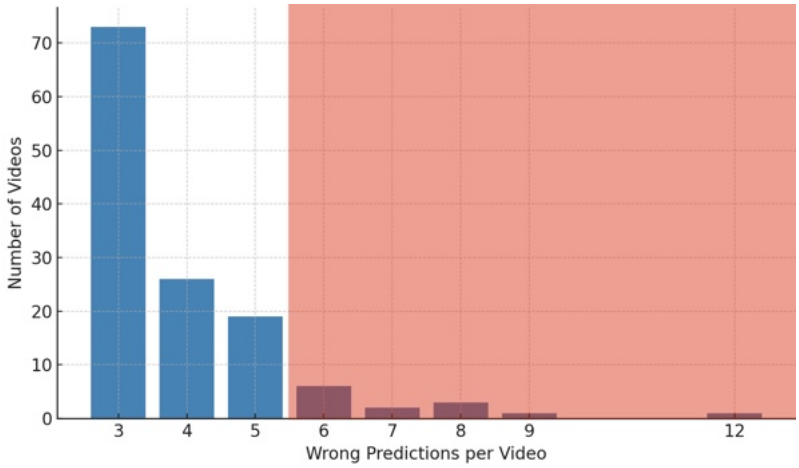


Figure 4.9: Distribution of Videos by Number of Wrong Predictions Including Misleading Videos

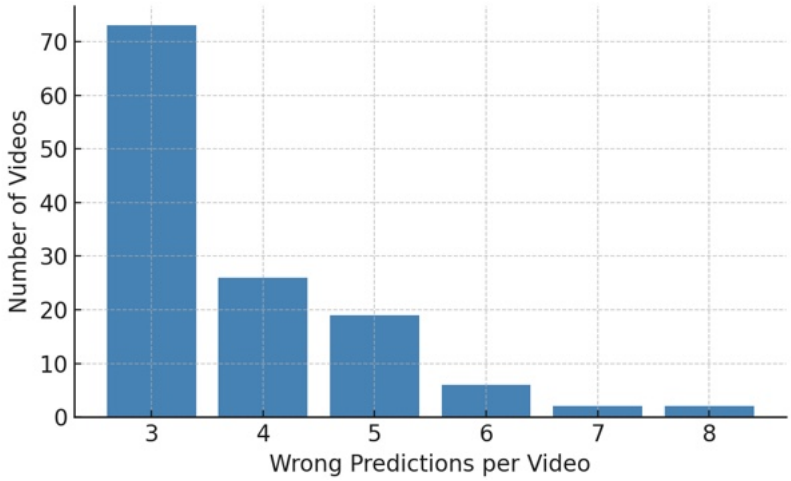


Figure 4.10: Distribution of Videos by Number of Wrong Predictions Excluding Misleading Videos

4.5 Analysis of the relations between labels and wrong predictions

To investigate whether the detector exhibits biases linked to specific facial attributes, we conducted a quantitative analysis correlating classification errors with high-level attribute labels. In particular, we computed standard performance metrics – precision, recall, F1-score, and error rate – for each category within five key features: gender, ethnicity, hair color, hair length, and ear visibility. These metrics provide complementary insights: *precision* (Prec) indicates the proportion of correct predictions among all predictions for a given label, *recall* (Rec) measures the

Table 4.9: Video-level classification report.

Class	Precision	Recall	F1-score	Videos
REAL	0.8848	0.9947	0.9366	950
FAKE	0.9605	0.6807	0.9278	946
Total				1896

ability to correctly detect all instances associated with that label, and the *F1-score* (F1) balances both. The *error rate* (ER) reflects the proportion of misclassified instances out of the total samples for each label. The complete results are reported in Table 4.10.

A few patterns clearly emerge. First, individuals labeled with `LONG` hair show the highest F1-score (0.76), compared to `SHORT` hair (0.69), suggesting that the model performs more reliably on longer hairstyles. Conversely, videos featuring subjects with `SHORT` hair present the highest error rate (30%), indicating a potential challenge for the model in processing these cases. Gender-wise, the model shows slightly better performance on `FEMALE` (F1-score: 0.74) compared to `MALE` (F1-score: 0.72), although the difference is marginal. Notably, the most frequent ethnicity in the dataset, `WHITE`, also has a relatively high error rate (26.9%), which may reflect overfitting to this majority class or hidden variance in the visual features within that group. On the other hand, labels with very few samples (e.g., `BALD`, `OTHER` for hair color, or `PONYTAIL`) display perfect or near-perfect scores. However, this should not be interpreted as strong model performance, as their sample sizes are too small to draw statistically reliable conclusions. Overall, the analysis reveals some signs of performance disparity across features. While no overwhelming bias is detected, attributes such as `SHORT` hair and `WHITE` ethnicity warrant closer scrutiny in further analysis to assess whether augmenting underperforming subgroups could improve fairness in DeepFake detection.

To deepen explore potential biases in the DeepFake detection pipeline, we computed three statistical correlation metrics between the presence of misclassifications (wrong predictions) and each attribute label: *Chi-square test statistic* (Equation 4.1), *p-value*, and *Mutual Information (MI)* (Equation 4.2). Given

Table 4.10: Label-wise classification metrics computed over the test video set.

Feature	Label	Videos	Prec (↑)	Rec (↑)	F1 (↑)	ER (↓)
gender	MALE	774	0.721	0.721	0.721	0.279
gender	FEMALE	1226	0.746	0.746	0.746	0.254
ethnicity	OTHER	16	0.938	0.938	0.938	0.063
ethnicity	ASIAN	66	0.803	0.803	0.803	0.197
ethnicity	WHITE	1890	0.731	0.731	0.731	0.269
ethnicity	AFRICAN	14	0.786	0.786	0.786	0.214
ethnicity	MIXED	14	0.857	0.857	0.857	0.143
hair_color	OTHER	4	0.750	0.750	0.750	0.250
hair_color	BLACK	1482	0.735	0.735	0.735	0.265
hair_color	GREY	112	0.741	0.741	0.741	0.259
hair_color	BROWN	102	0.686	0.686	0.686	0.314
hair_color	L. BROWN	26	0.731	0.731	0.731	0.269
hair_color	BLONDE	286	0.727	0.727	0.727	0.273
hair_length	LONG	1056	0.763	0.763	0.763	0.237
hair_length	SHORT	936	0.700	0.700	0.700	0.300
hair_length	BALD	4	0.750	0.750	0.750	0.250
hair_length	PONYTAIL	4	1.000	1.000	1.000	0.000
is_ears_visible	TRUE	1618	0.737	0.737	0.737	0.263
is_ears_visible	FALSE	382	0.723	0.723	0.723	0.277

Observed Frequence (O) and Expected Frequence (E) from which consequently O_{ij} and E_{ij} for cell (i, j), the *Chi-square* statistic is defined as

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{4.1}$$

and the resulting p-value estimates the probability that a discrepancy at least as large could arise under the null hypothesis of independence. The conventional decision rule adopts the threshold $p < 0.05$ to claim statistical significance.[43] Complementarily, MI quantifies how much information the attribute label X conveys about the correctness indicator Y:

$$MI(X,Y) = \sum_x \sum_y p(x,y) \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right) \tag{4.2}$$

The *Chi-square* test measures the discrepancy between the observed and expected

[43] R. A. Fisher, *Statistical Methods for Research Workers*, Oliver and Boyd, Edinburgh, United Kingdom 1925.

distributions of categorical variables, indicating whether the error rate is uniformly distributed across attribute labels. The *p-value* represents the probability that any observed difference is due to chance. A threshold of $p < 0.05$ is commonly adopted to denote statistical significance.[44] Lastly, *MI* quantifies how much knowing one variable (e.g., the label) reduces the uncertainty about the other (e.g., prediction correctness). Higher MI values imply stronger dependence. The results are summarized in Table4.11. Notably, the attribute `hair_length` stands out with a statistically significant p-value of 0.031793, suggesting that the likelihood of a wrong prediction is not evenly distributed across different hair length categories. This supports earlier classification metrics where `SHORT` hair length was associated with the highest error rate (30.04%). Expanding upon the previous analysis, which examined overall model

Table 4.11: Statistical correlation metrics between features and wrong predictions.

Feature	Chi-Square	p-value	MI
gender	1.449053	0.228680	0.000393
ethnicity	6.332719	0.175640	0.001902
hair_color	3.999698	0.779812	0.001166
hair_length	10.573959	0.031793	0.003107
is_ears_visible	1.190591	0.275210	0.000328

performance using Precision, Recall, F1-Score and Error Rate as well as statistical correlation metrics by attribute category, we now explore these relationships at a finer granularity. Specifically, we assess bias in model misclassifications by analyzing individual label values for each attribute.

The results, summarized in Table 4.12, indicate that the label `hair_length = SHORT` exhibits a statistically significant correlation with misclassifications ($p < 0.05$), suggesting a potential association between this attribute and increased prediction errors. While `ethnicity = WHITE` also shows statistical significance, this finding should be interpreted with caution. Given that this label is overwhelmingly dominant in the dataset, the higher error count may primarily reflect its prevalence rather than a specific model weakness toward this subgroup. Notably, `hair_length = LONG`, although above the threshold for statistical significance, stands out due to its

[44] Ibid.

performance in prior metrics, showing the highest F1-score (0.76) among hair length categories. This supports the idea that this label may be a particularly meaningful feature for DeepFake detection and warrants further investigation as a potentially informative trait rather than a source of model bias. These results emphasize the importance of considering label distribution when evaluating bias and suggest that more balanced datasets or stratified analyses are needed to isolate the true effect of features on prediction performance.

Table 4.12: Label-wise statistical correlation metrics with wrong predictions.

Feature	Label	Chi-Square	p-value	MI
gender	MALE	1.4491	0.2287	0.0000
gender	FEMALE	1.4491	0.2287	0.0143
ethnicity	OTHER	2.3948	0.1217	0.0000
ethnicity	ASIAN	1.2224	0.2689	0.0059
ethnicity	WHITE	4.4596	0.0347	0.0000
ethnicity	AFRICAN	0.0132	0.9084	0.0000
ethnicity	MIXED	0.5240	0.4691	0.0000
hair_color	OTHER	0.0000	1.0000	0.0000
hair_color	BLACK	0.0533	0.8174	0.0000
hair_color	GREY	0.0000	0.9979	0.0000
hair_color	BROWN	1.1377	0.2861	0.0000
hair_color	LIGHT BROWN	0.6108	0.4345	0.0000
hair_color	BLONDE	0.1414	0.7069	0.0139
hair_color	RED	0.0019	0.9654	0.0000
hair_length	LONG	3.5820	0.0584	0.0098
hair_length	SHORT	6.0163	0.0142	0.0000
hair_length	BALD	0.0000	1.0000	0.0000
hair_length	PONYTAIL	2.3158	0.1281	0.0043
is_ears_visible	FALSE	1.1906	0.2752	0.0148
is_ears_visible	TRUE	1.1906	0.2752	0.0000

4.6 Conclusions and future works

This study validates the feasibility of enriching DeepFake detection pipelines with automatically labeled facial attributes obtained via a semi-supervised learning process. This work contributes to the DeepFake detection community in three ways:

1. It presents a semi-supervised attribute labeling strategy applicable at scale.

2. It identifies performance disparities across attribute subgroups.

3. It highlights critical areas where bias mitigation should be prioritized in future detection pipelines. Our results demonstrate that such auxiliary information based on visual attributes can support improved detection.

Future works will be dedicated to better understand how such side information could be adopted during the model training phase to support DeepFake detection and, based on the result detailed in Section 4.5 should therefore enlarge the representation of short-haired subjects, enrich ear-occlusion scenarios, and incorporate fairness-aware regularisers to guarantee balanced performance across hairstyle and visibility conditions. While our current analysis employs classical significance testing (*chi-square* and MI), in future extensions we plan to apply multivariate analyses, such as logistic regression with interaction terms, generalized linear models or causal inference frameworks, on balanced datasets to better understand interaction effects between features and improve generalizability of the findings to model the interdependencies among attributes and error outcomes more precisely. To promote transparency and encourage further research, we make available all resources required to replicate our experiments[45].

[45]Public GitHub Repository created by the author: https://github.com/vstile/deepfake-attribute-detection (visited on 17 December 2025).

Chapter 5

Attribute-Aware Training Strategies

5.1 Introduction

5.1.1 Context and continuity: bias analysis

This section extends the study presented at *SGSOACS (Vienna)* on the relationship between DeepFake detection errors and facial attributes, we moved from simple correlation checks to *controlled exclusion experiments* and a shared evaluation pipeline on *accuracy*, *F1* and *AUC* with balanced splits (72/14/14) and fixed random seeds. In this phase we trained multiple models *excluding in training one or more attributes*, then evaluated *all* models on a *single complete test set* with no exclusions, to measure out-of-distribution generalization toward subgroups never seen during learning. Setup details are consolidated in Chapter 4, Figure 4.1 and summarized in Section 5.1.2; full code is available in the author's GitHub profile [1]. The overall workflow is: video input selection from *FaceForensics++*, frame extraction with `Haar cascades`, semi-supervised attribute labeling, exclusion of one or more labels, training a `VGG16`-based classifier, and test analysis with metric reporting and by-subgroup bias summaries and is represented in the Figure 5.1. In details starting from the *FaceForensics++* video corpus (REAL: `/youtube`, FAKE: `/Deepfakes`), frames are extracted and faces cropped with Haar Cascades (resize to 224×224, normalization). A semi-supervised `ResNet18` labeller assigns facial attributes

[1] The author's GitHub profile: https://github.com/vstile (visited on 21 February 2026)

(gender, hair color, hair length, ethnicity, ear visibility). One or more attribute values are then excluded to simulate distribution gaps during training, after which a `VGG16`-based classifier is trained on frame-level inputs and evaluated on a fixed test set. The analysis stage reports accuracy, F1, AUC, and subgroup confusion patterns, saving graphs, data, and model checkpoints.

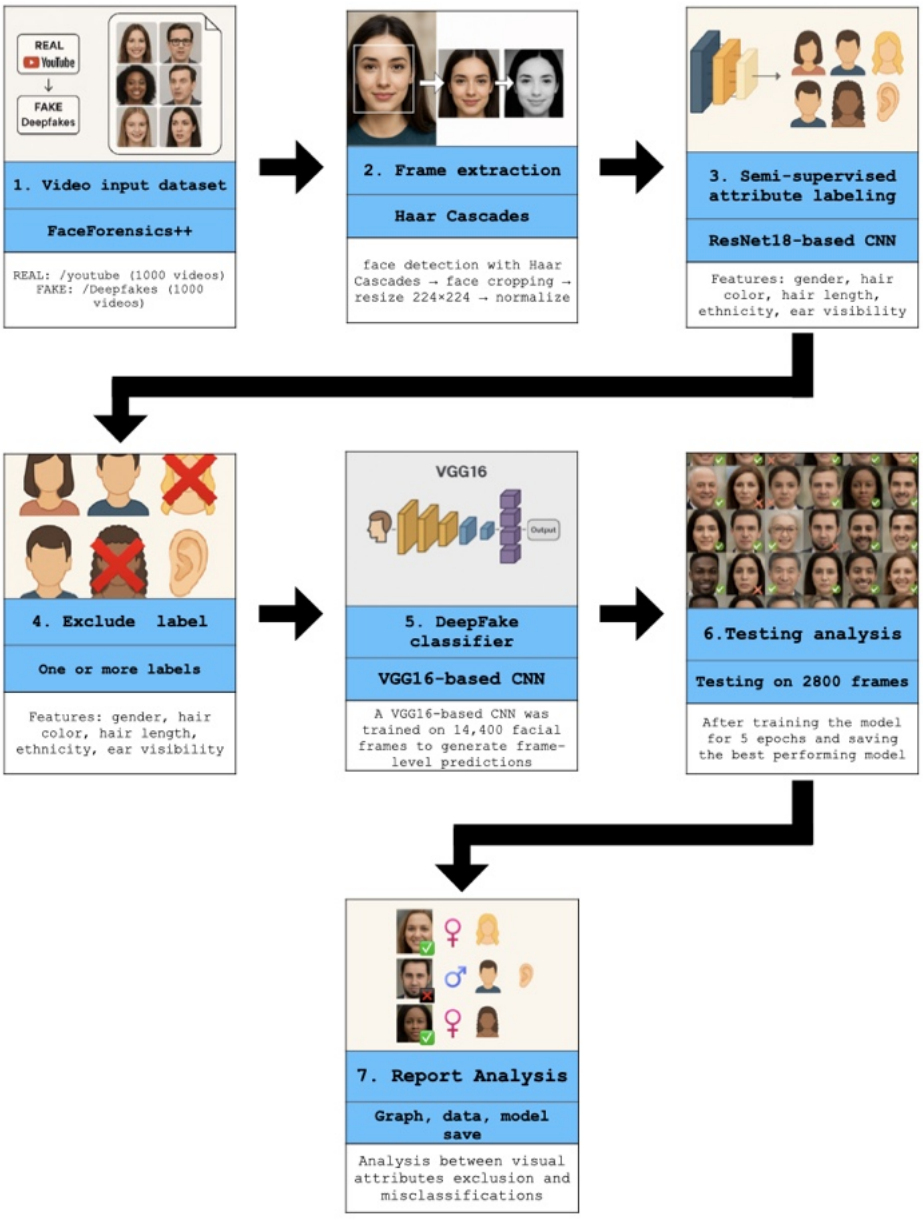


Figure 5.1: Overview of the attribute-aware exclusion pipeline. Steps: (1) video input, (2) frame extraction, (3) semi-supervised attribute labeling, (4) label exclusion, (5) DeepFake classifier, (6) testing analysis, (7) report and model save.

5.1.2 Experimental protocol

Experimental setup overview. Building on the frame-level detection and attribute-labeling pipeline introduced in Chapter 44 and summarized in Figure 4.1, this chapter adopts FaceForensics++ as benchmark and restricts the analysis to the `/youtube` (REAL) subset and its corresponding `/Deepfakes` (FAKE) manipulation subset under the `C40` compression level. Videos are decoded and uniformly sampled to extract frames; faces are detected and cropped with a classical detector and then resized to 224×224 with intensity normalization. DeepFake detection is performed at frame level using a `VGG16` backbone initialized from `ImageNet` weights with frozen convolutional blocks and a lightweight dense head, optimized with `Adam` and binary cross-entropy and selected on validation performance. In parallel, high-level facial attributes (e.g., gender, ethnicity, hair color, hair length, ear visibility) are obtained via a semi-supervised procedure that starts from a small manually annotated seed and propagates pseudo-labels to the remaining videos under a high-confidence criterion. The core evaluation in this chapter follows a controlled exclusion protocol (Figure 5.1): for each attribute value, all samples with that value are removed from training while the test distribution is kept fixed, enabling a direct assessment of subgroup-conditioned generalization gaps. Performance is reported globally and by subgroup using accuracy, F1, AUC, and confusion-derived rates, with minimum-support filtering to avoid unstable estimates for very small subgroups. Full scripts, configuration, and analysis artefacts are provided in the accompanying public repository to support reproducibility.

Design. For each attribute A (e.g., gender, hair length, ear visibility) and value $v \in \mathcal{V}_A$:

1. **Exclusion training**: remove from training all samples with $A{=}v$;
2. **Universal test**: evaluate the resulting model on the complete test set that includes *all* values of A, including v.

This design isolates the *generalization bias* induced by the systematic absence of a subgroup during learning while keeping the test distribution fixed. The operational

pipeline uses face cropping, normalization to 224×224, and a `VGG16` backbone with a dense head, optimized with `Adam` and binary cross-entropy with early stopping on validation AUC. Metrics include accuracy, F1-macro, AUC, and the confusion-derived True Positive Rate (TPR) (recall on REAL) and True Negative Rate (TNR) (specificity on FAKE). In our latest runs, the training set consisted of $N = 100{,}000$ balanced facial frames, and the classifier was trained for up to 5 epochs, with model selection on validation performance.

Furthermore we set a minimum subgroup size (n) that defines a support threshold for the analysis: any attribute subgroup with fewer than five instances (n≥5) is excluded from the plots, that is, we only report metrics for subgroups with (n≥5). This choice filters out very small subgroups whose performance estimates (e.g. accuracy, F1, AUC) are highly unstable, since a single misclassification can substantially change the measured values. By restricting visual comparison to subgroups with (n≥5), the resulting graphs emphasise patterns supported by a non-trivial amount of data and reduce the risk of over-interpreting fluctuations driven purely by sampling noise. The Python code implementing this procedure is available in the author's GitHub profile, where the threshold (n) is denoted by the parameter `MIN_N` [2].

Baseline operating point. For the no-exclusion model (`excl-none`) we obtain Accuracy = 0.806 and AUC = 0.823, but the confusion matrix reveals a skew toward the REAL class: TN=4715, FP=2285, FN=427, TP=6573, hence $\text{TPR} = 6573/(6573 + 427) = 0.939$ and $\text{TNR} = 4715/(4715 + 2285) = 0.674$. These values are summarized in Table 5.1.

This imbalance is expected when using a fixed decision threshold of 0.5: the AUC indicates good ranking ability, yet the chosen operating point favors recall on REAL at the expense of specificity on FAKE. In practice, threshold calibration on the validation set, for example choosing the threshold that maximizes Youden's J or equalizes TPR and TNR, would shift the operating point along the ROC curve

[2]The complete `Phython` scripts for subgroup metric computation and plotting, including the definition of `MIN_N`, are provided in the author's public GitHub repository: https://github.com/vstile/deepfake-attribute-detection/blob/main/notebooks/DeepFace_Detect_0.8_VGG16.ipynb (visited on 26 December 2025).

Table 5.1: Performance metrics for the `excl-none` model.

Metric	Symbol	Value
Accuracy	Acc	0.806
Area under ROC curve	AUC	0.823
True negatives	TN	4715
False positives	FP	2285
False negatives	FN	427
True positives	TP	6573
True positive rate	TPR	0.939
True negative rate	TNR	0.674

and reduce this asymmetry.[3] A second source of asymmetry arises at video level, where frame scores are aggregated with a rule such as "classify a video as FAKE if at least k frames are predicted FAKE". With $k = 3$ a detector can still miss a FAKE video when manipulations are temporally sparse or when face detection fails on a fraction of frames, while a REAL video rarely accumulates three false positives, which tilts errors toward FAKE $\rightarrow$ REAL. This behavior is consistent with prior observations that thresholding and protocol choices substantially affect reported performance in *FaceForensics++* and related benchmarks[4].[5] Moreover, high–quality or well–blended fakes can compress artifacts and reduce separability, increasing FAKE to REAL errors at conservative thresholds.[6] For fair comparison, thresholds for both frame level and video level should be selected on a held-out set according to the target criterion, for example maximum F1 or balanced accuracy, and all metrics should also be reported by subgroup when relevant.

5.1.3 Focus on `hair_length`

Distribution. In the test set, `SHORT` and `LONG` are widely represented, while `PONYTAIL` is rare. Metrics for `PONYTAIL` should be interpreted with caution due to sample size.

[3] Verdoliva, "Media Forensics and DeepFakes", op. cit.; Rana, Nobi, Murali, and Sung, "Deepfake Detection", op. cit.

[4] Rössler, Cozzolino, Verdoliva, Riess, Thies, and Nießner, "FaceForensics++", op. cit.

[5] Li, Yang, Sun, Qi, and Lyu, "Celeb-DF", op. cit.

[6] Beckmann, Hilsmann, and Eisert, *Fooling State-of-the-Art Deepfake Detection with High-Quality Deepfakes*, op. cit.

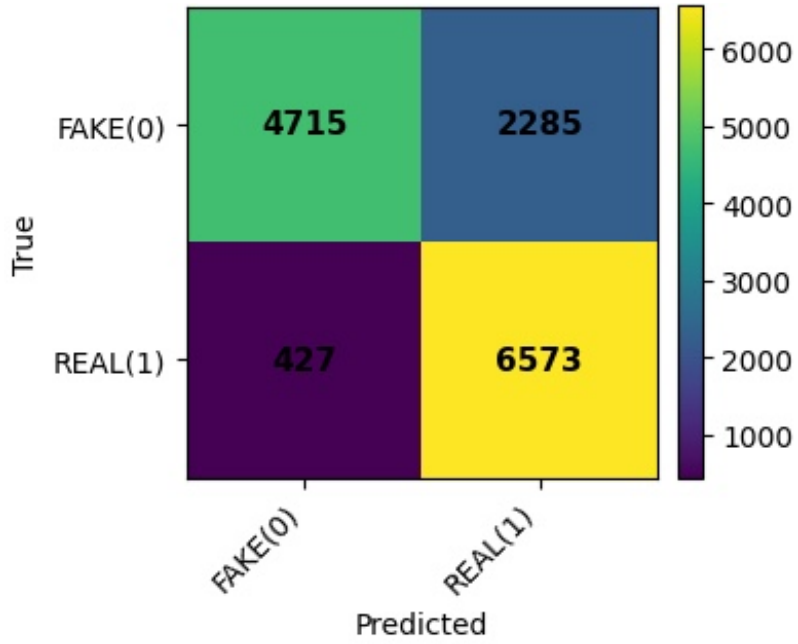


Figure 5.2: Confusion matrix for the *baseline* model trained without exclusions and evaluated on the full test set (n=14,000 frames). Accuracy = 0.806, TPR on REAL = 0.939, TNR on FAKE = 0.674. This serves as the reference for attribute-exclusion experiments.

Subgroup baseline and exclusions. Without exclusions, performance is slightly higher on `LONG` than `SHORT`, with `PONYTAIL` showing optimistic figures that reflect low counts. Excluding `SHORT` in training produces a noticeable decline when generalizing to `SHORT` at test time, consistent with a generalization gap. Excluding `LONG` is more tolerable: it improves robustness on `SHORT` with a manageable penalty on `LONG`. Overall, `hair_length` behaves as a *context feature*: it matters and should be balanced during training, but it is not the most critical factor in the system.

5.1.4 Focus on `is_ears_visible`

Why it is a special case. Ear visibility changes the *visual geometry* of the face: lateral occlusions, hairline shape, peri-auricular contours, and skin versus hair textures. It is a morphological indicator that can become a spurious decision signal if not properly covered in training.

Subgroup baseline. With no exclusions we observe similar accuracy and AUC for `is_ears_visible=1` and `is_ears_visible=0`, with a small advantage for visible

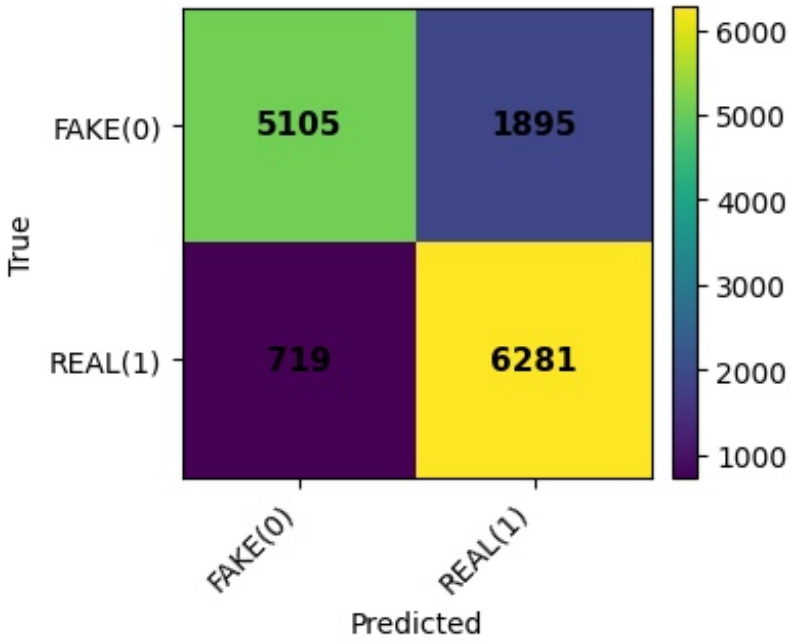


Figure 5.3: Confusion matrix for the model trained *excluding* samples with `is_ears_visible=0` and evaluated on the full test set (n=14,000 frames). Accuracy = 0.813, TPR on REAL = 0.897, TNR on FAKE = 0.729. Performance is comparable to baseline, with higher specificity and slightly lower sensitivity.

Table 5.2: Controlled exclusion on ear visibility, `VGG16`-based classifier, unified test set.

Experiment	**Seed**	**Accuracy**	**AUC (global)**	n_{test}
Baseline (no exclusion)	42	0.806	0.823	14,000
Excluding `is_ears_visible=1`	42	0.741	0.763	14,000
Excluding `is_ears_visible=0`	42	0.813	0.832	14,000

ears, which suggests a latent dependency on this detail.

Excluding `is_ears_visible=1` (never seeing visible ears in training). This is the *worst* case observed in our latest runs: *Accuracy* = 0.741 and *AUC* = 0.763. The model becomes biased toward REAL and lets many FAKE samples pass, indicating a collapse in specificity when the experience of visible ears is removed during learning.

Excluding `is_ears_visible=0` (never seeing non-visible ears in training). In this test, performance does not degrade; on the contrary, it is slightly higher than the baseline, with *Accuracy* = 0.813 (+0.007) and *AUC* = 0.832 (+0.009). This suggests that cases with non-visible ears provide relatively weak or noisy supervisory signal for this detector, so removing them from training does not harm, and may

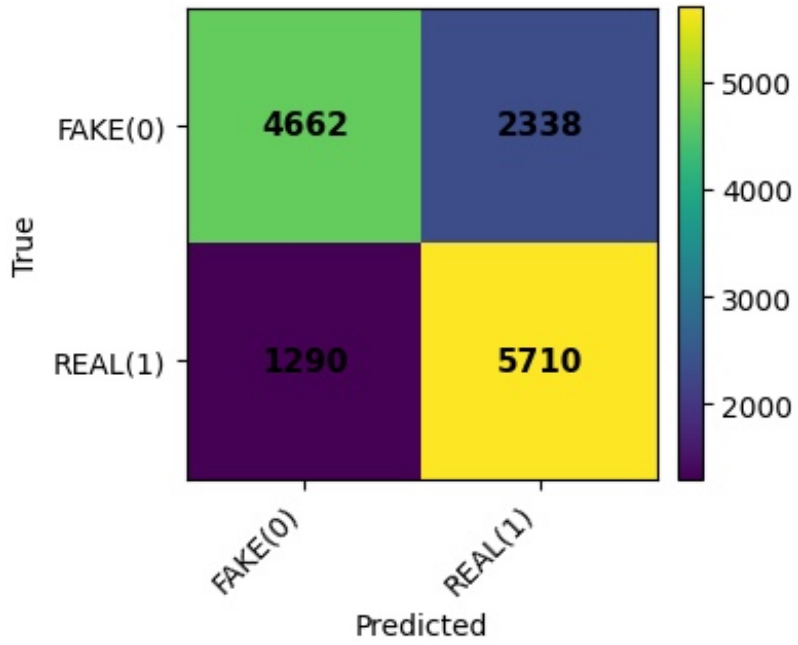


Figure 5.4: Confusion matrix for the model trained *excluding* samples with `is_ears_visible=1` and evaluated on the full test set (n=14,000 frames). Accuracy = 0.741, TPR on REAL = 0.816, TNR on FAKE = 0.666. This is the worst case, with a marked drop in both sensitivity and specificity.

even stabilise, the decision boundary. Crucially, this behaviour contrasts with the severe performance drop observed when excluding `is_ears_visible=1`, where ear visibility acts as a highly informative contextual cue.

Critical analysis. Both experiments converge on one point, `is_ears_visible` is the highest-impact attribute. Removing `is_ears_visible = 1` from training produces a marked drop in accuracy and AUC, with an unsafe rise of FAKE as false negatives. We observe a substantial degradation in AUC across all attributes, using the same evaluation protocol as in the previous test. This effect is clearly visible in the accuracy plots for *gender* (Figure 5.6) and *hair_color* (Figure 5.7), as well as in the AUC plots for *hair_length* (Figure 5.8) and *is_ears_visible* (Figure 5.9). For readability, the legend of experimental runs is reported in Figure 5.5. Removing `is_ears_visible=0` has a non-impacting effect and can even improve performance. In operational terms, coverage of ear occlusion is a critical requirement for robust detection.

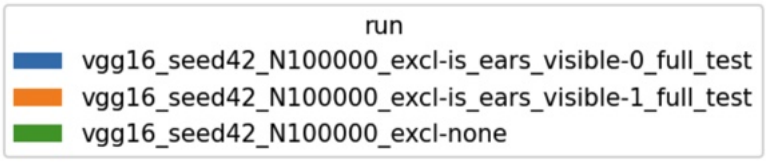


Figure 5.5: Legend of the experimental runs used in the comparative plots. Blue: training excludes samples with `is_ears_visible=0`. Orange: training excludes samples with `is_ears_visible=1`. Green: baseline with no exclusions. All runs use `seed=42`, `N=10,0000` frames, and are evaluated on the full test set.

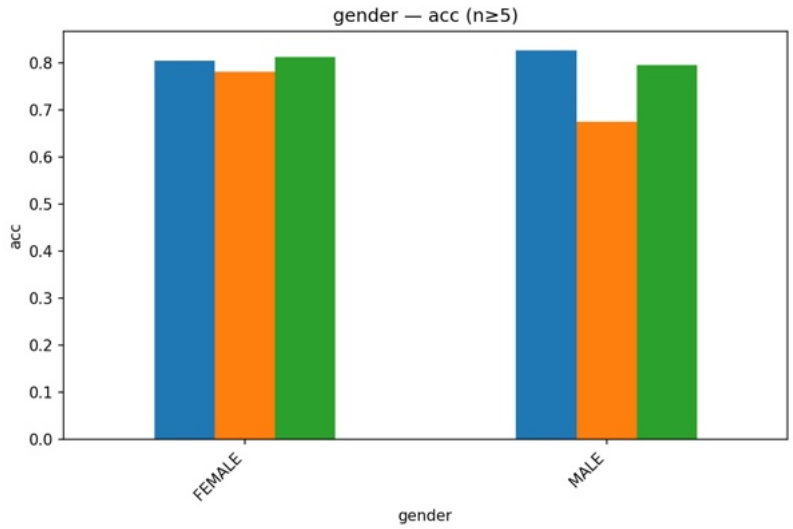


Figure 5.6: Accuracy by gender (n≥5). Bars show the three training regimes: blue excludes `is_ears_visible=0`, orange excludes `is_ears_visible=1`, green is the baseline without exclusions; excluding samples with visible ears reduces accuracy most for the MALE subgroup.

5.1.5 Global results and confusion-matrix analysis

Summary. Controlled exclusion experiments reveal three recurring patterns:

1. **Overall degradation** when a *decision-relevant* subgroup, such as the label `is_ears_visible = TRUE`, is excluded from training, visible as a drop in accuracy and AUC as shown in the Table 5.2;

2. **TPR/TNR rebalancing**, where the model becomes either more conservative or more permissive depending on the excluded subgroup;

3. **Distribution sensitivity**, with larger impacts when the excluded subgroup is numerically relevant or corresponds to a *structural* variation of facial appearance.

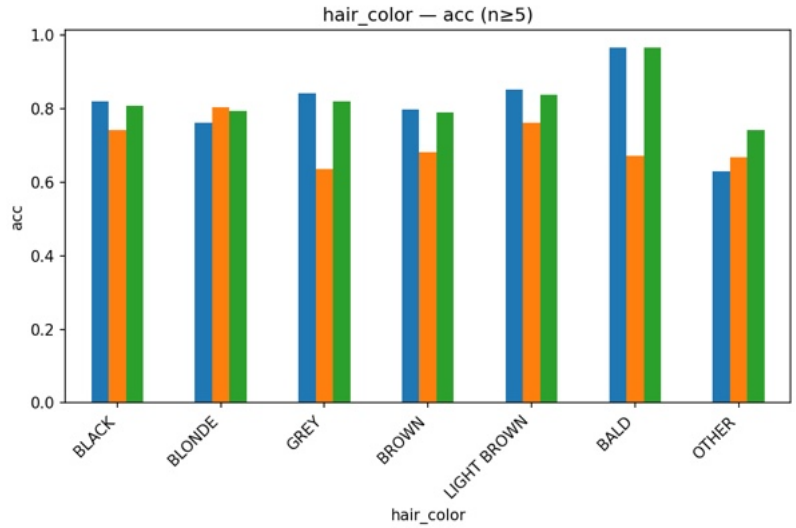


Figure 5.7: Accuracy by hair color (n≥5). Performance across hair colors under the three regimes (blue: exclude `is_ears_visible=0`; orange: exclude `is_ears_visible=1`; green: baseline). Removing samples with visible ears produces the broadest accuracy erosion across categories.

Two attributes emerge as particularly influential:

- **Hair length** (`hair_length`): a moderate yet consistent effect;
- **Ear visibility** (`is_ears_visible`): a markedly critical effect for correct FAKE/REAL discrimination.

5.1.6 Conclusions

Interactions between `hair_length` and `is_ears_visible`. Hair `LONG` or `SHORT` indirectly affects `is_ears_visible` because long hair often covers the ears. Many false negatives arise when `SHORT` coexists with `is_ears_visible=0`, this condition certainly highlights a particular case in that for these two conditions to occur we must be faced within oblique views, with backlighting, or with side fringes. This explains why balancing the length of the hair alone is not sufficient on its own but must always be analyzed in parallel with the visibility of the ears.

Operational recommendations (bias-aware training). The whole analysis in the Chapter 5 shows that some visual attributes, in particular `hair_length` and `is_ears_visible`, are not merely descriptive metadata but act as determinants of

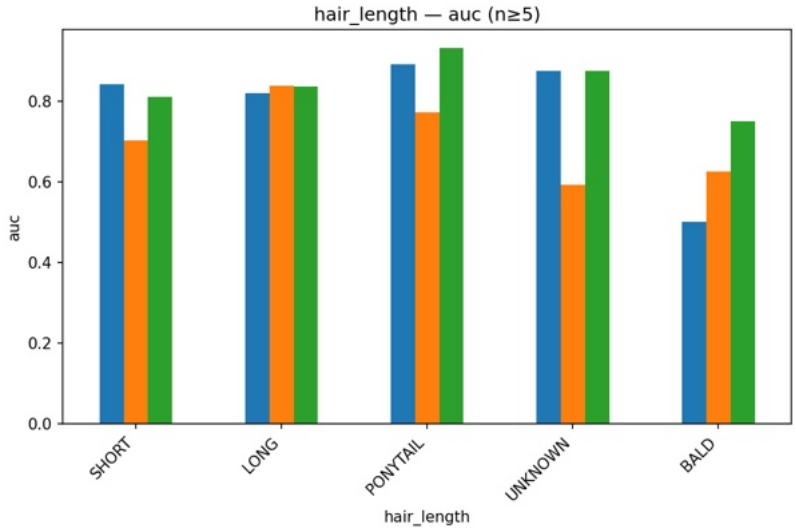


Figure 5.8: AUC by hair length (n≥5). Comparison of the three regimes (blue: exclude `is_ears_visible=0`; orange: exclude `is_ears_visible=1`; green: baseline). Excluding visible ears notably depresses AUC for `SHORT` and `UNKNOWN`, while the baseline attains the highest AUC for `PONYTAIL` and improves `BALD`.

model reliability. The label `is_ears_visible = TRUE` has the strongest impact on correct DeepFake recognition, excluding cases with visible ears during training leads to a sharp reduction in *accuracy* and *AUC*, with a dangerous drift toward REAL predictions on FAKE instances. Hair length has a *moderate* effect that is manageable through balanced data and minor training adjustments. So when specific conditions are under-represented during training, the detector systematically degrades on those groups, which creates both a security vulnerability and an equity issue, since certain physical traits or demographic profiles receive consistently worse treatment.

From an operational perspective, training data should therefore be curated with explicit coverage targets for the most influential attributes. In the case of `is_ears_visible`, this implies ensuring a broad and balanced set of examples for both values `0` and `1`, combined with realistic pose variation, lateral occlusions, headwear and hair coverage. In particular, additional emphasis should be placed on `is_ears_visible=0` to prevent the observed collapse in specificity that arises when the model is exposed mainly to uncovered ears in training. The same considerations apply at test and deployment time. Attribute distributions should be monitored and performance should be reported per attribute group, rather than only in aggregate, so that systematic blind spots can be identified. In high-risk applications, it may be

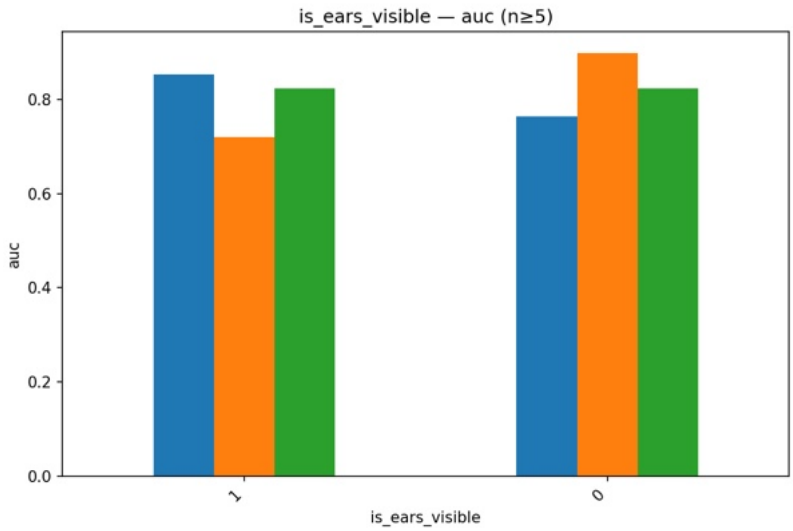


Figure 5.9: AUC by ear visibility (n≥5). When training excludes `is_ears_visible=1` (orange), AUC drops on cases with ears visible and rises on cases with ears not visible; excluding `is_ears_visible=0` (blue) has the opposite effect. The baseline (green) stays between the two.

appropriate to adopt attribute-aware calibration or decision thresholds, for example by tightening the acceptance criteria for configurations that historically exhibit higher error rates. This type of bias-aware training and evaluation supports both security, by reducing exploitable weaknesses, and fairness, by promoting a more equal treatment of faces across gender, appearance, and other sensitive characteristics.

Chapter 6

Conclusions and Future Work

This thesis explored whether semi-supervised facial attribute labeling can support a more detailed analysis of DeepFake detection systems. The work combined a standard face-centric pipeline on *FaceForensics++* with an automatic attribute labeling step and a series of experiments on misclassification patterns and attribute-aware training. The focus was not on proposing a new state-of-the-art detector, but on examining how conventional detector behaves when predictions are viewed through the lens of high-level visual attributes.

From an empirical point of view, the study suggests that attributes may have some value for post hoc analysis and monitoring. Therefore, integrating a semi-supervised labeling strategy into the pipeline of a detection system may inform an improvement in the state of the art. The semi-supervised procedure, based on manual annotation of a small seed of videos and a `ResNet18` labeller, produced a full set of labels for the `/youtube` subset of *FaceForensics++*. These annotations were then used to compute error statistics by subgroup and simple dependence measures such as chi-square and mutual information. Although the results are exploratory, they indicate that certain appearance factors, in particular hair length and ear visibility, are associated with variations in error rates and may therefore act as informative context for bias-aware inspection. Training detectors while withholding specific attribute values and then testing on a complete, fixed test set revealed moderate sensitivity to `hair_length` and a stronger sensitivity to `is_ears_visible`. In particular, removing cases with visible ears during training was associated with a noticeable drop in accuracy and

AUC and with an increase in FAKE videos misclassified as REAL. These findings do not establish causal relations, but they suggest that some visual attributes can influence the reliability of standard DeepFake detectors and may deserve explicit consideration during data curation and evaluation.

6.1 Limitations

All experiments are confined to a single dataset and compression level, with a strong dominance of one ethnicity and limited coverage of minority groups. The attribute set is relatively small and focused on coarse appearance cues. The analysis relies on univariate statistics and simple aggregation rules, which are suitable for an initial exploration but cannot capture more complex interactions among attributes. Finally, the detector architecture is deliberately simple, so the observations reported here may not transfer directly to more advanced multimodal or transformer-based systems.

Scope of validity. The evidence reported in this thesis is conditional on the specific benchmark configuration and evaluation protocol adopted. All experiments are confined to a single dataset and compression level [1], a face-centric preprocessing pipeline and the reported subgroup effects should therefore not be assumed to transfer to other benchmarks, manipulation families, or compression regimes without dedicated replication. The dataset is characterised by a strong dominance of one ethnicity and limited coverage of minority groups, hence estimates for under-represented subgroups are inherently less stable. The attribute set is relatively small and focused on coarse appearance cues. The analysis relies on univariate statistics and simple aggregation rules, which are suitable for an initial exploration but cannot capture interactions among attributes, nor support causal claims. Finally, the detector architecture is deliberately simple, so the observations reported here may not transfer directly to more advanced multimodal or transformer-based systems.

In light of these constraints, the main contribution of this thesis should be

[1] FaceForensics++ restricted to the `/youtube` vs `/Deepfakes` subsets at `C40` compression.

considered methodological and exploratory. The attribute-labeling pipeline, the released annotations and code, and the exclusion protocol may be seen as building blocks for more comprehensive studies on bias and robustness in DeepFake detection, rather than as definitive answers on fairness or security.

6.2 Future Work

Future work may extend this line by adopting richer attribute sets, more diverse datasets, and multivariate analysis, and by integrating attribute information directly into training or calibration procedures. Within such a broader agenda, the present results can be viewed as a first indication that facial attributes may help organise error analysis and may support more transparent and bias-aware evaluation of DeepFake detectors.

Several directions emerge from the present study and may be explored in future work:

1. A first line of extension concerns data and attributes. The current experiments are limited to a single benchmark, one compression level, and a relatively small set of coarse attributes. Replicating and extending the analysis on more diverse datasets, richer attribute taxonomies, and better-balanced demographic distributions could clarify to what extent the observed patterns generalise beyond *FaceForensics++*.

2. A second direction involves the analytical tools. The thesis deliberately relied on univariate statistics and simple aggregation rules, which are suitable for an initial exploration but cannot capture complex interactions among attributes. Future work may exploit multivariate models, for example logistic regression with interaction terms, generalized linear models, or causal-inspired analyses on appropriately balanced data. Such methods may offer a more precise view of how combinations of attributes relate to error patterns, while also quantifying uncertainty more rigorously.

3. A third line concerns the integration of attributes into the detection process

itself. In this thesis, attributes are used as side information for evaluation and diagnostic purposes. In future studies, attribute signals could be incorporated for selective training, calibration, or threshold selection, for instance through attribute-aware sampling, regularisation, or group-wise operating points. Such strategies may help the fairness and robustness of the DeepFake detectors .

6.3 Final remarks

This thesis introduced an attribute-aware methodology for interrogating DeepFake detectors, prioritising critical behaviour analysis over raw performance optimisation. By combining semi-supervised facial attribute labelling with controlled exclusion training, the proposed framework makes detector behaviour observable in terms of concrete appearance factors and dataset coverage. Two outcomes are particularly salient. First, attribute-conditioned inspection suggests a moderate but recurrent sensitivity to hair length, motivating explicit monitoring of hairstyle coverage in training and evaluation. Second, ear visibility emerges as a stronger contextual factor: when the training distribution lacks specific ear-visibility configurations, performance degrades on a fixed, complete test set and the detector exhibits an asymmetric failure mode that increases `FAKE` $\rightarrow$ `REAL` errors (false negatives). This asymmetry is operationally important in forensic and security settings, where false negatives are especially costly. Overall, the study supports the recommendation that robustness audits should track attribute coverage, report subgroup metrics, and consider attribute-aware curation and calibration to mitigate systematic blind spots. The released code, annotations, and evaluation protocol are intended to enable replication and extension to additional datasets, manipulation types, and detector families, turning these findings into a broader agenda for bias- and robustness-aware DeepFake detection.

Appendix 1: Multidisciplinary Application of Artificial Intelligence

The author, during the three years of the PhD program, has also carried out collateral research activities on the application of AI in multidisciplinary domains. Drawing on a background in industrial management engineering, these projects explored AI enabled support for Building Information Modeling (BIM), data governance, higher education and business organisation. Although distinct from the core topic of DeepFake detection, they share recurring themes such as data readiness, explainability, and role sensitive design. The following works are therefore included as complementary examples of how AI can be integrated into complex socio technical systems beyond the forensic setting.

Engineering and Building Information Modeling (BIM)

In civil engineering I combine AI with BIM, implemented via openBIM to ensure interoperability and data governance, thereby supporting AI-driven analytics, AI-powered workflows and digital twinning, and Big Data pipelines to transform static models into living digital twins. The work establishes end-to-end data workflows, from Industry Foundation Classes (IFC)-aware extraction and Internet of Things (IoT) streams to feature stores, model interrogation, and automated checking. It argues for machine-readable requirements and auditable analytics, and it demonstrates measurable gains in reporting, compliance, and time-to-insight through pilot studies

conducted with practitioners.

AI-Enhanced Building Information Modeling and Big Data Analytics for Civil Engineering Innovation

Authors: Vittorio Stile and Antonio Fontanella (2025, October 17)

Abstract. *The integration of Artificial Intelligence (AI), Building Information Modeling (BIM), and big data analytics is emerging as a critical enabler for the digital transformation of civil engineering. By combining predictive modeling, machine learning, and computer vision with large-scale data processing, BIM platforms can evolve into adaptive decision-support systems. This research introduces a prototype platform designed to optimize building design, resource allocation, and regulatory compliance, while simultaneously analyzing heterogeneous data sources generated throughout the construction lifecycle. The incorporation of big data analytics allows for the identification of patterns and correlations that enhance predictive accuracy and improve real-time monitoring. Preliminary applications in urban infrastructure projects demonstrate measurable gains in resource efficiency, risk mitigation, and project delivery times. Despite these promising results, broader empirical validation and sustained research funding remain essential to generalize findings across diverse construction scenarios. The study positions AI-BIM integration, supported by big data analysis, as a pivotal step toward intelligent, data-driven, and sustainable construction engineering.*

Comment. In this position paper, written with Antonio Fontanella (Ordine degli Ingegneri della Provincia di Napoli (OIN) and Ingegneria Specialistica Fontanella (ISF)), and presented to the 4th Conference on Creativity and Innovation in Digital Economy (CIDE) in Ploiești, we propose the integration of AI, BIM, and big data analytics as an enabler for the digital transformation of civil engineering, introducing a prototype platform that evolves BIM into an adaptive decision-support system with measurable gains in efficiency and governance.[2] The approach combines

[2] V. Stile and A. Fontanella, *AI-Enhanced Building Information Modelling and Big Data Analytics for Civil Engineering Innovation*, ENG, in: *Book of Abstract of the 4th International Conference*

openBIM practices, IFC parsing, and Information Delivery Specification (IDS)–based acceptance tests with an Extract, Transform, Load (ETL) pipeline and a feature store for heterogeneous sources (IFC, IoT streams, site logs). The system architecture comprises an AI engine (Large Language Models (LLMs), Computer Vision (CV), rule-based reasoning), a Big Data layer (ingestion, governance, analytics), and a BIM interface (Application Programming Interfaces (APIs) and scripting for Revit/BlenderBIM) orchestrated by Decision Support Services for predictive, compliance, and optimization modules. The methodology spans supervised prediction on IFC-derived graphs and BIM-rendered imagery, computer vision for inspection and progress, hybrid compliance checking (ontology/logic and LLM+DL pipelines) anchored to IDS, and search-based planning where A* baselines outperform conditional Generative Adversarial Network (cGAN) under tested coordination settings. Two pilots (building and urban corridor) demonstrate scripted querying, Quantity Take-Off (QTO) automation, and streaming connectors, with time reductions surfaced as Key Performance Indicators (KPIs) and fed back to learning pipelines. Benefits include efficiency (task-time reductions), cost/risk control (earlier deviation detection), and governance/compliance (machine-readable requirements with IDS), while challenges persist in IFC-to-AI data readiness, explainability and assurance for regulatory contexts, generalization under domain shift, and skills for sustained adoption. The paper emphasizes human-in-the-loop oversight for IDS curation, review of automated findings, and arbitration of optimization trade-offs. Conclusions are position-like and hypothesis-generating: the AI–BIM–Big Data stack is feasible and promising but requires broader empirical validation. Future work details pre-registered protocols, public datasets and ground truths, effect sizes with uncertainty, ablations, reproducibility artifacts, and user-acceptance measures (Unified Theory of Acceptance and Use of Technology (UTAUT)/Technology Acceptance Model (TAM)) within a closed-loop sensing–prediction–feedback operation.

Creativity And Innovation In Digital Economy, vol. Section 1: Innovative open business models and platforms, Section 1: Innovative open business models and platforms, Petroleum-Gas University of Ploiești Publishing House, Petroleum-Gas University of Ploiești (UPG), Ploiești, Romania Oct. 2025, ISBN: ISSN 2971–9798.

Business Organization and human–AI Collaboration

Through multidisciplinary collaboration with colleagues of *Universitas Mercatorum*, and drawing on my prior industry experience as an industrial management engineer, I developed a research stream on human–AI collaboration in small and medium-sized enterprises.

The impact BI and AI on traditional structures with legal and philosophical insights

Authors: **Vittorio Stile**, Vittorio Bonino, and Nunzia Cosmo

Abstract. *This research explores the transformative impacts of Business Intelligence (BI) and Artificial Intelligence (AI) technologies on organisational structures, with a special focus on legal and philosophical implications. As businesses increasingly integrate data analytics into their operations, the role of BI and AI extends beyond operational efficiency to shaping organisational culture and power dynamics. This study examines how these technologies influence decision-making processes, employee roles, and organisational ethics, under the framework of data protection regulations like the GDPR and considerations of algorithmic transparency. Through qualitative methods including focus groups and a comprehensive survey, the research captures diverse professional perspectives on the adoption and effects of BI and AI technologies. Results indicate a significant shift towards data-driven decision-making, increased transparency, and changes in power dynamics, with a notable impact on the efficiency and work roles of employees. The study also addresses the challenges of integrating these technologies into traditional organisational structures, highlighting issues such as resistance to change and the need for skill development. The discussion extends to the legal responsibilities concerning data privacy and the philosophical debates on autonomy and fairness, suggesting that the alignment of technology with ethical and legal standards is crucial. This research contributes to a better understanding of the multifaceted impacts of BI and AI technologies, recommending strategies for their ethical integration to foster more dynamic and compliant organisational environments.*

Comment. In this paper, written in collaboration with a doctoral colleague from the legal domain and a colleague from philosophy domain and presented to the 21st Conference of the Italian Chapter of AIS (itAIS), we conducted a mixed-methods study to examine how Business Intelligence (BI) and AI reshape organisational structures, combining a quantitative survey with qualitative focus groups to triangulate cultural, legal, and operational impacts.[3] Our survey shows broad uptake of data analytics, with 81.5% of organisations adopting these technologies and 59.1% using them for more than three years, while 74.1% reported a neutral-to-significant cultural impact, 51.9% perceived greater centralisation of decision-making, and 85.2% observed improved operational efficiency, almost half reported increased influence in decisions and 44.4% saw a change in their individual roles. Focus groups surfaced the same duality, transparency, faster reactions, and collaboration grew alongside tensions, skills gaps, and infrastructure challenges, with power shifting toward data-skilled teams. We interpret these findings within European Union (EU) governance frameworks such as the General Data Protection Regulation (GDPR) and the proposed European Union Artificial Intelligence Act (AI Act), and engage with Floridi's synthesis of beneficence, non-maleficence, autonomy, justice, and explicability to ground recommendations on accountability and responsible adoption.[4] While the work offers actionable guidance on governance and training, it also acknowledges limitations, including a sentiment-oriented questionnaire not based on validated scales and the inherent constraints of small focus groups, which future studies should address with larger, more representative samples and validated instruments.

[3] V. Stile, V. Bonino, and N. Cosmo, *The impact BI and AI on traditional structures with legal and philosophical insights*, ENG, in: *Proceedings of the 21st Conference of the Italian Chapter of AIS (itAIS 2024)*, vol. 21, AISeL - Springer LNISO, Università Cattolica del Sacro Cuore (UCSC), Piacenza, Italy Oct. 2024, ISBN: 979-12-82308-00-7, DOI: 10.979.1282308/007, URL: https://aisel.aisnet.org/itais2024/21.

[4] L. Floridi, *Etica dell'intelligenza artificiale: sviluppi, opportunità, sfide*, ita, ed. by M. Durante, Prima edizione, Scienza e idee 340, Raffaello Cortina Editore, Milano, Italy 2022, ISBN: 9788832854091.

Human-AI Collaboration in SMEs: A Role-Sensitive Framework for Cognitive Enterprise Hubs

Authors: Fabrizio Benelli, Franco Maciariello, Claudio Salvadori, Erdet Këlliçi, and **Vittorio Stile** (2025, October 18)

Abstract. *Traditional enterprise automation systems often lack the contextual intelligence and flexibility required in logistics-intensive environments, particularly for small and Medium-Sized Enterprise (SME). This paper proposes a five-phase implementation framework for Cognitive Enterprise Hub (CEH), emphasizing role-sensitive deployment and continuous alignment between human and AI. The model combines architectural planning, AI integration and cultural adaptation to support scalable and adaptive collaboration across federated ecosystems. Empirical validation is provided through two cross-sector case studies and a multi-role survey with 18 participants from IT, cybersecurity and telecommunications sectors. Findings reveal notable perceptual differences across organizational roles, especially between IT leaders, transformation strategists and frontline employees, regarding CEH impact on productivity, support and future opportunities. The study underscores that CEH success depends not only on technical orchestration but also on socio-cultural alignment. To address this, we offer practical deployment guidelines tailored for SME and logistics-driven operations. By integrating technical and organizational perspectives, this work advances the practical deployment of intelligent enterprise systems, positioning CEH as enablers of inclusive, cognitively enhanced coordination frameworks.*

Comment. In this paper, presented at the 22nd conference of the itAIS in Castellanza with colleagues of *Mercatorum University* from the 39th doctoral cycle, professor Këlliçi of *Tirana Business University* and expert from industry, we introduced a role-sensitive framework for CEH in SMEs, outlining a phased path from requirements analysis and semantic data integration to explainable AI services and pilot validation. A targeted survey highlighted perception gaps between executives, digital transformation leads, Information Technology (IT) managers, and knowledge workers, and showed how adoption depends on cultural readiness and participatory

onboarding.

AI-Enabled People & Culture: un framework socio-tecnico per la sostenibilità organizzativa

Authors: Fabrizio Benelli, Ida Giannetti, **Vittorio Stile**, and Franco Maciariello (accepted for XLI Convegno Nazionale AIDEA 2026, January 22-23)

Abstract. *L'obiettivo di questo paper è proporre un framework socio-tecnico di Responsible People Analytics che armonizzi Intelligenza Artificiale (IA) e competenze umane per la sostenibilità organizzativa in enti di piccole e medie dimensioni. È stato adottato un dataset (N = 287) integrato da quattro studi (Benelli et al., 2024; Benelli et al., 2025a; Benelli et al., 2025b; Maciariello et al., 2025); triangolazione SEM, Machine Learning (RF, XAI) ed Econometria Spaziale (SAR). L'adozione della IA incide positivamente sui People KPI solo in climi di fiducia e partecipazione (0.40; R²0.40); la governance etica (EPCIS 2.0, Hashing/Hyperledger) rafforza trasparenza e fiducia; spillover intra-team significativi (=0.31; p<0.01). Ci sono implicazioni manageriali che ricadono sulla progettazione di team e processi HR orientati alla fiducia, formazione per ruolo e data-governance votata alla semplificazione degli audit; usare XAI per una migliore accountability. I limiti di questo studio sono evidenze cross-sectional e prevalenza di PMI italiane; utili estensioni longitudinali e cross-settoriali. L'originalità del contributo sono l'integrazione multilivello SEM-ML-SAR che collega AI, KPI e aspetti culturali sotto un'unica logica di governance dei dati.*

Comment. This paper, submitted to the XLI Convegno Nazionale Accademia Italiana di Economia Aziendale (AIDEA) in Milano, proposes a socio-technical framework for *Responsible People Analytics* in SMEs, integrating Structural Equation Modeling (SEM), Machine Learning (Random Forest with XAI), and Spatial Econometrics (SAR) on a consolidated dataset ($N = 287$) from four studies. Results show that AI adoption improves People KPIs only in climates of trust and participation ($\beta \approx 0.40$, $R^2 \approx 0.40$); ethical data-governance mechanisms based on EPCIS 2.0 with selective hashing on Hyperledger strengthen transparency and trust; and intra-team

spillovers are significant ($\rho = 0.31,\ p < 0.01$). Managerial implications include trust-oriented HR processes, role-sensitive team design, targeted training, streamlined audit readiness, and the use of XAI to enhance accountability. Limitations concern cross-sectional evidence and a sample skewed toward Italian SMEs; longitudinal and cross-sector validations are suggested. The main contribution is a multilayer SEM–ML–SAR integration that links AI adoption, KPIs, and cultural factors under a unified data-governance logic.

Education and Learning Analytics

As a tenured high school teacher in Computer Science and Technologies, I have examined how AI reshapes higher education. Two lines stand out.

Enhance Student Well-being and Digital Literacy with Machine Learning and Spatial Analysis

Authors: Fabrizio Benelli, Erdet Këlliçi, Franco Maciariello, Claudio Salvadori, and **Vittorio Stile** (2025, October 26)

Abstract. *Advanced data analytics and machine learning can reshape the way schools cultivate both digital skills and student well-being. We analysed data from three Italian high-school classes ($N = 64$), combining random-forest and neural-network predictors with Spatial Autoregressive and Geographically Weighted Regression models. The approach captures how individual attributes, classroom geography and peer interactions jointly influence learning. Average grades rose from 5.34 to 6.15 and well-being scores from 0.48 to 0.95 over one semester. Spatial estimates ($\rho = 0.31,\ p < 0.01$) show that sitting next to high achievers yields a mean gain of 0.38 grade points, while local pockets of well-being amplify the effect of digital-literacy growth on performance. The results demonstrate that digital-literacy interventions, when delivered in spatially aware learning environments, produce measurable academic and affective benefits. The study offers a reproducible pipeline that blends machine-learning prediction with spatial econometrics and provides*

evidence to guide data-driven, equitable strategies for classroom design, teacher training and student support.

Comment. In this paper, presented at the 2nd Workshop on Education for Artificial Intelligence (edu4AI) at the 28th European Conference on Artificial Intelligence (ECAI) in Bologna with colleagues of *Mercatorum University* from the 39th doctoral cycle, professor Këlliçi of *Tirana Business University* and expert from industry, we combined Machine Learning (ML) with spatial econometrics to study how classroom spatial distribution relates to performance and well-being.[5] Results suggest positive peer effects by proximity to high achievers, and provide actionable implications for classroom layouts, teacher training, and equitable support.

Rethinking Higher Computer Science Education in the Age of AI: Insights from Computer Science Students in North Africa

Authors: Sonia Sahli, **Vittorio Stile**, and Danis Gillet (accepted for EDUCON 2026, April 27–30)

Abstract. *This study aims to examine the future of computer science programs in higher education in North Africa in the age of AI. The main objective is to identify students' expectations of the Bachelor program, specializing in computer systems development, regarding curriculum reform, with a focus on the professions of tomorrow in the age of AI. A survey collected 170 responses from final-year bachelor students asking them to suggest changes they would make to the program if they were responsible for educational policy in 2026. The results were divided into three key categories: a shift to more interactive teaching techniques, the addition of new subjects centered on AI, and the development of new educational modalities and more flexibility. These findings confirm the urgent need for program reforms aligned with the current and fast AI revolution, especially in computer science.*

[5] F. Benelli, E. Këlliçi, F. Maciariello, C. Salvadori, and V. Stile, *Enhance Student Well-being and Digital Literacy with Machine Learning and Spatial Analysis*, ENG, in: *Proceedings of the 2nd International Workshop on Education for Artificial Intelligence (EDU4AI 2025)*, vol. 4114, AI*IA SERIES, urn:nbn:de:0074-4114-x, CEUR Workshop Proceedings, The Engineering School of University of Bologna, Bologna, Italy Oct. 2025, Session S2: AI Literacy and Education, ISBN: ISSN 1613-0073, URL: https://ceur-ws.org/Vol-4114/.

Comment. In this paper, submitted to the 17th IEEE Global Engineering Education Conference (EDUCON) in Cairo in collaboration with researchers and professors from Tunisia and Switzerland, we analyzed expectations for future curricula, teaching modalities, and integrity safeguards. The study recommends a rebalancing of theory and practice, new AI-centered modules, and structured industry links with recognized certifications.

Physical Internet (PI)

The Physical Internet (PI) paradigm addresses fragmentation and inefficiency in logistics by adopting modularity, standardisation, and dynamic routing principles inspired by the Internet, with interoperable containers circulating across nodes and hubs under open protocols. In this work we take an extended, cross–domain view of PI, where intelligent orchestration spans logistics, energy infrastructures, and cyber–physical assets, enabled by AI, IoT, and security analytics. In the energy domain, we outline an *Energy Physical Internet* in which electricity is abstracted as "energy packets" that can be balanced and exchanged over interoperable microgrids, tracked via distributed ledgers and regulated through smart contracts, with expected gains in transaction costs, control latency, and resilience, consistent with European green–transition priorities. To coordinate heterogeneous domains, we propose a *secure cognitive orchestration* framework that integrates load forecasting, reinforcement learning (RL), and multi–agent mechanisms for decentralised optimisation, coupled with SIEM capabilities and human–AI interfaces to ensure transparency, dependability, and sustainable cognitive load for operators. This perspective positions PI as an enabling infrastructure that aligns operational efficiency, cybersecurity, and human oversight, and it motivates the following sections, which review the state of the art and detail our architectural proposals and research hypotheses.

Artificial Intelligence for Decentralized Orchestration in the Physical Internet: Opportunities, Business Trade-offs, and Risks in Road Freight Logistics

Authors: Fabrizio Benelli, Franco Maciariello, Erdet Këlliçi, and **Vittorio Stile** (2025, October 17)

Abstract. *The Physical Internet (PI) has been conceptualized as a disruptive response to inefficiencies and fragmentation in logistics. Through modularization, interoperability, and open hubs, PI envisions a hyperconnected and sustainable road freight ecosystem. Yet its realization critically depends on Artificial Intelligence (AI) as the principal enabler of decentralized orchestration. This paper develops a theoretical framework for AI-driven PI logistics. Our novelty lies in a unified orchestration lens that jointly tackles fairness, data-sharing governance, and cybersecurity in multi-agent PI systems at scale. We argue that orchestration will require multi-agent AI systems capable of decentralized load consolidation, routing optimization, and hub allocation. These agents interact with IoT-equipped vehicles and containers, generating high-frequency data streams for predictive modeling and adaptive decision making. Complementarily, blockchain infrastructures secure transactions, while smart contracts provide automated enforcement of collaborative agreements. We highlight implications for business performance: trade-offs between cost-to-serve, service-level adherence, and collaboration Return on Investment (ROI) must be considered alongside sustainability gains. Evidence from existing multi-agent last-mile simulations and large-scale PI pilot projects in France shows quantifiable improvements (e.g., reduced vehicle kilometers traveled, higher load factors), offering empirical support. Scenario Key Performance Indicators (KPIs) such as ΔV_{KT}, $\Delta load$ factor, and fairness indices are proposed to assess impact. Nevertheless, embedding AI in PI road freight introduces challenges: (i) ensuring algorithmic fairness to protect Small and Medium Enterprises (SMEs), (ii) implementing robust data governance and interoperability frameworks, and (iii) guaranteeing cybersecurity and resilience against adversarial threats. To address these, we propose research propositions on scalable orchestration, governance for trustworthy data and algorithms, and risk*

mitigation strategies. A conceptual framework and architecture diagram will be provided to integrate technical, organizational, and governance perspectives. By outlining opportunities, business trade-offs, and systemic risks, this paper establishes a conceptual agenda for research and policy, aligned with the European Union (EU) AI Act, digitalization, and green transition priorities.

Comment. In this work, written with colleagues of *Mercatorum University* from the 39th doctoral cycle, professor Këlliçi of *Tirana Business University* and presented to the CIDE 2025 in Ploiești, we propose an AI-driven, decentralized orchestration approach for PI nodes that supports dynamic task allocation, local decision making, and inter-node coordination under partial information.[6] We formulate orchestration as a multi-agent control problem, outline a cognitive loop for sensing–reasoning–acting across logistics assets, and discuss mechanisms for scalability (federated learning), resilience (fault-tolerant consensus), and interoperability (standardized data/intent interfaces).

Towards an Energy Physical Internet

Authors: Fabrizio Benelli, Franco Maciariello, Redvin Marku, and **Vittorio Stile** (2025, October 17)

Abstract. *The Physical Internet (PI) has emerged as a systemic paradigm for overcoming inefficiencies and fragmentation in logistics by introducing modular, interoperable, and hyperconnected infrastructures. Building on foundational PI work and European roadmaps (Montreuil; Ballot; ALICE-ETP), this paper extends the PI concept to the energy sector, advancing the notion of an "Energy Physical Internet" (EPI). Here, electricity is conceptualized as modular "energy packets" dynamically routed across interoperable microgrids and distribution hubs, enabling innovative platform-based business models for decentralized energy trading. The proposed*

[6] F. Benelli, E. Këlliçi, F. Maciariello, and V. Stile, *Artificial Intelligence for Decentralized Orchestration in the Physical Internet: Opportunities, Business Trade-offs, and Risks in Road Freight Logistics*, ENG, in: *Book of Abstract of the 4th International Conference Creativity And Innovation In Digital Economy*, vol. Section 2: Co-creation, living labs and innovation ecosystems, Section 2: Co-creation, living labs and innovation ecosystems, Petroleum-Gas University of Ploiești Publishing House, Petroleum-Gas University of Ploiești (UPG), Ploiești, Romania Oct. 2025, ISBN: ISSN 2971–9798.

EPI architecture relies on three mutually reinforcing pillars. First, IoT-enabled infrastructures provide real-time observability and control of distributed assets. Second, blockchain-based registries ensure tamper-proof traceability of energy provenance, carbon intensity, and transactions, fostering trust and regulatory compliance. Third, cryptocurrency-enabled conditional payments, implemented via Ethereum smart contracts, automate peer-to-peer settlements conditional on real-time balance and renewable performance. Pilot projects such as the Brooklyn Microgrid already illustrate the feasibility of blockchain-governed energy platforms. The model departs from hierarchical utility-driven systems by empowering prosumers, energy communities, and peer-to-peer marketplaces to act as autonomous yet interoperable nodes. Expected benefits can be assessed through KPIs such as transaction-cost reduction, settlement latency, and resilience indices. Moreover, governance touchpoints, including DSO/TSO coordination and rules for energy data spaces, are recognized as essential for scalability. Our contribution is theoretical and hypothesis-driven. We propose research propositions on (i) interoperability standards for modular energy exchange, (ii) AI-assisted routing for balancing energy packets, and (iii) token-based incentives for prosumer engagement. These propositions directly address known challenges in IoT security, blockchain scalability, and governance. We also announce the development of a conceptual framework and architecture diagram to guide empirical research. By positioning energy as a new frontier for the PI, this paper defines an agenda for interdisciplinary research at the intersection of logistics, energy economics, and digital platforms, aligned with EU Green Deal and Horizon Europe objectives.

Comment. In this work, written with colleagues of *Mercatorum University* from the 39th doctoral cycle, professor Marku of *Tirana Business University* and presented to the CIDE 2025 in Ploiești, we extend PI principles of our research on the PI to the energy domain, arguing for packetized, routable energy flows coordinated by digital twins and market-aware schedulers.[7] We present an architecture that couples

[7]F. Benelli, F. Maciariello, R. Marku, and V. Stile, *Towards an Energy Physical Internet: Open Business Models and Platforms for Electricity Distribution Enabled by IoT, Blockchain, and Conditional Payments*, ENG, in: *Book of Abstract of the 4th International Conference Creativity And*

distributed energy resources, storage, and flexible loads with predictive optimization to balance constraints on capacity, latency, losses, and carbon intensity, and we delineate research needs on pricing, congestion control, and cross-sector coupling with transport.

Secure Cognitive Orchestration Framework for Multi-Domain Physical Internet: Integrating AI-Driven Logistics, Energy Distribution, and Cybersecurity

Authors: Fabrizio Benelli, Franco Maciariello, and **Vittorio Stile** (accepted for IHSI 2026, February 11-13)

Abstract. *The convergence of distributed logistics networks, decentralized energy systems, and interconnected cyber-physical infrastructures demands novel orchestration paradigms capable of addressing operational complexity, cybersecurity vulnerabilities, and human-AI alignment challenges. This research presents the Secure Cognitive Orchestration Framework (SCOF), a multi-layer architecture integrating Physical Internet (PI) principles with artificial intelligence-based decision-making, Security Information and Event Management (SIEM) analytics, and adaptive human-machine teaming for cross-domain infrastructure coordination. Using a hybrid agent-based simulation methodology, we model 20 logistics hubs processing 1,800 daily shipments, 85 energy feeders serving 350 distribution transformers, and 230,000 cybersecurity event logs representing multi-stage attack scenarios. The AI orchestration layer employs reinforcement learning for decentralized routing optimization, LSTM networks for load forecasting, and multi-agent consensus mechanisms for coordinated decision-making. Empirical evaluation demonstrates that SCOF achieves 12-18% improvement in logistics load factor, 8-14% reduction in vehicle kilometers traveled, 11-16% enhancement in grid balancing performance, 35-55% faster anomaly detection, and 18% reduction in operator cognitive workload compared to conventional approaches. These findings*

Innovation In Digital Economy, vol. Section 4: New Pathways in Knowledge, Education and Law, Section 4: New Pathways in Knowledge, Education and Law, Petroleum-Gas University of Ploiești Publishing House, Petroleum-Gas University of Ploiești (UPG), Ploiești, Romania Oct. 2025.

validate SCOF as a replicable architectural framework enabling secure, intelligent, and human-centered Physical Internet operations applicable to smart logistics, power distribution networks, and cyber-physical critical infrastructures.

Comment. In this paper, submitted to the 9th International Conference on Human Intelligent Systems Integration (IHSI) in Florence, we design a secure, cognitive orchestration framework that unifies logistics and energy PI layers under a zero-trust cybersecurity posture. We specify a multilayer control plane combining intent-based policies, digital twins, and learning-enabled planners, while embedding identity, attestation, and anomaly detection along the data and actuation paths. We show how the framework coordinates cross-domain workflows (e.g., routing freight and scheduling charging) with verifiable compliance, resilience to adversarial behavior, and end-to-end observability.

Other Multidisciplinary Works

Distributed Artificial Intelligence and Health Governance: A Multidimensional Analysis of the Tensions Between Rules, Ethics and Innovation

Authors: Fabio Liberti, Francesco Avolio, Vito Saverio Cicoira, Nunzia Cosmo, Alfonso Laudonia, Franco Maciariello, and **Vittorio Stile** (2025, October 17)

Abstract. *Distributed Artificial Intelligence (DAI) is transforming healthcare systems by enabling collaborative, privacy-preserving data analysis across institutions. Technologies such as Federated Learning and Edge Computing allow the development of high-performing AI models without centralized data storage, addressing regulatory constraints while opening new pathways for innovation. However, the adoption of distributed AI introduces multidimensional tensions involving technological, economic, legal, and ethical domains. This paper proposes an integrated analytical framework to assess and navigate these tensions in the context of healthcare governance. The framework encompasses four key axes: (1) Technology,*

focusing on interoperability, resilience, and performance; (2) Economy, addressing cost-efficiency, value distribution, and sustainability; (3) Law, analyzing compliance with GDPR, AI Act, and sector-specific regulations; and (4) Ethics, highlighting fairness, transparency, and patient autonomy. Each axis is detailed through specific constructs and evaluation metrics. We apply this model to real-world healthcare scenarios, identifying critical trade-offs, such as those between security and cost, innovation and regulation, or algorithmic performance and interpretability. The resulting tension matrix offers a tool to visualize interdependencies and prioritize governance actions. By integrating interdisciplinary expertise, the proposed framework supports adaptive governance strategies tailored to the dynamic nature of AI systems and healthcare environments. Our approach facilitates responsible innovation by aligning technological capabilities with legal requirements, ethical principles, and economic viability. The paper concludes by highlighting future research directions and policy implications for sustainable AI adoption in digital health.

Comment. The paper, presented at the 22nd conference of the itAIS in Castellanza, Italy, proposes a governance framework for distributed AI in healthcare that integrates four axes, technology, economy, law, and ethics, with concrete constructs, metrics, and a scoring rubric, and complements this with a "tension matrix" that makes trade-offs explicit across dimensions. The contribution moves beyond principle-level guidance by targeting federated learning and edge settings, comparing against major frameworks, and illustrating use through three cases, radiology Federated Learning (FL), edge monitoring for chronic care, and collaborative, multi-party drug discovery. Strengths include operational measurability, attention to interoperability, privacy and liability, and a realistic treatment of organizational resistance and cross-border compliance. The limitations are acknowledged, secondary-source assessment, European focus, temporal drift as rules and technology evolve, and the need for primary validation and longitudinal evidence. Future work should develop jurisdictional adaptation matrices, automate the metric pipeline for continuous auditing, and deepen human-in-the-loop practices that bind explainability and clinical accountability to everyday workflow.

Competences for Society 5.0: Multidisciplinary Corporate Training for Inclusion, Safety and Competitiveness

Authors: Franco Maciariello, Francesco Avolio, Vito Saverio Cicoira, Nunzia Cosmo, Alfonso Laudonia, Ida Giannetti, Paola Liberanome, and **Vittorio Stile** (2025, October 29)

Abstract. *This desk-based review investigates how corporate training can support the transition from the efficiency-oriented paradigm of Industry 4.0 to the human-centric vision of Society 5.0. Drawing on philosophical, legal, economic, engineering and occupational-health literatures, as well as European and Italian policy sources, the study maps converging requirements for inclusive, technology-enabled learning. Analysis shows that (i) ethical frameworks position innovation as a means to social well-being; (ii) recent equality statutes and disability directives provide enforceable rights that training must operationalise; (iii) digital delivery of continuous learning yields cost savings of* $\approx 40\,\%$ *and productivity gains up to* $20\,\%$ *when coupled with clear feedback loops; (iv) AI-driven platforms and augmented-reality modules make mass personalisation feasible but introduce governance challenges; and (v) smart monitoring reduces classical hazards yet raises new psychosocial and privacy risks, especially for ageing workforces. The paper argues that effective programmes treat training budgets as strategic investments, co-design curricula with stakeholders, integrate universal-design and data-protection principles, and align organisational targets with EU human-centric policy goals. These findings provide a multidisciplinary blueprint for firms and policymakers seeking to reconcile digital ambition with social responsibility in the era of Society 5.0.*

Comment. This desk-based review, presented at the 35th International Conference Rethinking Services for Society 5.0 (RESER), examines how corporate training can support the shift from the efficiency-centred logic of Industry 4.0 to the human-centric vision of Society 5.0. By synthesising philosophical, legal, economic, engineering and occupational-health literatures, together with European and Italian policy sources, the study maps converging requirements for inclusive, technology-enabled learning. Results indicate that ethical frameworks cast innovation as a means to social

well-being; recent equality and disability regulations create enforceable rights that training must operationalise; digital delivery of continuous learning can cut logistics costs by ≈ 40% and raise productivity up to 20% when coupled with clear feedback loops; AI platforms and augmented-reality modules enable mass personalisation yet introduce governance and accountability challenges; and smart monitoring reduces classical hazards while raising psychosocial and privacy risks, especially for ageing workforces. The paper argues that effective programmes treat training budgets as strategic investments, co-design curricula with stakeholders, integrate universal-design and data-protection principles, and align organisational targets with EU human-centric goals. Overall, the contribution offers a multidisciplinary blueprint for firms and policymakers who seek to reconcile digital ambition with social responsibility in the era of Society 5.0.

AI-Driven Financial Risk Prevention: the Role of HR Analytics in Corporate Crisis Management Under Industry 5.0

Authors: Alfonso Laudonia, Francesco Avolio, Nunzia Cosmo, Ida Giannetti, Paola Liberanome, Franco Maciariello, and **Vittorio Stile** (2025, November 13)

Abstract. *This paper addresses a persistent research gap: existing financial-distress models rely almost exclusively on ex post accounting ratios, overlooking human-capital variables that often signal crises much earlier. We propose an AI-driven framework that integrates HR analytics with financial indicators under the Industry 5.0 paradigm. The pipeline combines machine learning, Natural Language Processing (NLP), federated learning and blockchain-based auditability, producing a legally compliant and explainable early-warning system. Using a harmonised dataset of 50 European SMEs (2019–2024), the model achieves an Area Under the Receiver-Operating-Characteristic curve (AUROC) of 0.92 and reduces false positives by nearly 50% compared to Altman-based triggers. A comparative analysis with traditional credit-scoring and Capital Adequacy, Asset quality, Management, Earnings, Liquidity (CAMEL) ratios confirms a 25–30% improvement in forecasting crises 6–12 months in advance. Governance mechanisms—AI committees, audit*

logs and human-in-the-loop oversight—embed EU regulatory requirements–AI Act, Data Act, Data Governance Act (DGA)–and ensure transparency, accountability and worker protection. The study provides actionable insights for managers, highlighting how turnover and skill gaps anticipate liquidity shocks, thereby linking workforce planning to financial resilience in Industry 5.0 organisations.

Comment. This multidisciplinary paper, co-authored with colleagues affiliated with three different PhD programmes at *Mercatorum University*, was presented at International Conference on Industry of the Future and Smart Manufacturing (ISM) 2025 in Valletta, Malta, addresses a persistent gap in SME crisis prediction by integrating Human Resources (HR) analytics with financial ratios into an explainable and privacy-preserving early-warning pipeline. The contribution is socio-technical: a federated XGBoost workflow with a permissioned audit trail, SHapley Additive exPlanations (SHAP)-based explanations for board-level decisions, bias monitoring with clear human oversight, and explicit alignment to GDPR, the EU AI Act and Italian Code of Business Crisis and Insolvency (CCII) duties. Methodologically, the study assembles a multi-company panel, engineers HR features such as turnover, skill gap, training Return on Investment (ROI) and sentiment, and operationalises thresholded alerts within governance dashboards. Empirical evidence indicates earlier and more stable warnings than classical ratio triggers, fewer false positives, and interpretable drivers of risk that managers can act on. Strengths include the tight coupling of modelling and governance, practical explainability, and a data-minimising federated strategy suited to cross-firm collaboration. Limitations concern the European SME focus, a restricted set of fairness attributes, and the need for stronger drift and privacy stress-tests in live settings. Future work should extend the panel to other sectors and jurisdictions, incorporate differential privacy, broaden bias auditing to intersectional groups, and quantify cost-benefit under alternative macroeconomic conditions.

Explainable Federated Learning for Secure Telemedicine: Protecting Patient Identity through Privacy-Preserving Deepfake Detection in Digital Health Platforms

Authors: Raimondo Fanale, Fabio Liberti, and **Vittorio Stile** (2025, October 17)

Abstract. *Secure telemedicine requires privacy-preserving deepfake detection in digital health systems to protect patient identification. Post-pandemic growth in the European telemedicine business from €45B to €380B has created a key vulnerability: medical deepfakes. Multiple European sources warn that deepfakes could destabilize digital healthcare infrastructure.*[8] *This study proposes Explainable Federated Learning (XFL) as the identity verification standard for telemedicine systems. Without such a norm, unavoidable breaches might severely erode public trust. Deepfake technology makes basic webcam checks and static ID images obsolete. Thus, federated, privacy-preserving detection systems need immediate regulation. These systems enable GDPR compliance and avoid medical identity fraud. Our investigation has three goals. We first highlight the fundamental healthcare privacy breaches in centralized deepfake detection methods. We believe federated options uniquely promote data sovereignty. Second, we will propose real regulatory frameworks with timetables, compliance standards, and penalties to mandate XFL adoption. Thirdly, we will analyze the economic effects of proactive vs reactive solutions and show that compulsory implementation is usually cheaper than medical identity fraud losses. We focus on integrating technical architecture, regulatory effect evaluation, and economic modeling. The technical design uses a three-level hierarchical federated topology. Tier-1 includes large hospital networks (over 500 beds) that provide computational power and diverse patient demographics. Domain-specific attack patterns are available on Tier-2 telemedicine systems. Tier-3 edge devices enable real-time verification. Design architecture enforces differential*

[8] European Data Protection Supervisor, *TechSonar Report 2023–2024: Emerging Technologies (including Deepfake Detection)*, https://www.edps.europa.eu/system/files/2023-12/23-12-04_techsonar_23-24_en.pdf, Report published 4 Dec 2023, Dec. 2023, (visited on 11/07/2025); European Union, *Regulation (EU) 2024/1689 (Artificial Intelligence Act), including Annex III: High-Risk AI Systems Categories*, https://eur-lex.europa.eu/eli/reg/2024/1689/oj, Official Journal of the European Union, 2024, (visited on 11/07/2025).

privacy ($\epsilon = 1.5$) to keep biometric data inside institutional boundaries while ensuring successful detection. We map GDPR Articles 25 (Privacy-by-Design) and 32 (Security-of-Processing) to technological standards to show how federated systems alone meet privacy and security requirements.[9] *Clinicians require interpretable explanations for decisions when consultations are possibly fraudulent to avoid medical-legal liability. Deepfake detectors are Class-IIa software under the Medical Device Regulation (MDR).*[10] *GDPR protects privacy, and the planned AI Act addresses high-risk applications.*[11] *Economic modeling includes market data from 15 European telehealth platforms, although estimates are theoretical until verified. Interestingly, 94% of European telemedicine systems lack deepfake detection. The other 6%? They rely on centralized systems, which may violate GDPR data minimization restrictions. Federated techniques detect deepfakes with 97.8% accuracy. While that's lower than 99.2% for centralized settings, federated solutions provide total anonymity. Now, this 1.4% accuracy difference? It pales in comparison to the mess that may follow a major biometric database leak, which could harm millions of people and raise ethical concerns, especially with AI's role in Africa.*[12] *Starting up these systems? Budget €2–3 million per key platform. That's only 0.5% of their annual revenue—peanuts compared to what they could lose if security fails. One public medical deepfake might cost over €500 million in direct losses. The damage to their reputation could undermine years of progress in promoting digital health. Three phases make up our quick and realistic plan. First, from 2025 to 2026, we're considering voluntary adoption with tax advantages and less liability for mistakes.*

[9] European Union, *Regulation (EU) 2016/679 (General Data Protection Regulation): Articles 25 (Data Protection by Design and by Default) and 32 (Security of Processing)*, https://eur-lex.europa.eu/eli/reg/2016/679/oj, Official Journal of the European Union, 2016, (visited on 11/07/2025); European Data Protection Board, *Guidelines 4/2019 on Article 25 Data Protection by Design and by Default*, https://www.edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-42019-article-25-data-protection-design-and_en, 2020, (visited on 11/07/2025).

[10] Medical Device Coordination Group, *MDCG 2019-11 Rev.1: Guidance on Qualification and Classification of Software under Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR*, https://health.ec.europa.eu/document/download/b45335c5-1679-4c71-a91c-fc7a4d37f12b_en, Revision 1 (June 2025), June 2025, (visited on 11/07/2025).

[11] European Union, *Regulation (EU) 2024/1689 (Artificial Intelligence Act), including Annex III: High-Risk AI Systems Categories*, op. cit.

[12] A. A. Tapo, A. Traore, S. Danioko, and H. Tembine, *Machine Intelligence in Africa: a survey*, in: *arXiv* (2024), Accepted for DSAI 2024, DOI: 10.48550/arXiv.2402.02218, arXiv: 2402.02218 [cs.CY], URL: https://arxiv.org/abs/2402.02218.

Phase two in 2027? New platforms should require it, while older ones may not. Finally, by 2028 and beyond, everyone requires it, and those who don't will face harsher penalties. These systems provide patients with plain-language and picture explanations, doctors with dashboards showing technical confidence levels and weird patterns, and regulators with solid, tamper-proof records to check compliance. Deepfake detection without explainability could make healthcare access harder for the elderly and tech-illiterate. Can federated learning become a requirement for ethical digital health?[13] *Infrastructure investments don't become obsolete, which is beneficial. Simply adding federated verification levels improves security and privacy. Meanwhile, policymakers require help crafting policies that promote innovation and protect rights.*[14] *Researchers will naturally ask: what's the best balance between privacy and utility in medical situations? How can we develop cross-border federation protocols that protect data sovereignty? How can we assure it works for various groups in real life? Our system has limitations, such as untested economic estimates and technical challenges for smaller clinics.*[15] *Our advice may also be outdated due to changing rules. However, waiting for something horrible to happen before acting is unacceptable when we can prevent it.*[16] *It's not just about technology—Europe must choose between leading the way in secure and private telemedicine or losing years of digital health advances due to security issues. Federated learning protects privacy and security. Technology, legislation, and economic benefits are clear. Not enough people want it before the market fails and people suffer. We'll see if we acted early or waited until prevention became fixing the harm.*

Comment. This extended abstract, written with colleagues of the PhD course in *Big Data and Artificial Intelligence* in *Mercatorum University*, and presented to the CIDE 2025 in Ploiești, Romania, argues that XFL should become the default

[13] World Health Organization, *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance*, World Health Organization, Genève, Switzerland 2021, ISBN: 978-92-4-002920-0, URL: https://www.who.int/publications/i/item/9789240029200 (visited on 11/07/2025).

[14] World Health Organization, *Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multimodal Models (LMMs)*, Genève, Switzerland, 2024, URL: https://www.who.int/publications/i/item/9789240084759 (visited on 11/07/2025).

[15] Ibid.

[16] Idem, *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance*, op. cit.

identity-verification layer for European telemedicine. The case is built on a clear risk framing (rapid post-pandemic growth and medical deepfakes), regulatory alignment (GDPR Arts. 25 and 32, MDR software Class IIa, AI Act high-risk), and an implementable three-tier topology spanning hospital hubs, telemedicine back-ends, and edge devices, with differential privacy $\epsilon = 1.5$. The authors claim that XFL preserves data sovereignty, attains competitive accuracy (97.8% versus 99.2% for centralised baselines), and requires investments of about 2 to 3 million EUR per major platform, which is small compared with breach losses. A staged roadmap is proposed: incentives in 2025 to 2026, requirements for new platforms in 2027, and universal mandate from 2028. Strengths include the coupling of technical design with governance artefacts and clinician- and patient-oriented explainability. Limitations concern preliminary economics, uneven readiness of smaller clinics, and adoption estimates. Overall, the work offers a coherent blueprint that integrates privacy by design, explainability, and federation to mitigate identity-fraud risks in digital health, while motivating further research on privacy and utility trade-offs, cross-border federation, and fairness in deployment.

Overall Output and Recognition

Beyond the works highlighted above, my research activity includes numerous co-authored outputs with colleagues, for a total of 14 submitted papers, 1 poster presentation, 3 conference proceedings, 4 books of abstracts, and 2 patents, along with multiple recognitions and invitations as an invited speaker.

These outcomes stem from a deliberate effort to strengthen transversal competencies, driven by a steady curiosity and a commitment to continuous learning and research. In the current landscape, where AI and my curriculum focus on *Big Data Management for the Digital Transition*, a multidisciplinary approach becomes almost natural, since these themes now cut across every sector. Working in such a stimulating, collaborative environment has been a source of real growth for me, allowing core elements of my professional practice to remain active, relevant, and constantly evolving.

Appendix 2: Scientific outputs arising from the PhD

Published in international peer-reviewed journals

1. Laudonia, A., Avolio, F., Cosmo, N., Giannetti, I., Liberanome, P., Maciariello, F., **Stile, V.** (2026). *AI-Driven Financial Risk Prevention: The Role of HR Analytics in Corporate Crisis Management Under Industry 5.0.*. In *Procedia Computer Science, Elsevier*. ISSN: 1877-0509

Under review in international peer-reviewed journals

1. **Stile, V.**, Caldelli, R., Guerrero-Contreras, G., Balderas-Díaz, S., and Medina-Bulo, I. *Facial Attribute-Aware DeepFake Detection through Semi-Supervised Facial Attribute Labeling*. In *Frontiers in Imaging, Sec. Imaging Applications*. ISSN: 2813-3315

Conference proceedings

1. **Stile, V.**, Caldelli, R., Guerrero-Contreras, G., Balderas-Díaz, S., and Medina-Bulo, I. *Analysis of DeepFake Detection through Semi-Supervised Facial Attribute Labeling*. In *Applied Computer Science: Proceedings of the 11th Spanish German Symposium (SGSOACS 2025), Communications*

in Computer and Information Science (CCIS), vol. 2831 (Springer Cham), Vienna, Austria Jun. 30 – Jul. 3 2025. ISBN: 978-3-032-14815-5

2. **Stile, V.**, Bonino, V., and Cosmo, N. *The impact BI and AI on traditional structures with legal and philosophical insights*. In *Proceedings of the 21st Conference of the Italian Chapter of AIS (itAIS 2024) vol. 21 (AISeL - Springer LNISO*, Piacenza, Italy Oct. 11-12 2024. ISBN: 979-12-82308-00-7

3. Maciariello, F., Avolio, F., Cicoira, V., Cosmo, N., Laudonia, A., Giannetti, I., Liberanome, P., and **Stile, V.** *Competences for Society 5.0: Multidisciplinary Corporate Training for Inclusion, Safety and Competitiveness*. In the *Rethinking services for society 5.0: Opportunities and Challenges, Conference Proceedings, 2025 RESER Annual Conference*, Aracne, Rome, Italy Oct. 27 2025. ISBN: 979-12-218-2272-4

4. Liberti, F., Avolio, F., Cicoira, V., Cosmo, N., Laudonia, A., Maciariello, F., and **Stile, V.** *Distributed Artificial Intelligence and Health Governance: A Multidimensional Analysis of the Tensions Between Rules, Ethics and Innovation*. In *Proceedings of the 22nd Conference of the Italian Chapter of the Association for Information Systems (ITAIS 2025)*, Libera Università Carlo Cattaneo, Castellanza (VA), Italy Oct. 17 2025. https://aisel.aisnet.org/itais2025/23/

5. Benelli, F., Maciariello, F., Salvadori, C., Këlliçi, E., and **Stile, V.** *Human-AI Collaboration in SMEs: A Role-Sensitive Framework for Cognitive Enterprise Hubs*. In *Proceedings of the 22nd Conference of the Italian Chapter of the Association for Information Systems (ITAIS 2025)*, Libera Università Carlo Cattaneo, Castellanza (VA), Italy Oct. 18 2025. https://aisel.aisnet.org/itais2025/24/

Poster presentation

1. **Stile, V.** *Towards Bias-Aware and Interpretable DeepFake Detection through Semi-Supervised Facial Attribute Labeling*, In *The 11th IEEE-EURASIP*

Summer School on Signal Processing (S3P-2025), Livorno, Italy Sep. 22 2025. DOI: 102025000023347

Conference book of abstract

1. **Stile, V.**, and Fontanella, A. *AI-Enhanced Building Information Modeling and Big Data Analytics for Civil Engineering Innovation*. In *Book of Abstract: Creativity and Innovation in Digital Economy*, Petroleum-Gas University of Ploiești PublishingHouse, Ploiești, Romania 2025. ISSN: 2971-9798

2. Benelli, F., Kelliçi, E., Maciariello, F., and **Stile, V.** *Artificial Intelligence for Decentralized Orchestration in the Physical Internet: Opportunities, Business Trade-offs, and Risks in Road Freight Logistics*. In *Book of Abstract: Creativity and Innovation in Digital Economy*, Petroleum-Gas University of Ploiești PublishingHouse, Ploiești, Romania 2025. ISSN: 2971-9798

3. Benelli, F., Marku, R., Maciariello, F., and **Stile, V.** *Towards an Energy Physical Internet: Open Business Models and Platforms for Electricity Distribution Enabled by IoT, Blockchain, and Conditional Payments*. In *Book of Abstract: Creativity and Innovation in Digital Economy*, Petroleum-Gas University of Ploiești PublishingHouse, Ploiești, Romania 2025. ISSN: 2971-9798

4. Fanale, R., Liberti, F., and **Stile, V.** *Explainable Federated Learning for Secure Telemedicine: Protecting Patient Identity through Privacy-Preserving Deepfake Detection in Digital Health Platforms*. In *Book of Abstract: Creativity and Innovation in Digital Economy*, Petroleum-Gas University of Ploiești PublishingHouse, Ploiești, Romania 2025. ISSN: 2971-9798

Presentations at referred conferences

1. **Stile, V.**, Caldelli, R., Guerrero-Contreras, G., Balderas-Díaz, S., and Medina-Bulo, I. *Analysis of DeepFake Detection through Semi-Supervised*

Facial Attribute Labeling. In the *11th Spanish German Symposium (SGSOACS 2025)*, Vienna, Austria Jul. 2 2025.

2. **Stile, V.**, Bonino, V., and Cosmo, N. *The impact BI and AI on traditional structures with legal and philosophical insights*. In the *21st Conference of the Italian Chapter of AIS (itAIS 2024)*, Catholic University of the Sacred Heart, Department of Economic and Social Sciences, Piacenza, Italy, Oct. 11 2024.

3. Liberti, F., Avolio, F., Cicoira, V., Cosmo, N., Laudonia, A., Maciariello, F., and **Stile, V.** *Distributed Artificial Intelligence and Health Governance: A Multidimensional Analysis of the Tensions Between Rules, Ethics and Innovation*. In *the 22nd Conference of the Italian Chapter of the Association for Information Systems (ITAIS 2025)*, Libera Università Carlo Cattaneo, Castellanza (VA), Italy Oct. 17 2025. https://aisnet.org

4. **Stile, V.**, and Fontanella, A. *AI-Enhanced Building Information Modeling and Big Data Analytics for Civil Engineering Innovation*. In the *4th International Conference on Creativity and Innovation in Digital Economy (CIDE 2025)*, Petroleum-Gas University, Ploiești, Romania Oct. 17 2025.

5. Benelli, F., Këlliçi, E., Maciariello, F., and **Stile, V.** *Artificial Intelligence for Decentralized Orchestration in the Physical Internet: Opportunities, Business Trade-offs, and Risks in Road Freight Logistics*. In the *4th International Conference on Creativity and Innovation in Digital Economy (CIDE 2025)*, Petroleum-Gas University, Ploiești, Romania Oct. 17 2025.

6. Benelli, F., Marku, R., Maciariello, F., and **Stile, V.** *Towards an Energy Physical Internet: Open Business Models and Platforms for Electricity Distribution Enabled by IoT, Blockchain, and Conditional Payments*. In the *4th International Conference on Creativity and Innovation in Digital Economy (CIDE 2025)*, Petroleum-Gas University, Ploiești, Romania Oct. 17 2025.

7. Fanale, R., Liberti, F., and **Stile, V.** *Explainable Federated Learning for Secure Telemedicine: Protecting Patient Identity through Privacy-Preserving*

Deepfake Detection in Digital Health Platforms. In *the 4th International Conference on Creativity and Innovation in Digital Economy (CIDE 2025)*, Petroleum-Gas University, Ploiești, Romania Oct. 17 2025.

8. Benelli, F., Maciariello, F., Salvadori, C., Këlliçi, E., and **Stile, V.** *Human-AI Collaboration in SMEs: A Role-Sensitive Framework for Cognitive Enterprise Hubs*. In *the 22nd Conference of the Italian Chapter of the Association for Information Systems (ITAIS 2025)*, Libera Università Carlo Cattaneo, Castellanza (VA), Italy Oct. 18 2025. https://aisnet.org

9. Benelli, F., Kelliçi, E., Maciariello, F., Salvadori, C., and **Stile, V.** *Enhance Student Well-being and Digital Literacy with Machine Learning and Spatial Analysis*. In the *2nd Workshop on Education for Artificial Intelligence (EDU4AI 2025)*, Bologna, Italy Oct. 26 2025.

10. Maciariello, F., Avolio, F., Cicoira, V., Cosmo, N., Laudonia, A., Giannetti, I., Liberanome, P., and **Stile, V.** *Competences for Society 5.0: Multidisciplinary Corporate Training for Inclusion, Safety and Competitiveness*. In the *35th RESER International Conference – Rethinking Services for Society 5.0: Opportunities and Challenges*, Università Tor Vergata, Rome, Italy Oct. 29 2025.

11. Laudonia, A., Avolio, F., Cosmo, N., Giannetti, I., Liberanome, P., Maciariello, F., and **Stile, V.** *AI-Driven Financial Risk Prevention: the Role of HR Analytics in Corporate Crisis Management Under Industry 5.0*. In the *7th International Conference on Industry of the Future and Smart Manufacturing (ISM 2025)*, Valletta, Malta Nov. 12 2025).

12. Benelli, F., Giannetti, I., **Stile, V.**, and Maciariello, F. *AI-Enabled People & Culture: un framework socio-tecnico per la sostenibilità organizzativa*. In the *41° Convegno Nazionale dell'Accademia Italiana di Economia Aziendale (AIDEA 2026): Le Intelligenze Aziendali per la competitività sostenibile e il bene comune*, Università Cattolica del Sacro Cuore, Milano, Italy Jan. 22-23 2026.

13. Benelli, F., Maciariello, F., and **Stile, V.** *Secure Cognitive Orchestration Framework for Multi-Domain Physical Internet: Integrating AI-Driven Logistics, Energy Distribution, and Cybersecurity*. In the *9th International Conference on Human Intelligent Systems Integration (IHSI 2026): Disruptive and Innovative Technologies*, Università di Firenze, Florence, Italy Feb. 11-13 2026.

14. Accepted for presentation: Sahli, S., **Stile, V.**, and Gillet, D. *Rethinking Higher Computer Science Education in the Age of AI: Insights from Computer Science Students in Tunisia*. In the *17th IEEE Global Engineering Education Conference (EDUCON 2026)*, Cairo, Egypt Apr. 27-30 2026.

Patents

1. **Stile, V.**, and Gasloli, A. *Sistema portatile per la digitalizzazione dei documenti*, 2025. `Patent application no. 102025000019540`

2. **Stile, V.**, Giordano, A., Chierchia, G. *Sistema portatile per la digitalizzazione dei documenti*, 2025. `Patent application no. 102025000023347`

Other

1. Invited speaker: **Stile, V.**, Bracale, M., Guida, F. *Bandi: istruzioni per l'uso* In *Investimenti e strategie brevetti, marchi, disegni e dintorni*, Sala Consiliare Comune di Gragnano, Gragnano (NA), Italy Jan. 16 2024.

2. Invited speaker: **Stile, V.**, and Fontanella, A. *Integrating AI and BIM: Innovations in Civil Engineering through Smart Design and Real-Time Analytics*. In *The 2nd World Conference on Construction and Building Technology (BuildTech Week 2025)*, Crowne Plaza Madrid Centre Retiro, Madrid, Spain May 12 2025. `DOI: 10.5281/ZENODO.17287941`

3. Presentation: **Stile, V.** *Recognition of Deepfakes Generated Through AI*. In the *Junior Faculty and Doctoral Consortium (JFDC) 2024*, Catholic University

of the Sacred Heart, Department of Economic and Social Sciences, Piacenza, Italy Oct. 10 2024. DOI: 10.5281/zenodo.17929426

4. Award: Amamou, M., Achouri, M. A., Sahli, S., and **Stile, V.** *Best Pitch Award with SmartExam an Edutech tool to help teachers save time on creating examinations* during *The IEEE Entrepreneurship Tunisia Workshop 2025*, Gammarth, Tunis, Tunisia Aug. 24-27 2025. https://entrepreneurship.ieee.org/2025_tunisia/

References

Afchar, D., Nozick, V., Yamagishi, J., and Echizen, I., *MesoNet: a Compact Facial Video Forgery Detection Network*, in: *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, IEEE, Hong Kong Dec. 2018, pp. 1–7, ISBN: 9781538665367, DOI: 10.1109/WIFS.2018.8630761, URL: https://ieeexplore.ieee.org/document/8630761/ (visited on 10/27/2025).

Amerini, I., Galteri, L., Caldelli, R., and Del Bimbo, A., *Deepfake Video Detection through Optical Flow Based CNN*, in: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, IEEE, Seoul, Korea (South) Oct. 2019, pp. 1205–1207, ISBN: 9781728150239, DOI: 10.1109/ICCVW.2019.00152, URL: https://ieeexplore.ieee.org/document/9022558/ (visited on 10/27/2025).

Anshul, A., Gopal, S., Rajan, D., and Chng, E. S., *Intra-modal and Cross-modal Synchronization for Audio-visual Deepfake Detection and Temporal Localization*, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Honolulu, Hawaii, USA Oct. 2025, pp. 13826–13836.

Beckmann, A., Hilsmann, A., and Eisert, P., *Fooling State-of-the-Art Deepfake Detection with High-Quality Deepfakes*, arXiv:2305.05282, May 2023, DOI: 10.48550/arXiv.2305.05282, URL: http://arxiv.org/abs/2305.05282 (visited on 10/27/2025).

Benelli, F., Këlliçi, E., Maciariello, F., Salvadori, C., and Stile, V., *Enhance Student Well-being and Digital Literacy with Machine Learning and Spatial Analysis*, ENG, in: *Proceedings of the 2nd International Workshop on Education for Artificial Intelligence (EDU4AI 2025)*, vol. 4114, AI*IA SERIES, urn:nbn:de:0074-4114-x,

CEUR Workshop Proceedings, The Engineering School of University of Bologna, Bologna, Italy Oct. 2025, Session S2: AI Literacy and Education, ISBN: ISSN 1613-0073, URL: https://ceur-ws.org/Vol-4114/.

Benelli, F., Këlliçi, E., Maciariello, F., and Stile, V., *Artificial Intelligence for Decentralized Orchestration in the Physical Internet: Opportunities, Business Trade-offs, and Risks in Road Freight Logistics*, ENG, in: *Book of Abstract of the 4th International Conference Creativity And Innovation In Digital Economy*, vol. Section 2: Co-creation, living labs and innovation ecosystems, Section 2: Co-creation, living labs and innovation ecosystems, Petroleum-Gas University of Ploiești Publishing House, Petroleum-Gas University of Ploiești (UPG), Ploiești, Romania Oct. 2025, ISBN: ISSN 2971–9798.

Benelli, F., Maciariello, F., Marku, R., and Stile, V., *Towards an Energy Physical Internet: Open Business Models and Platforms for Electricity Distribution Enabled by IoT, Blockchain, and Conditional Payments*, ENG, in: *Book of Abstract of the 4th International Conference Creativity And Innovation In Digital Economy*, vol. Section 4: New Pathways in Knowledge, Education and Law, Section 4: New Pathways in Knowledge, Education and Law, Petroleum-Gas University of Ploiești Publishing House, Petroleum-Gas University of Ploiești (UPG), Ploiești, Romania Oct. 2025.

Dang, H., Liu, F., Stehouwer, H., Liu, X., and Jain, A. K., *Detection of deepfake videos using multi-attentional convolutional neural networks*, in: *European Conference on Computer Vision*, Springer 2020, pp. 660–676.

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., and Ferrer, C. C., *The DeepFake Detection Challenge (DFDC) Dataset*, 2020, DOI: 10.48550/ARXIV.2006.07397, URL: https://arxiv.org/abs/2006.07397 (visited on 12/18/2025).

Durall, R., Keuper, M., and Keuper, J., *Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions*, in: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 2020, pp. 7887–7896, DOI: 10.1109/CVPR42600.2020.00791.

European Data Protection Board, *Guidelines 4/2019 on Article 25 Data Protection by Design and by Default*, https://www.edpb.europa.eu/our-work-tools/our-documents/

guidelines/guidelines-42019-article-25-data-protection-design-and_en, 2020, (visited on 11/07/2025).

European Data Protection Supervisor, *TechSonar Report 2023–2024: Emerging Technologies (including Deepfake Detection)*, https://www.edps.europa.eu/system/files/2023-12/23-12-04_techsonar_23-24_en.pdf, Report published 4 Dec 2023, Dec. 2023, (visited on 11/07/2025).

European Union, *Regulation (EU) 2016/679 (General Data Protection Regulation): Articles 25 (Data Protection by Design and by Default) and 32 (Security of Processing)*, https://eur-lex.europa.eu/eli/reg/2016/679/oj, Official Journal of the European Union, 2016, (visited on 11/07/2025).

— *Regulation (EU) 2024/1689 (Artificial Intelligence Act), including Annex III: High-Risk AI Systems Categories*, https://eur-lex.europa.eu/eli/reg/2024/1689/oj, Official Journal of the European Union, 2024, (visited on 11/07/2025).

Fisher, R. A., *Statistical Methods for Research Workers*, Oliver and Boyd, Edinburgh, United Kingdom 1925.

Floridi, L., *Etica dell'intelligenza artificiale: sviluppi, opportunità, sfide*, ita, ed. by M. Durante, Prima edizione, Scienza e idee 340, Raffaello Cortina Editore, Milano, Italy 2022, ISBN: 9788832854091.

Gong, Y. and Zhang, P., *Research on Mnist Handwritten Numbers Recognition based on CNN*, in: *Journal of Physics: Conference Series* 2138.1 (Dec. 2021), p. 012002, ISSN: 1742-6588, 1742-6596, DOI: 10.1088/1742-6596/2138/1/012002, URL: https://iopscience.iop.org/article/10.1088/1742-6596/2138/1/012002 (visited on 12/16/2025).

Guarnera, L., Giudice, O., Guarnera, F., Ortis, A., Puglisi, G., Paratore, A., Bui, L. M. Q., Fontani, M., Coccomini, D. A., Caldelli, R., Falchi, F., Gennaro, C., Messina, N., Amato, G., Perelli, G., Concas, S., Cuccu, C., Orrù, G., Marcialis, G. L., and Battiato, S., *The Face Deepfake Detection Challenge*, en, in: *Journal of Imaging* 8.10 (Sept. 2022), p. 263, ISSN: 2313-433X, DOI: 10.3390/jimaging8100263, URL: https://www.mdpi.com/2313-433X/8/10/263 (visited on 10/27/2025).

Guera, D. and Delp, E. J., *Deepfake Video Detection Using Recurrent Neural Networks*, in: *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, IEEE, Auckland, New Zealand Nov. 2018, pp. 1–6, ISBN: 978-1-5386-9294-3, DOI: 10.1109/AVSS.2018.8639163, URL: https://ieeexplore.ieee.org/document/8639163/ (visited on 06/10/2025).

Guerrero-Contreras, G., Balderas-Díaz, S., García-Pascual, A., and Munõz, A., *Self-Learning Systems for Enhanced Traffic Management in Urban Settings*, enc, in: (June 2024), DOI: 10.5281/ZENODO.11917270, URL: https://zenodo.org/doi/10.5281/zenodo.11917270 (visited on 11/28/2025).

Guerrero-Contreras, G., Balderas-Díaz, S., García-Pascual, A., and Muñoz, A., *Adaptive Vehicle Detection in Urban Environments: A Self-learning Approach*, en, in: *Ambient Intelligence – Software and Applications – 15th International Symposium on Ambient Intelligence*, ed. by P. Novais, P. B. D., I. Satoh, V. J. Inglada, S. R. González, E. Jove Pérez, J. Parra Domínguez, P. Chamoso, and R. S. Alonso, vol. 1279, Springer Nature Switzerland, Cham, Switzerland 2025, pp. 25–34, ISBN: 9783031831164 9783031831171, DOI: 10.1007/978-3-031-83117-1_3, URL: https://link.springer.com/10.1007/978-3-031-83117-1_3 (visited on 10/27/2025).

— *Adaptive Vehicle Detection in Urban Environments: A Self-learning Approach*, en, in: *Ambient Intelligence – Software and Applications – 15th International Symposium on Ambient Intelligence*, ed. by P. Novais, P. B. D., I. Satoh, V. J. Inglada, S. R. González, E. Jove Pérez, J. Parra Domínguez, P. Chamoso, and R. S. Alonso, vol. 1279, Springer Nature Switzerland, Cham, Switzerland 2025, pp. 25–34, ISBN: 9783031831164 9783031831171, DOI: 10.1007/978-3-031-83117-1_3, URL: https://link.springer.com/10.1007/978-3-031-83117-1_3 (visited on 10/27/2025).

Hazirbas, C., Bitton, J., Dolhansky, B., Pan, J., Gordo, A., and Ferrer, C. C., *Towards Measuring Fairness in AI: The Casual Conversations Dataset*, in: *IEEE Transactions on Biometrics, Behavior, and Identity Science* 4.3 (July 2022), pp. 324–332, ISSN: 2637-6407, DOI: 10.1109/TBIOM.2021.3132237, URL: https://ieeexplore.ieee.org/document/9634168/ (visited on 12/17/2025).

Katamneni, V. S. and Rattani, A., *Contextual Cross-Modal Attention for Audio-Visual Deepfake Detection and Localization*, arXiv:2408.01532, Aug. 2024, DOI: 10.48550/arXiv.2408.01532, URL: http://arxiv.org/abs/2408.01532 (visited on 12/17/2025).

Korshunov, P. and Marcel, S., *DeepFakes: a New Threat to Face Recognition? Assessment and Detection*, arXiv:1812.08685, Dec. 2018, DOI: 10.48550/arXiv.1812.08685, URL: http://arxiv.org/abs/1812.08685 (visited on 10/27/2025).

Levi, G. and Hassner, T., *Age and gender classification using convolutional neural networks*, in: *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Boston, MA, USA June 2015, pp. 34–42, ISBN: 9781467367592, DOI: 10.1109/CVPRW.2015.7301352, URL: http://ieeexplore.ieee.org/document/7301352/ (visited on 12/17/2025).

Li, Y., Chang, M.-C., and Lyu, S., *Celeb-DF (v2): A new dataset for deepfake forensics*, 2020, URL: https://cse.buffalo.edu/~siweilyu/celeb-deepfakeforensics.html (visited on 10/27/2025).

— *In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking*, in: *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, IEEE, Hong Kong Dec. 2018, pp. 1–7, ISBN: 9781538665367, DOI: 10.1109/WIFS.2018.8630787, URL: https://ieeexplore.ieee.org/document/8630787/ (visited on 10/27/2025).

Li, Y., Yang, X., Sun, P., Qi, H., and Lyu, S., *Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics*, in: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Seattle, WA, USA June 2020, pp. 3204–3213, ISBN: 9781728171685, DOI: 10.1109/CVPR42600.2020.00327, URL: https://ieeexplore.ieee.org/document/9156368/ (visited on 10/27/2025).

Medical Device Coordination Group, *MDCG 2019-11 Rev.1: Guidance on Qualification and Classification of Software under Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR*, https://health.ec.europa.eu/document/download/b45335c5-1679-4c71-a91c-fc7a4d37f12b_en, Revision 1 (June 2025), June 2025, (visited on 11/07/2025).

Neekhara, P., Dolhansky, B., Bitton, J., and Ferrer, C. C., *Adversarial Threats to DeepFake Detection: A Practical Perspective*, in: *2021 IEEE/CVF Conference on Computer*

Vision and Pattern Recognition Workshops (CVPRW), IEEE, Nashville, TN, USA June 2021, pp. 923–932, ISBN: 9781665448994, DOI: 10.1109/CVPRW53098.2021.00103, URL: https://ieeexplore.ieee.org/document/9522903/ (visited on 12/17/2025).

Rana, M. S., Nobi, M. N., Murali, B., and Sung, A. H., *Deepfake Detection: A Systematic Literature Review*, in: *IEEE Access* 10 (2022), pp. 25494–25513, ISSN: 2169-3536, DOI: 10.1109/ACCESS.2022.3154404, URL: https://ieeexplore.ieee.org/document/9721302/ (visited on 10/27/2025).

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M., *FaceForensics++: Learning to Detect Manipulated Facial Images*, in: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Seoul, Korea (South) Oct. 2019, pp. 1–11, ISBN: 9781728148038, DOI: 10.1109/ICCV.2019.00009, URL: https://ieeexplore.ieee.org/document/9010912/ (visited on 10/27/2025).

Sabir, E., Cheng, W., Jaiswal, A., AbdAlmageed, W., Masi, I., and Natarajan, P., *Recurrent convolutional strategies for face manipulation detection in videos*, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* 2019, pp. 1–9.

Stile, V., Bonino, V., and Cosmo, N., *The impact BI and AI on traditional structures with legal and philosophical insights*, ENG, in: *Proceedings of the 21st Conference of the Italian Chapter of AIS (itAIS 2024)*, vol. 21, AISeL - Springer LNISO, Università Cattolica del Sacro Cuore (UCSC), Piacenza, Italy Oct. 2024, ISBN: 979-12-82308-00-7, DOI: 10.979.1282308/007, URL: https://aisel.aisnet.org/itais2024/21.

Stile, V., Caldelli, R., Guerrero-Contreras, G., Balderas-Díaz, S., and Medina-Bulo, I., *Analysis of DeepFake Detection through Semi-Supervised Facial Attribute Labeling*, ENG, in: *Proceedings of the 11th Spanish-German Symposium on Applied Computer Science (SGSOACS 2025)*, vol. 2831, Communications in Computer and Information Science (CCIS), Springer Cham, Wien, Austria July 2025, pp. XX, 138, ISBN: 978-3-032-14815-5, URL: https://link.springer.com/book/9783032148155.

Stile, V. and Fontanella, A., *AI-Enhanced Building Information Modelling and Big Data Analytics for Civil Engineering Innovation*, ENG, in: *Book of Abstract of the 4th International Conference Creativity And Innovation In Digital Economy*, vol. Section

1: Innovative open business models and platforms, Section 1: Innovative open business models and platforms, Petroleum-Gas University of Ploiești Publishing House, Petroleum-Gas University of Ploiești (UPG), Ploiești, Romania Oct. 2025, ISBN: ISSN 2971–9798.

Tapo, A. A., Traore, A., Danioko, S., and Tembine, H., *Machine Intelligence in Africa: a survey*, in: *arXiv* (2024), Accepted for DSAI 2024, DOI: 10.48550/arXiv.2402.02218, arXiv: 2402.02218 [cs.CY], URL: https://arxiv.org/abs/2402.02218.

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., and Ortega-Garcia, J., *Deepfakes and beyond: A survey of face manipulation and fake detection*, in: *Information Fusion* 64 (2020), pp. 131–148.

Verdoliva, L., *Media Forensics and DeepFakes: An Overview*, in: *IEEE Journal of Selected Topics in Signal Processing* 14.5 (Aug. 2020), pp. 910–932, ISSN: 1932-4553, 1941-0484, DOI: 10.1109/JSTSP.2020.3002101, URL: https://ieeexplore.ieee.org/document/9115874/ (visited on 10/27/2025).

World Health Organization, *Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multimodal Models (LMMs)*, Genève, Switzerland, 2024, URL: https://www.who.int/publications/i/item/9789240084759 (visited on 11/07/2025).

— *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance*, World Health Organization, Genève, Switzerland 2021, ISBN: 978-92-4-002920-0, URL: https://www.who.int/publications/i/item/9789240029200 (visited on 11/07/2025).

Yang, X., Li, Y., and Lyu, S., *Exposing Deep Fakes Using Inconsistent Head Poses*, in: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Brighton, United Kingdom May 2019, pp. 8261–8265, ISBN: 9781479981311, DOI: 10.1109/ICASSP.2019.8683164, URL: https://ieeexplore.ieee.org/document/8683164/ (visited on 10/27/2025).

Zhang, N., Paluri, M., Ranzato, M., Darrell, T., and Bourdev, L., *PANDA: Pose aligned networks for deep attribute modeling*, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2014, pp. 1637–1644.

The PhD scholarship co-funded with resources from the European Union – NextGeneration EU National Recovery and Resilience Plan (PNRR), Mission 4, Component 2 *"From Research to Business"* – Investment 3.3 *"Introduction of innovative PhD programmes that meet the innovation needs of enterprises and promote the recruitment of researchers by companies"* – CUP D83C22001880003