Analyzing the Interplay Between Privacy, Explainability, and Fairness in Responsible Artificial Intelligence

Doctoral Dissertation submitted to the
Faculty of Informatics of the Università della Svizzera italiana
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

presented by

Fatima Ezzeddine

under the supervision of

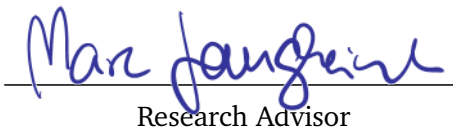
Marc Langheinrich and Silvia Giordano, Omran Ayoub

June 2026

Dissertation Committee

Prof. Gabriele Bavota	Università della Svizzera italiana
Prof. Paolo Tonella	Università della Svizzera italiana
Prof. Grégoire Montavon	Charité Berlin, Germany
Prof. Kévin Huguenin	Université de Lausanne, Switzerland
Prof. Francesco Regazzoni	University of Amsterdam, Netherlands
Prof. Luca Longo	University College Cork, Ireland

Dissertation accepted on 5 June 2026



Research Advisor

Marc Langheinrich

Co-Advisor

Silvia Giordano, Omran Ayoub

PhD Program Director

Prof. Walter Binder/ Prof. Stefan Wolf

I certify that except where due acknowledgement has been given, the work presented in this thesis is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; and the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program.

Fatima Ezzeddine
Lugano, 5 June 2026

To my family

On Fairness & Equality

"People are of two kinds: either your brothers in faith, or your equals in humanity."

On Privacy & Human Dignity

"Do not disclose whatever of it is hidden from you because your obligation is to correct what is manifest to you."

On Equal Treatment

"Do not behave with people in a way that you would dislike being treated yourself."

— Letters 31 and 53 (to his son Hassan and to Mālik al-Ashtar)
Imam Ali, Nahj al-Balāgha

Abstract

Machine learning (ML) models are increasingly adopted in domains where decisions have significant social and economic impact, such as education, recruitment, and finance. As their use expands, the responsible deployment of such models has become a central concern. Beyond achieving strong predictive performance, deployed models should be audited and understood by humans, avoid reproducing or amplifying unfair biases, and safeguard the sensitive information contained in their training data. Recent regulatory frameworks, such as the European General Data Protection Regulation and the Artificial Intelligence (AI) Act, further reinforce these demands by mandating transparency, accountability, and safeguards for privacy and fairness in automated decision-making. As a result, data collection, model training, and deployment must satisfy not only technical performance criteria but also legal requirements, such as *privacy, explainability, and fairness*, for Responsible AI. Although substantial effort has been devoted to addressing each desiderata separately, achieving them simultaneously is considerably challenging. These requirements are often studied in isolation, despite their complex, interdependent interactions that can reinforce or undermine one another. For instance, enhancing transparency and explainability through explainable AI may expose sensitive information, while strong privacy guarantees can obscure explanations and hinder fairness auditing. Likewise, fairness interventions may alter model behavior in ways that affect explainability or increase privacy risks. In this thesis, we investigate how explainability interacts with other core Responsible AI principles to analyze the interplay among: *(1) privacy and explainability*, focusing on how explanations can be exploited for privacy attacks, such as membership inference and model extraction, and on corresponding mitigation strategies, including differential privacy and data distillation. We further investigate the synergies between collective privacy and explainability and identify associated open questions. *(2) fairness and explainability*, analyzing how fairness-aware modeling influences explanations and how the fairness of explanations themselves can be enforced. By characterizing the empirical trade-offs and identifying the conditions under which these desiderata can be jointly satisfied or necessarily conflict, this thesis contributes to the development of models that balance multiple Responsible AI requirements.

Acknowledgements

Above all, I thank God for His guidance, blessings, and for giving me the strength and patience to overcome the challenges of this journey.

I would like to express my deepest gratitude to my advisors, *Prof. Silvia Giordano, Prof. Marc Langheinrich, and Dr. Omran Ayoub*, for their invaluable guidance, support, and constant encouragement throughout my PhD. Their expertise, patience, and thoughtful feedback have shaped not only this thesis but also my development as a researcher.

Thank you, *Prof. Giordano*, for your continuous support and for being a steady source of clarity and reassurance throughout this journey.

Thank you, *Dr. Ayoub*, from whom I learned immensely and whose mentorship has had a lasting impact on my work and growth, allowing me to grow both professionally and personally.

Thank you, *Prof. Langheinrich*, for teaching me to think beyond conventions and to approach problems with creativity, criticism, and openness.

I am sincerely thankful to the members of my thesis committee: *Prof. Paolo Tonella, Prof. Gabriele Bavota, Prof. Luca Longo, Prof. Francesco Regazzoni, Prof. Kévin Huguenin, and Prof. Grégoire Montavon* for their time, careful evaluation of this dissertation, and constructive feedback.

I would like to express my gratitude to the *Swiss Government* for awarding me the Swiss Government Excellence Scholarship. Their support made my doctoral studies possible and enabled me to pursue this research under exceptional academic and professional conditions. I am also sincerely grateful to *Ms. Maurizia Ruinelli and Ms. Arianna Imberti Dosi* for their support, guidance, and assistance throughout my stay.

My sincere gratitude also goes to the director of the Institute of Information Systems and Networking at SUPSI, *Prof. Tiziano Leidi*, to *Lara Ruggiu*, and to the entire team for their commitment to fostering a welcoming academic environment.

My sincere appreciation also goes to *Prof. Verena Zimmermann and Dr. Noé Zufferey* from the Security, Privacy and Society Group at ETH Zurich for their warm welcome during my research visit. Their mentorship and the enriching academic environment they fostered made my time there deeply valuable.

I am equally thankful for the opportunity to conduct a research visit to the University of Quebec at Montreal with *Prof. Sébastien Gambs and his group*. Their hospitality, insightful discussions, and collaborative spirit made my time in Montreal academically stimulating.

I would also like to extend my thanks to the professors I had the privilege of working with as a teaching assistant: *Prof. Cesare Alippi and Prof. Luca Gambardella*. Their trust and guidance made teaching an enriching and formative part of my doctoral journey and beyond.

My deepest gratitude goes to my family: *my father Issam, my mother Alia, my sister Mariam, my brothers Ibrahim and Ismail, my nephew Karim*, and to my extended family *Haytham, Rachal, Alaa*, for their unconditional love, endless encouragement, and belief in me. To my parents, thank you for teaching me the values of hard work, perseverance, and education.

My warm thanks go to my colleagues at the university: *Davide Andreoletti, Jacopo Talpini, Matteo Ciniselli, Martin Gjoreski, Tomasz Szandala, Natalia Russin, Enrico Verdolotti, Azza Bouleimen, Andrea Doresti, Ali Housseini, Obaida Ammar, Gianluca Nogara, Souheila Serbout, Lorenzo Di Folco, Giada Ferraio, and Francesca Fusco*, for the many conversations, shared experiences, and moments that made everyday work more enjoyable. Their companionship created a supportive and motivating atmosphere.

I am also grateful to *Claudia Boschung, Lorin Schöni, Neele Roch, Hannah Sievers, Adrienn Toth, Marine Bruttin, Enzo Petrocco, Amol Deshmukh, Jana Seich, Maria Rivera, and David Crevillen* for their academic and cultural exchanges and for the simple moments of connection. You made my time in Zurich memorable.

I am equally grateful to my friends near and far, *Maha Ezzeddine, Fatima Safieddine, Fatima Dhayni, Nour Daher, Zahraa Nasser, Nada Salman, Hiba Kobeissy, Zaynab Akil, Jana Eid, Hussein Jammal, Bilal Tahini, Hussein Jiche, Hadi Housseini and Israa Fakih*, whose encouragement and understanding helped me maintain balance and perspective during the demanding years of doctoral research.

Thank you, *Mirna Saad*, for your constant support, for being my everyday companion throughout this journey, and for always listening whenever I needed it. Your presence

made even the most challenging days lighter.

Thank you, *Hussein Fawaz*, for patiently learning photography along the way and for generously sharing your time and talent.

Thank you, *Douaa Akil*, for being my best friend. Your encouragement, care, and constant presence, whether through video calls, visits, or simply being there when I needed it most, made this journey immeasurably lighter. Your understanding, positivity, and the comfort you brought along the way have been a true gift.

Thank you, *Fatima Lakkis*, my partner in everything since our earliest university days. Thank you for standing beside me through every challenge, for believing in science and in this path as deeply as I do, and for supporting me with loyalty and sincerity.

Thank you, *Mariouma*, for being the sister who has believed in me from the very beginning, for standing by my side through every step of my life, for always seeing my potential even when I doubted it myself, and for always reminding me to take care of myself.

Lastly but not least, I would also like to honor my southern Lebanese ancestors and our homeland, whose resilience, heritage, and enduring spirit continue to shape who I am. I am grateful for the cultural richness, values, and sense of belonging that this heritage has planted in me.

Contents

Figures

Tables

Chapter 1

Introduction

Artificial Intelligence (AI) and Machine learning (ML) have become integral to automated decision-making, powering models that analyze complex data and decisions across domains, from medical diagnosis and financial forecasting to criminal risk assessment, thereby shaping outcomes that affect individuals and groups [4, 5, 6, 7]. However, growing concerns have emerged about their ethical implications, transparency, potential biases, and privacy risks to individuals and collectives [8, 9]. For instance, it is essential that ML model decisions remain auditable, transparent, and interpretable by humans [10, 11], avoid biased or discriminatory behavior [12], and ensure the privacy of their training data [13, 14]. These requirements have made the responsible, transparent, and accountable data collection and deployment of ML models a central focus of both research and policy, such as the European (EU) General Data Protection Regulation (GDPR) [15], the EU AI Act [16], and the Swiss Federal Act on Data Protection (nFADP) [17]. Such efforts have initiated a new focus toward embedding *privacy*, *explainability*, and *fairness* into the design and evaluation of data collection and model training under the umbrella of *Responsible AI*.

1.1 Research Motivation and Problem Definition

1.1.1 Privacy

The reliance of ML training on large, often sensitive datasets [18] makes *privacy* a central concern. Safeguarding these datasets is essential to protecting individual and group privacy while still enabling models to extract meaningful insights. The EU AI Act and the GDPR, as well as many national acts like the Swiss nFADP, emphasize data privacy through data governance and transparency (e.g., GDPR Articles 10 and 13). Specifically, privacy is a human right that gives individuals the right to control and pro-

tect their data from unauthorized access, disclosure, or inference throughout the data collection and AI lifecycle, from collection and training to deployment [19]. To protect privacy and sensitive information present in collected datasets, privacy-preserving methods have been proposed, which are generally categorized into *data release* and *model release* approaches [20]. Data release methods obfuscate datasets to prevent de-identification, while model release methods secure trained ML models by protecting internal representations that could encode sensitive information about the training dataset. These methods defend against privacy attacks, ensuring privacy in the deployment of Machine Learning as a Service (MLaaS). Privacy attacks such as model extraction attacks (MEA) can steal proprietary functionality of deployed models, membership inference attacks (MIA) aim to reveal whether an individual's data was used for training, and property inference attacks (PIA) reconstruct training samples, or infer sensitive attributes [21].

In addition to technical methods designed to ensure privacy by design, in order to mitigate privacy risks, a range of privacy-preserving methods have been proposed, including data anonymisation and differential privacy (DP) [22], the former relying on removing or masking identifiers, the latter providing a rigorous mathematical guarantee by bounding the influence of any individual record on model outputs or aggregate statistics within a privacy budget. However, such protections impose a *privacy-utility trade-off*: stronger privacy guarantees inherently introduce tighter bounds, which typically degrade model performance and generalization [23].

1.1.2 Explainability

ML models often operate as black boxes, with their internal decision-making processes inaccessible, opaque, or only partially understood, even to experts who train, deploy, or audit them. This opacity limits the ability to understand how models arrive at their decisions, undermining trust even when predictive performance is high. This lack of transparency underscores the growing need for *explainability* [24], as also emphasized by the EU AI Act through its 'right to explanation' and specific requirements for high-risk AI systems. Explainability through Explainable AI (xAI) addresses these limitations by enabling the interpretation of a model's internal reasoning, promoting transparency by revealing the rationale behind predictions, and aligning explanations with human reasoning [25, 26]. The degree of opacity varies with model complexity, ranging from relatively transparent models such as decision trees and linear regression to highly complex architectures such as deep neural networks, which involve millions of parameters and highly non-linear transformations that are inherently difficult to interpret. This variety in complexity often leads to an *explainability–utility trade-off*, because simpler, more interpretable models usually can't capture patterns as richly as complex black-box

models.

Among the widely adopted xAI techniques are *feature attribution* (FA) and *counterfactual explanations* (CF), which provide complementary perspectives on model behavior. FA addresses *why* a decision was made, CF explains *what* would need to change for a different decision, making them particularly valuable for actionable recourse. Specifically, FA methods quantify the contribution of each input feature to a prediction by assigning each feature an importance score indicating how strongly it influences the model's output [27, 28]. In contrast, CF provides hypothetical scenarios by identifying minimal changes to the input features that would lead the model to produce a desired alternative outcome [10].

1.1.3 Fairness

Finally, *fairness* aims to ensure that ML decisions are equitable across individuals and groups, e.g., free from bias or discrimination [29]. Decisions may be considered discriminatory if they are based on hard-to-change variables (e.g., gender, ethnicity, age, or disability). Bias can emerge from various sources, including imbalanced training data, biased labeling, feature selection, or model design, potentially leading to outcomes that disadvantage certain groups [30]. However, promoting fairness is essential not only for ethical and societal reasons but also to ensure compliance with emerging AI regulations and established non-discrimination laws, and to improve model reliability and generalization across diverse scenarios. In practice, achieving fairness involves mitigating disparities in data and model behavior to guarantee that comparable individuals receive comparable outcomes [31].

Specifically, fairness is formalized through two complementary notions: *individual fairness*, which requires that similar individuals receive similar outcomes, and *group fairness*, which requires that statistical outcomes are equitable across demographic groups, measured through metrics such as demographic parity, equalized odds, and predictive parity [32]. Yet fairness interventions introduce a fundamental *fairness-utility tradeoff*, where enforcing stricter fairness constraints typically comes at the cost of reduced predictive performance, making the calibration of this tradeoff a central challenge for fair ML [29].

1.1.4 Interactions, Tensions, and Synergies

Privacy, explainability, and fairness are often studied as separate desiderata, with primary attention given to model utility, such as the *privacy-utility*, *explainability-utility*, and *fairness-utility* tradeoffs. In practice, they act as coupled constraints on the same underlying model, same data distribution, and same deployment context. Each dimen-

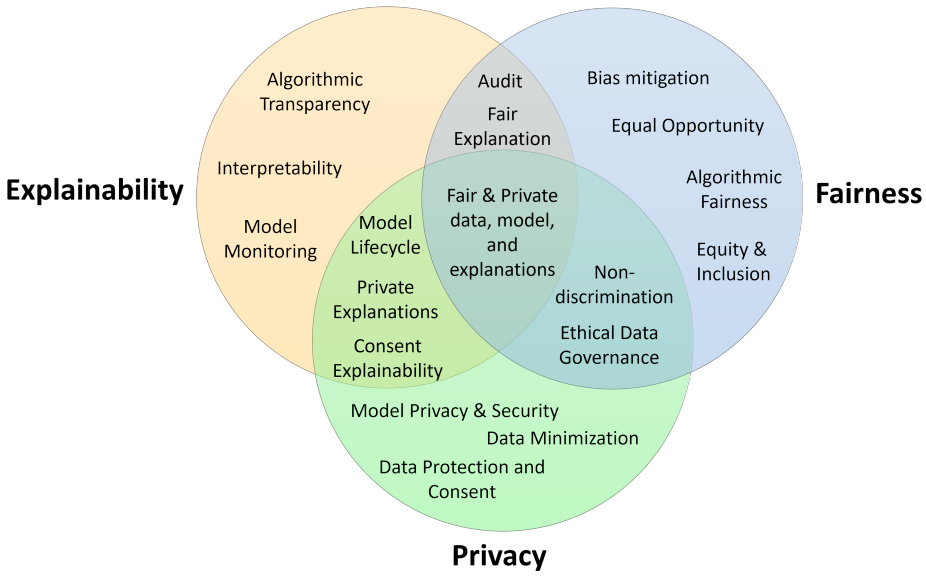


Figure 1.1. Triad of Privacy-Fairness-Explainability

sion affects what a model learns, how it represents individuals and groups, and what information it exposes to users. Consequently, interventions designed to improve one desiderata inevitably propagate to the others, because they shape the model's representations, decision boundaries, and information flows. A privacy intervention may alter group-level behavior, a fairness constraint may reshape the features that explanations highlight, and an explainability technique may expose sensitive correlations. These interactions can sometimes be mutually reinforcing, but more often they introduce new tensions that are invisible when each objective is treated in isolation. Understanding this triadic interplay is, therefore, essential for developing Responsible AI systems that must satisfy all three requirements simultaneously. Figure 1.1 illustrates the triadic interplays: *privacy and explainability*, *fairness and explainability*, *privacy and fairness* and *privacy, explainability, and fairness*. In the following, we describe each tension and highlight the challenges that arise when different desiderata intersect.

A tension exists between *privacy and explainability* [33]: explanations may inadvertently expose sensitive information, such as CFs and FA that can reveal protected attributes, decision boundaries, and that can be exploited for MEA or MIA [34, 35]. Conversely, privacy-preserving methods such as DP introduce statistical noise that can reduce the reliability of explanations [36]. This tension is structural, as explanations (e.g., CFs, gradients, SHAP values) reveal the local sensitivity of outputs to inputs. These sensitivities are precisely what DP tries to obscure. Privacy and explainability can also be complementary, as xAI outputs can serve as tools for auditing and reveal-

ing the influence of sensitive attributes, thereby enabling their removal or masking to enhance privacy.

Fairness and explainability often exhibit a bidirectional and structurally embedded tension. On the one hand, explainability methods are indispensable for auditing fairness, as they help reveal whether and how sensitive attributes, or their proxies, influence model predictions. Without explanations, it becomes challenging to diagnose disparate treatment, detect discrimination, or understand the mechanisms through which bias propagates through a model. On the other hand, fairness interventions can substantially alter behavior and the quality of explanations. Many fairness-enhancing techniques, such as regularization-based constraints, adversarial debiasing, or representation learning, modify the model's internal reasoning, shift decision boundaries, or reshape latent representations. These shifts can propagate to post-hoc explanations, affecting their stability, fidelity, and semantic coherence. For example, fairness regularizers may suppress correlations with sensitive attributes, but in doing so, they can also impact FA or produce explanations that are less faithful to the underlying predictive mechanism. Additionally, model-level fairness does not guarantee explanation-level fairness. A classifier may satisfy group fairness metrics while still producing explanations that differ systematically across demographic groups, for instance, by attributing decisions to different features for otherwise similar individuals, or by generating less stable or less informative explanations for minority groups. Such disparities can create new forms of opacity or discrimination at the explanation layer, even when the model's outputs appear fair. This motivates the need to evaluate not only the fairness of predictions but also the robustness, consistency, and equity of the explanations themselves.

Privacy and fairness represent another area of tension, as efforts to enhance privacy can inadvertently compromise fairness, and vice versa [37, 38]. The US Census Bureau illustrates the practical challenges of adopting DP for Housing in 2020 [39]. DP effectively protected sensitive information, but the injected noise later produced unequal budget allocations across school districts. Similarly, the Apple Card incident in 2019 [40] drew public attention when users reported that the credit limit model offered lower limits to women than to men with comparable financial backgrounds, although gender was excluded as an input variable for privacy reasons, this exclusion hindered post-hoc bias detection, as gender-based disparities emerged indirectly through correlated features [11]. Additionally, initial research showed that fairness-aware models are more vulnerable to privacy attacks, such as MIA [41].

Taken together, these pairwise tensions demonstrate that *privacy, explainability, and fairness* do not operate as independent desiderata but constitute a triadic interdependent constraint. Interventions aimed at improving one dimension inevitably propagate through the others, often in counterintuitive ways. Investigating this triadic interplay is therefore crucial for developing methods that remain robust, mutually com-

patible, and ethically coherent in real-world deployments where all three requirements must be satisfied jointly under responsible AI. This recognition directly motivates the research questions (RQ) of this thesis, which aim to systematically investigate how privacy, explainability, and fairness interact and how these interactions can be managed or leveraged in practice.

1.2 Thesis Goal

At the outset of my Ph.D. studies, we explored multiple research directions to identify the most impactful path forward. Since this thesis aims to advance responsible AI by addressing privacy, explainability, and fairness jointly, it is grounded in the hypothesis that these three dimensions are not independent objectives, but interacting constraints: interventions aimed at strengthening one can systematically reshape model behavior in ways that affect the others, sometimes improving them but often introducing new trade-offs. Specifically, this thesis focuses on two underexplored pairwise interplays: *privacy and explainability* and *fairness and explainability*. We illustrate key patterns in how interventions in one dimension propagate to others and examine these trade-offs in depth, propose methods to quantify and mitigate them, and offer insights into pursuing jointly responsible AI desiderata. Particularly, the aim is to focus on *how explainability interacts with other key responsible AI requirements, such as privacy and fairness, and how we can design methods that balance the objectives*. In particular, we address the following research questions (RQ):

- *Privacy and Explainability*
  - *RQ1: To what extent do different types of explanations expose models to privacy attacks, and what factors influence the magnitude of this leakage?*
  - *RQ2: How do different privacy defense mechanisms influence the effectiveness of explanation-based privacy attacks, and what is their effect on explanation quality?*
  - *RQ3: To what extent can explainability play a beneficial role in enhancing privacy protection?*
- *Fairness and Explainability*
  - *RQ4: How do model fairness-enhancing techniques impact explanations and what methods can enhance fairness with minimal impact on explanations?*
  - *RQ5: How effective are multi-objective optimization techniques at generating explanations that balance quality and fairness, and do they provide a better quantification of the interplay?*

Although the *privacy–fairness* interaction is not addressed in this thesis, it is highlighted as an important avenue for future work, and the insights gained from the studied tradeoffs provide guidance for reasoning about broader triadic interactions. Building on these findings, the thesis also identifies additional open questions that require further investigation, including:

- *Privacy and Fairness*
  - *RQ6: To what extent do fairness-enhancing techniques affect the performance of privacy attacks, and vice-versa?*
- *The Triadic Interplay*
  - *RQ7: To what extent model fairness-enhancing techniques mitigate discriminatory effects in model predictions and explanations while preserving privacy jointly?*

1.3 Contributions and Publications

This thesis contributes by (i) quantifying the extent to which CFs can be exploited for MEA and MIA, (ii) analyzing mitigation frameworks to protect against these attacks, and characterizing the resulting privacy–explainability trade-off, (iii) investigating the concept of collective privacy and defining future research direction along its legality and need of explainability, (iv) proposing a method for preserving explanation quality while satisfying fairness desiderata, and (v) distinguishing between fair CFs and fair models, and introducing a technique for generating CFs that satisfy fairness constraints.

Specifically, in the first part of the thesis, we examine the interdependence between *privacy and explainability*, focusing on privacy risks arising when explanations are exposed to end users through MLaaS. We focus on how CFs can be exploited for performing MEA and MIA. Specifically, we propose a knowledge-distillation-based method to perform MEA that leverages a realistic deployment setting and CFs (supporting *RQ1*). To mitigate MEA vulnerability, we propose generating CFs with DP guarantees during training, thereby reducing the attack surface while preserving their practical utility (*RQ2*). Building on this foundation, we extend the analysis to MIA, with a particular focus on shadow–based MIAs and how CFs can amplify privacy leakage (*RQ1*). We also evaluate mitigation strategies grounded in active learning and DP, quantifying their impact on predictive performance and explainability (*RQ2*). We then broaden the scope to a new dimension of privacy, collective privacy, arguing for regulatory attention and outlining open questions concerning the relationship between collective harms and the potential benefits of explainability (*RQ3*).

In the second part of the thesis, we investigate the interplay between *fairness and explainability*, focusing on training fair models that preserve both predictive performance and the quality of the explanations they produce, particularly with respect to FA (*RQ4*). We then examine the fairness of explanations themselves by introducing a method for generating CFs that satisfies individual, group, and hybrid fairness at the CF explanation level (*RQ5*).

The contributions of this thesis extend the current landscape of responsible AI research along many aspects. Existing work on the interplay between privacy has examined how privacy-preserving mechanisms impact model and data utility. In contrast, the originality of the thesis lies in advancing a different perspective: transparency through explanations itself can serve as a surface for privacy attacks. This thesis investigates how CFs, originally introduced to enhance transparency, can be exploited to facilitate MEA. This reveals a tension for responsible AI: mechanisms intended to increase transparency may inadvertently increase privacy threats. The thesis further shows that such vulnerability is not intrinsic. The thesis is the first to propose integrating DP into the GAN-based CF generation process, enabling the production of explanations that preserve privacy while retaining practical utility. Achieving this outcome requires rethinking how DP is employed, how noise is introduced during the generative process, and how the trade-off between privacy protection and explanation quality is formally balanced. Moreover, this research direction extends to MIA, where prior work has examined how FA may amplify such attacks, yet no analysis has considered CFs in a shadow-based MIA. Therefore, this thesis provides the first systematic study of MIA that exploits CFs for shadow-based MIA and introduces a joint defense strategy that combines DP with active learning. In this formulation, active learning is repurposed not to reduce labeling costs but to minimize the number of training points exposed, thereby limiting the model's exposure to individual samples. This reframing illustrates how methods developed for one objective can serve a fundamentally different purpose when viewed through the lens of privacy preservation.

The thesis also addresses an aspect of the interplay between fairness and explainability that has received limited attention in the literature, the impact of fair model learning on explanations, and the fairness of explanations. The thesis advances a technique that safeguards the quality of FA through gradient-based methods even when fairness constraints are imposed at the model level, thereby demonstrating that fairness-aware optimization need not compromise explanatory reliability. Furthermore, we investigate when a model may satisfy fairness constraints at the level of predictions, but the CF generation method offers recourse options that are inequitable across protected groups. Although this distinction between fair predictions and fair recourse has been noted in the literature, it has not been formally operationalized. Particularly, in this thesis, we address the problem of generating CFs under hybrid fairness constraints:

individual-level fairness, which requires that similar individuals receive similar explanations, and group-level fairness, which requires equitable access to recourse across demographic groups.

The innovation of this thesis, what the new knowledge enables in practice, is demonstrated through several key outcomes that translate methodological contributions into practical value. Beyond its academic contributions, the thesis offers actionable guidance for practitioners navigating the legal and operational demands of transparency, privacy, and fairness in real-world deployments. *For MLaaS providers*, the findings inform the design of a framework to assess and mitigate the privacy risks posed by explanation interfaces. The DP-based methods provide concrete mechanisms for reducing susceptibility to model extraction and membership inference attacks while preserving explanation quality, enabling providers to make informed decisions about the degree of privacy protection required and the corresponding impact on explanation quality. *For researchers and engineers* working at the intersection of privacy, fairness, and explainability, the proposed evaluation framework and methods enable systematic quantification of how privacy-preserving interventions affect explanation fidelity and stability. *For organizations subject to emerging regulatory requirements*, the fair CF framework offers a practical method to ensure equitable recourse across protected groups, directly addressing obligations under the GDPR and the EU AI Act regarding non-discriminatory algorithmic decision-making. Finally, the conceptual analysis of collective privacy provides a foundation for future research into privacy harms affecting groups rather than individuals. By distinguishing among individual, interdependent, and collective privacy, and by identifying concrete use cases and regulatory gaps, the thesis outlines a research direction for an area that is only beginning to receive sustained attention. It further highlights the direct synergies between collective privacy and explainability, suggesting that explanation interfaces may serve as a mechanism for communicating group-level risks and auditing collective governance.

Beyond these specific contributions, the thesis operates at the intersection of three research communities that interact only minimally. Privacy research has focused on protecting data, explainability research on making models transparent, and fairness research on ensuring equitable outcomes, with each community developing its own methods, metrics, and assumptions.

The key publications representing the contributions of this thesis are as follows:

1. **Fatima Ezzeddine**, Obaida Ammar, Silvia Giordano, and Omran Ayoub. *Fair Recourse for All: Ensuring Individual and Group Fairness in Counterfactual Explanations.* IEEE Transactions on Artificial Intelligence Journal, 2026.

2. **Fatima Ezzeddine**, Silvia Giordano, and Omran Ayoub. *Exploiting Explanations for Model Extraction via Knowledge Distillation and Mitigation with Private Coun-*

terfactuals. Frontiers in Artificial Intelligence Journal, 2026.

3. **Fatima Ezzeddine**, Osama Zammar, Silvia Giordano, and Omran Ayoub. *Explanations Leak: Membership Inference with Differential Privacy and Active Learning Defense.* International Joint Conference on Neural Networks (IJCNN), in IEEE World Congress on Computational Intelligence (WCCI 2026), Nominated for Best Student Paper Award 🏅.

4. **Fatima Ezzeddine**, Noé Zufferey, Omran Ayoub, Silvia Giordano, and Verena Zimmermann. *Sok: Beyond the Individual: A Call for Collective Privacy Protection and Regulation.* 2026, Submitted Conference.

5. **Fatima Ezzeddine**, Rinad Akl, Ihab Sbeity, Silvia Giordano, Marc Langheinrich, and Omran Ayoub. *On the interplay of Explainability, Privacy and Predictive Performance with Explanation-assisted Model Extraction.* World Conference on Explainable Artificial Intelligence, Late-breaking Work, Demos and Doctoral Consortium Joint, CEUR Workshop Proceedings (xAI 2025).

6. Tomasz Szandala, **Fatima Ezzeddine**, Natalia Rusin, Silvia Giordano, and Omran Ayoub. *Fair-FLIP: Fair Deepfake Detection with Fairness-Oriented Final Layer Input Prioritising*, Swiss Data Science Conference (SDS 2025).

7. **Fatima Ezzeddine**, Omran Ayoub, and Silvia Giordano. *Knowledge Distillation-Based Model Extraction using Private GAN-based Counterfactual Explanations*. Workshop on Regulatable ML at Conference of Neural Information Processing (NeurIPS 2024).

8. **Fatima Ezzeddine**, Mirna Saad, Omran Ayoub, Davide Andreoletti, Martin Gjoreski, Ihab Sbeity, Marc Langheinrich, and Silvia Giordano. *Differential Privacy for Anomaly Detection: Analyzing the Trade-off Between Privacy and Explainability.* World Conference on eXplainable Artificial Intelligence (xAI 2024).

9. **Fatima Ezzeddine** *Privacy Implications of Explainable AI in Data-Driven Systems* Late-breaking Work, Demos and Doctoral Consortium Joint, CEUR Workshop Proceedings. *Best Doctoral Proposal Award* 🏆 (xAI 2024).

10. **Fatima Ezzeddine**, Omran Ayoub, Davide Andreoletti, and Silvia Giordano. *SACFACT: Soft Actor-Critic Reinforcement Learning for Counterfactual Explanations.* World Conference on eXplainable Artificial Intelligence (xAI 2023).

Other publications that are not included as a main component of this thesis.

1. **Fatima Ezzeddine**, Omran Ayoub, Silvia Giordano, Gianluca Nogara, Ihab Sbeity, Emilio Ferrara, and Luca Luceri. *Exposing influence campaigns in the age of LLMs: a behavioral-based AI approach to detecting state-sponsored trolls*. EPJ Data Science Journal, 2023.

2. **Fatima Ezzeddine**, Omran Ayoub, Davide Andreoletti, Massimo Tornatore and Silvia Giordano. *Vertical Split Learning-based Identification and Explainable Deep Learning-based Localization of Failures in Multi-Domain NFV Systems*. IEEE 9th Conference on Network Function Virtualization and Software Defined Networks (IEEE NFV/SDN 2023).

3. Davide Andreoletti, Cristina Rottondi, **Fatima Ezzeddine**, Omran Ayoub, Silvia Giordano. *ML-based network pruning for routing data overhead reduction in wireless sensor networks*. 18th Wireless On-Demand Network Systems and Services Conference (WONS 2023).

4. Hussein Fawaz, **Fatima Ezzeddine**, Silvia Giordano, and Omran Ayoub. *Towards Better-Calibrated ML Models for Reliable Network Intrusion Detection via Calibration-Aware SHAP-based Feature Selection*, IEEE International Workshop on Generative and eXplainable AI for Networking (GenXNet'25) in conjunction with WiMob 2025.

1.4 Organization

This thesis is organized as follows:

- Chapter 1: The present chapter introduces the interdependent dimensions of privacy, explainability, and fairness, and outlines our aim in investigating the tensions. We present the research questions that guide this thesis, and state the goals we seek to achieve. Finally, we provide an overview of the thesis structure, summarize the main contributions, and list the associated publications.

- Chapter 2 Present the technical background that forms the basis for the remainder of the thesis.

- Chapter 3 Describe the proposed approach for conducting MEA by exploiting CFs. We further describe the proposed method for generating privacy-preserving CFs under DP constraints, enabling the investigation of MEA risks in settings where privacy guarantees are enforced.

- Chapter 4 Investigates how CFs can be exploited for shadow-based MIA and provides a comprehensive analysis of mitigation strategies based on DP and active learning in the context of MIA.
- Chapter 5 Analyze the interplay between DP and xAI for anomaly detection. We analyze how DP influences model performance and the quality of the explanations it produces, shedding light on the inherent tradeoffs.
- Chapter 6 Introduces the concept of collective privacy, discusses its regulatory aspects, and poses research directions on how explainability techniques can support and enhance this emerging notion.
- Chapter 7 Present the proposed method for maintaining fairness in predictive image models while ensuring consistency in the FA explanations they generate, and evaluate its effectiveness.
- Chapter 8 Describes the proposed approach for generating fair CFs at individual, group, and hybrid levels. We analyze its performance and discuss its implications for fairness-aware explainability.
- Chapter 9 The thesis concludes with a synthesis of the main findings and a discussion of future research directions that arise from this work.

Chapter 2

Preliminaries

This chapter presents the key concepts that form the background of the remainder of the thesis. We introduce the essential background on privacy, explainability, and fairness, followed by the core notions for anomaly detection, knowledge distillation, active learning, and reinforcement learning.

2.1 Machine Learning

Machine Learning (ML) is a branch of AI concerned with developing computational methods that learn patterns, regularities, and decision rules from data. Rather than being explicitly programmed with fixed instructions for every possible situation, ML systems infer patterns from examples and use them to make predictions, classifications, recommendations, or decisions on previously unseen data. This data-driven paradigm has become central to modern intelligent systems, enabling progress in domains such as computer vision, natural language processing, healthcare, cybersecurity, finance, recommender systems, and autonomous decision-making. At its core, ML relies on the idea of generalization. A model is trained on a finite set of observations, usually referred to as the training data, and is expected to perform well on new data drawn from the same or a related distribution. The learning process typically involves selecting a model class, defining an objective function that measures prediction error, and optimizing the model parameters to minimize this error. However, achieving good performance on training data alone is not sufficient. A central challenge in ML is avoiding overfitting, in which a model memorizes training examples rather than learning robust patterns that generalize to unseen cases. For this reason, model evaluation is usually performed using validation and test data, along with performance metrics suited to the task, such as accuracy, precision, recall, F1 Score, area under the curve, or mean squared error.

ML methods are commonly categorized according to the type of supervision avail-

able during training. In supervised learning, models are trained using input-output pairs, where each example is associated with a target label or value. Classification and regression are the most common supervised tasks. In unsupervised learning, the goal is to discover hidden patterns in unlabeled data, for example, through clustering, dimensionality reduction, or representation learning. Reinforcement learning considers a different setting in which an agent interacts with an environment and learns a policy by receiving rewards or penalties for its actions. Deep learning has become a dominant ML paradigm, using neural networks with multiple layers to learn increasingly abstract representations from large-scale data.

Formally, let $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^{n}$ denote a dataset composed of n examples, where $x_i \in \mathscr{X}$ represents the input features and $y_i \in \mathscr{Y}$ represents the corresponding target label or value. Given a learning task, such as classification or regression, the objective of ML is to learn a function $f_\theta : \mathscr{X} \to \mathscr{Y}$, parameterized by θ, that maps inputs to outputs while generalizing well to unseen data. This is commonly achieved by minimizing an empirical risk over the training dataset:

$$\theta^* = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \mathscr{L}\big(f_\theta(x_i), y_i\big) + \lambda \Omega(\theta) \tag{2.1}$$

where $\mathscr{L}$ is a loss function that measures the discrepancy between the model prediction and the true target, $\Omega(\theta)$ is a regularization term used to control model complexity, and λ determines the strength of regularization. After training, the learned model f_{θ^*} is evaluated on previously unseen data to estimate its ability to generalize beyond the examples observed during training.

2.2 Privacy

Privacy refers to the ability of individuals or groups to control how information about them is collected, used, and shared [42], to seclude themselves or their information, thereby selectively expressing themselves. Researchers address privacy issues along two major directions: *data release*, *model release*, and *statistical release*. To this end, a fundamental distinction must be drawn between them, as each release poses distinct privacy risks and therefore requires different protection methods. Specifically:

- **Data release:** The objective is to release or publish a dataset, or a transformed version of it, so that external parties, such as researchers or analysts, can analyze it directly with the released records without compromising the privacy of the individuals present in the dataset. The central challenge lies in preventing the disclosure of sensitive information via explicit identifiers, quasi-identifiers, information

that can be exploited for linkage attacks, and in avoiding the de-identification or membership inference of individuals whose data appears in the dataset.

- **Model release:** The objective is to protect the privacy of the training data used to train an ML model while still enabling its use. Instead of exposing raw records, one releases the model parameters or provides query access to its predictions. However, even without direct access to the underlying data, privacy risks persist: information about training samples may be implicitly encoded in the model's parameters or revealed through its outputs, enabling privacy attacks.

- **Statistical Release:** The objective is to publish aggregate statistics in a way that maintains their analytical utility while implementing safeguards that prevent the re-identification of any individual whose data contributed to the underlying statistics.

A wide range of techniques has been developed to protect privacy, from classical anonymization methods to more advanced approaches. These methods aim to enable researchers, auditors, and practitioners to work with released datasets while limiting the privacy risk. Despite these methods, several challenges persist, mainly regarding preserving data quality and analytical utility, since stronger privacy protections often require altering, masking, or perturbing the data in ways that degrade its usefulness for practical tasks. To balance data utility with privacy, several privacy notions have been proposed. In this thesis, the most important one is *differential privacy* (DP), but we also briefly review *k-anonymity*, *l-diversity*, and *t-closeness*. These earlier notions help motivate why DP is preferred in modern ML. We present these concepts in the setting of tabular datasets, where privacy research was initially developed.

2.2.1 Anonymisation

Let D denote a dataset, and let $\{A_1, \ldots, A_n\}$ be its set of attributes. Each tuple corresponds to a single individual, but it may also represent another entity. Assume that explicit identifiers such as names, social security numbers, or phone numbers have been removed. A *quasi-identifier* $\{A_i, \ldots, A_j\}$ is a set of attributes that can be linked with external information to uniquely identify at least one individual. The privacy notions below, prior to DP, are based on controlling the disclosure of such quasi-identifiers.

k-Anonymity relies on the notion of equivalence classes formed by records sharing identical values across a set of quasi-identifiers. A dataset is k-anonymous if every value combination of every quasi-identifier appears at least k times. In practice, k-anonymity

is enforced through *generalization* or *suppression*. Generalization replaces specific values with coarser categories, such as age ranges or truncated ZIP codes, whereas suppression removes tuples that would otherwise require excessive generalization.

Definition 2.2.1 (k-Anonymity)**.** A release of data is k-anonymous if every combination of values of quasi-identifiers can be indistinctly matched to at least k individuals, where k is the size of the smallest class.

l-Diversity l-diversity defines privacy in terms of an *equivalence class*, which is a group of tuples whose non-sensitive attributes are generalized to the same values, making them indistinguishable with respect to those attributes. "Well-represented" means that at least l distinct sensitive values appear in each block.

Definition 2.2.2 (l-Diversity)**.** A block of tuples is l-diverse if it contains at least l "well-represented" values for the sensitive attribute S. A dataset is l-diverse if all blocks are l-diverse.

t-Closeness Let Q denote the distribution of a sensitive attribute in the full database, and let P denote its distribution within an equivalence class. *t-closeness* is proposed to ensure that these two distributions remain similar.

Definition 2.2.3 (t-Closeness)**.** An equivalence class satisfies t-closeness if the distance between P and Q is bounded by a threshold t. A dataset satisfies t-closeness if all equivalence classes satisfy t-closeness.

Despite its simplicity and practical relevance, k-anonymity has important limitations. It is vulnerable to *homogeneity attacks*, where all records in an equivalence class share the same sensitive value, and to *background knowledge attacks*, where external information enables the inference of sensitive attributes. l-diversity addresses homogeneity attacks by requiring diversity of sensitive values within each group, but it remains insufficient against background knowledge, skewness, and similarity attacks. It is also difficult to enforce in high-dimensional datasets or when multiple sensitive attributes are present. These limitations motivated *t-closeness*, which requires the distribution of sensitive values within each group to remain close to the global distribution. However, it can be difficult to satisfy with multiple sensitive attributes, and its effectiveness depends on the chosen distance metric, which may not capture all privacy risks. Such limitations motivated the need for Differential Privacy (DP).

2.2.2 Differential Privacy

Differential Privacy (DP) [22] provides a formal mathematical framework for releasing statistical information about a dataset while controlling the disclosure risk to any

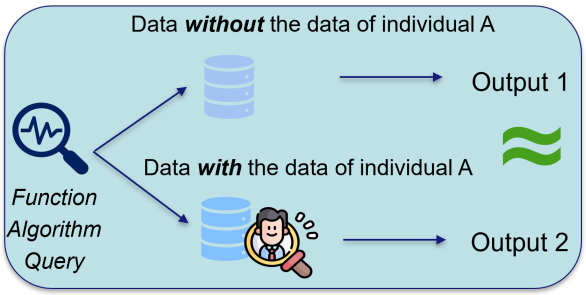


Figure 2.1. Differential Privacy

individual. Its goal is to allow meaningful insights to be shared without revealing sensitive details about specific records. To achieve this, DP introduces noise calibrated to the sensitivity of the underlying computation (function, algorithm, or query), ensuring that the resulting statistic remains useful while providing a provable bound on what an adversary can infer about any individual. Instead of requiring zero disclosure, DP limits the additional privacy risk an individual incurs from inclusion in a dataset (as shown in Fig. 2.1), for example, in statistical releasing.

To formally define DP, let $\mathscr{A} : \mathscr{D} \to \mathscr{R}$ be a randomized algorithm that maps datasets from domain $\mathscr{D}$ to outputs in range $\mathscr{R}$. Two datasets $D, D' \in \mathscr{D}$ are called *adjacent* if they differ in exactly one data point.

Definition 2.2.4 (ϵ-DP)**.** A randomised function (algorithm or query) $\mathscr{A} : \mathscr{D} \to \mathscr{R}$ with domain $\mathscr{D}$ and range $\mathscr{R}$ satisfies ϵ-DP if for any two adjacent datasets $D, D' \in \mathscr{D}$ and for any set of outcomes $S \subseteq \mathscr{R}$,

$$\Pr[\mathscr{A}(D) \in S] \leq e^{\epsilon} \Pr[\mathscr{A}(D') \in S]. \tag{2.2}$$

This notion is also known as *pure DP*. In practice, it is often too restrictive for real-world applications, and a relaxed form, called *approximate DP*, is commonly used.

Definition 2.2.5 ((ϵ, δ)-DP)**.** A randomised function (algorithm or query) $\mathscr{A} : \mathscr{D} \to \mathscr{R}$ with domain $\mathscr{D}$ and range $\mathscr{R}$ satisfies (ϵ, δ)-DP if for any two adjacent datasets $D, D' \in \mathscr{D}$ and for any set of outcomes $S \subseteq \mathscr{R}$,

$$\Pr[\mathscr{A}(D) \in S] \leq e^{\epsilon} \Pr[\mathscr{A}(D') \in S] + \delta. \tag{2.3}$$

To guarantee the DP property, in a *data release* setting, one does not usually publish the raw database directly. Instead, one releases the output of a query or a statistic computed on the dataset after adding carefully calibrated random noise sampled from

either a Laplace or a Gaussian distribution. The amount of noise depends on how much the query output can change when a single individual's data is removed. This is captured by the notion of *sensitivity*. Intuitively, sensitivity measures the maximum influence an individual can have on a query's output. Once this quantity is known, noise can be added to mask that contribution and guarantee DP.

Definition 2.2.6 (Sensitivity)**.** Let $f : \mathscr{D} \to \mathbb{R}^k$ be a query function. The global ℓ_1-sensitivity of f is $\Delta_1(f) = \max_{D\sim D'} \|f(D)-f(D')\|_1$, and the global ℓ_2-sensitivity of f is $\Delta_2(f) = \max_{D\sim D'} \|f(D)-f(D')\|_2$ where the maximum is taken over all adjacent datasets D and D'.

Laplace mechanism. Given a query $f : \mathscr{D} \to \mathbb{R}^k$, one releases

$$\mathscr{A}(D) = f(D) + (Y_1, \ldots, Y_k), \tag{2.4}$$

where the random variables Y_i are drawn independently from a Laplace distribution with mean 0 and scale

$$b = \frac{\Delta_1(f)}{\epsilon}. \tag{2.5}$$

$$Y_i \sim \mathrm{Lap}(0, b), \qquad \text{with density} \quad p(x|b) = \frac{1}{2b}\exp\left(-\frac{|x|}{b}\right). \tag{2.6}$$

Gaussian mechanism. The Gaussian noise mechanism achieving (ϵ, δ)-DP, for a function $f : \mathscr{D} \to \mathbb{R}^m$, is defined as

$$\mathscr{A}(D) = f(D) + \mathcal{N}(0, \sigma^2 I_m), \tag{2.7}$$

where $\sigma \geq \frac{\sqrt{2\ln(1.25/\delta)}C}{\epsilon}$ and $C = \max_{D\sim D'} \|f(D)-f(D')\|_2$ is the ℓ_2-sensitivity of f. The DP concept forms the foundation for Chapters 3, 4, and 5.

2.2.3 Attacks on Machine Learning Models

Model release introduces privacy risks that differ from those of direct data release. Even if a model is not explicitly disclosing records, it may still leak information about its training data through its parameters or outputs. This creates a new attack surface: adversaries can exploit access to a model to infer properties of the training set, recover functionality, or determine whether specific individual data was used during model training. In the following, we present Model Extraction Attacks (MEA) [43] and Membership Inference Attacks (MIA) [44].

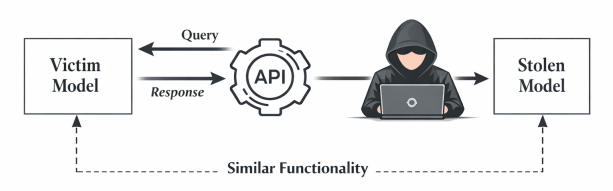


Figure 2.2. Model Extraction Attack

Model extraction attacks (MEA)

In a model extraction attack (MEA), a malicious entity, or an adversary, seeks to reconstruct a deployed model by training a surrogate model that, either exactly or approximately, replicates the behavior of a released/deployed target model by issuing queries and observing the corresponding outputs. In such a setting, an adversary typically has only black-box access to MLaaS prediction Application Programming Interface (API), yet attempts to infer the target's decision function through systematic querying. Such attacks raise both security and privacy concerns: they can facilitate intellectual-property theft, enable the circumvention of usage or access controls, and serve as precursors to other privacy attacks that exploit the extracted model to infer sensitive information about the training data, and to perform other privacy attacks.

As presented in 2.2, to perform MEA, an attacker or adversary queries the deployed model (the victim), collects responses of input-output pairs, and then trains a surrogate as the threat model (stolen model) on these samples. In the simplest setting, the adversary samples inputs from the data distribution, queries the target model, and uses the returned labels or confidence scores. More advanced methods adaptively select queries, for example, by focusing on uncertain regions near decision boundaries or by synthesizing inputs that reveal more information about the model. If confidence scores or logits are available, extraction becomes easier because these outputs provide more information than labels alone. The extracted threat need not exactly match the original parameters, it is often sufficient that it reproduces the target model's behavior with high fidelity. High fidelity refers to how closely a threat model replicates the target model's behavior, even when its internal parameters or structure differ. In other words, a high-fidelity extracted model produces outputs that are nearly indistinguishable from those of the original model across inputs.
The MEA concept forms the foundation for Chapter 3.

Membership inference attacks (MIA)

In a *membership inference attack* (MIA), an adversary aims to determine whether a given data point was part of a model's training set. MIA exposes several privacy vulner-

abilities, as it can reveal sensitive attributes indirectly, for instance, when membership implies belonging to a group with a particular condition or label, and they undermine the confidentiality of training data by disclosing who is represented in the dataset [45, 46]. MIAs also weaken anonymization, since linking an individual to the training set can effectively de-identify them even when explicit identifiers have been removed.

To perform MIA, the adversary exploits the fact that models often behave differently on training (members) and unseen samples (non-members), for example, by assigning higher confidence or lower loss to training samples. Two main categories of methods for MIA exist: *threshold-based attack* [47] and *shadow-model-based attacks* [44, 48]. Formally, let M be a learning algorithm, let S be a training set, and let $f_S = M(S)$ denote the trained model. Given a candidate example $z = (x, y)$ and black-box or white-box access to f_S, the adversary outputs a bit indicating whether it believes $z \in S$. Thus, an MIA is a hypothesis test on membership in the training set.

Definition 2.2.7 (Membership Inference Attack)**.** Let S be a training dataset, let f_S be a model trained on S, and let $z = (x, y)$ be a candidate sample. A membership inference attacker is a function

$$\mathscr{M}(f_S, z) \in \{0, 1\},$$

where output 1 indicates the prediction that z was used in training, and output 0 indicates the prediction that it was not.

Definition 2.2.8 (Threshold-Based MIA)**.** Let $s(f_S, z) \in \mathbb{R}$ be a score assigned to sample z by model f_S, such as prediction confidence or negative loss. A threshold-based MIA assumes that training samples tend to have higher confidence or lower loss, so a threshold on that score can differentiate members from non-members and allows predicting that a point z is a member of the training set if for some threshold $\tau \in \mathbb{R}$ computed by the attacker. Equivalently,

$$\mathscr{M}_\tau(f_S, z) = \begin{cases} 1 & \text{if } s(f_S, z) \geq \tau, \\ 0 & \text{otherwise.} \end{cases}$$

Definition 2.2.9 (Shadow-Based MIA)**.** Let f_S be a target model. A shadow-based MIA trains one or more *shadow models* on auxiliary datasets sampled from a distribution intended to approximate that of the target training data. Using the outputs of these shadow models on member and non-member samples, the attacker constructs an attack model $\mathscr{A} : \mathscr{O} \to \{0, 1\}$ where $\mathscr{O}$ denotes a representation of the target model's output (e.g., posterior probabilities, logits, or loss values). The final membership decision for a candidate sample z is then:

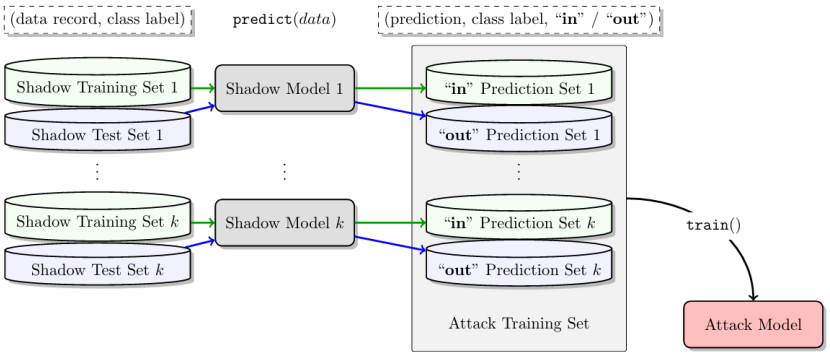


Figure 2.3. Shadow-based Membership Inference Attack

$$\mathcal{M}(f_S, z) = \mathcal{A}(\phi(f_S, z)), \tag{2.8}$$

where $\phi(f_S, z)$ is the feature representation extracted from the target model's response on z.

Specifically, shadow-based attacks are more flexible than thresholding because they learn, from auxiliary data, how members and non-members differ in the target model's output space. In practice, they often use confidence vectors or losses as attack features and a binary classifier as the attack model (summarized in 2.3 and as proposed by Shokri et al. [44]). The attacker first trains several shadow models to imitate the target model's behavior. For each shadow model, the attacker knows which samples were used for training and which were not. By querying the shadow models on both member and non-member samples, the attacker collects labeled model outputs along with their true membership status. These outputs, often given as posterior probability vectors, are then used to train a binary *attack model* that learns to distinguish between the output patterns of members and non-members. Once the attack model has been trained, the attacker queries the target model on a candidate sample z, extracts a feature representation $\phi(f_S, z)$ from the target output, and feeds it to the attack model. If the output pattern resembles those observed for training points of the shadow models, the sample is predicted to be a member, otherwise, it is predicted to be a non-member. Intuitively, the attack succeeds because ML models often behave differently on data they have seen during training, for example, by assigning higher confidence to such samples.

The MIA concept forms the foundation for the methods developed in Chapter 4.

2.2.4 Differentially Private Machine Learning

A standard way to protect ML models against training-data leakage is to make the training algorithm itself differentially private, rather than ensuring a DP guarantee at the data level [19]. A widely used algorithm is *Differentially Private Stochastic Gradient Descent* (DP-SGD) (Algorithm 1), introduced by Abadi et al. [19]. DP-SGD modifies standard mini-batch SGD so that the contribution of each training example is bounded with respect to the model output, and then adds random noise to obscure its effect on the weight updates.

More precisely, at each iteration, gradients are computed per example in the sampled mini-batch. Each gradient is then *clipped* to have ℓ_2 norm at most a fixed threshold C, which limits the sensitivity of the update to any single example. The clipped gradients are averaged, and Gaussian noise with variance proportional to $C^2\sigma^2$ is added before updating the parameters. The privacy loss across iterations is tracked using a privacy accountant, yielding a final (ϵ, δ)-DP guarantee. The main hyperparameters that control privacy and utility are the clipping norm C, the noise multiplier σ, the batch size, and the number of training steps. Larger noise and smaller effective sampling rates improve privacy, but usually reduce model utility. As a result, DP-SGD provides formal privacy guarantees against a wide range of attacks, although strong privacy typically comes at the cost of lower model utility. The pseudo-code of DP-SGD is presented in 1.
The DP-SGD concept serves as the foundation for the methods developed in Chapters 3 and 4.

2.3 Explainable Artificial Intelligence

Explainable Artificial Intelligence (xAI) refers to methods and techniques that make the behavior, predictions, and decision-making processes of AI systems more understandable to humans. To enhance transparency, xAI techniques provide tools to open complex black boxes and shed light on how AI decisions are made [49, 50], thereby promoting transparency and accountability in real-world settings [51]. By addressing essential *How?, What?, and Why?* questions about AI models, xAI serves as a valuable tool for stakeholders, users, researchers, and model developers, helping address the growing ethical and legal issues associated with them and ensuring the trustworthiness and responsible use of models.

Understanding an AI system with xAI relies on its training data, process, and model. Therefore, xAI can be applied throughout the entire AI development pipeline. Specifically, it can be applied at different stages of modeling, including before (pre-modeling explainability), during (explainable modeling), and after (post-modeling explainabil-

Algorithm 1 Differentially Private SGD

Input: Training examples $\{x_1, \ldots, x_N\}$, loss function $\mathscr{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\mathscr{L}(\theta, x_i)$
Parameters: learning rate η_t, noise multiplier σ, expected batch size L, clipping norm C
Initialize θ_0 randomly
for $t \in [T]$ **do**
    Sample a mini-batch L_t by including each example independently with probability L/N
    Compute per-example gradients
    For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathscr{L}(\theta_t, x_i)$
    Clip gradients
    $\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) \,/\, \max\Big(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\Big)$
    Add noise and average
    $\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\Big(\sum_{i\in L_t}\bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\Big)$
    Update parameters
    $\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$
end for
Output: θ_T and the overall privacy cost (ϵ, δ) computed using a privacy accountant

ity). In this thesis, the primary emphasis will be on post-modeling (post-hoc), since AI models are often developed with only predictive performance in mind.

2.3.1 Post-hoc Explainability

Post-hoc explainability is a technique for gaining insight into the decision-making process of a trained ML model. In this context, post-hoc means that the model's interpretability is addressed after its training, regardless of its complexity or the algorithms used. The approach primarily involves querying the model with diverse input datasets to observe its behavior across scenarios. Through these interactions, we can effectively map the model's decision boundaries, thereby elucidating which factors influence its predictions. Visualizations and explanations can then be applied to make these insights more accessible and human-friendly, ultimately enabling a better understanding of the model's predictions. These visual aids are essential for making the insights gained more accessible to data scientists, end users, and domain experts who wish to understand why the model makes specific predictions. By going through this process, post-hoc explainability serves a vital role in improving model transparency and building trust in its performance.

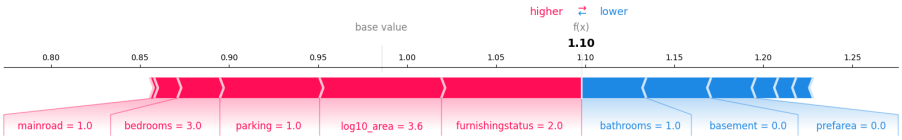


Figure 2.4. A SHAP-based explanation illustrating the contribution of each feature to this specific prediction.

2.3.2 Types of Explanations

Feature Attribution

Feature Attribution (FA) explanations assign a quantitative score to each input feature in a given model. The primary goal of calculating FA is to identify which features have influential effects on the model's predictions and which have a relatively smaller impact. These importance scores help practitioners and data scientists gain insights into which factors are most critical in influencing those decisions. Features that, when modified, cause more substantial shifts in the model's output are considered more important because they have a greater influence on the final prediction.

For deep neural network (DNN) models, many feature-based explanation functions are gradient-based techniques that analyze the flow of gradients through the model. Approaches such as Layer-wise Relevance Propagation (LRP) [52], Integrated gradients (IG) [53], and DeepLIFT (Deep Learning Important FeaTures) [54] exist. The key idea behind LRP is to back-propagate relevance through the layers of a DNN, assigning importance scores to neurons by distributing the output of the final layer backward through the network. IG is another attribution method that computes the integral of gradients along a path from a baseline input to the actual input. DeepLIFT is another FA method that compares each neuron's activations to a reference baseline state.

For model-agnostic FA frameworks, perturbation methods include Shapley explanations (SHAP) [55], local Interpretable Model-agnostic Explanations (LIME) [28], and many others [56, 57]. SHAP is based on cooperative game theory and marginal contribution calculations. LIME is another method that generates a dataset of perturbed samples and their corresponding black-box model predictions, then trains an inherently interpretable model to learn a good local approximation of the black-box model's predictions. Figure 2.4 shows the SHAP force plot for a single prediction, showing how individual input features push the model output toward the final prediction.
The concept of FA serves as the foundation for Chapters 5 and 7.

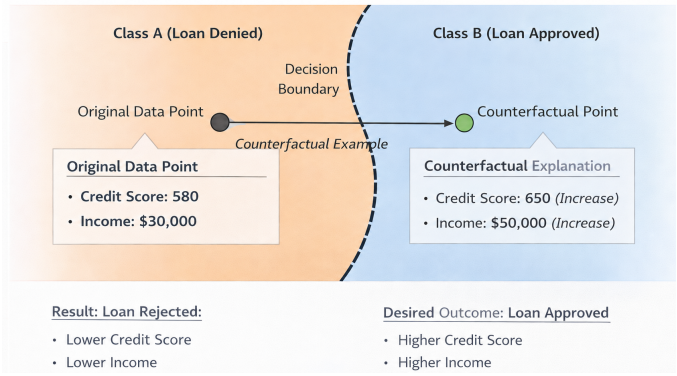


Figure 2.5. CF example involving navigating the decision boundary to find the optimal change in input for a desired outcome.

Counterfactual explanations

A CF explanation (Fig. 2.5) is a form of example-based explanation that provides a hypothetical scenario that illustrates how a different decision or outcome of an underlying ML model could have been achieved through minimal changes to the input data [58]. CFs provide users with an actionable explanation by showing how altering the input would affect the model's output [59]. Additionally, CFs allow the identification of the variables that should have differed in a given input to observe a different outcome, thus making it possible to assess the influence of specific factors [58]. CF has already been shown to improve the interpretability of ML models across several domains, such as healthcare and finance [57, 60, 61]. For instance, Figure 2.5 presents an example from a loan prediction model, where a CF indicates which features and by how much should be changed to alter the prediction from loan denied to loan approved.

The generation of CFs can be viewed as an optimization problem that minimizes a cost function encoding desirable CF explanation properties, such as actionability [62, 63]. The cost function may be model-dependent, i.e., it depends on the model's inner workings, such as gradients for differentiable models [64, 65] or tree structures for tree-based models [66, 67], or model-agnostic [68]. Numerous works focus on the CF generation problem investigating various methodologies such as mixed integer programming [69], heuristic searches [70, 71], gradient-based approaches [72, 73], and advanced meta-heuristics such as genetic algorithms [74, 75], Generative Adversarial Networks (GAN) [76] and Reinforcement Learning (RL) [77, 78].

GAN-based Counterfactual Explanations

We now describe GAN-based CFs to form the basis of chapter 3. The method aims to generate CFs for real-time recourse and interpretability using Residual GANs (CounteRGAN) as a CF generator [79]. CounteRGAN is a novel residual GAN that trains the generator to produce residuals that are intuitive to the notion of perturbations used in CF searches. The search process seeks to maximize the value function with respect to the discriminator D and minimize it with respect to the generator G. Where C_t is the target classifier to be explained, Reg(G, x_i) is a regularization term, and x_i are samples drawn from the entire data distribution (Eq. 2.9).

$$V_{\text{CounteRGAN-bb}}(D,G) = \frac{\sum_i C_t(x_i)\log D(x_i)}{\sum_i C_t(x_i)} + \frac{1}{N}\sum_i \log(1-D(x_i+G(x_i))) + \text{Reg}(G,\{x_i\}), \tag{2.9}$$

CF generators aim to identify in-distribution data points within the training data by optimizing metrics of proximity, plausibility, and sparsity [80]. Specifically, *sparsity* measures how many features of the CF data point are different from the original data point, while *proximity* measures the overall distance between the CF and the original data point, such as Euclidean and Gower distance, where a low proximity value indicates that the CF is similar to the original point. This can be useful for providing actionable feedback, as it suggests that the model recommends only a few changes likely to have a significant impact on the output. *Plausibility* ensures that the generated CFs are realistic and fit in the training data distribution.
The concept of CF serves as the foundation for the methods developed in Chapters 3, 4, and 8.

2.4 Fairness

In the context of ML, fairness is the formal study and application of techniques to ensure that algorithmic decisions do not result in unjust or prejudicial treatment of individuals based on protected characteristics (such as race, gender or age). Fairness is broadly partitioned into two distinct categories: Individual Fairness and Group Fairness. These categories correspond to the socio-legal concepts of Equality and Equity, respectively.

2.4.1 Individual Fairness (Equality of Treatment)

Individual fairness is rooted in the principle of **Equality**, specifically procedural justice. It posits that any two individuals who are similar with respect to a specific task should receive similar algorithmic outcomes.

Formally, let $\mathscr{X}$ be the set of individuals and $M : \mathscr{X} \to \Delta(\mathscr{Y})$ be a learning algorithm mapping individuals to a distribution over outcomes. Given a distance metric $d : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ that measures the similarity between individuals, the algorithm satisfies individual fairness if, for any two individuals $x, y \in \mathscr{X}$:

$$D(M(x), M(y)) \leq d(x, y) \tag{2.10}$$

where D is a distance metric between distributions (e.g., Total Variation distance).

2.4.2 Group Fairness (Equity of Outcome)

Group fairness is grounded in the principle of **Equity** and distributive justice. It focuses on parity in outcomes across protected demographic groups (e.g., race, gender, or age), aiming to eliminate the correlation between a protected attribute and the model's predictions.

In its most common form, known as **Statistical Parity**, group fairness is achieved if the prediction $\hat{Y}$ is independent of the protected attribute A. For a binary classifier and a binary protected attribute $A \in \{0, 1\}$, the condition is:

$$P(\hat{Y} = 1 \mid A = 0) = P(\hat{Y} = 1 \mid A = 1) \tag{2.11}$$

2.4.3 Fairness Metrics

The literature identifies numerous metrics to quantify the group fairness of a model, primarily focusing on performance comparisons across protected attributes such as, e.g., gender, ethnicity, or physical features [81]. In this section, we introduce four fairness metrics that broadly capture key aspects of fairness evaluation: True Positive Parity (TPP), False Positive Parity (FPP), Positive Predictive Value (PPV), and Negative Predictive Value (NPV). They provide a comprehensive view of model performance across protected categories, highlighting error distributions.

Let $a, b, c \in A$ represent different protected attribute values (e.g. ethnicity white, black, asian, etc.), and let TP_a, TN_a, FP_a, FN_a be each attribute's respective true/false positives/negatives.

Equality of Odds. Equality of Odds metric requires that *True Positive Parity* (TPP) and *False Positive Parity* (FPP) rates be consistent across groups [82]. TPP (eq. 2.12), also known as Equality of Opportunity, ensures that individuals in the positive class have the same probability of being correctly classified across groups. On the contrary, FPP (eq. 2.13) checks that the rate of false positives is not excessive for certain groups. In other words, skewed TPP means that certain groups are recognized more accurately than others, whereas skewed FPP means that some groups are more often mislabelled as positive.

$$TPP_a = \frac{TP_a}{TP_a + FN_a} \quad (2.12) \qquad FPP_a = \frac{FP_a}{FP_a + TN_a} \quad (2.13)$$

Predictive Value Parity. Predictive Value Parity checks whether the fraction of correct predictions among all positive or negative predictions is equal across groups [83]. *Positive Predictive Value* (PPV) measures how reliably a positive prediction corresponds to a true positive (eq. 2.14), while *Negative Predictive Value* (NPV) measures how reliably a negative prediction corresponds to a true negative (eq. 2.15). A skew in PPV implies that positive predictions are more accurate for some groups, while a skew in NPV indicates that negative predictions are more reliable for specific groups.

$$PPV_a = \frac{TP_a}{TP_a + FP_a} \quad (2.14) \qquad NPV_a = \frac{TN_a}{TN_a + FN_a} \quad (2.15)$$

Any imbalance in parity, the ratio of lowest to highest value (eq. 2.16), across any of these metrics signals bias in prediction. In our work, we aim to bring all fairness metrics as close as possible to 1.0 while maintaining accuracy close to the original.

$$metric\_parity = \frac{\min(metric_a, metric_b, metric_c, \dots)}{\max(metric_a, metric_b, metric_c, \dots)} \quad (2.16)$$

2.4.4 Model De-Biasing Techniques

Researchers primarily employ methods for model de-biasing at three phases in a machine learning pipeline, namely, *(1) pre-processing data*, *(2) in-processing model training*, and *(3) post-processing model's outputs*. Each method has inherent trade-offs regarding effectiveness, complexity and computational resources.

Pre-Processing Techniques

Pre-processing methods focus on refining input data to reduce unfair patterns prior to model training. Traditional approaches include:

- **Reweighting and Resampling** [84]: Assigning different importance weights or sampling rates to protected and non-protected groups to ensure balanced representation.
- **Latent Feature Transformation** [85]: Transforming the feature space to obfuscate sensitive attributes whilst preserving predictive signals.
- **FairTest** [86]: introduces a framework for detecting unfair or discriminatory treatment in data-driven systems by broadening group fairness definitions, providing guidelines for applying fairness principles at the stage of model and dataset preparation.

The main drawbacks of pre-processing techniques are that they may inadvertently distort key characteristics or oversimplify complex correlations and fail to fully eliminate bias [87].

In-Processing Techniques

In-processing involves integrating fairness constraints or penalties into the model training process (e.g., by adding them to the model's loss function). This approach explicitly requires the model to account for fairness requirements during optimisation. This may lead to a reduction in overall accuracy when fairness and performance objectives conflict [88]. Moreover, their application includes a relatively high computational overhead during training, extending the model's training time. These methods often employ:

- **Fairness-Constrained Optimisation** [89, 90]: Adding fairness constraints (e.g., demographic parity) to the loss function.
- **Adversarial Fairness** [91, 92]: Training a predictor simultaneously attempting to confuse an adversary predicting sensitive attributes.

Post-Processing Techniques

Post-processing alters model predictions after training, offering a relatively simpler to enforce fairness with respect to the above-mentioned techniques. Example techniques include:

- **Group-Specific Thresholding** [82, 93]: Adjusting decision thresholds per demographic group to achieve targets like equalised odds. Alternatively, we can modify the threshold globally when we do not want to provide a protected attribute during inference.
- **Reject-Option Classifiers** [84]: Overriding uncertain predictions based on sensitive attributes.

The concept of fairness serves as the foundation for the methods developed in chapters 7 and 8.

2.5 Anomaly Detection

This section introduces the concept of AD and the theoretical background relevant to the Anomaly Detection (AD) algorithms considered in this work, namely Local Outlier Factor (LOF) and Isolation Forest (iForest).

Anomalies refers to data points that deviate significantly from the expected patterns or behaviors observed within a dataset. This deviation can indicate various underlying

factors, ranging from irregularities to errors in data collection or processing. AD refers to the process of identifying these anomalies [94]. In many applications, practitioners are particularly interested in finding data points that deviate from their immediate neighbors locally (local anomalies) or globally, referring to data points that deviate significantly from the overall distribution of the data set rather than just its immediate neighbors (global anomalies). To address both local and global anomalies, our study incorporates LOF [95] (for local AD), and iForest [96] (for global AD), which are two unsupervised models that proved scalable and efficient [97].

Isolation forest

The core principle behind iForest [96] lies in constructing decision trees using randomly sampled data to isolate outliers. Given the intrinsic sparsity of anomalies relative to normal data points, their isolation pathways within the iForest exhibit demonstrably shorter lengths, therefore, anomalies are isolated faster within the constructed trees. In other words, the fewer branches that need to be traversed in the tree to isolate a data point (indicating a shorter path), the higher the likelihood that it is an anomaly. To build a tree, start by randomly selecting a feature and a split value between its minimum and maximum. It then partitions the data into two sets based on the chosen split value. This process is recursively repeated to build the decision tree structure. A predefined maximum tree depth can be set to ensure consistency and avoid excessively deep trees. Afterward, an anomaly score is computed as the average path length of a data point across all trees in the forest. Lower scores indicate a higher likelihood of being an anomaly, since deeper paths, indicating greater effort to isolate, suggest greater normality, whereas shallower paths, signifying easier isolation, point toward potential anomalies. Since iForest focuses on random partitioning, faster anomaly isolation, and independence from local density distributions, it is a strong choice for detecting global anomalies that deviate significantly from the overall data distribution.

Local outlier factor

LOF [95] algorithm identifies anomalies by comparing the local density of a data point with the density of its k-nearest neighbors. Local density comparison assesses how much an individual point deviates from its surrounding environment. LOF first identifies the k-nearest neighbors for each data point. Then, it computes a local reachability density (LRD) that estimates the density of the area around a given data point relative to its neighbors. Then, LOF calculates a score for each data point by comparing its LRD to the average LRD of its k-nearest neighbors. Points with an LRD significantly lower than that of their neighbors are considered potential outliers, and higher LOF scores

indicate higher local density and greater normality. In contrast, lower scores indicate potential anomalies that deviate significantly from their surrounding data points. LOF focuses on identifying local anomalies and excels at this task by considering local density measures. This approach makes LOF well-suited for identifying local anomalies that deviate from the norm within their local context.
The concept of AD serves as the foundation for the methods developed in Chapter 5.

2.6 Data and Model Distillation

2.6.1 Data Distillation and Active Learning

Modern ML methods, and especially deep learning models, typically require large amounts of labeled training data to achieve strong generalization performance. However, collecting, storing, and processing such datasets can be costly in terms of annotation effort, memory usage, computational resources, and energy consumption. These challenges have motivated growing interest in methods that reduce the effective size of the training set while preserving as much useful information as possible. One important line of research in this direction is *data distillation*, as the process of compressing the knowledge in a large dataset into a much smaller set of informative samples while maintaining the ability to train a model that performs competitively on the original task. Data distillation is closely related to several neighboring concepts, such as *Active*

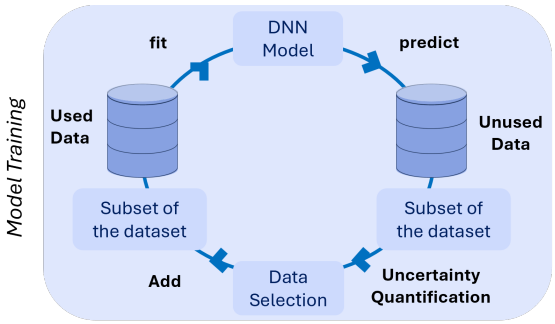


Figure 2.6. Active Learning

learning (AL), as both aim to identify highly informative data points. In this sense, the two paradigms are complementary: one seeks to reduce the amount of data needed for training, and the other seeks to reduce the amount of labeling needed to obtain that data. AL addresses this issue by selecting the most informative unlabeled instances for annotation, aiming to achieve high model performance with as few labeled examples as possible. AL is a learning paradigm in which the model can interactively query an

oracle, typically a human annotator, to obtain labels for carefully selected samples. Instead of labeling data uniformly at random, the learner aims to identify those instances that are expected to provide the greatest improvement to the model if labeled. The central intuition is that some samples are substantially more informative than others, particularly when they represent underexplored regions of the input space or reveal model uncertainty.

The standard AL process is iterative (Fig. 2.6). Initially, a small labeled set is used to train a preliminary model, while a larger pool of unlabeled data remains available. At each round, a query strategy is applied to rank or select unlabeled instances based on an informativeness criterion. The selected instances are then sent to the oracle for annotation, added to the labeled set, and the model is retrained or updated. This cycle continues until the labeling budget is exhausted or the target performance level is reached. A variety of query strategies have been proposed in the AL literature. Among the most widely used are *uncertainty-based methods* which select samples for which the current model is least confident. Common uncertainty criteria include least confidence, margin sampling, and entropy-based selection.

2.6.2 Knowledge Distillation

Knowledge distillation (also known as model distillation), is a process that involves the transfer of knowledge from a high-complex model, known as a *teacher model*, to a reduced-complexity model, known as a *student model*, that remains operable within real-world constraints [98]. While KD constitutes a specific instance of model compression, it serves as a means to extract essential insights, patterns, and expertise embedded in the larger models to create a deployable model. The student model follows a training procedure with the primary objective of imitating the performance of the teacher model. This knowledge transfer mechanism involves the student model learning not only the surface-level predictions made by the teacher model but also the deeper patterns, generalizations, and decision-making strategies embedded within the teacher architecture. To achieve this, the student model is guided during training by incorporating two primary sources of information: the actual target labels or predictions for the dataset at hand, and the soft labels or probability distributions generated by the teacher model in response to the same dataset.

KD necessitates the availability of a well-trained teacher model, a trainable student model, and the specification of a student loss function for assessing predictions against ground-truth labels (L). A distillation loss function, accompanied by a temperature parameter ($temp$), is employed to bridge the knowledge gap by comparing the soft predictions of the student to the softened teacher labels. The student and the distillation losses are weighted via an alpha (α) which is essential for balancing task-specific

accuracy and knowledge transfer. Subsequently, the losses incurred are computed, incorporating a weighted combination of the student loss (weighted by α) and the distillation loss (weighted by $1-\alpha$). This weighting mechanism allows for a fine-tuned balance between preserving task-specific accuracy (student loss) and incorporating the knowledge distilled from the teacher model (distillation loss). The Kullback-Leibler (KL) divergence, denoted as $KL(P \parallel Q)$ (Eq. 2.17), is a mathematical measure employed in KD to assess the difference between two probability distributions, P and Q, where the summation is taken over all classes or categories.

Suppose P represents the soft probabilistic predictions of a teacher model, while Q represents the corresponding probabilistic predictions of a student model (expressed as the output made by a softmax activation function). During training, KL divergence as the distillation loss encourages the student model to capture the nuances and uncertainties present in the teacher's predictions. In summary, the loss to be optimized by the student is shown by Eq. 2.18.

$$distillation\_loss = KL(P \parallel Q) = \sum_i P(i) \log\left(\frac{P(i)}{Q(i)}\right) \tag{2.17}$$

$$loss = \alpha \cdot student\_loss + (1-\alpha) \cdot distillation\_loss \tag{2.18}$$

The KD serves as the foundation for the methods developed in Chapter 3.

2.7 Reinforcement Learning

Reinforcement Learning (RL) [99] draws inspiration from behavioral psychology, in which an RL agent interacts with an environment to learn a policy that supports a specific goal. Specifically, RL is a paradigm in ML where an agent learns to make decisions by interacting with an environment and receiving feedback in the form of rewards or penalties. Rather than being given the correct labeled action, the agent learns effective behavior through trial and error, gradually improving its policy (i.e., its strategy) to maximize cumulative reward over time. Typically, this environment is represented as a Markov Decision Process. The state space, referred to by S, and the action space, referred to by A, are defined, and a policy $\pi : S \rightarrow A$ that maps states to actions determines the behavior of the learning agent. The policy $\pi(a_t = a|s_t = s)$ is the probability of performing a certain action $a \in A$ when the agent is in a state $s \in S$ at time step t. The objective of the RL agent is to maximize the expected sum of rewards R_t, known as the return, over a specific timeline T (Eq. 2.19). The return is computed from the rewards obtained at each time step, denoted by r_k, using a discount factor $\gamma \in [0, 1]$. The discount factor determines the significance of future rewards versus immediate rewards.

$$R_t = \sum_{k=t+1}^{T} \gamma^{k-t-1} r_k \tag{2.19}$$

The action-state value function $Q_\pi(s,a)$, also known as Q value of policy π, is defined as the expected return when starting from state s and taking action a, following policy π, as shown in 2.22.

$$Q_\pi(s,a) = \mathbb{E}_\pi(R_t | s_t = s, a_t = a) \tag{2.20}$$

Actor-Critic RL Actor Critic (AC) is an RL architecture that is trained by optimizing both a policy (actor) and an estimate of the value function (critic) simultaneously [100]. The actor takes actions according to the current policy, while the critic evaluates the value of the actor's actions and provides feedback on their quality. The policy is updated by the actor based on the feedback received from the critic. Compared with other RL algorithms, AC better handles scalability issues and slow learning in sparse environments (where the agent receives low or infrequent feedback or reward from the environment for its actions) [101]. In fact, AC combines the strengths of value-based and policy-based RL methods, namely the ability to learn from high-dimensional input spaces and to learn stochastic policies for continuous action spaces. By exploiting both methods, AC algorithms can learn more efficiently and converge to optimal solutions faster than traditional RL methods. AC algorithms are also well-suited to tasks that involve continuous action spaces, as they can handle complex, high-dimensional action spaces more effectively [102, 103].

However, the AC algorithm also has several limitations, including sample inefficiency and policy entropy regularization. The former is the requirement for a large number of samples to accurately estimate the value function and update the policy, the latter is that the model may converge to a deterministic policy, which can result in a lack of exploration and a suboptimal solution. To tackle these issues, the Soft Actor-Critic (SAC) was introduced [104]. In an SAC architecture, the actor aims to maximize both the expected reward and the entropy of the actions taken. As entropy measures the uncertainty or unpredictability of a random variable, its maximization strikes a better balance between exploitation and exploration, thereby preventing the learning of non-optimal policies.

The SAC architecture is composed of three distinct networks: a state value function V that is parameterized by ψ, a soft Q-function Q that is parameterized by θ, and a policy function π that is parameterized by ϕ. As described in [104], the approximation of V and Q facilitates the convergence of the algorithm. The optimization is as follows:

1) Train the Value network V by minimizing the squared difference between the prediction of the value network and the expected prediction of the Q function, plus the entropy of the policy function π (measured here by the negative log), across all states sampled from the experience replay buffer (past experiences). More formally, we minimize Eq. 2.21:

$$J_V(\psi) = \mathbb{E}_{s_t \sim D}\left[\frac{1}{2}(V_\psi(s_t) - \mathbb{E}_{a_t \sim \pi_\phi}[Q_\phi(s_t, a_t) - \log \pi_\phi(a_t|s_t)])^2\right] \tag{2.21}$$

2) Train the Q network by minimizing the squared difference between the prediction of the Q function and the immediate (one-time-step) reward plus the discounted expected Value of the next state (computed with $\hat{Q}$), across all states sampled from the experience replay buffer (Eq. 2.22).

$$J_Q(\theta) = \mathbb{E}_{(s_t, a_t) \sim D}\left[\frac{1}{2}(Q_\theta(s_t, a_t) - \hat{Q}(s_t, a_t))^2\right] \tag{2.22}$$

3) Train the Policy network π by minimizing the Kullback-Leibler (KL) divergence between the policy function's distribution and the distribution obtained by normalizing the exponentiation of the Q function with another function Z (Eq. 2.23).

$$J_\pi(\phi) = \mathbb{E}_{s_t \sim D}\left[D_{KL}\left(\pi(.|s_t) \middle|\middle| \frac{exp(Q_\theta(s_t, .))}{Z_\theta(s_t)}\right)\right] \tag{2.23}$$

The RL serves as the foundation for the methods developed in Chapter 8.

Part I

Privacy and Explainability

Chapter 3

Exploiting Counterfactual Explanations for Model Extraction Attacks

In this chapter, we investigate how CFs can be exploited to perform MEA. We study a realistic MLaaS setting in which explanations are provided alongside model predictions and show how an adversary can use them to approximate the behavior of a target model. We then analyze the impact of incorporating DP into the CF generation as a mitigation strategy. Through experiments, we examine the trade-off between privacy protection, attack effectiveness, predictive performance, and explanation quality. This chapter is based on the content available in the following works:

1. **Fatima Ezzeddine**, Omran Ayoub, and Silvia Giordano. *Knowledge Distillation-Based Model Extraction using Private GAN-based Counterfactual Explanations.* Workshop on Regulatable ML at Neural Information Processing Conference (NeurIPS), 2024.

2. **Fatima Ezzeddine**, Rinad Akl, Ihab Sbeity, Silvia Giordano, Marc Langheinrich, and Omran Ayoub. *On the interplay of Explainability, Privacy and Predictive Performance with Explanation-assisted Model Extraction.* Conference on eXplainable Artificial Intelligence (xAI 2025), Late Breaking Work.

3. **Fatima Ezzeddine**, Omran Ayoub, and Silvia Giordano. *Exploiting Explanations for Model Extraction via Knowledge Distillation and Mitigation with Private Counterfactuals.* Frontiers in Artificial Intelligence Journal, 2026.

3.1 Introduction

Recent years have witnessed a growing trend in employing Deep Neural Networks (DNNs) as learning algorithms for ML-based applications. In particular, DNNs have gained substantial popularity due to their remarkable success in diverse domains [105]. However, the complexity of their training process is resource-intensive, requiring substantial data acquisition and computational power [106], thereby hindering their spread adoption and accessibility. One approach to mitigating these complexity challenges is to offer ML models via MLaaS platforms [35, 107, 108]. MLaaS platforms streamline access to complex models by hosting and training them in the cloud, allowing third-party practitioners to query pre-trained models via APIs [34, 108], while the private dataset used to train the model remains inaccessible to users.

Numerous platforms offer MLaaS, such as Google Cloud, Microsoft, IBM, and Amazon Web Services [109, 110, 111, 112, 113]. The API's input-output format for these services is publicly accessible, meaning that users know the format of the input data required by the service (i.e., the ML model) and can therefore receive decisions from the model. This opens the door for malicious users (or, more precisely, attackers) to exploit these interfaces to conduct privacy-breaching attacks, such as MEA [43].

Several mitigation strategies for these security- and privacy-breaching attacks are already employed [114, 115]. However, these strategies face new challenges as emerging transparency requirements demand explanations for ML model decisions. In response, MLaaS platforms are shifting towards providing explanations alongside the predictions made by their deployed models [110, 111, 112, 113]. The need to provide users with model explanations is linked to the need for transparent decision-making processes in data-driven systems [116]. The urge to enhance transparency through model explanations stems from a dual objective: on the one hand, to satisfy emerging regulatory and compliance requirements, and, on the other hand, to provide users with actionable insights derived from the model's decision-making process [116, 117, 118, 119, 120].

Model explanations are generally derived using xAI techniques [28, 121] may unintentionally reveal information that attackers can exploit to refine their strategies, thereby opening new avenues for existing attack methodologies and posing additional risks to current security and privacy safeguards [47, 122, 123]. In this chapter, we investigate how explanations generated by xAI frameworks, such as CFs, can be exploited to conduct MEA in MLaaS scenarios. We focus specifically on CFs because they provide unique insights. CFs reveal how minimal alterations to original data instances can lead to different model outcomes, consequently helping users interpret the dynamics underlying the model's prediction shift from the decision boundary [57]. However, due to their inherent proximity to the decision boundary and their reflection of the training

data, CFs exhibit a dual nature: while they serve as valuable interpretative tools, they can also be a valuable asset, revealing additional information that can be exploited to uncover model vulnerabilities. In this context, we first propose a novel approach based on knowledge distillation (KD) for high-fidelity MEA, leveraging CFs as a surrogate training dataset. Our method derives a threat model that closely replicates the target model's functionality. Moreover, we propose a mitigation approach that integrates differential privacy (DP) into the CF generator's training pipeline to reduce the risk of providing CF explanations. Our analysis evaluates the effectiveness of the proposed attack method relative to established baselines and assesses the impact of DP on both the quality of CF and the performance of the MEA, with the aim to address *RQ1 and RQ2*. The contribution of this chapter is summarized as follows:

- **KD-Based MEA with CFs**: We focus on leveraging CF through a novel method founded on KD to perform MEA in MLaaS with a specific focus on MEA facilitated by KD as extraction techniques and using CFs as informative proxies for the original training data. To substantiate the effectiveness of this method, we simulate an adversarial scenario in which an attacker employs the proposed KD-based method to train a high-fidelity substitute model of the target model.

- **Private Counterfactual Explanations**: Acknowledging the critical importance of preserving the privacy of training data, we introduce the integration of DP into the GAN-generated CF explanation pipeline. This enhancement is designed to generate CFs that deviate from the statistical properties of the sensitive dataset, thereby offering a layer of protection against potential privacy breaches. We also investigate the trade-off associated with integrating DP into the generated CFs, assessing its impact on both privacy and utility.

3.2 Related Works

Several works have proposed strategies for performing MEA. Tramer et al. conduct successful MEAs against different ML models, including decision trees and support vector machines, by using equation-solving, path-finding algorithms, and learning-theoretic techniques [43]. MEAs can also be conducted against deep neural networks (DNNs) using active learning with unannotated public data [124]. In the natural language processing domain, Krishna et al. investigate MEAs against BERT-based APIs and explore the use of transfer learning to improve attack effectiveness [125]. Other works have focused on querying strategies for MLaaS systems, aiming to query the target model effectively and extract accurate surrogate models [126, 127, 128]. For example, Orekondy et al. train a knockoff network using predictions for queried images and pro-

pose a reinforcement-learning-based approach that improves query-sample efficiency and attack performance [126]. Juuti et al. propose a method to generate synthetic queries and optimize training hyperparameters, and further introduce a mechanism to detect effective DNN MEAs [127]. More recent works have considered related risks and defenses in emerging settings. Pang et al. introduce watermarking mechanisms for text generated by large language models (LLMs) to address copyright concerns and track content origin [129]. Zhang et al. propose a metric to quantify MEA risk by measuring the distance between the target model and its surrogate in weight space [130]. Sun et al. explore obfuscation techniques and propose an efficient model-obfuscation method to protect against extraction attacks [131]. They analyze the distribution of consecutive API queries and raise an alarm when this distribution deviates from normal behavior. The types and amounts of internal information that can be extracted from deployed ML models, including the architecture, optimization procedure, and training data, have also been investigated [128]. In addition, Jagielski et al. study how to improve the query efficiency of MEAs through learning-based methods, with a particular focus on the fidelity of the extracted model [132].

Other works have investigated MEAs that exploit explanations provided by MLaaS systems [133, 134, 135, 136, 137]. For instance, Oksuz et al. investigate how Local Interpretable Model-agnostic Explanations (LIME) can be exploited to infer decision boundaries by sending adaptive queries that generate new data samples near those boundaries [134]. Miura et al. analyze how gradient-based explanations reveal the decision boundary of a target model by modeling a data-free MEA against gradient-based XAI methods, while also exploiting a generative model to reduce the number of required queries [138]. Yan et al. propose methodologies that perform MEA by jointly minimizing task-classification and task-explanation losses [135]. Similar to our work, several studies have investigated how counterfactual explanations (CFs) can be exploited to conduct MEAs. Aivodji et al. model an attack that leverages both the target model's predictions and CFs to directly train an attack model [133]. Wang et al. propose a strategy that reveals the decision boundary of a target model and then performs MEA by considering the CF of the CF as pairs of training samples used to directly train the extracted model [34].

Our work aligns with this line of research by exploring how CF explanations can be exploited for MEA. However, our contribution lies in modeling a novel MEA approach that more effectively leverages the insights provided by CFs compared to approaches that directly train a substitute model on extracted CFs. Moreover, in contrast to previous studies, our work assumes that the attacker has no prior knowledge of the training set, requiring the attacker to query the model with zero prior knowledge and motivating a new strategy for performing MEA [34, 133].

In terms of mitigation strategies, few works have proposed methodologies for gen-

erating explanations while preventing the disclosure of sensitive information about the decision boundary, training set, or model architecture. For instance, Patel et al. employ differential privacy (DP) algorithms to construct feature-based model explanations [139]. Yang et al. generate differentially private CFs using functional mechanisms [140]. Pentyala et al. generate recourse paths by leveraging differentially private clustering and demonstrate that constructing a graph over cluster centers can yield private recourse paths as CFs [141].

This work also exploits DP by integrating it into the CF generation process. With this approach, we eliminate the need for a separate method to generate private CFs. We specifically focus on GAN-based CF generators due to their ability to generate high-quality CFs compared to other CF generation methods. To the best of our knowledge, our work is the first to integrate DP into CF-based generative methods. We demonstrate the robustness of the proposed DP-based method against MEA and provide a detailed analysis that sheds light on the impact of CFs on the training set distribution, using well-defined metrics.

3.3 Problem Formulation and Methodology

3.3.1 KD-Based MEA

A DNN model f_θ trained on a dataset $\mathscr{D} : \mathscr{X} \in \mathbb{R}^d \rightarrow \mathscr{Y} \in \mathbb{R}$ takes a numerical input query $x \in \mathscr{X}$ and predicts the output $y \in \mathscr{Y}$. f_θ a trained model with θ be the weight matrix, b be the bias vector. The output of the DNN can be computed as $f_\theta = \theta \cdot X + b$. The output of the network y are confidence scores, which are expressed in terms of probabilities, essentially indicating how confident the ML model is about each possible class or category in its training data. The CF explanation method, which trains a CounteRGAN generator E: $f_\theta \times \mathscr{X} \rightarrow \mathscr{X}$ generates a perturbed instance $c \in \mathscr{X}$ for the input instance x such that $f_\theta(c)$ has a different output while minimizing an objective function g. Formally, searching for a CF can be framed as $argmin\ g(x,c)$ $s.t\ f_\theta(c) \neq f_\theta(x)$. Where $g : \mathscr{X} \times \mathscr{X} \rightarrow \mathscr{R}_+$ is a cost metric measuring the changes between the input x and c, and y is the desirable target that is different from the original prediction f_θ as shown in Fig. 3.1. That is, we seek to find CFs that belong to a target class y while remaining proximal to the original instance.

Figure 3.2 illustrates the steps of the proposed scenario. First, a user submits a query, containing input data representing a data record x, for which they seek a prediction. Upon receiving the query, the MLaaS API forwards it to the ML service, which computes the prediction $f_\theta(x)$ and generates a corresponding CF explanation $c = E(x)$, where $f_\theta(c)$ yields a different prediction. Finally, the service returns both

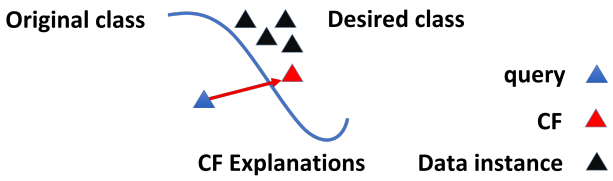


Figure 3.1. CF example involving navigating the decision boundary to find the optimal change in input for a desired outcome.

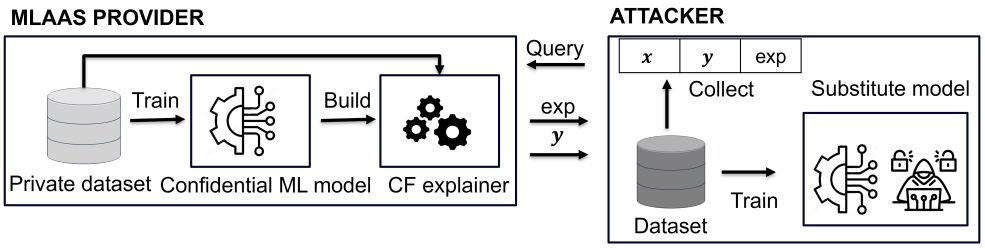


Figure 3.2. Scenario of MEA where MLaaS provides explanations alongside the prediction

the prediction and the CF explanation to the user. Within the scope of this work, our primary emphasis is on a specific type of attack, high-fidelity MEA, specifically directed toward CFs. The objective of this attack is to derive a model threat_model t_Υ that accurately replicates the functionality of the target model.

The MEA is formulated as follows: for a set of queries $\mathscr{Q}$ and a set of corresponding CF explanations $\mathscr{C}$, an attacker trains a threat model t_Υ that performs equivalently on an evaluation set $\mathscr{T}$. The goal is to maximize an agreement function (the agreement metric will be detailed later), between f_θ and t_Υ as shown in Eq. 3.2, with a minimal number of queries sent to the API.

We now describe our proposed KD-based MEA and train a threat model t_Υ. We then describe our proposed methods for integrating DP into the CF generation process. Figure 3.3 illustrates our proposed methodology of the KD method along with the CF generation process. An owner of a private dataset trains a DNN classifier (Step 1 in Fig. 3.3), and a CounteRGAN CF generator (Steps 2A, 2B in Fig. 3.3) and deploys it on an MLaaS platform.

To make the MEA scenario more realistic and challenging for the attacker, we assume that the attacker has no prior knowledge of the training set. The attacker generates random samples and collects both the query output of the target model and the corresponding CFs provided for those outputs (Step 3 in Fig. 3.3). After successfully collecting a set of CFs (Step 4 in Fig. 3.3), the attacker employs our KD-based MEA ap-

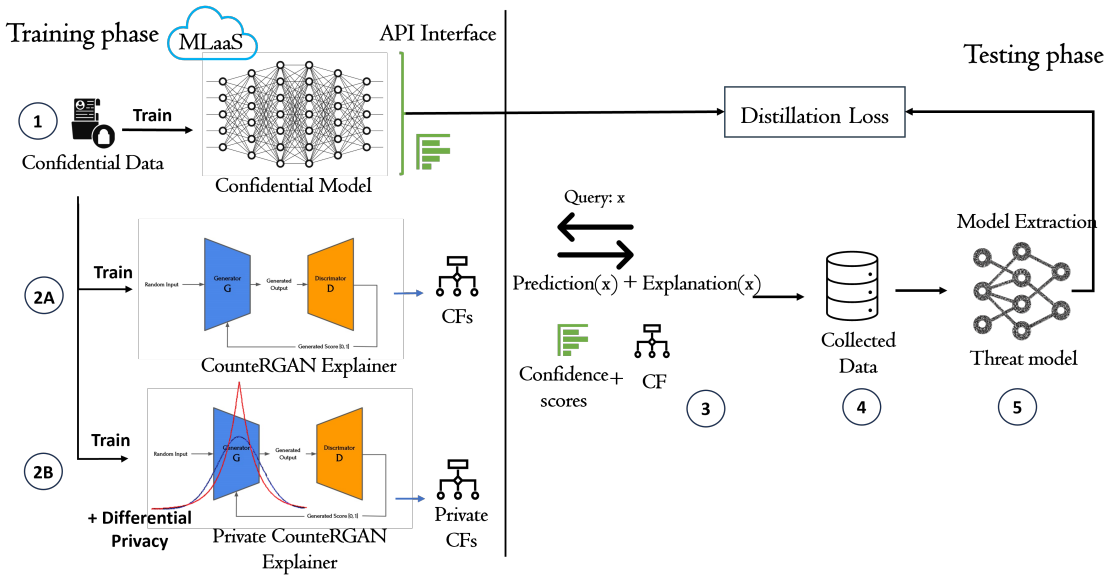


Figure 3.3. Methodology Framework, describing the KD-based MEA, in the deployment phase, in addition to the CF generation phase.

proach on the collected dataset as input data to train t_Υ. The rationale for selecting KD as the extraction method lies in its unique ability to address the problem not only as a standard supervised classification task but also to precisely imitate the probability distribution. Specifically, we propose training t_Υ by minimizing a loss that combines the threat-model classification loss with the distillation loss. To emphasize the importance of mimicking the target model's output probability distribution in the threat model, we use the Jensen-Shannon (JS) metric (3.1) as the distillation loss. The choice of JS is to address divergence issues and ensure a symmetric output, so we assume the attacker replaces the KL divergence with JS. Afterward, the attacker trains the threat model by minimizing the loss of KD (Step 5 in Fig. 3.3) until the convergence of the agreement on a separate validation set.

$$\mathrm{JS}(P\|Q) = \frac{1}{2}\mathrm{KL}\left(P\left\|\frac{P+Q}{2}\right.\right) + \frac{1}{2}\mathrm{KL}\left(Q\left\|\frac{P+Q}{2}\right.\right) \tag{3.1}$$

3.3.2 Private CFs generation with Differential Privacy

We are now aligned with the service provider and present a strategy to mitigate MEA by integrating DP into the CF generation process. Our methodology does not expose statistical information about the training set.

To this end, we integrate DP into the employed CF generator, CounteRGAN, during

the optimization process as follows. First, we employ the Adam differential private optimizer (DP Adam). The DP Adam process often requires multiple training iterations to achieve privacy guarantees. We then repeatedly add noise across several rounds, thereby strengthening the overall privacy guarantees by injecting noise into the generator's gradients, in addition to a gradient clipping step. This process involves clipping gradients for a random subset of examples, clipping each gradient's norm, computing the average, adding calibrated noise, and finally performing the traditional backpropagation step to update the weights. Note that the gradient clipping step is essential as it controls the amount of noise added and prevents large updates that could reveal too much information about individual data points. The key assurance provided by this approach is that the generator does not memorize or reveal details of the training data, and therefore does not generate CFs that are very similar to the training points. While DP guarantees ensure the privacy of the CFs, the added noise can degrade their quality and affect the training process. We also aim to examine the effects of incorporating DP into the CF generation process on MEA performance and the quality of the provided CFs, considering metrics such as proximity, prediction gain, and plausibility (described later in sec 3.4), as DP has an impact on *performance-utility trade-off* and *explainability-utility* as well [36]. Additionally, we aim to demonstrate how private CFs fit with the distribution of the training set.

3.4 Experimental settings

Datasets We perform our experiments using three classification datasets, namely, *Give Me Some Credits* [142], *Credit Card Fraud* [143], and *Housing* [144]:

- Give Me Some Credits (GMSC) [142]. The dataset is collected to forecast the likelihood that an individual will experience financial distress over the next 2 years, based on their financial and demographic information. The dataset comprises 150,000 applicants, of whom 139,974 are classified as good and 10,026 as bad. We follow the preprocessing described in [69]. Specifically, starting from the raw training data, we first build a balanced dataset by retaining all minority-class samples and downsampling the majority class to match their counts, resulting in 20,052 instances with 10,026 samples per class. We impute missing values in MonthlyIncome and NumberOfDependents, add quantile-binned versions of selected variables, and apply a preprocessing pipeline that includes median imputation with indicator variables for missing values, power transformation, feature expansion, and feature selection, resulting in 73 features as input.

- Credit Card Fraud (Credit) [143]. The dataset contains transactions conducted

by cardholders using credit cards in September 2013, represented by 30 features. In this dataset, transactions over two days reveal 492 fraudulent instances out of a total of 284,807. In the preprocessing pipeline, the Amount feature is scaled using RobustScaler, and the Time feature is normalized to the [0,1] range. We then construct a balanced dataset by retaining all fraud transactions and randomly sampling an equal number of non-fraud transactions, yielding approximately 1000 instances with a balanced class distribution.

- Housing [144]. The dataset was created from the 1990 U.S. Census, have 20640 samples consisting of 8 features and a target variable representing the median house value for California districts, in dollars. The target variable is then split into two classes using a threshold set to the median. Specifically, we convert the original regression target into a binary classification task by thresholding at the median target value. This produces a balanced binary target. In the current implementation, the input features are standardized using the feature-wise mean and standard deviation.

Baseline scenarios We compare our proposed approach for performing MEA, KD-based MEA, to 2 baseline approaches, presented in [133], referred to as *Direct Train* and *DualCF* [34]. Direct Train trains a threat model as a standard supervised classification task using the raw training data, without incorporating any specific optimization techniques for MEA, and DualCF trains a threat model using the CF of the CF directly. We examine three scenarios pertaining to the MLaaS provider's inclusion or exclusion of CF explanations alongside the predictions of the ML model, 1) *No CF:* In this scenario, the MLaaS does not provide the user with any CFs. 2) *CF:* In this case, the MLaaS provides CFs using counteRGAN. *3) Private CF:* The MLaaS offers private CFs by implementing our proposed approach for generating private CFs.

We begin by conducting an analysis aimed at 1) demonstrating how CFs can be exploited for MEA and (2) assessing how our proposed KD-based approach can further strengthen MEA. We also 3) evaluate private CFs under both direct training and KD. To perform this analysis, we simulate the following six scenarios:

- *Direct No CF.* We train a model directly on the randomly generated data points, and the focus shifts to examining the potential of direct queries and analyzing the power of CFs and KD.

- *Direct CF.* The MEA is performed following the baseline [133]. We train a model directly on CFs, this dimension is dedicated to evaluating the standalone power of CFs without the incorporation of KD, providing insights into the capabilities of CFs separately.

- *KD-based No CF.* We employ our proposed KD-based approach on the randomly generated data points, and the focus shifts to examining the potential enhancement when KD is utilized independently of CFs.

- *KD-based CF.* We employ our proposed KD-based approach, trained with CFs, to analyze the influence of KD when applied alongside CF.

- *Direct Private CF.* We train a model directly on CFs extracted from the DP generator, the focus shifts to examining the potential of direct queries and to analyzing the power of KD and CFs alongside.

- *KD-based Private CF.* We employ our proposed KD-based approach on CFs extracted from the differentially private generator, and the focus shifts to examining the potential effect of integrating DP in the CF generation.

We then extend our analysis to the second baseline, *DualCF*. These scenarios are designed to (1) evaluate the standalone effectiveness of DualCFs as a baseline compared to our KD-based approach, and (2) investigate the impact of integrating DualCFs with KD, yielding four distinct scenarios:

- *DualCF.* We train a model directly on DualCF, the aim is to evaluate the standalone power of DualCFs compared to KD, providing insights into their capabilities separately.

- *Private DualCF.* We train a model directly on DualCFs extracted from the DP generator, the focus shifts to examining the effect of DP and to analyzing it with DualCFs alongside.

- *KD-based DualCF.* We employ our proposed KD-based approach, trained with DualCFs, to understand the influence of KD when applied alongside DualCF.

- *KD-based Private DualCF.* We employ our proposed KD-based approach on DualCFs extracted from the DP generator, the focus shifts to examining the potential effect of integrating DP on the MEA.

To better represent the experiments and to compare scenarios before examining the numerical results, table 3.1 explicitly lists, for each scenario, the queried input source, the data used to train the extracted model, whether CFs are used, whether the CFs are private, whether KD is used, and which extraction family the scenario belongs to.

Table 3.1. Summary of experimental scenarios used in the MEA evaluation.

Scenario	CF	Private CF	DualCF	KD	Direct
Direct No CF	×	×	×	×	✓
Direct CF	✓	×	×	×	✓
KD-based No CF	×	×	×	✓	×
KD-based CF	✓	×	×	✓	×
Direct Private CF	×	✓	×	×	✓
KD-based Private CF	×	✓	×	✓	×
DualCF	✓	×	✓	×	✓
Private DualCF	×	✓	✓	×	×
KD-based DualCF	✓	×	✓	✓	×
KD-based Private DualCF	×	✓	✓	✓	×

Target model We adhere to consistent training procedures for the target model (to be deployed and queried by the API). We train a DNN comprising 16 hidden layers to create a complex model. The layer configurations consist of 64, 32, 16, 32, 64, 128, 64, 32, 128, 64, 128, 64, 128, 64, 32, and 16 neurons per layer. The activation function is set to Gelu for all layers, while the output layer employs softmax to produce model outputs as confidence score probabilities. We use the Adam optimizer to minimize the cross-entropy loss, with default parameter initialization. Each target model is trained using 80% of the training dataset. For all datasets, the model weights were initialized using LeCun uniform initialization with a random seed, a dropout rate of 0.2, and the Adam optimizer with a learning rate of 0.05, a batch size of 32, and 1000 training epochs [1]

Threat model We specify consistent training procedures for the threat model, assuming that attackers have no knowledge of the target model architecture but can build a DNN with a standard architecture. The attacker aims to train the threat model either using our proposed KD-based MEA approach or directly on the data points, using a DNN with 3 hidden layers of 16, 32, and 64 neurons consecutively, with ReLU activations and a softmax output layer. In the scenario we are considering, we assume that the attacker does not possess any knowledge about the distribution of the training data used for training the target model. Consequently, the attacker generates 1000 random data points for each dataset. The data point values are randomly generated, with val-

[1] Experiments were computed with a machine of intel Core i7, a GPU of RTX 3070, and 8 GB of RAM.

ues drawn from a range of -3 to 3 for each feature. We specify -3 and 3 as the ranges for random values, aiming to obtain randomized data points that are as dissimilar as possible to the training datasets. The threat model is trained for 1000 epochs, Adamax as optimizer with a learning rate of 0.01. For the KD-based MEA training, we vary α from 0 to 0.5 and report the highest agreement across 1000 epochs. We report the average of 10 runs with randomly selected subsets for each experiment.

Counterfactual generator To generate CFs using CountRGAN, we train the discriminator with 3 layers comprising 32, 16, and 1 neurons, respectively. Each hidden layer is followed by a dropout layer with a factor of 0.2. We use a ReLU activation function for the hidden layers and a sigmoid activation function for the output layer. The CountRGAN schedule consists of two discriminator steps and four generator steps, with a total of 2000 training iterations and a minibatch size of 64. The generator, which takes a datapoint as input and produces the corresponding CF, is trained with three hidden layers comprising 64, 48, and 32 neurons, using ReLU activation for the hidden layers and a linear activation function for the output layer. We specify the optimizer as Adam with a learning rate of 0.005. For the DP generator, we change the optimizer to one that supports DP. We integrate the DP Adam optimizer from the TensorFlow privacy library. To ensure a good privacy budget, we set the DP Adam optimizer's l2_norm_clip to 1 and the noise_multiplier to 3.

Evaluation metrics The evaluation metrics comprise two sets. The first set measures the effectiveness of MEA, while the second set of metrics assesses the quality of the CFs.

Effectiveness of MEA: A commonly employed metric for assessing the effectiveness of MEA is the *Agreement* measure. *Agreement* measures the degree of alignment between two models, i.e., the similarity in the predictions between the two different ML models. Agreements assess the output similarity between the predictions of an extracted model t_Υ and those of the target model f_θ for a given set of data records (see Eqn. 3.2). Since our goal is to replicate the behavior of a target model with an extracted model, higher agreement indicates a more successful MEA and, hence, a more effective MEA strategy.

$$agreement(f_\theta, t_\Upsilon) = \sum_{x_i \in \mathcal{T}} 1_{f_\theta = t_\Upsilon} \tag{3.2}$$

Success of explainer and quality of CFs: To measure the quality of extracted CFs, and to analyze the CF generator prediction classes shift, we employ three commonly used metrics, namely, *Prediction Gain*, *Plausibility*, and *Proximity*.

The *Prediction Gain* measures the change in the classifier's predicted probability for a specific target class t when comparing the CF explanation to the original data point. In other words, it quantifies the extent to which the generator succeeds in shifting the classifier's prediction toward the target class using the generated CFs. In our case, as the classifier outputs probabilities, the Prediction Gain ranges from 0 to 1, with higher values indicating a stronger shift toward the target class t.

$$PredictionGain = f_\theta(t \mid CF) - f_\theta(t \mid input) \tag{3.3}$$

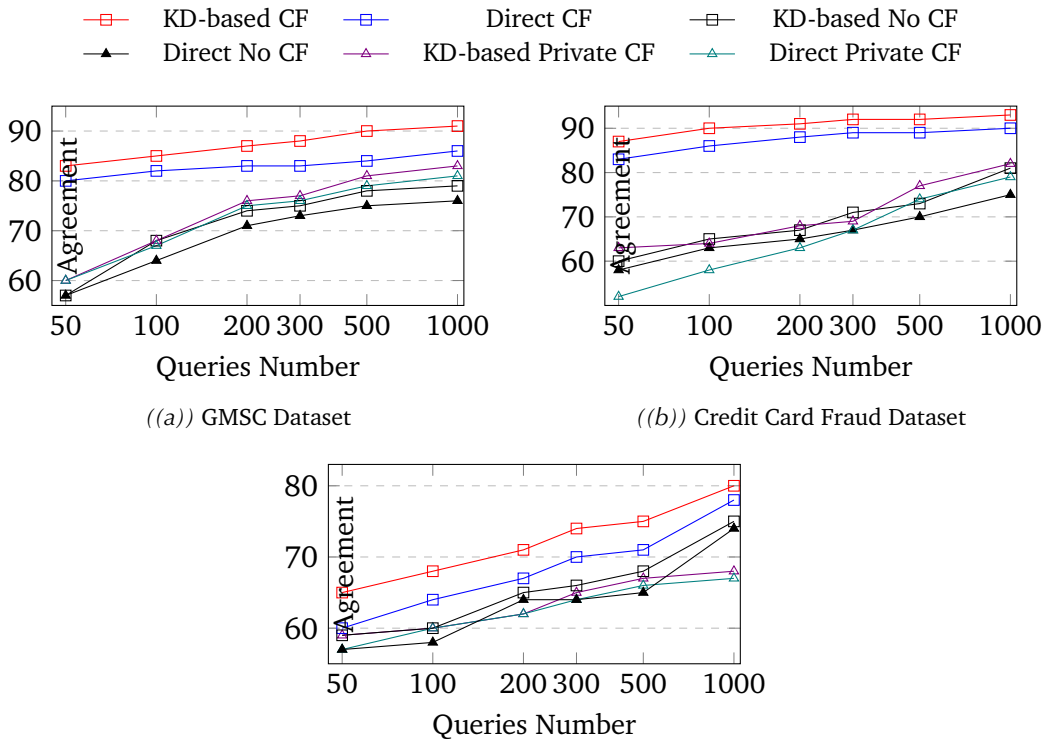
Plausibility quantifies how closely a data instance fits a known data distribution. In this work, we leverage DP to prevent CFs from revealing statistical information about the training set. Accordingly, realism as a plausibility measure assesses how well both standard and private CFs conform to the underlying data distribution. Realism also enables comparison of a data point with its corresponding CF to assess the authenticity of CFs and the effects of DP-induced perturbations. Following [79, 145], we train a denoising autoencoder on the noised training set and compute realism as the reconstruction error, measured by the mean squared error of the autoencoder (Eq. 3.4). Lower realism values indicate a closer fit to the data distribution and better plausibility.

$$Realism = \frac{1}{N}\sum_{i=1}^{N} \|input_i - reconstruction_i\|^2 \tag{3.4}$$

Proximity is a commonly used metric that allows for measuring the quality of a CF. It considers both the number of altered features and the magnitude of their changes (the amount of modification in each feature) relative to the original data instance. We compute *Proximity* by taking the L1 norm of the absolute difference between the data point and its corresponding CF (Eq. 3.5. Lower *Proximity* values (desired) indicate that only a small subset of input features has been perturbed to achieve a different outcome, suggesting higher-quality CFs.

$$\text{Proximity} = \frac{1}{N}\sum_{i=1}^{N} \|\text{input}_i - \text{CF}_i\|^1 \tag{3.5}$$

By computing this comprehensive set of metrics, we aim to explore how DP affects the feasibility of taking actionable steps and how this influence contributes to the prediction gain.



((a)) GMSC Dataset

((b)) Credit Card Fraud Dataset

((c)) Housing Dataset

Figure 3.4. Agreement values achieved by the MEA approach across the GMSC, Credit Fraud, and Housing datasets.

3.5 Numerical Results

3.5.1 Agreement for Model Extraction Attack

We start our discussion by analyzing the *agreement* achieved by the employed MEA approaches. During the threat model's training phase, the attacker interacts with the target model by querying it with randomly generated data points. In our experimentation, we systematically increase the number of queries between 50 and 1000 to conduct a comprehensive analysis of their impact on MEA.

Figures 3.4 (a), (b) and (c) show the *agreement* achieved by the six scenarios (i.e., with *KD-based MEA* and *Direct* as MEA methodology, and while leveraging CFs, private CFs or without CFs), with respect to the number of queries made to the API across the GMSC, Credit Card Fraud, and Housing datasets, respectively. First, we notice that results show a consistent pattern across all scenarios, as *agreement* initially increases significantly with the number of queries until a point where additional queries cease to

yield a notable increase. Note that this pattern is less noticeable for *KD-based CF* and *Direct CF*, which already exhibit relatively high *agreement* even with a low number of queries. As a result, these methods show a less substantial improvement in *agreement* compared to the others. We analyze deeper this observation in the following discussion. *Takeaway 1: Additional Queries lead to a saturation point in the MEA, up until further queries may not necessarily contribute to conducting a more effective MEA, which is in line with the findings reported in the literature.*

We now focus on analyzing the extent to which CFs can strengthen the adversary's ability to perform MEAs, comparing the *agreement* achieved by each of *KD-based MEA* and *Direct*, when CFs are and are not utilized (i.e., comparing *Direct No CF* to *Direct-CF*, and *KD-Based No CF* to *KD-based CF*). This comparison focuses on the use of CFs for MEA, independent of the methodology applied. In both cases, when CFs are used, and across all three datasets, the MEA is significantly more effective, achieving an *agreement* higher than its counterpart. For instance, in GMSC, *Direct-CF* achieves an *agreement* ranging between 80 and 85, up to 25% in some cases than *Direct No CF*, which achieves an *agreement* ranging between 55 and 75 (never higher than that of *Direct-CF*). For Credit Card Fraud, this difference is further amplified when CFs are used, as *Direct-CF* shows an *agreement* ranging between 83 and 90 while *Direct-no-CF* achieves an *agreement* ranging between 58 and 75. Similar results are observed across the Housing, where exploiting CFs allows for an additional agreement of around 10%. Moving to *KD-based MEA*, i.e., comparing *KD-based CF* to *KD-Based No CF*) across the three datasets, we notice that exploiting CFs permits achieving an agreement of up to 27% more (for GMSC at 50 queries) and, in the worst case, an additional 4% on *agreement* (for Housing at 1000 queries), than when CFs are not used. *Takeaway 2: Utilizing CF has a significant impact on leveraging model's CF for performing MEA regardless of the methodology used (KD-based MEA or Direct).*

We now switch to comparing the effectiveness of our proposed approach (*KD-based MEA*) to that of *Direct* across the three cases of leveraging CFs or not (i.e., no CF, CF, and Private CF), across the three datasets. We start with the case where no CFs are used. Results show that across all datasets, *KD-based MEA* outperforms its *Direct* counterpart. More specifically, *KD-based No CF* maintains a 3% to 6% higher *agreement*, across all datasets, than *Direct no CF*. Another way of looking at the performance is in terms of number of queries required by each approach to achieve a particular performance, which allows to understand better the effectiveness of a given approach. In this consideration, using CFs significantly reduces the number of instances required for MEA to reach a specific high agreement level. For example, across GMSC, achieving 80% agreement with CFs took just 50 queries, while accomplishing the same level without CFs required over 1000 queries. This pattern holds across datasets, such as the Credit Card Fraud Dataset, where 83% agreement needed only 50 CF queries com-

pared to over 1000 queries without CFs. *Takeaway 3: The use of CFs consistently leads to a substantial reduction in the number of queries needed to achieve high agreement levels across diverse datasets.*

We now consider the case where CFs are used (i.e., *KD-based CF* vs. *Direct CF*) with the aim of analyzing the extent to which the KD-based approach strengthens the MEA Results show that across all datasets, and for any number of queries, *KD-based CF* is more effective than *Direct CF*, maintaining an *agreement* of 2% to 6% higher than that of *Direct CF*. Investigating these results in terms of the number of queries, we notice that *KD-based CF* can be significantly more effective than *Direct CF*. Specifically, *KD-based CF* achieves an *agreement* of 82% and 85% with only 50 and 100 queries, a performance that is achievable by *Direct CF* with x6 number of queries (300) and x10 number of queries (500), respectively. A similar finding can be extracted across the Credit Card Fraud, as KD-based CF achieves an *agreement* of 90% with only 100 queries while *Direct-CF* uses 1000 queries to attain such an effective MEA. Similar results can also be observed by analyzing the performance achieved by these two scenarios with the Housing. *Takeaway 4: The KD-based MEA consistently results in higher agreement levels of MEA compared to the baseline approach. This holds in various scenarios and is independent of whether the MLaaS provides users with CF explanations, across diverse datasets and query numbers. Importantly, when CFs are employed, MEA requires significantly fewer instances to reach a specific high agreement level. This highlights the efficiency of incorporating CFs in conjunction with KD to enhance the performance of the MEA.*

The experimental results and takeaways collectively highlight the interplay between CFs and KD within the context of MEA. A central observation is that the KD-based MEA consistently outperforms the baselines when CFs are exploited. This behavior aligns with the theoretical strengths of KD, where soft probability distributions provide a rich representation of training datapoints than hard labels, and CFs, by construction, lie close to the decision boundary, where these soft representations are most informative. As a result, the extracted model accurately reconstructs the target model's decision-making, leading to high agreement and improvements across datasets and query budgets.

We now turn our attention to the scenario in which the MLaaS provider offers users private CFs, i.e., comparing *KD-based Private CF* to *Direct Private CF*.

We analyze the impact of generating private CFs through integrating DP in the CF generator, i.e., by incorporating DP during the generation process of explanations, on mitigating MEA. This analysis aims to quantify the degree of protection offered by providing users with private CFs (as opposed to CFs that are not generated with DP), such as preventing attackers from leveraging vulnerabilities associated with non-private CFs. To this end, we compare the performance of each of *KD-based CF* and *Direct CF* to its private CF counterparts (*KD-based Private CF* and *Direct Private CF*, re-

spectively). In GMSC, results show that for the relatively low number of queries (50 to 200), the difference is vast between each scenario with private CF and its counterpart. For instance, *KD-based Private CF* achieves an agreement ranging between 60% and 75% while that *KD-based CF* obtains an agreement between 83% and 87%. We observe that the incorporation of DP shows that our proposed strategy can maintain agreement levels comparable to scenarios without CFs and less than when incorporating traditional CFs, with an agreement range of 60% to 83% with *KD-based CF* and 60% to 81% with direct training, compared to 57% to 79% with *KD-based MEA no CF* and a range of 57% to 76% with *Direct No CF*. For the credit card fraud dataset, a similar trend is observed. With *KD-based Private CF*, the agreement range is between 63% and 82% and 52% and 79% for a direct-DP train, compared to *KD-based no CF* of 60% to 81% and a range of 58% to 75% with *Direct No CF*. For the housing dataset, the agreement of *KD-based CF* ranges from 59% to 68% with KD and 57% to 67% with *Direct Private CF*, compared to 59% to 75% with *KD-based no CF*, and direct no CF ranges from 57% to 74%. It is also worth noting that *KD-based Private CF* is slightly better than direct training across all datasets. For *KD-based MEA*, when DP is integrated, the noise introduced during the training process to achieve privacy guarantees could interfere with the teacher-student model's ability to transfer knowledge much more effectively. The observed lower in agreement of MEA with private CFs may be attributed to the privacy-preserving nature of DP. DP introduces intentional noise into the data to protect individual privacy, making it more challenging for attackers to accurately extract sensitive information since CFs may not be very close to the decision boundary as before. *Takeaway 5: The incorporation of DP can play a crucial role in maintaining agreement levels comparable to scenarios without CFs, as they show that across the three datasets, mitigation can be achieved.*

Takeaways show that when DP is introduced into the CF generation process, the behavior changes substantially. DP injects calibrated noise into the gradients of the CF generator, which impacts the generated CFs. This degradation manifests in lower plausibility with respect to the training distribution and increased perturbation variance (discussed later in details). These distortions weaken the alignment between CFs and the true decision boundary, thereby impacting the effectiveness of KD. More specifically, because DP enforces privacy by bounding the sensitivity of each gradient update, it systematically suppresses fine-grained, instance-level information, precisely the information CF generators rely on to produce minimal, directionally accurate perturbations.

We now turn to the results of the second MEA baseline, *DualCF*, comparing it with our proposed *KD-based* approach in terms of agreement, including evaluations involving Private CFs. To reduce redundancy and highlight the most relevant insights, we limit our presentation to query sizes of 200 and 1000. Table 3.3 presents the agree-

Table 3.2. Agreement values achieved by the MEA approach in comparison with DualCF baseline scenarios across the (a) GMSC, (b) Credit Fraud, and (c) Housing datasets, with respect to the number of queries made.

Dataset	Query	Direct CF	Dual CF	KD-based CF	KD-based DualCF	Direct Private CF	DualCF Private	KD-based Private CF	KD-based Private DualCF
Housing	200	67	70	71	79	62	62	62	71
	1000	78	79	80	82	67	68	68	76
Credit Fraud	200	88	89	91	92	63	64	68	69
	1000	90	90	93	93	79	80	82	85
GMSC	200	83	83	89	89	75	77	76	82
	1000	86	87	91	92	81	82	83	84

ment of the *Direct CF*, *DualCF*, *KD-based CF*, *KD-based DualCF*, along with their respective scenarios involving Private CFs.

We first compare the MEA agreement of *KD-based CF* to *DualCF*, across all datasets and query sizes. *KD-based CF* demonstrates clear improvements over *DualCF*: on Housing, agreement reaches 71% versus 70% with 200 queries and 80% versus 79% with 1000 queries. On Credit Card Fraud, 91% versus 89% with 200 queries and 93% versus 90% with 1000 queries. Similar trend on GMSC, 89% versus 83% with 200 queries and 91% versus 87% with 1000 queries. With regard to *KD-based Private CF*, it maintains equivalence or superiority in nearly all cases. On Housing, both methods yield identical agreement at 62% with 200 queries and 68% with 1000 queries. On Credit Card Fraud, *KD-based Private CF* leads with 68% versus 64% for *DualCF Private* at 200 queries and 82% versus 80% at 1000 queries, respectively. On GMSC, *KD-based Private CF* slightly outperforms *DualCF Private* at 1000 queries with 83% versus 82%, with only one exception at 200 queries where *DualCF Private* attains 77% and *KD-based Private CF* reaches 76%. *Takeaway 6: KD produces more accurate and generalizable MEA by effectively capturing underlying fidelity and structural patterns in the data compared to DualCF, as shown across multiple datasets, query sizes, and privacy conditions, and evidenced by higher agreement scores.*

We now analyze the *agreement* of *KD-based CF* compared to *KD-based DualCF*, and *KD-based Private CF* compared to *KD-based Private DualCF*. A consistent trend is evident across all datasets and query sizes: integrating *DualCF* with KD further increases *agreement* relative to *KD-based CF*. In the absence of DP, *KD-based DualCF* achieves the highest overall performance. For the Housing dataset, agreement improves from 71% with *KD-based CF* to 79% with *KD-based DualCF* at 200 queries, and from 80% to 82% at 1000 queries. On the Credit Card Fraud dataset, *agreement* increases from 91% to 92% at 200 queries and remains at 93% at 1000 queries. For GMSC, *agree-*

ment remains at 89% at 200 queries and increases from 91% to 92% at 1000 queries. The same trend is observed for *KD-based Private DualCF* compared to *KD-based Private CF*. On Housing, *agreement* improves from 62% with *KD-based Private CF* to 71% with *KD-based Private DualCF* at 200 queries, and from 68% to 76% at 1000 queries. On Credit Card Fraud, *agreement* increases from 68% to 69% at 200 queries and from 82% to 85% at 1000 queries. Similarly, on GMSC, *agreement* improves from 76% to 82% at 200 queries and from 83% to 84% at 1000 queries. These results demonstrate that integrating *DualCF* with KD consistently enhances *agreement* across all datasets and query sizes, with particularly notable improvements under DP. *Takeaway 7: Combining KD with DualCF yields consistently higher agreement over KD-based CF alone, across all datasets, query sizes, and privacy settings. This combination leverages the representational strength of KD while benefiting from the structural alignment of DualCF, resulting in enhanced agreement and MEA.*

We now analyze the effect of query size on the *agreement* of *DualCF* and *KD-based DualCF* across the 3 datasets. A consistent trend emerges: agreement tends to be stronger with larger query counts across all datasets and privacy scenarios. For example, with *DualCF*, *agreement* reached 62% with 200 queries and 68% with 1000 queries on the Housing dataset, from 64% to 80% on credit card fraud, and from 77% to 82% on GMSC. A similar trend is evident in the Private Scenarios: *agreement* of MEA on Housing improves from 70% to 79%, Credit Card Fraud from 89% to 90%, and GMSC from 83% to 87%. These findings suggest that larger query sizes enhance agreement for *DualCF*.

We proceed to compare the *agreement* of *DualCF* to *DirectCF*, and *DualCF Private* to *Direct Private CF*, across the two query sizes on all datasets. As expected, the results consistently indicate that *DualCF* outperforms *DirectCF* across all evaluated scenarios. Similarly, *DualCF Private* demonstrates superior performance relative to *Direct Private CF*, under privacy constraints. For instance, on the Housing dataset, *DualCF* achieves 70% *agreement* compared to 67% for *DirectCF* at 200 queries, and 79% versus 78% at 1000 queries. On the Credit Card Fraud and GMSC datasets, *DualCF agreement* either matches or exceeds *DirectCF*: at 1000 queries, both methods yield 90% *agreement* on Credit Card Fraud, while GMSC shows a marginal advantage for *DualCF* (87% vs. 86%). These findings are consistent with prior results reported in the original research in [34], reinforcing the conclusion that *DualCF* is a more effective MEA approach than *DirectCF* across varying query sizes and privacy conditions.

From a security and privacy perspective, these findings reveal that DP alone does not eliminate the risk of MEA facilitated by explanations. Instead, DP reduces the attack surface but remains vulnerable when attackers can issue many queries. This suggests that MLaaS providers should consider DP as one component of a broader defense strategy, complemented by rate limiting or query auditing.

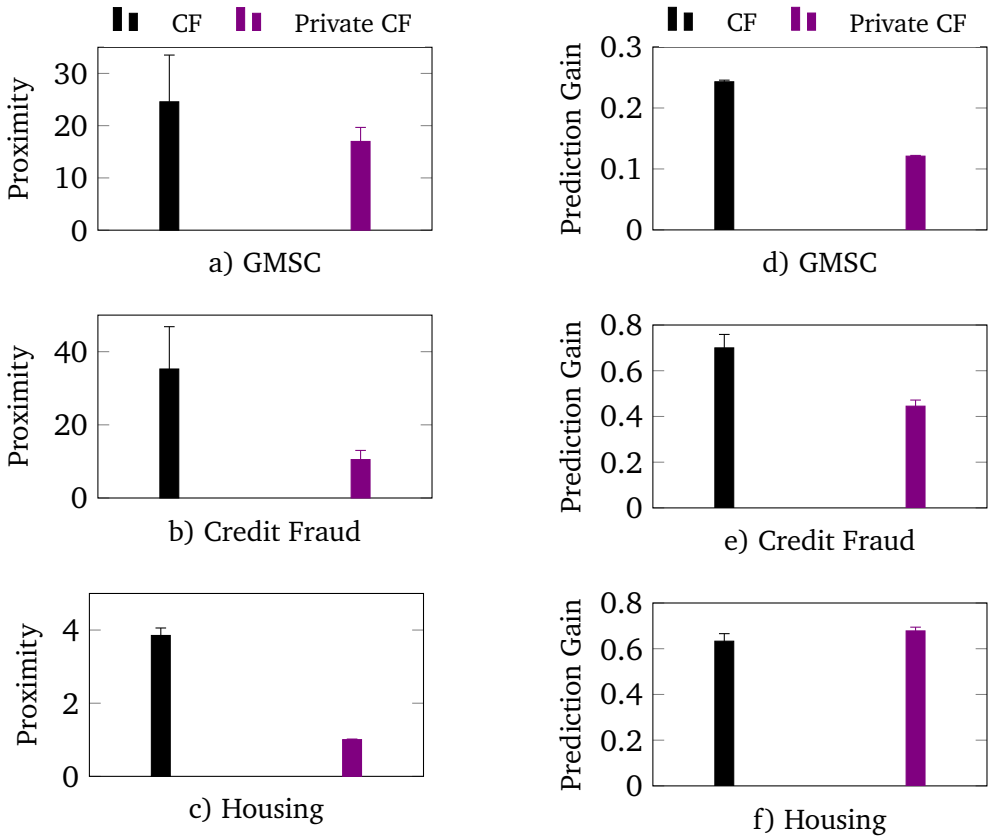


Figure 3.5. Proximity and Prediction Gain (± std) for CFs and Private CFs across datasets. Actionability is dataset-dependent since the ranges are based on the feature values

3.5.2 Impact of Incorporating DP in CF generator on quality of explanations

We now switch to analyzing the impact of integrating DP into CF generation process, i.e., into the CounteRGAN CF generator, on the quality of CFs in terms of the metrics introduced earlier [2]. Figure 3.5 shows CounteRGAN's *proximity* and *prediction gain* of the generated CFs with and without incorporating DP, across the three datasets[3].

In terms of *proximity*, which measures the degree of perturbation (or modification) of the CF compared to the initial point, results show that the generator generates CFs

[2]Note that this analysis is relative to the quality of the generated CFs, and not to the methodology adopted to perform MEA.

[3]The focus here is solely on comparing the quality of the generated CFs with and without the integration of DP. This analysis is independent of the specific MEA approach employed and regardless of the number if queries.

with lower values of *proximity* when DP is integrated than when not, across all datasets. For instance, for GMSC dataset, the *Proximity* of CFs with DP is 16.981 ($\pm$ 0.158) while that without DP is 24.5 ($\pm$ 0.364). Similarly, for Credit Fraud and Housing, *proximity* is lower when integrating DP from 35.269 $\pm$ 0.328 to 10.507 $\pm$ 0.238, and from 3.852 $\pm$ 0.053 to 1.004 $\pm$ 0.016, respectively. This means that the practical usefulness of the CFs or perturbations in guiding actionable decisions is reduced, suggesting that the privacy constraints imposed might compromise the effectiveness of CF analyses in providing actionable guidance for influencing desired outcomes. *Takeaway 8: CFs produced with DP integration consistently exhibit lower perturbation (proximity) values compared to those without DP. This suggests that integrating DP leads to CFs that require fewer modifications to preserve privacy. We did not optimize or fine-tune the level of privacy; these results show/quantify the existing tradeoff between the extent to which we can protect privacy and the impact on the quality in terms of proximity.*

Results show that CounteRGAN achieves a higher *prediction gain* (a higher probability shift, desired) when DP is not integrated within the CF generator than when not for GMSC (0.243 $\pm$ 0.011 versus 0.121 $\pm$ 0.01) and Credit Fraud datasets (0.700 $\pm$ 0.084 versus 0.445 $\pm$ 0.06) while for Housing dataset, CounteRGAN shows a prediction gain slightly higher in the case where DP is integrated rather than when not (0.678 versus 0.633). Unlike previous observations, the prediction gain of CFs and private CFs is comparable in this case. *Takeaway 9: The impact of DP integration on the prediction gain is dataset-specific, either maintaining it or, in some cases, restricting the explainer's progress towards the target class. This can be monitored and evaluated based on the use case and dataset.*

Although higher prediction gain and lower proximity are preferable when generating CFs, the results suggest that CF generation with DP has taken a different trajectory than without DP, leading to a reduction in prediction gain. This shift in the CF generation approach is reflected in the probability, ultimately contributing to the observed decrease in agreement, thereby motivating the smaller agreement for MEA.

Finally, we analyze the plausibility of the CFs in terms of *realism* of the data points, which enables us to quantify how well a data instance fits a data distribution of a dataset, including CFs in the scenarios with CF and with Private CF, and therefore to investigate the impact of incorporating DP in the generator. Combining this analysis with the previous ones, we can gain a deeper understanding of the impact of DP on MEA and the CFs. Figure 3.6 shows the *realism* of the data points used as initial queries and the CFs generated by the generator across the three datasets. We purposely present the initial queries *realism* to show the plausibility of the initial points compared to their corresponding CFs. Results show that *realism* of CFs when DP is not integrated in the generator is the lowest in comparison to that of initial queries and that of CFs when DP is employed. Specifically, in GMSC, the initially generated points have an average

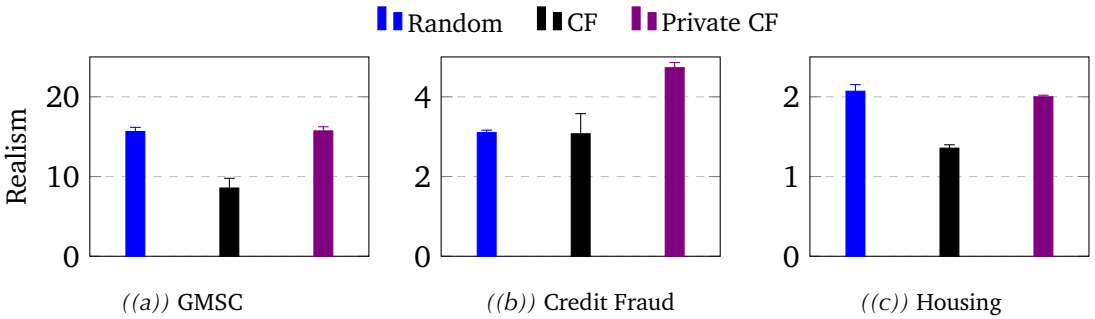


Figure 3.6. Realism scores (± standard deviation) for Random points, CFs, and Private CFs across the (a) GMSC, (b) Credit Fraud, and (c) Housing datasets.

realism of 15.6 (± 0.03), and the CF generator has produced corresponding CFs with a realism of 8.5 (± 0.14). Similarly, in Credit Fraud, the Realism of initial data points is 3.104 (± 0.019) while that of the corresponding CFs is slightly lower at 3.07 (± 0.16), respectively. In Housing, the realism of initial data points is 2.07 (± 0.04) while their corresponding CFs had a realism of 1.35 (± 0.03). This indicates that the CF generator, when DP is not integrated, generates CFs with lower realism from the original queried random points and aligns them more closely with the distribution of the training data. *Takeaway 10: The CF generator without DP integration has produced more plausible data points (lower realism) which means that the CFs fit better with the distribution of the training data.*

As shown in Fig. 3.6, initial random points are inherently unrealistic (do not exhibit high realism). Our private CF generation approach ensures this unrealistic nature is preserved when queried with a random data point. This is crucial because our experiments showed that traditional CFs, without privacy protections, generate more realistic outputs from random data, potentially revealing private information. Moreover, results also show that the private CFs preserve a realism very similar to that of random points. For instance, in GMSC, the private CFs have preserved their initial average level of realism of 15.723 (± 0.033). Similarly, in Housing, they preserve a realism of 2.0 (± 0.01). In Credit Fraud, the realism of the CFs deviates by one degree from the distribution. This means that incorporating DP, realism can be preserved, indicating the effectiveness of the Private CFs in maintaining the quality and distribution of the original data. *Takeaway 11: The CF generator, with DP integration, guarantees that the generated CFs maintain a similar level of plausibility to the original query data points.*

3.5.3 Classical Performance Metrics

Table 3.3 reports the classification performance of the extracted models using our proposed KD-based MEA and the state-of-the-art approaches, Direct and DualCF, compared to the performance of the original models, across all datasets, query budgets, and DP privacy variants (evaluation performed on a separate test set).

First, as expected, increasing the query budget from 200 to 1000 improves the performance of all extraction approaches. This trend holds across accuracy, precision, recall, and F1-score, confirming that additional queries provide higher performance and better extraction for the deployed model. In other words, increasing the number of queries from 200 to 1000 generally improves performance across the various MEA approaches, and the extracted models have comparable performance to the original model. Second, the KD-based approaches consistently achieve the highest performance across all metrics and datasets compared to Direct and DualCF. In particular, KD-based CF and KD-based DualCF, including its private variant, stand out as the most robust methods, maintaining superior predictive quality even under limited query budgets. This demonstrates that incorporating KD into MEA not only improves fidelity but also enhances the generalization ability of the extracted model. More importantly, the relative ranking of the approaches remains stable across scenarios. The KD-based approach, especially KD-based DualCF and KD-based Private DualCF, consistently achieves the highest performance, demonstrating that it is not only effective but also robust across datasets and query counts. Third, precision, recall, and F1-score also remain closely aligned with accuracy and reflect a consistent overall improvement in predictive quality. Interestingly, in some cases, the extracted model slightly outperforms the original model on ground-truth test labels. This is plausible because the regularization effect of distillation has been observed in prior KD teacher–student settings. However, this does not imply higher agreement (fidelity) than the deployed model, rather, it indicates that the extracted model can sometimes generalize better on the test distribution. Fourth, Precision, recall, and F1-score closely mirror the trends observed in accuracy, indicating that the improvements introduced by KD-based MEA are consistent across evaluation dimensions. Interestingly, in several cases, the extracted model slightly outperforms the original model on ground-truth test labels. This behavior aligns with the regularization effects commonly observed in teacher–student distillation frameworks. Crucially, this does not imply higher agreement to the deployed model, rather, it indicates that the extracted model can sometimes generalize better on the test distribution.

Table 3.3. Accuracy, precision, recall, and F1-score for all scenarios across all datasets. The original model is trained on the original dataset and is reported as a separate baseline.

Dataset	Query	Metric	Original Model	Direct CF	Dual CF	KD-based CF	KD-based DualCF	Direct Private CF	DualCF Private	KD-based Private CF	KD-based Private DualCF
Housing	200	Accuracy	88	66	70	70	79	61	62	62	72
		Precision	88	65	70	71	79	61	61	63	73
		Recall	88	64	70	70	78	60	61	62	72
		F1-score	88	64	70	71	78	61	61	62	72
	1000	Accuracy	88	77	79	80	82	67	68	68	76
		Precision	88	77	77	80	81	68	65	69	77
		Recall	88	77	78	80	80	66	66	68	76
		F1-score	88	77	77	80	81	67	65	68	76
Credit Fraud	200	Accuracy	91	88	89	91	92	63	64	68	69
		Precision	91	87	88	90	91	62	65	67	70
		Recall	91	89	90	92	91	64	63	69	68
		F1-score	91	88	89	91	91	63	64	68	69
	1000	Accuracy	91	90	90	93	93	79	80	82	85
		Precision	91	89	91	92	94	78	81	83	84
		Recall	91	91	89	94	92	80	79	81	86
		F1-score	91	90	90	93	93	79	80	82	85
GMSC	200	Accuracy	87	82	83	89	89	75	77	76	82
		Precision	88	81	84	88	90	74	78	77	81
		Recall	87	83	82	90	88	76	76	75	83
		F1-score	87	82	83	89	89	75	77	76	82
	1000	Accuracy	87	85	87	91	92	81	82	83	84
		Precision	88	84	88	90	93	80	83	82	85
		Recall	87	86	86	92	91	82	81	84	83
		F1-score	87	85	87	91	92	81	82	83	84

3.5.4 Interplay of Explainability, Privacy, and Predictive Performance

Complementing previous results, DP can affect predictive performance and explanation quality, and can be applied at both levels. Now, we focus on analyzing the mitigation framework that integrates DP at the model and at the explainer, and investigate the interplay between i) *model's accuracy*, as DP influences the model's inference capability, ii) *privacy*, as employing DP provides resilience against attacks, and iii) *explainability*, as noise added to the model or explainer may impact the quality of explanations [146]. We aim to quantify this interplay and to extract insights into the choice of where to employ DP (at the model, at the explainer, or at both) and the degree of noise to use to balance predictive performance, explainability, and privacy.

We investigate *(1) to what extent does applying DP at the model or at the explainer, or both, effectively mitigate MEA facilitated by CFs?*, *(2) How does noise level in DP influence the effectiveness of MEAs that leverage CFs?*, and *(3) In what ways does the quality*

of CF explanations differ when DP is applied at the model compared to the explainer? We focus on identifying a mitigation strategy against attacks that exploit the model's explanations. Specifically, we explore the application of DP to the ML model, the explainer, and both simultaneously.

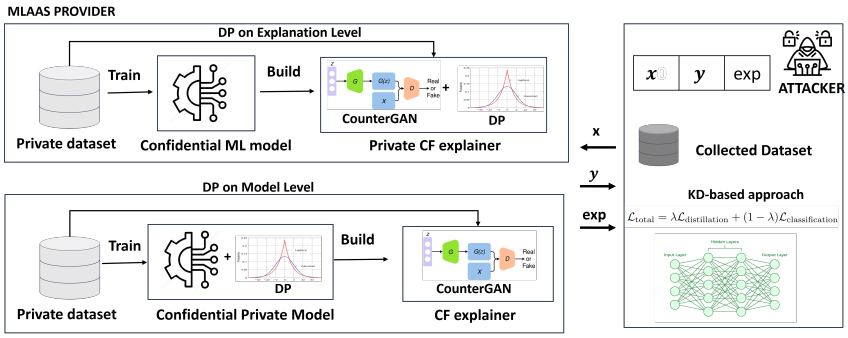


Figure 3.7. MEA within an MLaaS provider, depicting two different scenarios where DP is employed at the model or at the explainer to counter potential attacks.

As a mitigation against MEA (Figure 3.7, we employ two strategies: *1) DP-Model with DP-SGD:* where we apply DP-SGD during model training on $f(x;\theta)$. *2) DP-Explainer (DP in CounterGAN):* as proposed previously. We then perform MEA leveraging CFs under different DP settings, i.e., the approach adopted and the privacy parameter's noise level ϵ, and evaluate the adversary's MEA success and CF quality. Specifically in step 1, $f_{\text{baseline}}(x;\theta_{\text{base}})$ is first trained on D without DP. We also train a DP-protected model $f_{\text{DP}}(x;\theta_{\text{DP}})$ using DP-SGD with privacy parameter ϵ. Similarly, in step 2, we also train a private CounterGAN $\hat{x} = G_{private}(x;\phi)$ to generate the private CFs by varying the noise level. The attacker in step 3, apply MEA to extract $f_{\mathscr{A}}(x;\theta_{\mathscr{A}})$ using the KD-based method using either CFs generated by $G(x;\phi)$ or $G_{private}(x;\phi)$. For the comparative analysis, we consider four distinct scenarios:

- *(1) No DP:* Baseline scenario that does not incorporate DP at any level, allowing the evaluation of the unprotected model performance and vulnerability.
- *(2) DP-Model:* Only the target model employs DP. This protects the model from adversarial replication while the explanation generator remains unprotected.
- *(3) DP-Explainer:* DP is applied to the explanation generator. This scenario assesses the impact of DP explanations on their utility without directly affecting the target model.
- *(4) DP-Model-Explainer:* Both the target model and the explanation generator are protected with DP, aiming to balance model performance, explanation quality,

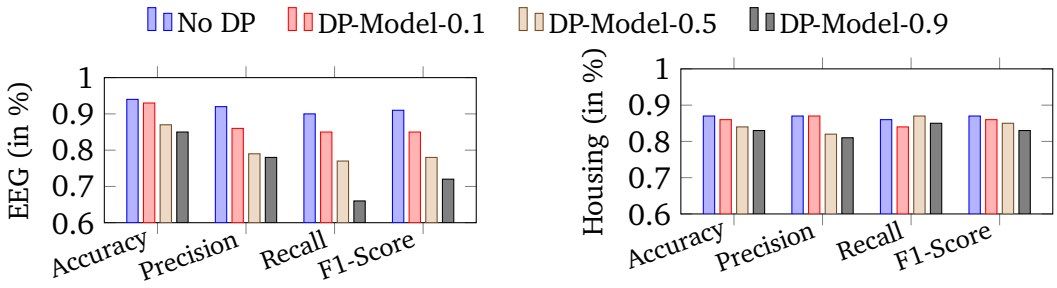


Figure 3.8. Model Performance achieved by the ML model across the two datasets for varying noise scales.

and resistance to MEA.

Datasets, Target and Threat Model

We evaluate on 2 datasets: *Housing* (previously introduced) and a new dataset, *EEG Eye State* [147]. The EEG Eye State dataset comprises EEG measurements recorded with a Neuroheadset, containing 14,980 data points and 14 features. The target variable is a binary label indicating whether the eye is closed or open. We focus on these two datasets because the DNNs, particularly when trained with DP-SGD, maintain reasonable predictive performance. *On the other datasets (GMSC and Bank), performance degraded substantially under DP-SGD, making the resulting analysis less informative.*

The target model f_θ is a DNN with 16 hidden layers of 64, 32, 16, 32, 64, 128, 64, 32, 128, 64, 128, 64, 128, 64, 32, and 16 neurons per layer with a GELU activation function and a softmax activation function in the output layer. We employ the Adam optimizer in both cases, whether DP is used or not. The model is trained without DP and with noise levels of 0.1, 0.5 and 0.9 for DP cases and with varying learning rates (0.001, 0.002, and 0.01), and we vary the l2_norm_clip to between 1, and 1.5 (l2_norm_clip bounds the sensitivity of the gradients by limiting the influence of any single training example on the overall gradient update, which is a crucial step before adding noise). Note that the more noise, the higher the privacy. The target models were trained on 80% of their respective datasets, and the model with the highest accuracy was selected. We tune different KD-based approach hyperparameters, specifically, alpha within the range of 0.1 and 0.5, and temperature within the range of 1 and 10. For scenarios where DP is employed, we applied DP with noise levels of 0.1, 0.5, and 0.9 to the generator. We varied the learning rate to 0.05, 0.005, 0.01, and 0.001, and the l2_norm_clip to 1, 1.5, and 3.

ML Model Predictive Performance

Fig 3.8 reports the predictive performance metrics of the models while varying the noise level across the two datasets used in our evaluations. As previously mentioned, we consider three noise levels when applying DP, 0.1, 0.5 and 0.9, and we refer to each case as DP-Model-noise level. As expected, the results across the two datasets indicate a decline in predictive performance metrics as the noise level increases. For instance, in the EEG dataset, accuracy, precision, recall, and F1-score are 94, 92, 90, and 91, respectively, when no DP is applied. However, at the highest noise level considered (0.9), these metrics drop to 85, 78, 66, and 72, respectively. Similar results are observed across the Housing dataset, where predictive performance metrics decline as the applied noise level increases.

Effectiveness of Differential Privacy in Mitigating MEA

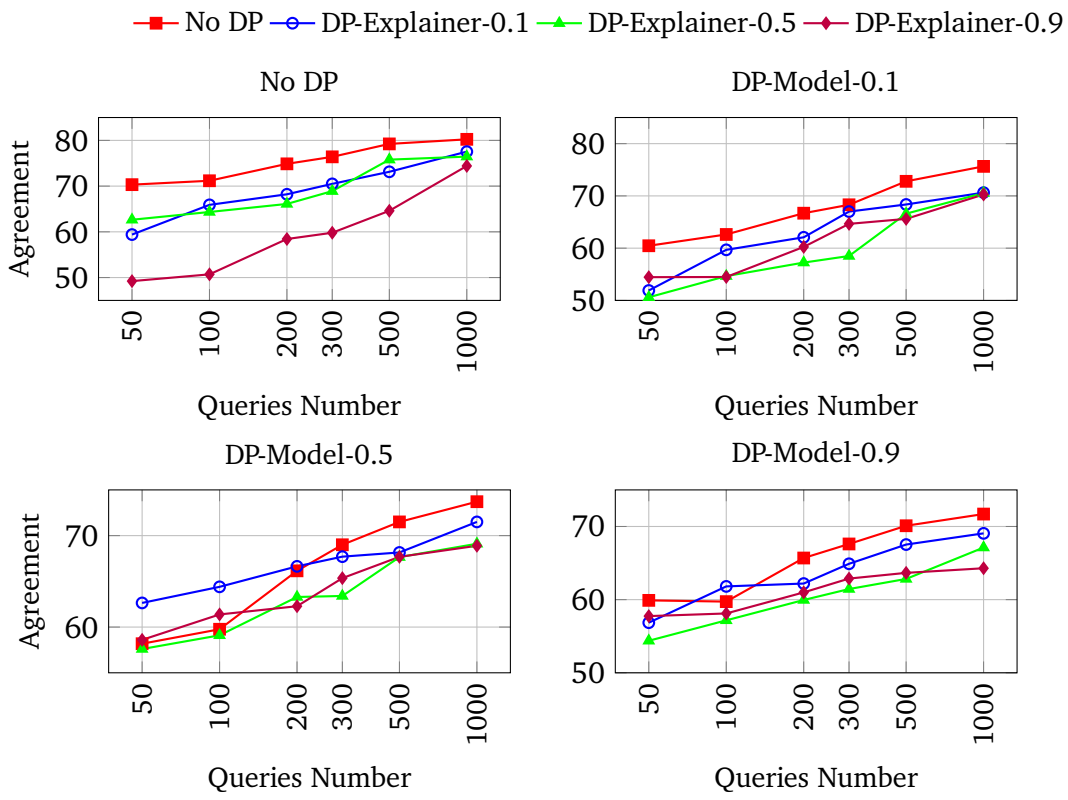


Figure 3.9. Housing dataset MEA agreement for various DP strategies, noise levels and number queries.

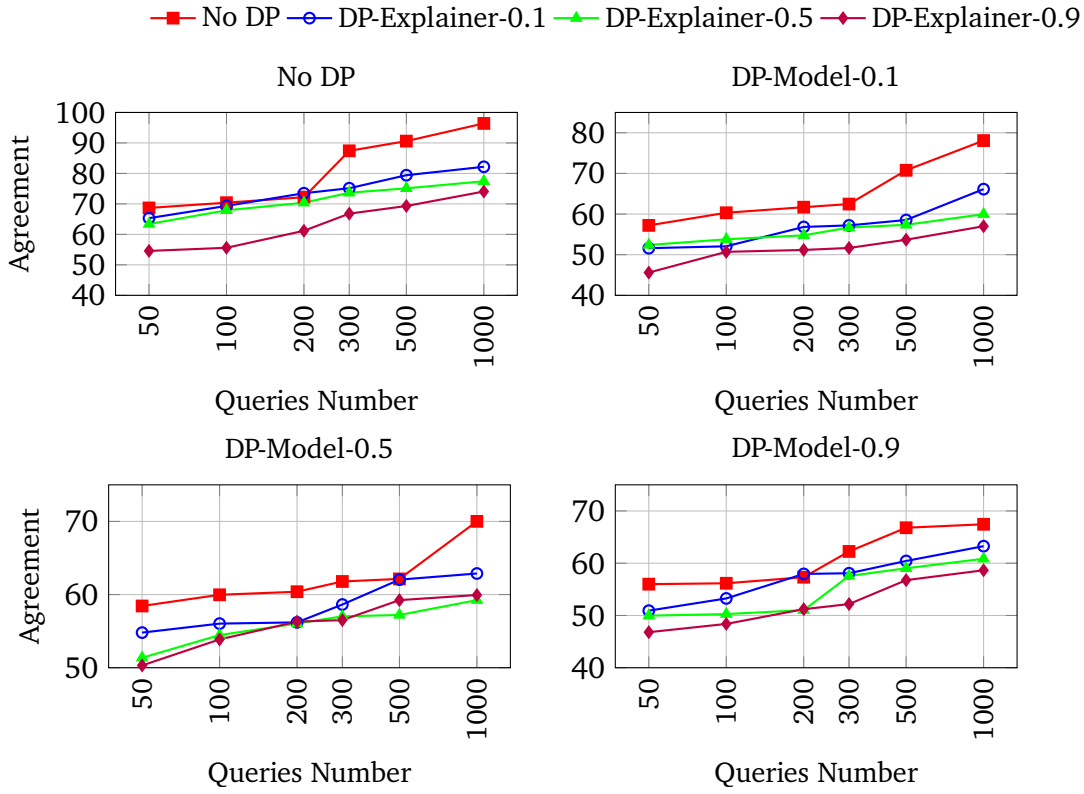


Figure 3.10. EEG dataset MEA agreement for various DP strategies, noise levels and number queries.

We considering the 3 scenarios of application of DP, namely, DP-Model, DP-Explainer and DP-Model-Explainer, and the baseline No DP scenario. Additionally, when we incorporate DP at explainer, we refer to each case as DP-Explainer-noise level (i.e., DP-Explainer-0.1). This evaluation will allow us to address RQ1 and RQ2. Figures 3.9 show the *agreement* observed by MEA across the various combinations of applying DP for varying noise levels and number of queries across the *Housing* dataset. We start with No DP, which allows us to quantify solely the impact of employing different levels of noise at the explainer on the success of the MEA. Results show a general trend where the MEA is more successful as the number of queries used increases across all cases (i.e., independent of the noise level applied). Comparing the *agreement* when employing different noise levels, results show that employing more noise, as expected, provides more defense against MEA. Specifically, with a noise level of 0.9, *agreement* ranges between 50 and 72 when the number of queries increases up to 1000. In contrast, when employing noise levels of 0.5 and 0.1, *agreement* falls within the ranges of 62–75 and 60–78, respectively. In the absence of DP at the explainer, *agreement* starts

at 70 with 50 queries and reaches 80 when 1000 queries are used. We now focus on the cases where DP is employed at the model level. Generally, results show a similar trend across all cases, where *agreement* increases with the number of queries used to perform the MEA. Comparing the *agreement* achieved when employing different noise levels in each case, results show, as expected, that employing higher noise levels at the explainer implies better protection against MEA. For instance, when employing DP-Model with a noise level of 0.1, the highest *agreement* observed is 70 when DP-Explainer is employed (which is a DP-Model-Explainer case), compared to 76 without DP-Explainer. Similarly, with a noise level of 0.5 at the model, the agreement consistently remains lower than in the No DP case, reaching a maximum of 70.63 versus 75 to when DP is only applied at the model. Similar trends were observed to DP-Model-0.9. Figure 3.10 show the agreement observed by MEA on the EEG dataset across the various cases. The results show similar trends to the one observed with the Housing Dataset. When no DP is applied to the model, the *agreement* improves with more queries ranging between 68% and 96%, with the highest agreement observed when DP is not applied at all.

Impact of Differential Privacy on Quality of Explanations

Figure 3.11 shows the prediction gain achieved by the explainer across the Housing and EEG datasets for varying noise levels. In the Housing dataset, a clear trend emerges as the DP-Explainer noise increases, the prediction gain decreases, which means that employing more noise decreases the CF probability toward the desired class. For example, for No DP, the prediction gain starts at 0.488. However, for DP-Explainer with noise level of 0.9 is applied, it drops dramatically to 0.055. This decline is observed consistently across all model noise levels. Moreover, for DP-Model noise levels (0.5 and 0.9) are introduced, the prediction gain prediction observed is less than that of no DP and DP-model 0.1, regardless of the DP-Explainer noise. Similarly, the EEG dataset follows a comparable pattern. In scenarios without DP applied to the model, the prediction gain is lower and ranges from 0.568 to 0.222 as the DP-Explainer noise is higher. When the model is subjected to DP noise at levels of 0.1, 0.5, and 0.9, the prediction gains are consistently lower. We now focus on analyzing the impact of incorporating DP on realism. Across both datasets, increasing the DP-Explainer noise consistently results in higher realism scores, indicating less realistic CFs and degradation in CF quality. In the Housing dataset, even without any DP-model noise, the realism score ranges from 0.113 to 1.116 at a DP-explainer noise of 0.9. This degradation is further amplified when additional DP-Model noise is introduced, e.g., with a DP-Model noise of 0.1, the realism score ranges from 0.356 to 3.289 as DP-explainer noise is higher, and similar patterns are observed for DP-Model-0.5 and 0.9. The EEG dataset exhibits a comparable pattern, although the No DP realism scores are generally higher.

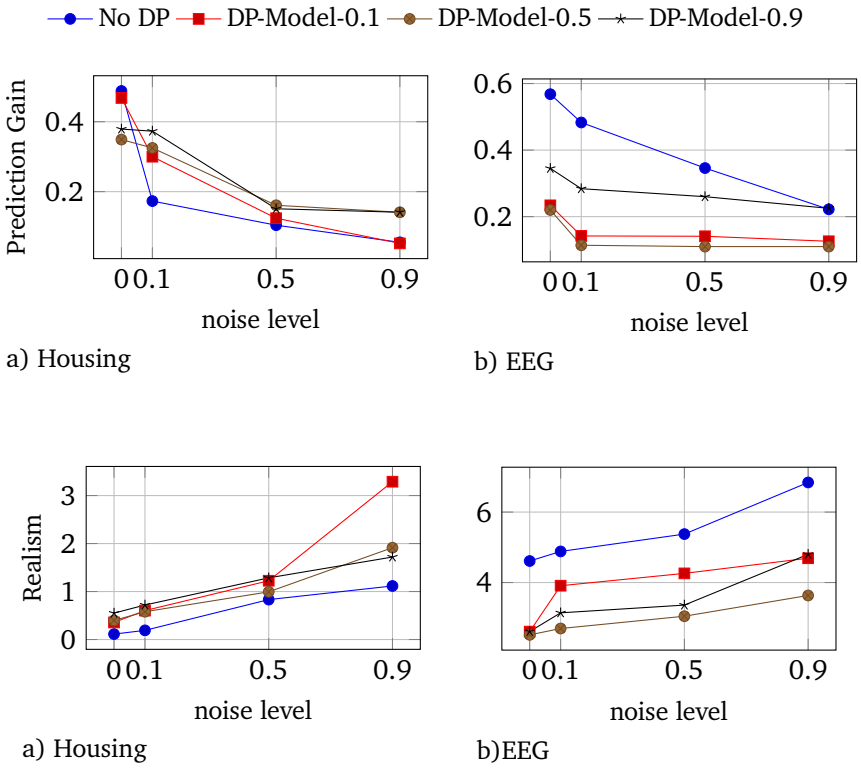


Figure 3.11. Prediction Gain and Realism achieved by explainers in the Housing and EEG datasets under various DP-Explainer noise levels.

Discussion on Performance-Privacy-Explanations Interplay: Results indicate that introducing DP mechanisms affects model performance, although the extent of this impact varies according to the specific use case and dataset. Similarly, the quality of the generated CF explanations is influenced by the privacy parameters applied. Experiments reveal that even slight amounts of noise, whether introduced at the DP-Model or within the DP-Explainer, can alter CF quality. In terms of the effectiveness of DP interventions in the context of MEA. Analysis shows that introducing minimal noise at the model level generally offers resistance to MEA. In contrast, higher noise levels provide a more robust defense, albeit at the cost of reduced model performance. When examining the impact of noise on the CFs, we observe that small increases in noise slightly reduce the MEA success rate, but larger increases yield a more pronounced protective effect. Notably, when both the model and the explainer are simultaneously subjected to DP, a synergistic improvement in resistance to MEA is observed.

3.6 Limitations and Future Directions

We discuss the contribution, limitations of our current approach, and potential future research directions:

- Privacy–Performance Trade-off: Integrating DP into the CF generator mitigates MEA, but it also introduces limitations by degrading explanation quality, including proximity, prediction gain, and plausibility in some cases. As future work, a more systematic study of the privacy–utility trade-off is needed, including evaluating a wider range of privacy budgets and optimization settings for the DP-based CF generator to better balance protection and explanation quality.

- Focus on Deep Learning Applications: This work focuses on MEA for DNNs, where knowledge distillation is particularly effective due to the expressive capacity of DNNs and their ability to learn rich data representations. This focus limits the applicability of our findings to other model families, such as traditional baselines including tree-based or ensemble methods. Since KD proved effective in the DNN setting, a natural future direction is to explore how KD-based extraction strategies can be adapted to non-DNN algorithms and whether similar performance gains can be achieved.

- This chapter paper evaluates the quality of CFs using quantitative metrics; these metrics primarily capture the degree of distortion or the distance between the original point and its modified counterpart. In other words, they measure how much the explanation changes, rather than whether the resulting counterfactual remains useful, understandable, or actionable for human users. Since CFs are intended to support human decision-making, it would be valuable to complement the current evaluation with user studies that assess whether users perceive the generated CFs as meaningful, actionable, and helpful in practice.

3.7 Conclusion

In this chapter, we investigated how counterfactual explanations (CFs) can be leveraged to conduct model extraction attacks (MEAs) via knowledge distillation (KD), and how differential privacy (DP) can be integrated into the CF generation process to mitigate such attacks. Our results show that CFs provide meaningful insights into the decision boundary of the target model, enabling an attacker to train a substitute model even without prior knowledge of the underlying data distribution. At the same time, we demonstrated that incorporating DP into the CF generator can substantially reduce the vulnerability of MLaaS systems by degrading the quality of the extracted CFs and

limiting the attacker's ability to reconstruct the target model. Experimental results show that exploiting CFs makes MEA easier, not only increasing their success in terms of agreement but also substantially reducing the number of queries needed to conduct the MEA. Moreover, results show that our proposed KD-based MEA approach outperforms the baseline approach that directly trains the model on the CFs.

Beyond demonstrating the feasibility of the proposed MEA and the effectiveness of the DP-based mitigation, our findings highlight several practical implications. First, they underscore that explanation mechanisms, particularly CF generators, should be treated as potentially sensitive outputs, as they may inadvertently reveal structural information about the model. Second, the results emphasize the importance of carefully calibrating privacy mechanisms: while DP protects the training data, it also affects the interpretability and actionability of the generated CFs, creating a privacy–utility tension that MLaaS providers must navigate. Third, our analysis suggests that the security and privacy of explanation-enabled services cannot be evaluated solely in terms of prediction accuracy or explanation fidelity, instead, robustness against extraction must be considered as a core design requirement. We also evaluate employing DP at the ML model level, at the explanation level, and at both simultaneously to investigate their respective impacts on MEA resilience. Our analysis demonstrates and quantifies a fundamental trade-off between privacy protection and utility. This work represents a crucial step toward the dual objectives of preserving privacy and maintaining the integrity of explanation quality. Overall, this work provides a foundation for understanding how CF explanations can be leveraged in MEA scenarios and for integrating privacy-preserving mechanisms into CF generation pipelines.

Chapter 4

Exploiting Counterfactual Explanations for Membership Inference Attacks

In this chapter, we extend the privacy attack analysis and analyze the role of CFs in MIA. We study whether explanations can amplify privacy leakage by inferring whether a specific data point was in the model's training set. We then evaluate mitigation strategies based on DP and AL, assessing their impact on attack performance, model utility, and CF quality. This chapter further illustrates the privacy risks introduced by explanation access and the trade-offs involved in protecting against them. This chapter is based on the content available in the following work:

1. **Fatima Ezzeddine**, Osama Zammar, Silvia Giordano, and Omran Ayoub. *Explanations Leak: Membership Inference with Differential Privacy and Active Learning Defense.* International Joint Conference on Neural Networks (IJCNN), in IEEE World Congress on Computational Intelligence (WCCI) 2026.

4.1 Introduction

Membership inference attacks (MIAs), which aim to determine whether a specific record was included in the model's training set [35], are another privacy threat to standard MLaaS deployments, where models are exposed via a remote API that provides queriable prediction functionality. Integrating explanations into MLaaS, in which models return explanations alongside predictions, amplifies MIA privacy concerns. Recent research has further demonstrated that providing explanations such as FA alongside predictions can amplify MIA risks [35]. This is because explanations, and particularly,

often reveal features, dependencies, and boundary regions on which the model relies, they provide adversaries with additional signals that can be systematically leveraged to infer sensitive information about the underlying training dataset, thereby increasing the effectiveness and precision of MIA privacy attacks [35, 47, 148].

MIA is particularly critical because it directly undermines the confidentiality of training data. By revealing whether a specific individual's record was used to train a model, MIAs create a tangible link between the deployed model and identifiable individuals. Such leakage significantly increases the risk that the model itself may be considered personal data under the GDPR, since it can enable the re-identification of individuals [149]. If CF explanations are available, attackers can further strengthen MIAs by combining standard model outputs (e.g., predicted probability distributions) with additional information revealed by the explanations. In particular, CFs can reveal how far an instance lies from the model's decision boundary and can be exploited by the intuition that training samples are more confidently classified and thus are often located farther from the decision boundary than non-members. Consequently, an adversary may infer membership when the distance between a given instance and its CF exceeds a pre-defined threshold [47].

In this context, it is essential to systematically analyze how CFs can be exploited to strengthen MIA and quantify the extent to which they enhance its effectiveness. At the same time, developing robust defense mechanisms is equally critical, where privacy-enhancing technologies naturally emerge as candidate solutions. However, introducing privacy-preserving mechanisms is not without cost. Applying DP may affect not only the model's predictive performance but also the usefulness of the generated explanations. Therefore, it is necessary to carefully evaluate how privacy protection impacts explanation utility, alongside model predictive performance. This tension gives rise to a fundamental interplay between three competing objectives: (i) resilience against privacy attacks such as MIAs, (ii) predictive performance of the underlying ML model, and (iii) the quality of the provided explanations (addressing RQ1 and RQ2).

In this chapter, we address this three-way tension, with the broader goal of contributing to the responsible deployment of xAI in socially sensitive high-stakes domains such as finance and public services. It is imperative to ensure that efforts to enhance transparency do not expose individuals to privacy harms. To this end, we first investigate how CFs can be exploited to perform MIAs, in which an adversary has API query access to both predictions and CFs. In particular, we focus on *shadow-based MIAs*, a practical, realistic, and well-established attack methodology, compared with *threshold-based MIAs* [150], which rely on predefined thresholds of membership indicators such as loss or confidence. Specifically, in *shadow-based MIAs*, attackers approximate the target model's behavior using surrogate models trained on auxiliary data. This assumption reflects realistic attack capabilities in deployment. Second, we propose a novel defense

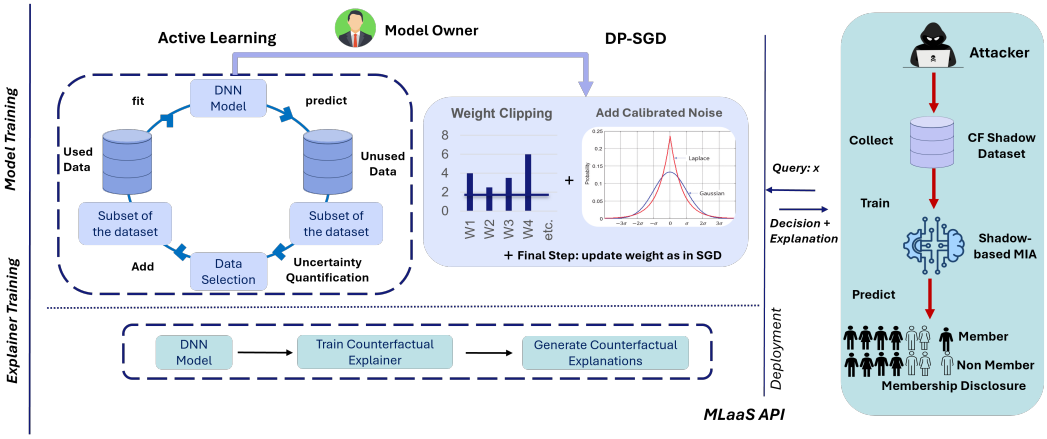


Figure 4.1. Scenario illustrating model training procedures performed, alongside the attacker's strategy of performing MIA.

mechanism that integrates DP with Active Learning (AL). Our intuition to use AL as a technique to further enhance privacy is discussed in more detail in Sec. 4.3 as a data-distillation strategy [151, 152, 153, 154, 155]. Our approach seeks to jointly combine DP and AL in a manner to strengthens resistance to MIAs while preserving both predictive performance and explanation quality (see Fig. 4.1). This design enables an evaluation of the trade-off between privacy preservation, measured by the attacker's ability to correctly infer membership, and explainability, quantified by CF quality.

To the best of our knowledge, this is among the first works to conduct a comprehensive study on shadow-based MIAs that leverage CFs and to propose a defense mechanism that combines DP and AL. Moreover, we systematically analyze the interplay between model predictive performance, MIA attack success, and CF quality (addressing RQ1 and RQ2). Specifically, the contributions can be summarized as:

- We systematically quantify how CFs can be leveraged as the adversary knowledge source for shadow-based MIA.

- We introduce a novel defense framework that integrates DP with AL to limit memorization and reduce exposure to training data.

- We empirically characterize the three-way trade-off between privacy leakage, predictive performance, and CF quality, offering a unified analysis of privacy, utility, and explanation fidelity.

4.2 Related Works

Explanation-based MIAs have emerged as powerful methods for performing threshold-based membership inference attacks (MIAs) [11]. For instance, Shokri et al. demonstrate that back-propagation-based explanations strengthen MIA by leveraging the intuition that higher variance in feature attributions indicates greater uncertainty in the logits, signaling that an instance is less likely to be a training member [35]. Pawelczyk et al. introduce a recourse-based MIA using counterfactual explanations (CFs), exploiting the intuition that training points should be farther from decision boundaries than test points; membership is then predicted when the distance to a CF exceeds a threshold [47]. Liu et al. further extend this understanding by introducing robustness-based MIA, where perturbations guided by attribution maps reveal that members exhibit confidence drops when important features are altered [156].

Other efforts have focused on defending against MIAs [157, 158]. Yeom et al. propose regularization-based defenses and attribute the success of MIAs to overfitting [157]. Shokri et al. suggest overfitting-prevention techniques such as dropout and L2 regularization [44]. Although helpful, these methods are not specifically designed for MIA defense and therefore may lack robustness. Other methods have been designed specifically to defend against MIAs. Nasr et al. introduce an adversarial regularization method that adds a membership-inference gain term to the loss function, effectively creating a min–max game between the target model and the membership inference adversary [158]. Li et al. propose a regularizer between the softmax output distributions of the training and validation sets to prevent MIA [159]. Another type of defense is adversarial example-based, where inputs or outputs are intentionally altered, often through perturbations, to mislead an attacker. For instance, Jia et al. propose adding crafted noise to the model outputs, thereby transforming them into adversarial examples that mislead the attacker's classifier and prevent MIA [160].

Several other works have explored the use of differential privacy (DP) to defend against MIA. Abadi et al. propose the differentially private stochastic gradient descent algorithm (DP-SGD) [19]. Rahman et al. reveal that even DP-SGD can still be compromised by MIA methods under certain conditions [161]. To mitigate the drawbacks of employing DP during model training or data generation, other works have explored incorporating DP at different stages. For instance, Ye et al. modify and normalize confidence-score vectors via a DP mechanism, thereby safeguarding privacy by obscuring membership information and reconstructing data [162]. Yuan et al. provide an in-depth analysis of privacy risks in neural network pruning and show that pruned models can remain susceptible to MIAs [163]. In the interplay between XAI and privacy, the demonstrated tension between them has prompted research into methods that balance explainability with privacy guarantees [164]. Naidu et al. demonstrate that applying

DP can significantly degrade visual explanations in deep neural networks [36, 164]. Despite these challenges, privacy-preserving XAI approaches have emerged. Nori et al. integrate DP into Explainable Boosting Machines, achieving promising accuracy with formal privacy guarantees, although the approach remains model-specific [165].

In contrast to prior work, this is the first work to investigate how CFs can be exploited within a shadow-based MIA framework, positioning CFs as a primary source of adversarial knowledge. Moreover, our work is the first to examine the combined use of DP and AL as a defense strategy against MIAs. While prior studies have primarily applied AL to reduce labeling costs or to generate synthetic datasets, we integrate AL with DP and assess how this combination strengthens privacy in MLaaS settings. The objective is to offer a comprehensive analysis of their interaction and its impact on model utility, membership leakage, and CF quality.

4.3 Reference Scenario and Methodology

4.3.1 Reference Scenario: Shadow-based MIA

Figure 4.1 illustrates our reference scenario. Consider a DNN model f_θ trained on a dataset $\mathscr{D}$: $\mathbf{X} \in \mathbb{R}^d \rightarrow \mathbf{Y} \in \mathbb{R}^t$. The model f_θ takes as input a numerical query $\mathbf{x} \in \mathbf{X}$ and predicts an output $\mathbf{y} \in \mathbf{Y}$. Let θ denote the weight matrix and $\mathbf{b}$ the bias vector of the DNN. The vector $\mathbf{y} = f_\theta(\mathbf{x})$ represents confidence scores (probabilities), indicating the model's probability toward each class. The model f_θ is deployed as an MLaaS, and the model owner trains a CF explainer E. The CF explainer E is trained on the dataset $\mathscr{D}$ to generate CF instances. The MLaaS returns $E(\mathbf{x}), f_\theta(\mathbf{x})$ when queried by x. An attacker aims to determine whether specific data points belong to $\mathscr{D}$ via MIA by exploiting the MLaaS API's output, as described in [44].

Specifically, the attacker collects a shadow dataset by repeatedly querying the API and distributes it into several datasets $\{\mathscr{D}_1, \mathscr{D}_2, \ldots, \mathscr{D}_S\}$, each drawn from the same distribution $\mathscr{P}$ as $\mathscr{D}_{\text{train}}$. For each dataset $\mathscr{D}_s$, the adversary trains a set of *shadow models* $f_s(\cdot; \theta_s)$, aiming to approximate the behavior of the target model $f(\cdot; \theta)$. The adversary fully controls $\mathscr{D}_s$ and its membership status, so for every sample $x \in \mathscr{D}_s$, it is known that x is a *member*, while for $x \notin \mathscr{D}_s$ (but drawn from $\mathscr{P}$), it is a *non-member*. Each shadow model f_s returns a posterior vector $\boldsymbol{p}_s(x) = f_s(x; \theta_s) \in \mathbb{R}^K$, where K is the number of classes. Using the shadow models posteriors $\boldsymbol{p}_s(x)$ for both *member* and *non-member* samples, the adversary trains an *attack model* $g(\cdot; \phi)$ to infer membership:

$$g\big(\boldsymbol{p}_s(x); \phi\big) = \begin{cases} 1, & \text{if } x \in \mathscr{D}_s, \\ 0, & \text{otherwise.} \end{cases}$$

We evaluate the MIA across two configuration settings that differ in how the shadow dataset is constructed: i) No-CF setting and ii) CF-enabled setting. The collected dataset is then used to train the shadow models, which indicate whether the MLaaS provider offers CF explanations. In both cases, the attacker must first assemble a shadow dataset that approximates the distribution of the original training data. Specifically:

No-CF setting The attacker does not have access to CF explanations, we follow the standard assumption in the literature: the shadow dataset contains model outputs for inputs drawn from the same distribution as the training data.

$$\mathscr{D}_{\text{shadow}} = \{f_\theta(\mathbf{x}_i)\}_{i=1}^{M}, \tag{4.1}$$

CF-enabled setting. The attacker leverages CF explanations provided by the MLaaS system, queries the CF explainer E for a set of input instances, and receives both the model outputs and CF explanations, thereby forming a shadow dataset. Access to CFs provides the attacker with additional side information $E(x)$ that increases the mutual information between the model's outputs and the membership variable.

$$\mathscr{D}_{\text{shadow}} = \{E(\mathbf{x}_i),\, f_\theta(\mathbf{x}_i)\}_{i=1}^{M}. \tag{4.2}$$

Therefore, the shadow model trained on $\mathscr{D}_{\text{shadow}}$ resulting in attack model $g(p_s(x) \,\|\, \phi(E(x), x)$, Note that while baselines such as CF-distance threshold exist, we do not include them because their performance depends strongly on the chosen CF generator and on feature scaling/normalization, making fair reproduction and threshold calibration ambiguous.

4.3.2 Defense with Differential Privacy and Active Learning

To defend against MIA, we formulate two main objectives (i) to limit the model's tendency to memorize training data during training, and (ii) to reduce the amount and sensitivity of data exposed. Therefore, we propose a defense mechanism in which, during model training, the model owner leverages both AL and DP. Specifically, we employ AL, which is traditionally used to reduce annotation costs by selecting only the most informative samples, as a technique to train the DNN and reduce training data by constructing a smaller yet high-utility training data subset. Moreover, we employ DP, specifically DP-SGD at the model level, with rigorous privacy accounting to achieve a target (ε, δ)-DP budget and provide formal, instance-level privacy guarantees against MIAs. We refer to this approach as *with DP*, i.e., where AL and DP are jointly employed, and consider an alternative approach, namely, *without DP*, where DP is not employed.

The goal is to evaluate the advantage of jointly leveraging AL and DP as a defense mechanism to mitigate CF-based shadow-based MIAs. Furthermore, we evaluate both model utility and the quality of the generated CFs, with the ultimate aim of analyzing how the proposed defense affects predictive performance and CF quality.

To initiate the AL process, we construct a labeled subset $\mathscr{S} \subset \mathscr{D}$ through random sampling. Initially, a small subset $\mathscr{S}_0 \subset \mathscr{D}$ is randomly selected to serve as the starting point for labeling. At iteration t, the current model f scores unlabeled data from $\mathscr{D} \setminus \mathscr{S}$. An acquisition function (e.g., entropy or uncertainty sampling) measures how *informative* each sample might be for improving the model. We select the most informative subset $\mathscr{S}_t$ from the pool using the least confidence metric. After adding the samples in $\mathscr{S}_t$, we incorporate them into the training set, i.e., $\mathscr{S} \leftarrow \mathscr{S} \cup \mathscr{S}_t$. The updated training set is then used in the next model retraining step $t+1$. These steps are repeated until a stopping condition is met, or a maximum number of iterations on a validation set is reached. *To integrate DP into the model training*, we update the model parameters following DP-SGD, where each weight in θ is updated during back-propagation step, where gradients computed using standard optimizers (e.g., SGD, Adam), are clipped to a fixed norm C, and Gaussian noise is chosen to satisfy a target (ϵ, δ)-DP guarantee. We consider the following scenarios:

- *Baseline*: Standard training without any additional techniques serving as a reference model for evaluating the benefits of DP and AL for preventing MIA.
- *Only-AL*: Incorporates AL to strategically select training samples with the aim of reducing the required training dataset size while maintaining predictive performance.
- *Only-DP*: Integrates DP into the training process of the model. It assesses the trade-offs between privacy and performance when DP is applied.
- *DP-Post-AL*: Applies AL first to strategically select training samples, and DP is then subsequently applied during the training phase. This setup assesses whether applying DP after AL can improve privacy preservation.
- *AL-Guided-DP*: Follows the same incremental training process as the default AL model, but with DP incorporated during each training step.

4.4 Experimental Settings

Datasets. We consider two datasets for conducting our experiments, *EEG* [166] and *Indoor Location* (In-Location) [167]. We selected these two datasets for their suitability

for training DNN models in terms of performance and practical relevance, as well as for the variation in the dimensionality of their input features. The EEG dataset comprises 14,980 data records, 14 numerical features, and 2 classes. The dataset contains brainwave and signal readings, with the objective of training a model to predict whether an eye is open or closed. In-Location comprises 20,000 data records, 529 features related to Wi-Fi signals, and multiple labels, and is used to predict the building.

Model Architectures and Training Protocol. Preprocessing consists of removing missing values, removing outliers, and normalizing with a standard scaler. For both datasets, we split the data into 45% for training the target model, 45% for training shadow models, and 10% for validation. For the EEG *Baseline* model, we employed a DNN with 4 hidden layers containing 32, 16, 32, and 16 nodes, respectively. Each hidden layer was followed by a ReLU activation. The In-Location *Baseline* model was constructed with 7 layers, with 600, 700, 600, 300, 600, 300, and 128 nodes, respectively, followed by a ReLU activation function, and an additional dropout layer was used to prevent overfitting. These models were optimized using the Adam optimizer to minimize the cross-entropy loss, with a learning rate of 0.01.

Defense Mechanisms. For scenarios in which only DP was applied during training (i.e., *Only-DP*), we retained the default model architecture. Models were trained with privacy budgets (ε) of 1, 3, 5, and 10, with lower budgets corresponding to stronger privacy guarantees, and with δ = 1e-5 and a grad normalization of 1.5. In the *Only-AL* approach, the model during the first iteration is trained using 10% of the available training set. With each iteration, an additional 50 data records are added; the process is repeated for 50 iterations, and the model achieving the highest accuracy is selected. The training subset used to train the optimal model is then saved as the *Best AL Dataset*, on which the model achieves the best predictive performance. For model training in the *DP-Post-AL*, we use the *Best AL Dataset* and train models with DP.

Membership Inference Attack Setup. MIA is implemented by performing hyperparameter tuning on the number of shadow models (DNNs) and attack models (XGBoost) for their outputs on both datasets. After hyperparameter tuning, the XGB attack model is configured with 150 estimators, a maximum depth of 15, and a learning rate of 0.01. The shadow dataset is collected by querying a separate subset of the dataset via MLaaS. The results are presented as the average of 5 runs.

Counterfactual Explanation Setup. All CFs were obtained by using the Nearest Instance Counterfactual Explanations (NICE) [168] with the following criteria: Distance Metric of Heterogeneous Euclidean-Overlap Metric, Normalization of Min-Max, and the algorithm was set to optimize proximity and sparsity when using the *In-Location* dataset. CF quality was measured using well-known metrics, such as average proximity and sparsity. Each data point from the testing set was explained individually, and its proximity and sparsity values were used to compute the average proximity and spar-

sity across all data points. We gain insight into how effectively the CF explainer can produce *concise* (i.e., small-distance, few-feature changes) yet *impactful* (i.e., changing the predicted outcome) CFs under different privacy conditions.

Evaluation Metrics Setup. Relative to MIA, we propose two distinct metrics. Let $\mathscr{D}$ denote the full dataset with $|\mathscr{D}| = N$ records. Let $\mathscr{S} \subseteq \mathscr{D}$ be the set of records used to train a given model (e.g., depending on whether AL is applied). Let $\hat{\mathscr{S}} \subseteq \mathscr{D}$ be the set of records that an adversary predicts as *members* via MIA.

Micro-level defended record ratio (per model) At the micro level, defended records quantify the fraction of *true members* whose membership remains hidden, i.e., training points that the attacker *fails* to infer as members, where $\mathscr{D}_{\text{protected}} = \mathscr{S} \setminus \hat{\mathscr{S}}$ and true positive rate (TPR) among all actual members, the fraction the attacker correctly detects.

$$Micro = \frac{|\mathscr{D}_{\text{protected}}|}{|\mathscr{S}|} = 1 - \text{TPR} \tag{4.3}$$

Macro-level defended record ratio (overall exposure) At the macro level, we additionally account for records that were *never* used for training and are therefore automatically unexposed to MIA. It describes the fraction of all records whose membership is not successfully revealed, i.e., records that are either saved (never trained) or protected despite being trained, where $\mathscr{D}_{\text{saved}} = \mathscr{D} \setminus \mathscr{S}$.

$$Macro = \frac{|\mathscr{D}_{\text{saved}}| + |\mathscr{D}_{\text{protected}}|}{|\mathscr{D}|} = 1 - \frac{|\mathscr{S}|}{|\mathscr{D}|} \cdot \text{Recall}_{\text{member}}. \tag{4.4}$$

4.5 Experimental Results

We present the numerical results along four complementary dimensions. First, we evaluate predictive performance under the different defense mechanisms. Second, we assess MIA effectiveness with and without CFs to quantify the additional privacy risk introduced by explanations and the robustness of the defenses. Third, we report micro- and macro-level defended record ratios to provide a granular view of privacy exposure. Finally, we analyze the impact of the defense strategies on CF quality.

4.5.1 Model Predictive Performance

Table 4.1 summarizes the predictive performance of the different models across all approaches and both datasets. Across both datasets, as expected, *Baseline*, which does not implement any privacy measure, achieves the highest performance across all metrics

Table 4.1. Predictive performance (Accuracy, Precision, Recall) under privacy budgets ε for EEG and In-Location. *Baseline and Only-AL are trained without DP and therefore have no associated ε.*

Data	Method	Accuracy				Precision				Recall			
		ε				ε				ε			
		1	3	5	10	1	3	5	10	1	3	5	10
EEG	Baseline	96				96				96			
	Only-AL	95				95				95			
	Only-DP	82	87	88	90	82	87	88	90	82	87	88	90
	AL-Guided-DP	82	87	90	91	82	88	90	92	82	87	90	91
	DP-Post-AL	66	72	83	86	69	74	84	86	66	72	83	86
In-Location	Baseline	100				100				100			
	Only-AL	100				100				100			
	Only-DP	40	75	81	86	24	79	85	88	40	75	81	86
	AL-Guided-DP	58	57	56	91	46	56	67	92	58	57	56	91
	DP-Post-AL	47	45	52	53	23	29	50	54	47	45	52	53

(e.g., accuracy, precision, and recall of 96% and 100% across EEG and In-Location, respectively), while *Only-AL*, which leverages only AL to reduce the training set size while ensuring performance matches that of *Baseline*, closely follows (*Only-AL* achieves this level of performance using only a subset of the dataset).

When DP is incorporated as part of the defense mechanism, namely in *Only-DP*, *AL-Guided-DP*, and *DP-Post-AL*, a consistent pattern emerges across both datasets. Stronger privacy guarantees (lower ε) lead to a noticeable performance degradation, whereas performance steadily improves as ε increases, corresponding to weaker privacy constraints.

In the EEG dataset, *AL-Guided-DP* and *Only-DP* exhibit comparable performance and consistently outperform *DP-Post-AL*, which yields the lowest scores. Specifically, *AL-Guided-DP* achieves accuracy between 82% and 91%, with precision ranging from 82% to 92%, and recall following a similar pattern. *Only-DP* attains accuracy between 82% and 90% across different ε values. For In-Location, DP exerts a stronger impact on performance than in EEG, although the overall trend remains similar. *Only-DP* and *AL-Guided-DP* alternate in achieving the best predictive results, with comparable performance levels. Importantly, *AL-Guided-DP* attains this performance while relying on fewer training samples (since it employs AL), which is beneficial from a privacy perspective as it reduces potential exposure to MIAs.

Regarding the subset size selected by AL, for EEG, *Only-AL* (and *DP-Post-AL*) uses

77.45% of the data. *AL-Guided-DP* selects 78.97%, 74.48%, 56.49%, and 66.98% for $\varepsilon = 1, 3, 5$, and 10, respectively. For In-Location, *Only-AL* uses only 10% of the data, while *DP-Post-AL* requires 29.13%, 21%, and 40.35% for $\varepsilon = 1, 3$, and 5. Overall, AL can substantially reduce the required training data; however, when combined with DP, larger subsets are often selected, reflecting the additional uncertainty introduced by DP noise.

Takeaway 1: AL preserves utility, matching baseline performance while using fewer training samples. Introducing DP leads to a utility loss at low ε. While *AL-Guided-DP* does not substantially outperform *Only-DP*, it achieves comparable performance with fewer samples. In contrast, *DP-Post-AL* consistently yields the lowest performance across both datasets.

4.5.2 Membership Inference Attack

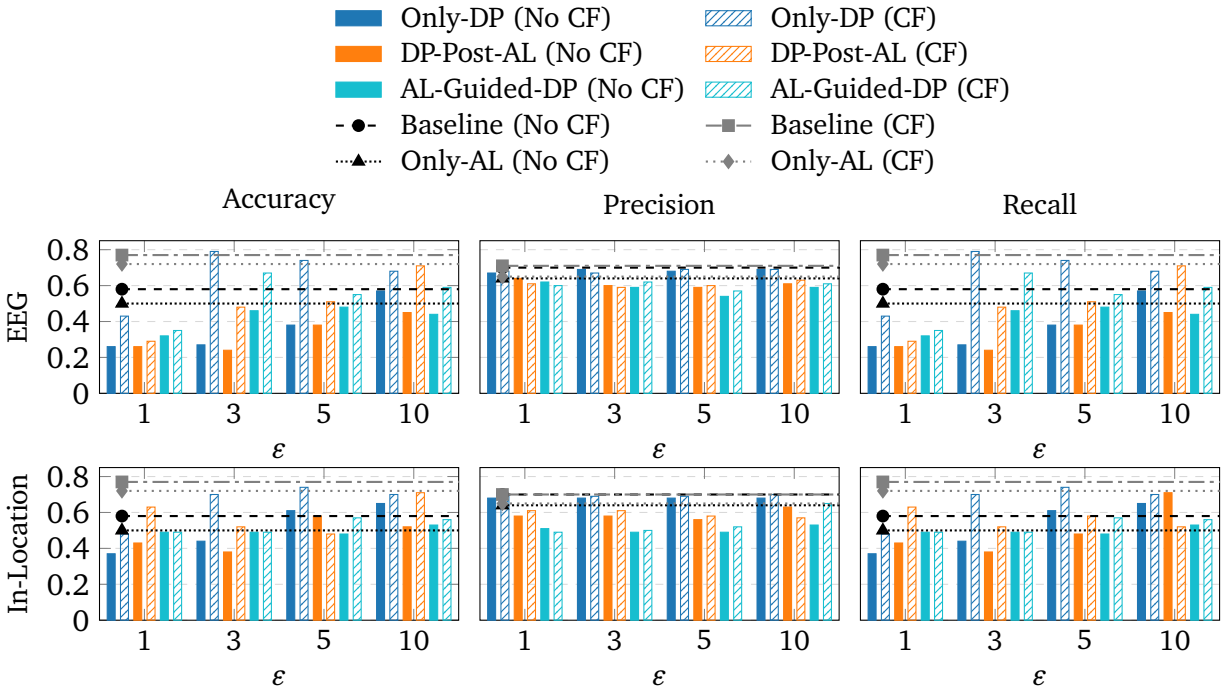


Figure 4.2. Accuracy, precision, and recall across baselines, privacy budgets ε with and without using CFs for MIA.

Figure 4.2 summarizes MIA performance across all approaches and privacy budgets ε of 1,3,5 and 10 (lower values indicate stronger privacy protection). Results show a clear trend across all cases: *leveraging CFs to perform MIA allows for substantially*

amplifying attack success. For instance, in the EEG dataset, *Baseline (No CF)* attains 58% accuracy/recall and 70% precision while *Baseline (CF)* attains 77% of accuracy/recall, and 71% of precision. Another example is *Only-AL (No CF)*, accuracy/recall increases from 50% to 72% (precision 64-65%) when leveraging CFs (*Only AL (CF)*. When employing DP, the impact of CFs on MIA success becomes even more evident. Comparing *Only-DP (no CF)* to *Only-DP (CF)*, accuracy/recall increases from 26% to 43% at $\varepsilon = 1$, from 27% to 79% at $\varepsilon = 3$, from 38% to 74% at $\varepsilon = 5$, and from 57% to 68% at $\varepsilon = 10$ when CFs are leveraged, while precision remains essentially close (67–69%). Results in In-Location show the same trend: leveraging CFs consistently strengthens MIA. For example, *Only-DP*, leveraging CFs raises accuracy/recall from $37, 44, 61, 65\%$ to $48, 70, 74, 70\%$ at $\varepsilon = 1, 3, 5, 10$. For *DP-Post-AL*, accuracy increases from $43, 38, 58, 52\%$ to $63, 52, 48, 71\%$, except at $\varepsilon = 5$. *AL-Guided-DP* shows little change at $\varepsilon = 1, 3$, a moderate gain at $\varepsilon = 5$, and at $\varepsilon = 10$ its largest precision increase (53% to 65%) with a slight accuracy rise (53% to 56%). *Takeaway 2:* While DP introduces a utility–privacy trade-off (Takeaway 1), leveraging CFs significantly weakens privacy protection across all settings a consistently amplifies MIA performance, often dramatically increasing attack success, even under DP. In several cases, CFs nearly double accuracy/recall at moderate privacy budgets, demonstrating that explanations can substantially undermine the privacy gains achieved by DP alone.

We now quantify the impact of ε on the defense against MIA. As expected, a larger ε weakens privacy: for example, on EEG, *Only-DP (No CF)* accuracy/recall range from 26% to 57% and from 43% to 68% *Only-DP (CF)*. Comparable monotonic trends appear in *DP-Post-AL* and *AL-Guided-DP*. We move to comparing *Only-AL to DP-Post-AL* in order to better understand the effect of employing DP after AL. For EEG, *Only-AL (No CF* has an accuracy/recall $\approx 50\%$, which reaches 72% when CFs are exposed. In contrast, *DP-Post-AL* starts from lower attack success at small privacy budgets (26% and 34% at ε 1 and 3), indicating stronger protection than *Only-AL*. However, CFs significantly erode this advantage: accuracy rises to 29% and 48% at $\varepsilon = 1$ and 3, and at larger ε of 10 reaching 71%. For In-Location, Similar trends are observed.

Takeaway 3: Consistent with the utility–privacy trade-off observed in Takeaway 1, increasing ε systematically weakens protection and enables stronger MIAs. However, beyond this expected trend, releasing CFs introduces an additional attack surface that substantially boosts member detection, primarily by increasing recall. Even when DP reduces attack success at low privacy budgets, this advantage is significantly diminished once CFs are exposed. While combining AL with DP can partially mitigate attack success, this comes at the cost of retaining more training data.

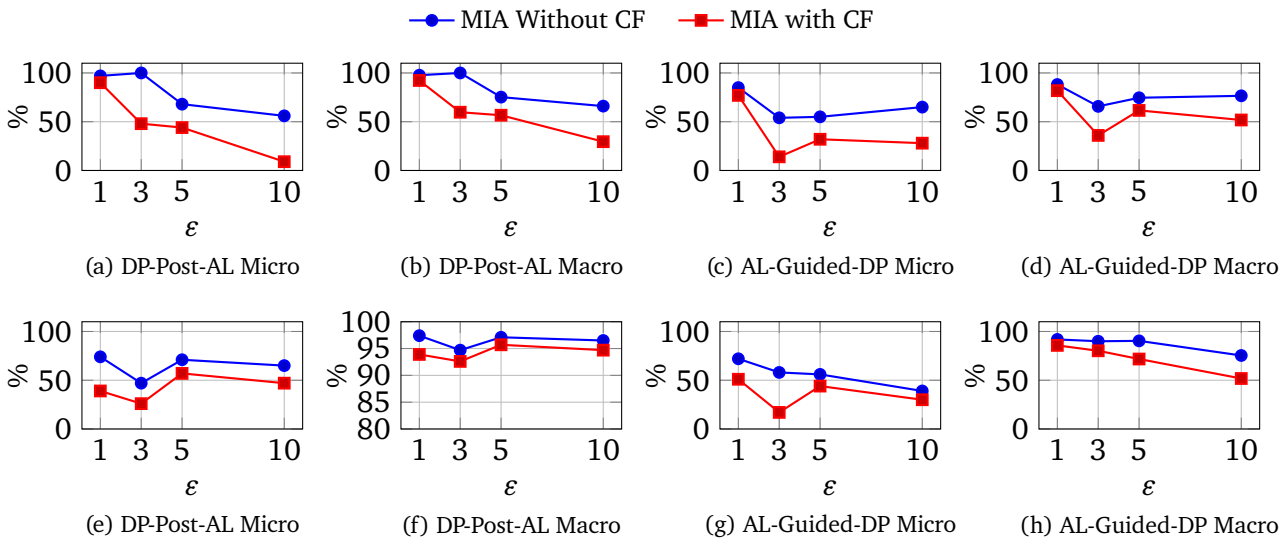


Figure 4.3. Percentage of micro & macro defended records under *DP-Post-AL* & *AL-Guided-DP*, with & without CFs. Top row: EEG; bottom row: In-Location.

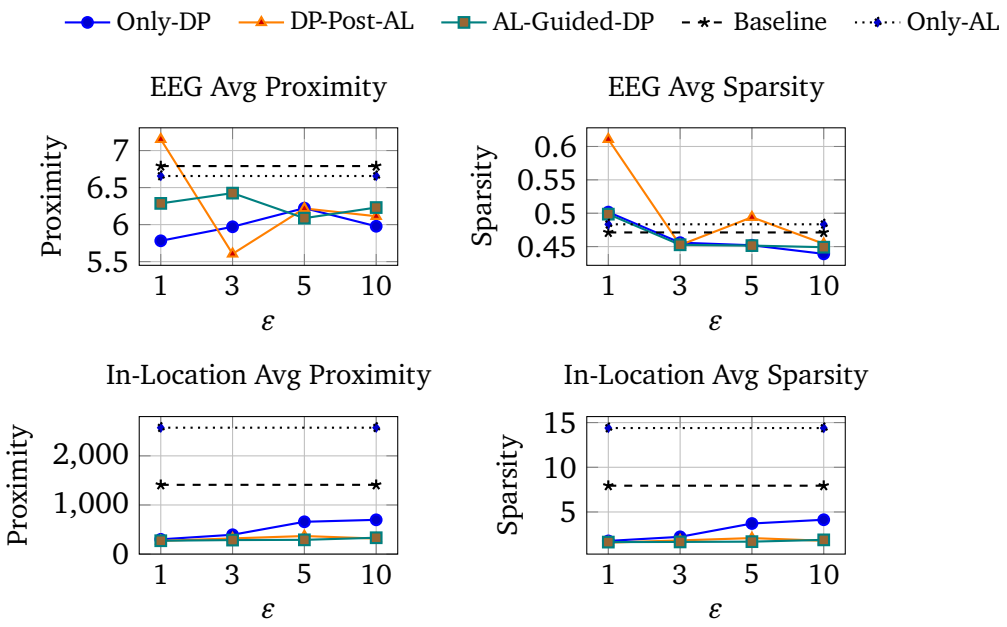


Figure 4.4. Sparsity and proximity of extracted CFs for EEG and In-Location across differential privacy budgets (ε).

4.5.3 Defense at Micro- and Macro-Level

In this part, we focus our analysis on the two main approaches, *DP-Post-AL* and *AL-Guided-DP*, and examine their behavior through the micro and macro defended record ratios, as presented in 4.3. The micro metric captures the fraction of *true training records* whose membership remains hidden, directly quantifying how well individual members are protected. In contrast, the macro metric reflects overall dataset exposure by additionally accounting for records that were never used for training and are therefore automatically unexposed to MIA. Consequently, macro values can remain high even when the protection of actual training points deteriorates.

Across all settings, leveraging CFs consistently reduces the proportion of defended points at both levels, with the effect being particularly pronounced at the micro level. On EEG, under *DP-Post-AL*, micro protection drops from 68% to 44% at $\varepsilon = 5$, and from 56% to 9% at $\varepsilon = 10$ when CFs are leveraged. A similar trend is observed for *AL-Guided-DP*, where micro defended rates decrease from 54% to 14% at $\varepsilon = 3$. These results indicate that CFs substantially increase correct member inference, directly exposing training data.

At the macro level, defended ratios also decrease when CFs are used, but the reduction is often less because macro protection includes automatically saved (non-trained) records. As a result, an approach may appear robust at the macro level while failing to adequately protect actual training members. For instance, in EEG at $\varepsilon = 3$, *AL-Guided-DP* decreases from 65.74% to 35.95% at the macro level, whereas the corresponding micro-level performance is considerably stronger. For EEG, *DP-Post-AL* achieves higher macro defended ratios than *AL-Guided-DP* at small ε, consistent with stronger privacy guarantees at low budgets, although this advantage is not uniform at larger ε. Similar patterns emerge for In-Location, in which CFs consistently erode micro-protection across ε.

Takeaway 4: Beyond amplifying MIA accuracy (Takeaway 2) and weakening protection as ε increases (Takeaway 3), CFs critically erode *micro-level* protection, directly exposing true training members. While macro-level defended ratios may remain relatively high due to automatically saved records, the micro analysis reveals a substantial loss of protection for the most sensitive subset, the actual training data. These results indicate that aggregate exposure metrics can mask individual-level vulnerability, and that releasing CFs can significantly undermine the practical privacy guarantees of DP-based approaches.

4.5.4 CF Quality Analysis

Figure 4.4 reports proximity and sparsity across scenarios. For EEG, proximity remains stable under DP (5.78–6.22 for *Only-DP*; 5.60–7.15 for *DP-Post-AL*), indicating that DP does not substantially degrade CF quality. Notably, *AL-Guided-DP* exhibits slightly improved proximity compared to *Only-AL*, which may be attributed to its use of a larger training subset. By incorporating more data points during training, *AL-Guided-DP* may learn more stable decision boundaries, resulting in CFs that require smaller input changes. For In-Location, proximity values are generally higher than the *Baseline*, reflecting the dataset's higher dimensionality. Sparsity follows a similar pattern: under tight privacy budgets (e.g., $\varepsilon = 1$), DP slightly reduces sparsity, indicating stronger noise effects. In summary, proximity remains largely stable across privacy levels, whereas sparsity is mildly affected and exhibits greater sensitivity in high-dimensional settings.

Building on our analysis, *Takeaway 5:* While DP introduces a utility–privacy trade-off (Takeaway 1) and CFs substantially amplify MIA success (Takeaways 2–4), the impact of DP on CF quality remains limited. Proximity remains stable across privacy budgets, and sparsity is only mildly affected, particularly in high-dimensional datasets.

These findings reveal a fundamental tension: explanations can significantly weaken privacy guarantees, yet privacy-preserving mechanisms can be applied without substantially degrading explanation quality. This creates a critical design challenge for MLaaS systems, which must carefully balance predictive utility, privacy robustness, and explanation fidelity. Beyond technical performance metrics, future work should examine the practical usefulness of CFs under privacy-preserving settings through empirical user studies. In particular, it is important to assess whether explanations that remain stable under DP also retain their interpretability and decision-support value for end users in high-stakes domains.

4.6 Conclusion

In this chapter, we investigate how counterfactual explanations (CFs) expand the attack surface of Machine Learning as a Service (MLaaS) systems by enabling effective shadow-based membership inference attacks (MIAs). We propose a defense framework that integrates Differential Privacy (DP) and Active Learning, and analyze the resulting interplay between privacy protection, predictive performance, and explanation quality. Our results demonstrate that while DP can mitigate membership leakage, its protection can be substantially weakened when CFs are exposed. At the same time, we show that privacy-preserving training does not necessarily degrade explanation quality, revealing a fundamental tension between transparency and privacy resilience.

Chapter 5

Differential Privacy for Anomaly Detection: Analyzing the Trade-off Between Privacy and Explainability

In this chapter, we examine the interplay between DP and explainability in anomaly detection (AD). We investigate how adding privacy-preserving noise affects both the predictive performance of AD models and the quality of their SHAP-based explanations. In particular, we analyze changes in explanation fidelity, attribution stability, and feature-importance patterns across different datasets, models, and privacy budgets. This chapter is based on the content available in the following work:

1. **Fatima Ezzeddine**, Mirna Saad, Omran Ayoub, Davide Andreoletti, Martin Gjoreski, Ihab Sbeity, Marc Langheinrich, and Silvia Giordano. Differential Privacy for Anomaly Detection: Analyzing the Trade-off Between Privacy and Explainability. World Conference on eXplainable Artificial Intelligence (xAI 2024).

5.1 Introduction

Within the realm of data-driven decision-making, anomalies, which are data points exhibiting statistically significant deviations from expected patterns, have a multifaceted impact. Anomalies indicate errors or inconsistencies in the data, but they can also reveal novel or critical situations. Deviations are subsequently flagged for further investigation as they can be highly informative, often signaling underlying issues or emerging trends. For instance, in the context of cyber-security, an anomaly might indicate a secu-

rity breach or an attempted attack. In healthcare, it could highlight rare illnesses. The timely identification of anomalies is crucial for maintaining security, efficiency, and safety in various fields, such as cyber-security, healthcare, network monitoring, and transportation [169, 170]. Therefore, developing highly effective Anomaly Detection (AD) systems is of paramount importance, as they represent a critical tool to address the challenge of detecting these anomalies [94]. By leveraging diverse statistical and ML techniques, AD establishes a statistical baseline for normal behavior within a dataset.

Recently, a main requirement of AD systems emerged, in addition to that of high predictive performance, which is the need of providing stakeholders with relevant information on why and how a specific data point is considered an anomaly, such as to enhance the transparency of AD systems, with the final aim of fostering trust in these systems [171, 172, 173]. In addition to that, privacy guarantees are also essential for AD in scenarios, as data owners may delegate the AD task to a third party in possession of the technical expertise needed for effective AD. This third-party data access introduces privacy concerns, especially when dealing with sensitive data such as healthcare information that includes confidential patient medical histories [174, 175, 176]. In this context, we are confronted with the challenge of ensuring both transparency and privacy in AD systems. To ensure the privacy of data and the transparency of decisions, DP [22] and xAI have been the main methods attracting the attention of the research community, respectively.

A conflict however arises between ensuring transparency (through xAI) and privacy (through DP) due to their opposing goals. XAI aims to provide insights into model behavior for transparency, while privacy-preserving solutions obscure data to prevent data leakage. Specifically, xAI techniques are proposed to demystify the inner workings of complex ML models and AD through different types of explanations, such as, e.g., FA, which scores the contribution and impact of each feature on the model's output, permitting data owners to identify which features in the data were most influential in identifying an anomaly. DP, on the contrary, works by injecting calibrated noise into the data before it is released, introducing a quantifiable privacy guarantee, and allowing control of the level of information revealed about individual data points.

The intersection of xAI and privacy-preserving techniques presents complex, nuanced challenges. Therefore, there is a critical need to precisely quantify how privacy-preserving techniques, such as DP, affect the explainability of ML or AD systems [177]. In this chapter, we investigate the complex relationship among DP, AD, and XAI. As a xAI framework, we rely on SHapley Additive exPlanations (SHAP) [178]. We focus on SHAP since it is widely applied for feature importance in tabular data. To investigate this interplay in the context of AD, and address RQ2, we formulate two sub-questions and address them as follows:

1. *To what extent does increasing DP noise affect the fidelity and stability of SHAP values in AD?* We explore XAI in DP-AD, specifically, we investigate SHAP with a focus on understanding how the SHAP values of AD models are affected by varying levels of DP noise on AD. By extensively analyzing quantitatively and qualitatively the output of SHAP, we shed light on the potential trade-offs between privacy protection and model explainability.

2. *To what extent does the application of DP impact the performance and explainability of AD algorithms designed for specific anomaly types (local vs. global)?* Leveraging the inherent specialization of AD algorithms towards distinct anomaly types (local vs. global), this research delves into the potential for DP and its impact on their performance and explainability under varying privacy constraints.

AD provides a particularly suitable case study for analyzing the privacy-explainability trade-off investigated in this thesis. First, AD is commonly deployed in high-stakes, privacy-sensitive domains, where abnormal events may indicate attacks, rare medical conditions, failures, or safety-critical situations. In these contexts, predictive performance alone is insufficient: stakeholders need explanations to understand why a given instance is flagged as anomalous, while data owners require privacy guarantees given the sensitive nature of the underlying data. Second, AD often relies on unsupervised or semi-supervised settings, where labels are scarce, and explanations can play an important role in validating model outputs. This makes AD a representative setting in which the tension between transparency and privacy becomes particularly visible. Within the broader scope of this thesis, this chapter complements the previous analyses on CFs and privacy attacks by studying a different form of explanation, namely feature attribution (FA) explanations through SHAP, and a different learning setting, namely AD. Therefore, the chapter broadens the empirical investigation of the privacy-explainability interplay beyond CF-based attacks and defenses, showing that the trade-off also arises when privacy-preserving mechanisms are applied to models whose explanations are used for diagnostic and decision-support purposes.

The chapter is organized as follows. Section 5.2 discusses related work. Section 5.3 details the problem and objectives. Section 5.4 presents the experimental setup and evaluation settings. Section 5.5 showcases the obtained results and analyzes their implications, offering valuable insights for practitioners navigating the trade-off between privacy and explainability in AD models. Section 5.6 concludes the chapter.

5.2 Related Work

In this section, we first discuss studies that have applied either privacy-preserving or explainability techniques to AD and then studies that have investigated the impact of privacy on the model's explainability. The related works highlight the growing tension between privacy and explainability. While research has explored privacy-preserving techniques and explainable models in many contexts, their intersection with AD remains largely unaddressed. This is particularly crucial because AD often handles sensitive data and requires different considerations than traditional classification and regression tasks. In this chapter, we analyze this intersection to reveal insights into the trade-offs involved, guiding the development of future robust, explainable, and privacy-preserving AD methods that empower informed decision-making.

5.2.1 Privacy-Preserving Anomaly Detection

AD methods play a crucial role in diverse sectors such as finance, healthcare, transportation, and smart grids [170, 176, 179, 180]. Most existing methods rely primarily on unsupervised ML techniques (e.g., [94, 181, 182, 183, 184]), with some supervised approaches as well (e.g., [185]). Numerous AD algorithms have been proposed in the literature, and each of them comes with strengths and weaknesses and is suitable for specific contexts and data types. For example, Isolation Forest (iForest) and Local Outlier Factor (LOF) are well suited for tabular data, Deep Learning-based AD with auto-encoders is suitable for images and time series data [94, 96, 186]. Several recent studies on AD have proposed employing privacy-preserving techniques to address the critical challenge of balancing effective AD with individual data protection. These techniques mitigate public concern over data sharing, reduce the risk of re-identification from anonymized data, facilitate compliance with data privacy regulations like the General Data Protection Regulation (GDPR), and foster collaboration between organizations, leading to more comprehensive and effective AD across various sectors [176, 187]. To achieve AD while safeguarding data privacy, several methods have been employed such as DP [188, 189, 190], homomorphic encryption and training on encrypted data [191, 192, 193], Secure Multi-Party Computation [194], and even novel, custom-designed approaches [195], such as generating synthetic data [175], or approaches for specific cases such as body movement and power systems [190, 196].

Specifically, among the various techniques, several studies have explored the integration of AD algorithms and DP, e.g., [188, 197]. Notably, Du et al. [188] show how DP can improve the utility of AD and novelty detection, with a focus on detecting poisoning samples in backdoor attacks. Jiang et. al [170] propose a privacy-preserving social network model that utilizes restricted local DP to sanitize user information collection.

Moreover, Chukkapalli et al. in [198] propose an approach for privacy-preserving AD in smart farming by adding noise to individual farm data. Giraldo et. al [189] examine how AD can be combined with DP to provide robust privacy and security for individuals. In addition to Degue et al. [176] DP is employed with AD in correlated data to analyze the trade-off between privacy level and detection accuracy of multivariate Gaussian signals.

Other studies have used homomorphic encryption to train AD models with encrypted data [192, 194, 199]. For instance, Alabdulatif et al. [192] focus on cloud-based models and propose an AD model that preserves data privacy with reliance on ciphertext, while, Mehnaz et al. [191] present a framework for efficient AD on real-time-series encrypted data. Guo et al. [199] propose an AD scheme for encrypted video bitstreams with format-compliant encryption. Zhang et al. [194] propose a semi-centralized privacy-preserving secure multiparty computation protocol for the PCA-based AD. Other approaches have been proposed based on synthesizing data and generating samples, such as [175]. Mayer et al. [175] analyze many approaches to generating synthetic data and assess the utility of the resulting datasets for AD in supervised, semi-supervised, and unsupervised settings. Prioritizing privacy has been the main focus of these recent AD studies. However, explainability in AD is also emerging as a crucial area of research, which we explore in the following subsection.

5.2.2 Explainable Anomaly Detection

Yuan et al. [200] discuss crucial challenges of AD such as trustworthiness, explainability, and robustness. In this context, several studies explore methods for extracting valuable insights for anomaly explanation [172] and, as detailed in [201], for other applications. For instance, Panjei et al. [201] categorize various types of explanations and analyze existing techniques for interpreting anomalies, paving the way for more meaningful AD analysis. Roshan et al. [202] demonstrate the use of XAI to explain the results of the autoencoder AD model, and in [173] leverages the Kernel SHAP method to explain network anomalies. Moreover, Ravi et al. [203] explore the feasibility and compare the performance of several state-of-the-art XAI frameworks on Convolutional Autoencoders for building AD systems in the visual domain. Finally, Tritscher et al. [204] highlight the growing interest in categorizing and analyzing these explainable methods based on their access to training data and the specific AD model. These works have demonstrably investigated interpretability and leveraged XAI techniques for AD, but have not considered privacy concerns.

5.2.3 Impact of Privacy on Explainability

Recent years have seen a surge in studies investigating the interplay of privacy with XAI [164, 205, 206] as both interpretability and privacy represent a requirement for deploying ML models and datasets. Some studies investigate the inherent trade-off between these concepts, while others propose novel methods to generate privacy-preserving explanations [207, 208, 209]. Bozorgpanah et al. [205] investigate the impact of various privacy-preserving techniques such as masking, and DP noise addition on the effectiveness of regression-based explainability methods utilizing Shapley values and show different behaviors in Shapley for different models by computing correlation metrics. Also, the authors in [164] show the impact of DP on the interpretability of Deep Neural Networks particularly in medical imaging application classification, and show significant visual differences in explanation with DP. Another study [210] investigates the use of example-based explainability models for retinal image analysis. The authors propose leveraging Generative Adversarial Networks (GANs) to generate synthetic examples that provide explanations for model predictions while preserving the privacy of the original retinal images. Nori et al. [165], a method for adding DP to Explainable Boosting Machines enables the training of interpretable classification and regression models with state-of-the-art accuracy while preserving privacy. Harder et al. [211] addresses the challenge of balancing interpretability and privacy in ML models by proposing a novel approach using simple models with locally linear maps to approximate complex models. This method achieves high classification accuracy while providing differentially private explanations for the classifications. Montenegro et al. [207] propose privacy-preserving GANs for privatizing case-based explanations in classification tasks. This GAN incorporates a counterfactual module, enabling the generation of both factual and counterfactual explanations while safeguarding privacy. Moreover, Jetchev et al. [209] introduces a novel, privacy-preserving algorithm for calculating Shapley values on decision tree ensembles within a secure multi-party computation framework, ensuring data privacy. Veugen et al. [208] propose the generation of privacy-friendly explanations by leveraging local foil trees.

5.3 Problem Formulation and Approach

We aim to investigate the effectiveness of AD techniques, specifically iForest and LOF, in identifying two types of anomalies within a dataset: local and global outliers. This investigation is conducted within a privacy-preserving setup, where we employ DP techniques. Crucially, we also aim to analyze how the SHAP explanations of the AD models change when DP is employed. Estimating these explainability changes will allow us to assess the impact of the privacy-preserving mechanisms on the reasoning behind the

identified anomalies. To achieve our goals, we will employ Privacy-Preserving AD with DP on the training data level. This means we will introduce calibrated noise directly into the training data before applying the AD algorithms. After injecting noise into the data, we employ SHAP to analyze how features contribute to changes in anomaly scores under different ε-DP guarantees. Figure 5.1 summarizes the steps followed for the con-

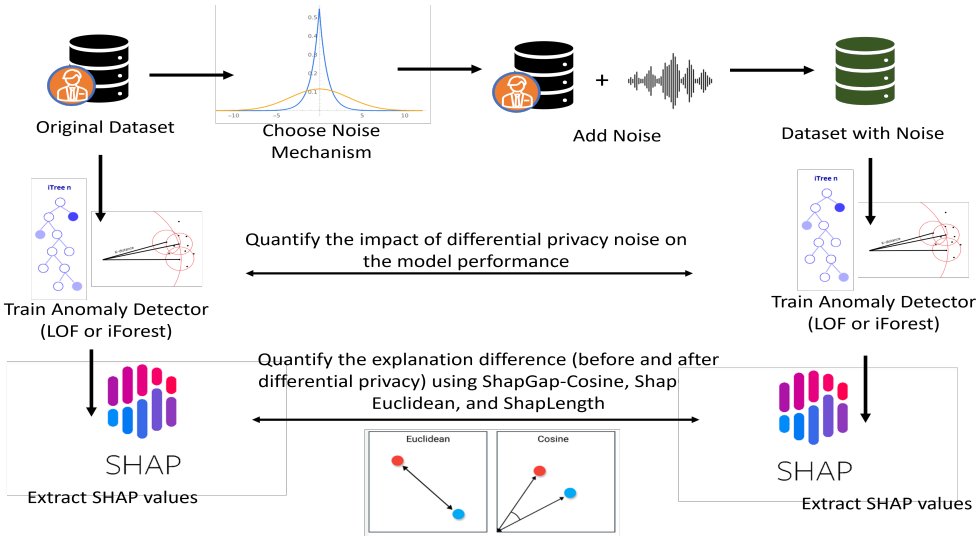


Figure 5.1. Overall scheme of the experimental setup

ducted analysis. Let D be a dataset $\mathscr{D}$: $\mathscr{X} \in \mathbb{R}^d \in \mathbb{R}$. The AD algorithm is trained on D in an unsupervised manner to estimate the anomaly score y_i for each datapoint $x_i \in \mathscr{X}$. To ensure a robust level of privacy protection measured by ε-differential privacy (ε-DP), we aim for small values of ε to minimize the influence of individual records on the overall performance. We introduce noise into the data using either a Laplacian or Gaussian distribution, with varying ε values (0.01, 0.1, 1, and 5). Subsequently, we retrain the AD models on the data augmented with noise. This process enables us to assess the impact of DP on both the performance and explainability of the models. Specifically, we compare the AD performance of models trained on noisy and original data using diverse metrics, focusing on understanding how DP affects the features driving AD with SHAP. Moreover, we explore the trade-off between privacy and explainability by employing SHAP.

5.4 Experimental Setting

This section describes the settings of our experiments, including the datasets, the procedures to train and evaluate the AD models, and the metrics used to quantify the

change in explainability.

Datasets Description We consider three datasets that are widely used for AD:

1. Mammography [212]: The mammography dataset corresponds to radiological scans to diagnose breast cancer. It consists of 6 features and 11183 records, of which 10923 are non-anomalous and 260 are anomalous.

2. Thyroid dataset [213]: The thyroid dataset corresponds to thyroid diseases. It comprises 21 features and contains 7,200 records, of which 6,666 are non-anomalous and 534 are anomalous.

3. Campaign (bank) dataset [213] includes banking information about individuals. It has 62 features and contains 41188 records, of which 36548 are non-anomalous and 4640 are anomalous.

Anomaly Detection Model training We perform hyperparameter tuning of the AD models using a Grid search with cross-validation (for a separate subset of the data). The hyperparameters that undergo an optimal selection aremax_features, n_estimators for iForest, and n_neighbors for LOF. It is important to emphasize that all the datasets are labeled. However, as iForest and LOF are unsupervised AD algorithms, we use the labels only for evaluating their performance, as illustrated in Figure 5.1.

Evaluation metrics

AD Performance Metrics We evaluate the AD model's predictive performance and output consistency. As for the former, we use precision, because it focuses on the ratio of true anomalies identified among all flagged data points, and Area under the ROC curve (AUC), because it is suitable for imbalanced datasets as metrics. Precision measures the proportion of true positives among all predicted positives, while AUC evaluates the model's ability to distinguish between positive and negative classes. As for the latter, we compute the fidelity score, which measures the degree of agreement between the predictions of two different models on the same set of inputs (i.e., how closely the outputs of one model mirror those of another). In our case, the fidelity score quantifies the agreement between the AD model's outputs before and after applying DP (as shown in Figure 5.1).

Quantitative analysis of SHAP difference metrics To quantify SHAP value changes between scenarios with and without DP, we use three recently proposed metrics: the

ShapGAP-Euclidean and ShapGAP-Cosine distances [214], and ShapLength [215]. While initially proposed for comparing SHAP values between black-box models and their corresponding surrogate white-box models, we leverage these metrics for a different purpose. Specifically, we compute the *ShapGAP* and *ShapLength* between the SHAP values of data points before and after adding DP noise, with the aim of quantifying the impact of DP on the explainability of AD models. More details about these metrics are provided in the following:

1. **ShapGAP-Euclidean Distance [214] (Eq. 5.1)**: ShapGAP-Euclidean provides a magnitude-based measure of the difference between the SHAP values generated for the same data point from two models S (without and with DP) across n data points x_i of a dataset D. It is useful for understanding the overall magnitude of changes.

$$ShapGAP_{L2}(D) = \frac{1}{n}\sum_{i}^{n} ||S_{\text{without DP}}(x_i) - S_{\text{with DP}}(x_i)||_2 \tag{5.1}$$

2. **ShapGAP-Cosine Distance [214] (Eq. 5.2)**: ShapGAP-Cosine computes the magnitude and directional relationship of the difference between two SHAP values of the same point from two models S (without and with DP) across n data points x_i of a dataset D. The ShapGAP-Cosine can range between 0 and 2, where higher values indicate higher dissimilarity. It is useful for capturing how similar the directions are, even if the magnitudes differ.

$$ShapGAP_{Cos}(D) = \frac{1}{n}\sum_{i}^{n}(1 - \frac{S_{\text{without DP}}(x_i) \cdot S_{\text{with DP}}(x_i)}{||S_{\text{without DP}}(x_i)||_2 ||S_{\text{with DP}}(x_i)||_2}) \tag{5.2}$$

3. **ShapLength [215]**: ShapLength is a model-agnostic and computationally efficient metric for assessing how human-understandable a model is. It builds upon the p%-complete explanation property, which finds the smallest set of features whose SHAP values sum exceed a defined threshold. Shap Length represents the number of features included in this p%-complete explanation. A higher ShapLength indicates a model that relies on a larger number of features or complex interactions, making it harder to explain and interpret.

5.5 Experimental Results

In this section, we quantitatively evaluate the impact of DP on the effectiveness (Subsection 5.5.1) and explainability of the AD models. Explainability is assessed through quantitative and qualitative evaluations, in subsection 5.5.2 and 5.5.4, respectively.

5.5.1 Impact of Differential Privacy on Anomaly Detection Models

Table 5.1. AUC and precision of iForest across the mammography, thyroid, and bank datasets, without DP and with varying DP-ε values, considering the two noise-adding mechanisms: Gaussian and Laplace.

Dataset	Metric	Without Privacy	Laplace				Gaussian			
			ε				ε			
			5	1	0.1	0.01	5	1	0.1	0.01
mammography	AUC	74	73	72	66	53	76	72	69	54
	Precision	90	91	90	88	84	92	90	89	85
thyroid	AUC	89	54	56	51	50	54	53	53	48
	Precision	90	58	60	56	55	58	58	58	53
bank	AUC	64	58	57	52	52	57	56	51	49
	Precision	68	64	63	58	58	63	62	58	55

Table 5.2. AUC and precision of LOF across the mammography, thyroid, and bank datasets, without DP and with varying DP-ε values, considering the two noise adding mechanism: Gaussian and Laplace.

Dataset	Metric	Without Privacy	Laplace				Gaussian			
			ε				ε			
			5	1	0.1	0.01	5	1	0.1	0.01
Mammography	AUC	74	74	74	74	66	74	74	73	70
	Precision	91	91	91	91	88	91	91	91	89
Thyroid	AUC	56	56	56	53	51	56	57	53	51
	Precision	60	60	60	58	56	60	61	58	56
Bank	AUC	59	59	59	59	59	59	59	59	59
	Precision	65	65	65	64	65	65	64	64	64

We start by analyzing the impact of employing DP on the performance of iForest and LOF. Table 5.1 reports the AUC and precision of iForest across the three considered datasets (mammography, thyroid, and bank), and across the two noise-adding mechanisms (Gaussian and Laplace) for varying values of ε. When DP is not employed, iForest achieves an AUC of 74%, 89%, and 64% for mammography, thyroid, and bank datasets, respectively. A precision of 90% for both the mammography and thyroid datasets, and 68% for the bank dataset. As DP is introduced, iForest generally exhibits decreased performance compared to the non-DP models. AUC and precision decrease already for high values of ε (i.e., less privacy). As expected, the difference is higher

for smaller values of ε (i.e., more privacy). By decreasing ε in the Laplace case, the AUC decreases from 73% to 53% for the mammography dataset, from 54% to 50% for the thyroid dataset, and from 58% to 52% for the bank dataset, and a similar decrease happens when decreasing ε with Gaussian noise. This is except for one case, that is for the mammography dataset for $\varepsilon = 5$ with Gaussian.

Table 5.2 reports the AUC and precision achieved by LOF across three considered datasets. Results show that in the case without applying DP, LOF achieves an AUC of 74%, 56%, and 59% for mammography, thyroid, and bank datasets, respectively, and a precision of 91%, 60%, and 65%. As DP is introduced, unlike with iForest, we observe a similar value for both the AUC and the precision across all datasets and all values of ε except for $\varepsilon = 0.01$.

This suggests that while iForest initially outperformed LOF without DP, LOF proved to be more robust and resilient to DP, maintaining its effectiveness under such constraints better than iForest, potentially due to its focus on k-nearest neighbors and local data density. This investigation suggests a trade-off between local and global outlier detection capabilities under DP. We explore this trade-off further with respect to SHAP explanations.

5.5.2 Impact of Differential Privacy on SHAP explanations

Figures 5.2, 5.3, 5.4 and 5.5, 5.6 and 5.7 report the results of ShapGap-Cosine and ShapGap-Euclidean distances along with the fidelity accuracy and ShapLength metrics of iForest and LOF, across the three datasets and for the various values of ε considered. Each row of the plot focuses on a specific metric on the x-axis. The y-axis consistently displays both ShapGap-Cosine and ShapGap-Euclidean distances across 5 runs for each ε experiment. Each point on the plots corresponds to the average of the corresponding ShapGap distance across all data points of each dataset for one single run. In an ideal scenario, the AD models with and without DP should agree to 100%, producing identical outputs for all data points.

iForest:

Figures 5.2, 5.3 and 5.4, presents the results for iForest across the three datasets. Results show that ShapGAP-Euclidean and ShapGap-Cosine distances across all datasets increase as the privacy guarantee increases (i.e., ε decreases). This difference means that the vectors of the features in the explanations extracted using SHAP before and after the application of DP change in both magnitude (captured by the Euclidean) and direction (captured by the Cosine). These findings reveal that as the ε decreases, the magnitude of SHAP values tends to deviate more from those obtained in the absence

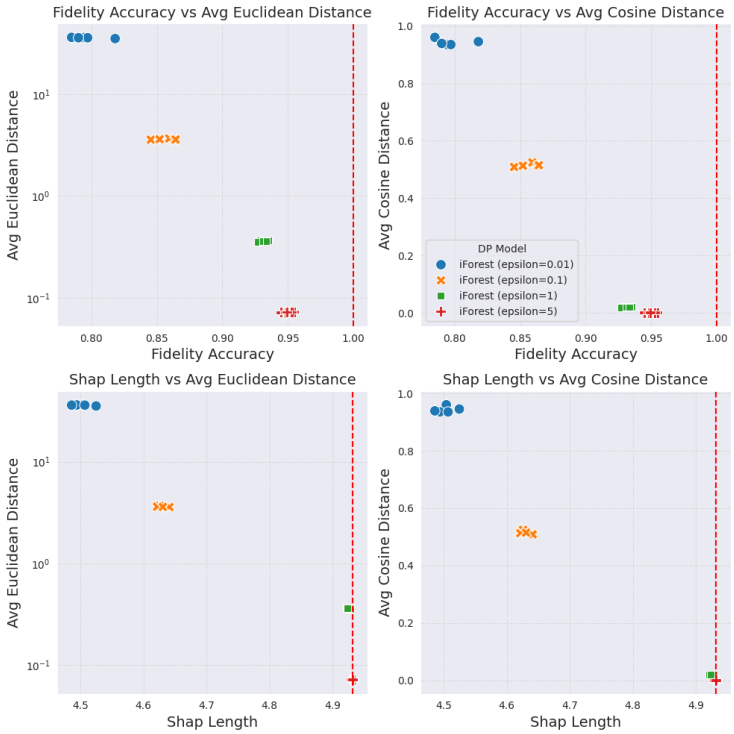


Figure 5.2. Fidelity Accuracy of iForest and average ShapGap-Euclidean distance, ShapGap-Cosine distance and ShapLength computed across the explanations extracted using SHAP for the various iForest models and the various values of epsilon on the Mammography dataset. The vertical dashed line represents the without DP metric presented at the x-axis.

of privacy constraints, for instance, a ShapGap-Cosine value close to 1 indicates high dissimilarity in the magnitude and direction of the SHAP value vectors before and after applying DP with ε of 0.01. Conversely, ShapGap-Euclidean has no upper bound, with higher values signifying greater dissimilarity. Specifically, the ShapGap-Euclidean metric for the mammography dataset ranges between 0 and 10, while for the thyroid and bank dataset, it ranges between 0 and 100. In contrast, the ShapGap-Cosine metric (Figures bottom and up right) scored between 0 to 1 for both datasets depending on ε. The results also show a consistent trend between the value of ε and the fidelity accuracy (Figures up right), as ShapGap-Cosine with smaller ε values correspond to lower fidelity accuracy, progressively decreasing from 100% relative to the ideal scenario when DP is not applied. This trend is evident for both distances (Figures up) and across all the datasets.

Another notable trend is the negative correlation between fidelity scores and dis-

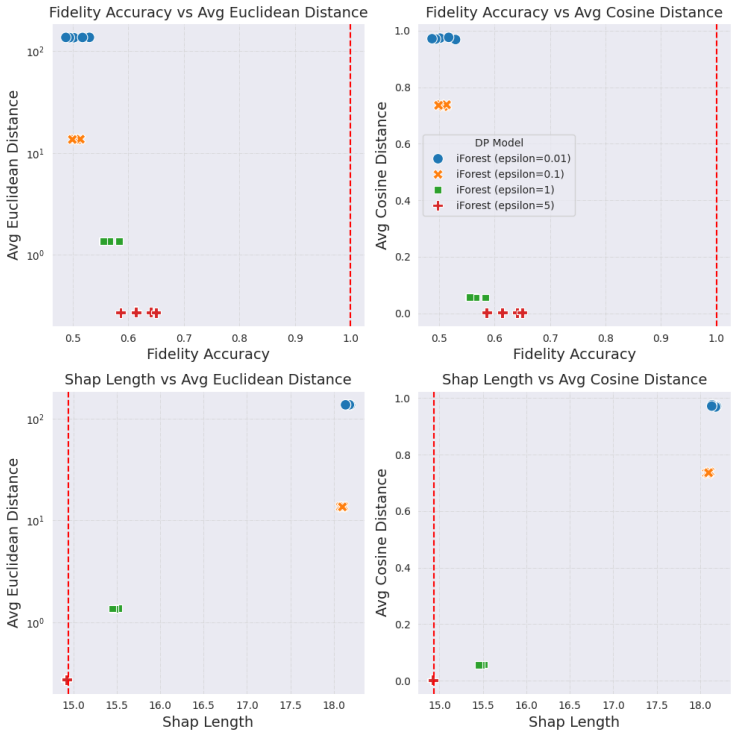


Figure 5.3. Fidelity Accuracy of iForest and average ShapGap-Euclidean distance, ShapGap-Cosine distance and ShapLength computed across the explanations extracted using SHAP for the various iForest models and the various values of epsilon on the Thyroid dataset. The vertical dashed line represents the without DP metric presented at the x-axis.

tance values (i.e., higher distances imply lower fidelity scores, and vice versa). In other words, when the model's reasoning aligns more closely with the original non-DP model (higher fidelity scores), the distances between SHAP values are minimized, indicating a stronger agreement in the influence of features on predictions. This trend is observable for both ShapGap-Euclidean and cosine distances, showing that both the magnitude and direction of the SHAP value vectors are closely linked to the model's reasoning and fidelity. *This result suggests a correlation between the difference in explanations and the model's ability to faithfully represent the original predictions, as indicated by the fidelity scores*.

We now discuss the complexity captured by the value of ShapLength. The findings vary based on the privacy budget (ε) value for DP and depending on the dataset. For the mammography dataset (Figure 5.2 bottom), observations at ε values of 1 and 5 indicate stability and consistency in ShapLength, maintaining an average value of 4.95,

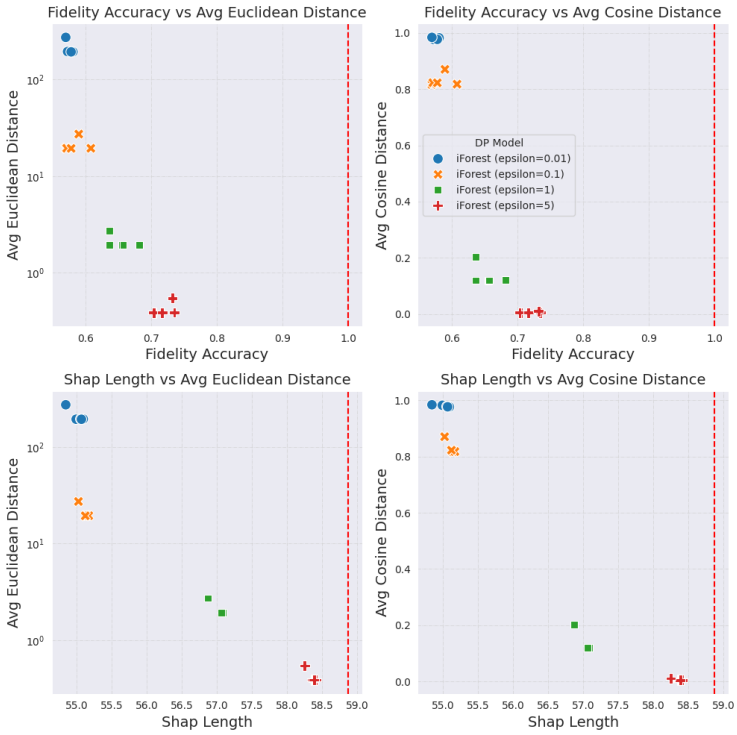


Figure 5.4. Fidelity Accuracy of iForest and average ShapGap-Euclidean distance, ShapGap-Cosine distance and ShapLength computed across the explanations extracted using SHAP for the various iForest models and the various values of epsilon on the Bank dataset. The vertical dashed line represents the without DP metric presented at the x-axis.

identical to that of the model without DP. *This implies that the implementation of DP, while ensuring a moderate level of privacy protection, does not impact the ShapLength, preserving the explainability complexity of the model as in the non-DP setting*. However, at lower ε values, specifically 0.1 and 1, there is a noticeable reduction in ShapLength to 4.5, indicating a slight deviation in model complexity compared to the model without DP. This indicates that DP with stricter privacy has reduced the model's complexity. For the thyroid dataset (Figure 5.3b bottom), we observe a different outcome: the ShapLength is highly affected by the application of DP, as with the decrease of ε, the ShapLength is increasing. When observing it with the SHAP Gap distances, smaller ShapLength aligns with larger Euclidean and Cosine distances. The bank dataset (Figure 5.4c bottom) exhibits a similar outcome in ShapLength compared to the mammography, as we observe a decrease in SHAP Length with smaller ε values even with $\varepsilon = 5$ where ShapLength has also decreased.

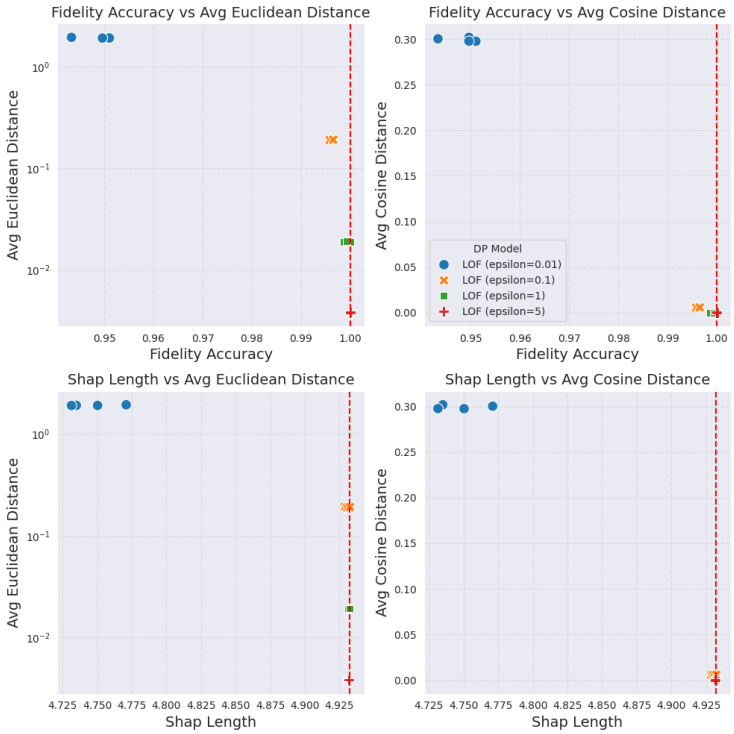


Figure 5.5. Fidelity Accuracy of LOF and average ShapGap-Euclidean distance, ShapGap-Cosine distance, and ShapLength computed across the explanations extracted using SHAP for the various LOF models and the various values of ε on the Mammography dataset. The vertical dashed line represents the without DP metric presented at the x-axis.

Our investigation reveals a privacy-explainability trade-off in applying DP to iForest models with SHAP explanations. Stricter privacy leads to increased divergence in SHAP values (magnitude and direction), decreased fidelity to the original model, and simpler models with less detailed explanations (reduced ShapLength) and are dependent. Conversely, relaxed privacy settings show better-preserved interpretability with explanations closer to the non-DP model and stable model complexity. *This negative correlation between explanation divergence and fidelity scores suggests a link between interpretability and the model's ability to faithfully represent the original model's predictions and explanations.*

LOF:

Figures 5.5, 5.6 and 5.7 shows the ShapGap metrics for the performance of the LOF model, highlighting the association between model fidelity and privacy levels achieved

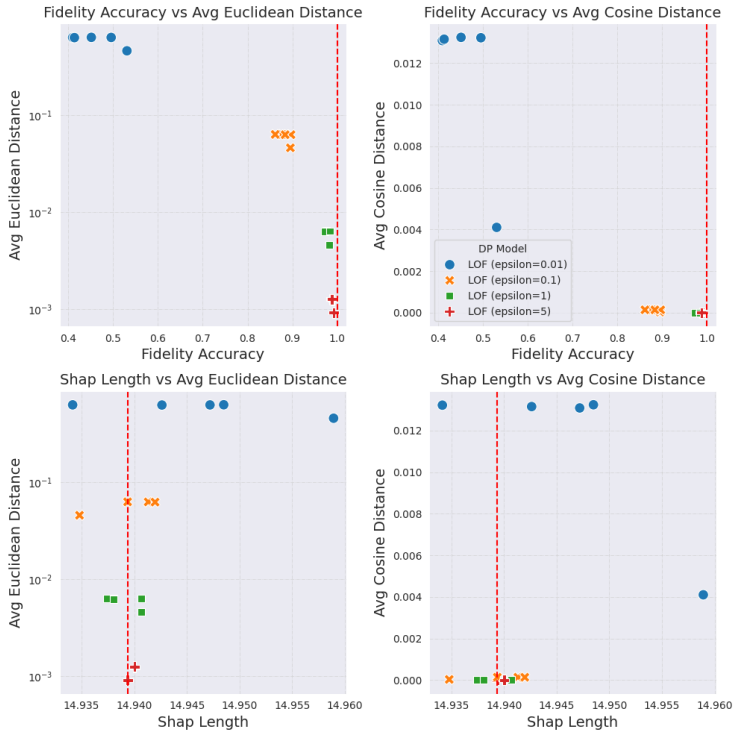


Figure 5.6. Fidelity Accuracy of LOF and average ShapGap-Euclidean distance, ShapGap-Cosine distance, and ShapLength computed across the explanations extracted using SHAP for the various LOF models and the various values of ε on the Thyroid dataset. The vertical dashed line represents the without DP metric presented at the x-axis.

through DP (Figure, located up). Across the mammography, thyroid, and bank datasets, model fidelity remains impressively high, ranging from 90% to 100% for ε values of 0.1, 1, and 5, nearly mirroring the fidelity seen in the AD model without DP. However, a notable fidelity reduction occurs at $\varepsilon = 0.01$, with drops by approximately 5%, 40%, and 10% for the mammography, thyroid, and bank datasets, respectively. Despite this high fidelity, we observe shifts and increases in Euclidean distances (Figures up left), indicating alterations in the scale of the underlying data that fidelity scores do not capture. Conversely, the ShapGap-Cosine distances (Figures up right), remain largely stable across 0.1, 1, and 5, ε values, suggesting that the directionality of the SHAP value vectors stays consistent despite the application of DP except for ε of 0.01. Specifically, for the mammography dataset, it increases to around 0.3. For the thyroid, and bank dataset, the ShapGap-Cosine also increases but stays within a very low cosine measure of around 0.07 and 0.012 respectively, indicating a low magnitude and

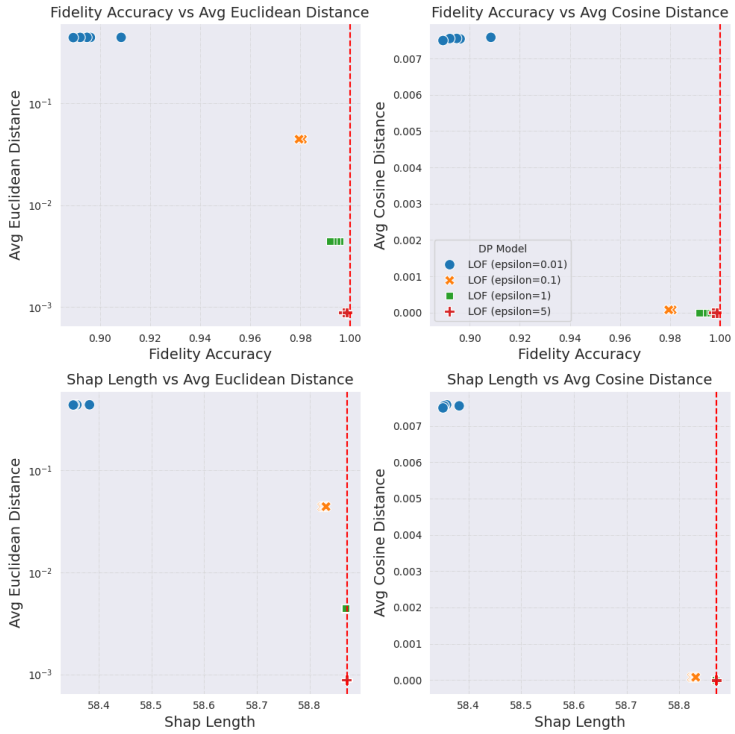


Figure 5.7. Fidelity Accuracy of LOF and average ShapGap-Euclidean distance, ShapGap-Cosine distance, and ShapLength computed across the explanations extracted using SHAP for the various LOF models and the various values of ε on the Bank dataset. The vertical dashed line represents the without DP metric presented at the x-axis.

direction of change within SHAP values. *This phenomenon suggests that while the overall magnitude of model explanations is influenced by DP, the vector direction reflecting feature contributions towards predictions remains minimally unaffected with DP for LOF.*

Analyzing model complexity through the ShapLength (Figures bottom), we find that in both the mammography and bank datasets, ShapLength remains relatively unchanged for ε values of 0.1, 1, and 5, suggesting minimal variations in model complexity. However, at a ε of 0.01, we notice a decline in ShapLength, which decreases from 4.925 to 4.725 in the mammography dataset and from 58.8 to 58.4 in the bank dataset. This indicates that the complexity of the model decreases slightly under more strict privacy conditions. In contrast, the thyroid dataset demonstrates a different pattern: ShapLength stays closely aligned with the small variance in complexity without DP, except at a ε of 0.01, where we note a relatively small increase in the variance of complexity as the privacy level increases across the different runs (from 14.935 to

14.96).

The observed inconsistencies across datasets in fidelity and ShapGap-Euclidean and Cosine distances likely stem from the inherent randomness introduced by DP and the chosen AD algorithms. While DP guarantees privacy, its added noise can alter various data characteristics, therefore the noise level and privacy level should be carefully chosen in a way that the overall distribution of the data is not largely affected. Therefore, this means that fidelity scores can remain elevated despite variations in Euclidean distance within higher ε. This also suggests that the data distribution undergoes modifications caused by the DP noise, but these modifications are limited such that the overall statistical properties remain largely preserved, thus maintaining high fidelity and low ShapGap-Cosine.

As anticipated, these findings indicate that iForest is more sensitive to data points that significantly deviate from the overall distribution of the data, while LOF more easily detects deviations within specific data regions. Since DP focuses on preserving overall data statistics, it can alter the distribution of the data, impacting how anomalies appear in the global picture. This, in turn, affects the SHAP values in iForest, as features contributing to isolation might be masked by the DP noise. LOF focuses on Local outliers and analyzes how different a data point is from a local perspective. DP's impact on local neighborhoods might be less significant compared to its effect on the entire data distribution. Additionally, LOF's SHAP values might focus on features relevant to the local anomaly score, which might be less sensitive to global distribution changes caused by DP.

5.5.3 ShapGap Distribution Analysis

To further explore the distribution of SHAP divergence across data points, we report the findings using box plots for ShapGap-Euclidean and Cosine. In Figure 5.8, we visualize the distribution of ShapGap-Cosine for iForest for all data points across the three datasets. We observe that for ε values of 1 and 5 across the three datasets, the ShapGap-Cosine has a lower spread in comparison to smaller ε of 0.1 and 0.01. This low spread indicates that most of the data points have a small cosine distance between 0 and 0.25, which means closer to the non-DP scenario. However, for smaller ε values (0.1 and 0.01), we see a wider spread of the ShapGap-Cosine distribution. The values range between 0 and 2 for $\varepsilon = 0.01$ with the mammography dataset, between 0.25 and 1.75 for thyroid, and between 0.6 and 1.4 for bank. This wider distribution suggests a significant portion of data points exhibiting high ShapGap, deviating from the behavior observed without DP. Regarding Euclidean distances, Fig. 5.9 we visualize the distribution of ShapGap-Euclidean for iForest across all data points for each dataset. We observe that for ε values of 0.1, 1, and 5 across the three datasets, the

ShapGap Euclidean distance has a small distribution compared to ε of 0.01. This indicates that the data points with small Euclidean distances (ranging between 0 and 20 for mammography, from 0 to 25 for thyroid, and from 0 to 50 for the bank), are closer to the non-DP scenario. However, for smaller ε values (0.01), we note a larger distribution of ShapGap Euclidean measures reaching up to 500 in the bank dataset, which is extremely large.

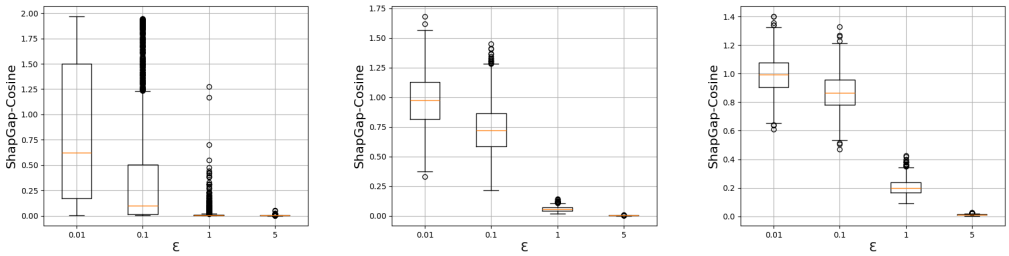


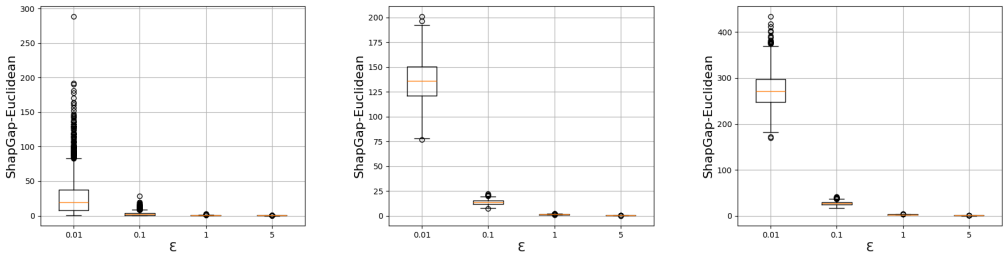
Figure 5.8. Distribution of iForest ShapGap-Cosine distances across (a) Mammography, (b) Thyroid, and (c) Bank datasets for the various ε values.

Figure 5.10 shows the distribution of ShapGap Cosine across all data points with LOF. For ε values of 0.1, 1, and 5, the distribution of ShapGap Cosine distance is relatively small (less than 0.2 for all the datasets), indicating that the SHAP values with and without DP are similar for most of the data points, suggesting minimal impact on the data due to DP at these privacy levels. For ε equal to 0.01, the distribution of ShapGap Cosine distance is larger (ranging from 0 to 1.75), indicating greater differences between SHAP values with and without DP. Regarding Euclidean distances, Figure 5.11 visualizes the distribution of ShapGap Euclidean for LOF across all data points. We observe that for ε of 0.1, 1, and 5, for the 3 datasets, the ShapGap Euclidean concentrates around at most between 0 and 1, and there is no diverged distribution. While only for the very small ε we observe the distribution of ShapGap Euclidean diverges covering a range between 0 and 6 as maximum, which is still relatively very small compared to the magnitude change scale of iForest.

These findings suggest that moderate privacy levels have minimal impact on SHAP with LOF. However, iForest appears to be more sensitive to the DP noise, evidenced by the large distribution of ShapGap distances and reflected by the distance distribution across the data points.

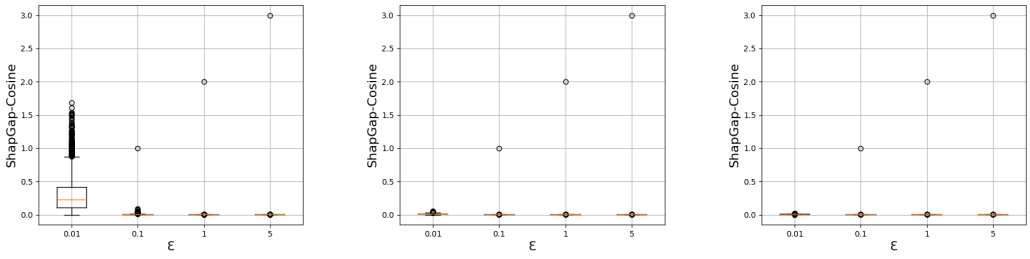
5.5.4 Impact of Differential Privacy on SHAP summary plots

After having quantitatively analyzed the impact of DP on SHAP values, we now examine the visual interpretation and analysis of SHAP summary plots to illustrate how DP noise



(a) Mammography (b) Thyroid (c) Bank

Figure 5.9. Distribution of iForest ShapGap-Euclidean distances across (a) Mammography, (b) Thyroid, and (c) Bank datasets for the various ε values.



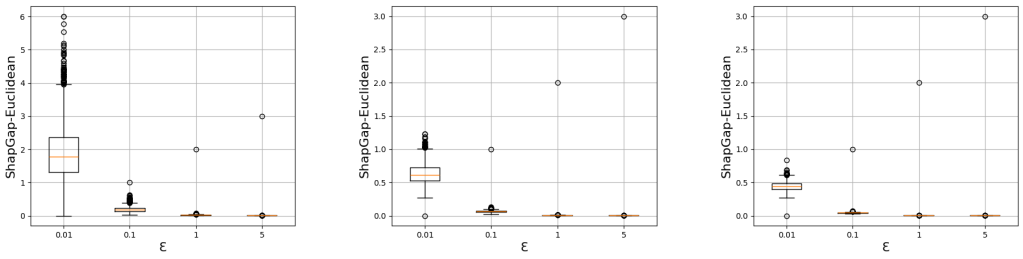
(a) Mammography (b) Thyroid (c) Bank

Figure 5.10. Distribution of LOF ShapGap-Cosine distances across (a) Mammography, (b) Thyroid, and (c) Bank datasets for the various ε values.

influences the interpretability of SHAP explanations visually. Figure 5.12 and Figure 5.13 present the summary plots relative to the mammography dataset for the iForest and LOF models respectively[1].

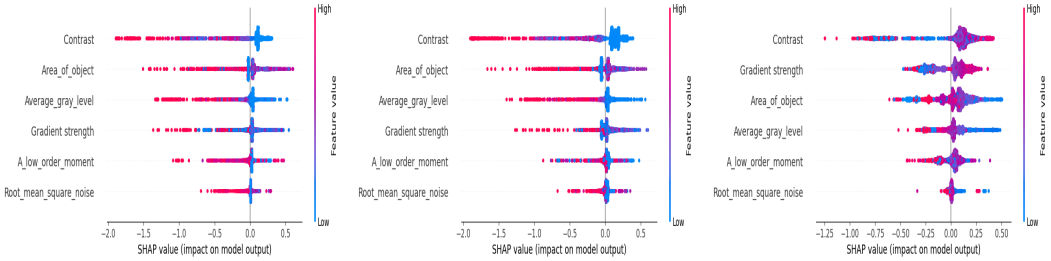
Figure 5.12 shows that the feature *Contrast* emerges as the most critical feature for the model's prediction, while a *low_order_moment* and *Root_mean_square_noise* remain the two least important regardless of the level of privacy ε introduced. This fact indicates that, for some features, the level of importance remains consistent even after applying DP, regardless of the value of ϵ. Instead, concerning other features, such as *Gradient_Strength, Area_of_object, and average_gray_level*, their importance accord-

[1]The summary plots display SHAP values for each feature and data point, indicating their impact on classifying normal or abnormal. On the x-axis, SHAP values show a feature's influence on predictions, with positive or negative values indicating a tendency towards an abnormal or normal prediction, respectively. The y-axis ranks features by importance, and point colors signify feature values—red for high and blue for low



(a) Mammography (b) Thyroid (c) Bank

Figure 5.11. Distribution of LOF ShapGap-Euclidean distances across (a) Mammography, (b) Thyroid, and (c) Bank datasets for the various ε values.



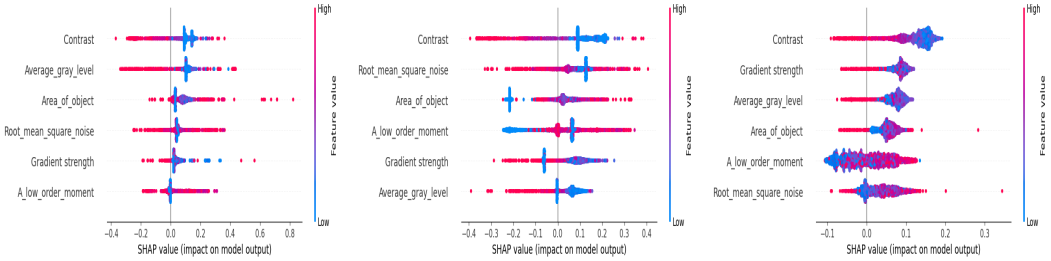
(a) without DP (b) ε=5 (c) ε=0.01

Figure 5.12. Summary plot for iForest for the mammography dataset with various ε

ing to SHAP varies with ϵ. For example, for a low level of privacy protection (e.g., $\epsilon = 5$), the consistency concerning the scenario without DP is maintained considering both feature importance and feature contribution, as observable by the similar color distributions in the various scenarios. If, instead, stronger privacy guarantees are set in place (e.g., $\epsilon = 0.01$), both the order and contribution of the three features significantly change. Indeed, the clear distinction between blue and red fades, indicating that the significant level of noise obscures the clarity of SHAP values, rendering the interpretation of output trends more difficult [2] Figure 5.13 shows the summary plots obtained for the LOF model. We observe that the *contrast* feature is consistently the most influential, irrespective of the value ϵ-DP. Specifically, we note that higher values of the contrast feature continue to be highlighted in blue and have a positive impact on the output, indicating their significance, while lower values are consistently highlighted in red and have negative SHAP values. Instead, for $\varepsilon = 0.01$, the SHAP values

[2]As similar trends in the SHAP summary plots are observed in the two other datasets, we omit to show them and the relative discussion. To illustrate the visual changes, we show summary plots for only two ε values (0.01 and 5).

are no longer distinguishable by color, signifying a less clear impact of feature values on the output of the model. Concerning the other features, there is a variation in their order of influence between the various ε, yet the underlying rationale behind their values remains consistent across most ε values. However, for a stricter privacy budget of $\varepsilon = 0.01$, there is a notable departure from this consistency, as the distinction between blue and red becomes less clear, thus reducing considerably the interpretability of the model through this plot. These findings align with the ShapGap and ShapLength



(a) without DP (b) ε=5 (c) ε=0.01

Figure 5.13. Summary plot for LOF for the mammography dataset with various ε

measures, demonstrating that iForest exhibits a greater sensitivity to DP perturbations compared to LOF in terms of SHAP interpretability. This is evidenced by the more pronounced shift observed in the SHAP summary plots for iForest with decreasing privacy budgets. While both models experience a decline in interpretability for stricter privacy constraints, LOF retains a relatively clearer distinction between influential and non-influential features even at lower ε values, specifically at ε equal to 0.01, where iForest's interpretability obscures significantly. Another finding is that while applying DP safeguards individual data privacy through noise injection, this mechanism hindered the interpretability of SHAP summary plots. DP's impact manifests in different ways such as distorted features importance and misleading interpretations. Firstly, the introduced randomness leads to fluctuations in SHAP feature attributions, making it difficult to accurately detect their true impact on model predictions. Secondly, the noise obscured data patterns, diminishing the overall precision of the AD and SHAP values and affecting the extraction of meaningful insights.

5.6 Conclusion

In this chapter, we investigate the impact of differential privacy (DP) on the performance and explainability of anomaly detection (AD) models. We compare the performance of Isolation Forest (iForest) and Local Outlier Factor (LOF) under various DP

noise conditions and across multiple datasets. The results show that while iForest initially outperforms LOF without DP, LOF exhibits greater robustness to DP. Furthermore, we analyze the impact of DP on explainability by comparing them across different distance metrics, both with and without DP applied. For explainability, we use SHapley Additive exPlanations (SHAP). We observe a correlation between the DP parameter (ε) and the magnitude and direction of changes in SHAP values across the metrics. Notably, the impact of DP on SHAP values manifested diversely across datasets and with the different AD techniques. This implies that distinctive data characteristics might affect the sensitivity of SHAP values to DP noise. These findings underscore the trade-off between privacy and explainability when employing DP alongside SHAP values in AD.

Although the empirical analysis focuses on anomaly detection, the findings offer broader insights for other ML applications where privacy-preserving mechanisms and post hoc explanations are combined. The results suggest that the impact of DP on explainability is not uniform, but depends on the model class, the data distribution, the privacy budget, and the explanation method. This observation generalizes beyond AD: whenever DP or other noise-based privacy mechanisms alter the learned decision function or the input distribution, feature-attribution explanations may also change in magnitude, direction, ranking, or stability. Consequently, explanation quality should not be assumed to remain unchanged after applying privacy-preserving transformations. Instead, privacy-preserving ML systems should be evaluated jointly along three dimensions: predictive utility, privacy protection, and explanation fidelity. At the same time, the specific numerical trends observed in this chapter should be interpreted as application-dependent rather than universal. Different domains, models, and explanation methods may exhibit different levels of robustness to privacy noise.

For future work, we aim to explore techniques to mitigate the effect of DP on SHAP values while upholding adequate privacy guarantees. Additionally, we aim to evaluate the effects of DP on other explainability techniques utilized with deep learning-based AD methodologies.

Chapter 6

Collective Privacy and Explainable AI

In this chapter, we broaden the discussion of privacy by focusing on collective privacy and its relationship with explainable AI. We review and distinguish related notions such as individual, group, interdependent, and collective privacy, and discuss the social and regulatory challenges that arise when privacy harms affect groups or communities rather than isolated individuals. We then examine how explainability may support collective privacy by making group-level inferences and risks more visible. Finally, we identify open challenges in designing and regulating AI systems that account for collective privacy concerns. This chapter is based on the content available in the following work:

1. **Fatima Ezzeddine**, Noé Zufferey, Omran Ayoub, Silvia Giordano, and Verena Zimmermann. Sok: Beyond the Individual: A Call for Collective Privacy Protection and Regulation. 2026, Submitted Conference.

6.1 Introduction

The emergence of Big Data, ML, AI, and pervasive digitalization has initiated an era of data-driven decision-making [216, 217]. Large datasets and advanced analytics have revolutionized organizations and industries by enabling them to extract meaningful patterns, generate accurate predictions, and derive actionable insights about individual people as well as collective behavior, market trends, operational efficiency, and risk management [218, 219, 220]. Whereas this transformation holds potential for societal and economic progress, it simultaneously introduces a challenge to data privacy [221, 222, 223], with implications for end users and system designers responsible for guiding users interactions with the system and the related data collection and han-

dling. Researchers, organizations, service providers, and AI engineers collect and analyze data from online interactions, social media profiles, financial transactions, health records, and other personal data streams to inform decision-making, thereby significantly increasing the risk of privacy breaches and the misuse of information. Such data collection raises significant concerns about how private data are handled, stored, and shared [224, 225].

These concerns have led to significant global efforts in regulatory and ethical considerations, leading to the development of regulations such as the GDPR, the EU AI Act, the Montreal AI Declaration, and national AI legislation in countries like the United States, Switzerland, Korea, and China [15, 16, 17, 226, 227, 228]. The aim is to balance innovation and individual privacy [229], preventing the identification of natural persons by their personally identifiable information (PII), and attributes [23]. However, such protection mechanisms focus solely on protecting individuals and do not safeguard the privacy of collective individuals. A collective is a group of individuals that forms based on individual shared attributes and common social values expressed as mutual interaction or interdependence between individuals, making individuals not isolated entities, such as a family or indigenous community with similar genetic or mobility information that can be identified even if the data does not hold any PII. The privacy of a collective is referred to as *group privacy* (term often used interchangeably with interdependent, etc) [230, 231, 232, 233] (Detailed in Sec. 6.2).

Collective privacy concerns the privacy of individuals within a group and the group as a whole, especially when the group holds sensitive collective information, raising a distinct concern about *collective privacy* rather than solely *individual privacy*. Collective privacy focuses on privacy harm even if no individual's privacy is identified. Moreover, collective privacy vulnerabilities arise when data intended to represent a collective instead reveals private information about the collective or its members. Collective privacy requires careful consideration of how information about group interactions, social values, genetics, and affiliations is collected, used, and shared. For example, human mobility can detect gatherings, genetic research can expose familial or population genetic information, and Large Language Models (LLMs) can reproduce, infer, or even amplify sensitive information about collectives, such as cultural practices, health trends, gender bias, or political affiliation, often without consent from any group [234, 235].

An ongoing discussion concerns whether collective privacy is a collection of individual privacy or an independent phenomenon [231]. We argue that collective privacy is an independent phenomenon because, even if individuals are not directly identified, a collection of data points can reveal patterns that not only identify individuals but also infer group characteristics, including existence, and potentially lead to discrimination or other harm at the collective level, not just the individual level. For instance, data analysis can yield inferences about a group's behavior, beliefs, or health status, even

without explicitly identifying individuals. Also, we argue that interdependent privacy (i.e., when data of a given individual provides information about another one) falls under collective privacy since collectivism is constructed without the knowledge of all individuals, not just about what individuals reveal, but also about what others reveal about them intentionally or non-intentionally, what machines assume about individuals, and what groups represent. Accounting for collective privacy makes the design of privacy-enhancing technologies particularly challenging with respect to effectiveness and usability, as the decisions/actions of some individuals can affect the privacy of others and/or the entire group. Therefore, it is essential not to view privacy as a static, individualistic concept, but to consider its interdependence and collective nature.

The relevance of collective privacy to this thesis, and especially to explainability, lies in recognizing that contemporary ML models increasingly engage in automated group profiling, and the opacity of these systems amplifies risks at the collective level. This thesis seeks to clarify existing definitions of collective privacy, thereby illuminating its conceptual foundations, and to advance future privacy research by fostering future human-centric approaches that strengthen user trust through the integration of xAI. Despite the growing importance of these issues, the relationship between collective privacy and algorithmic explainability remains markedly understudied, particularly regarding how collective-level inferences are generated, represented, and acted upon. In this context, xAI have potential to play a critical role: it can reveal how models construct group-level inferences (e.g. with clustering), identify the features that drive these inferences, and expose whether the system inadvertently encodes or amplifies sensitive collective attributes. By rendering algorithmic reasoning transparent, xAI can contributes to accountability, facilitates the auditing of group-based harms, and enables both individuals and collectives to understand and contest how they are represented within data-driven systems (discussing RQ3). *However, substantial work remains to develop xAI methods capable of detecting, explaining, and mitigating risks specific to collective privacy. Addressing this gap constitutes an important, largely unexplored frontier for future research, which falls outside the scope of this thesis.*

Goals and Scope of the chapter In this chapter, we first provide an overview of current definitions of collective privacy and discuss existing privacy regulations that primarily protect individual rights but remain limited in their ability to address collective privacy concerns. We aim to contribute to the ongoing discourse where safeguarding individual privacy should be complemented by collective privacy, contributing to raising awareness about the critical need to evolve privacy preservation and safeguard collective interests. Additionally, the chapter presents a call for a shift towards recognizing collective rights and discusses potential ethical and legal mechanisms like collective consent and data governance. Specifically, our contribution is summarized as:

1. We provide a systematization of the collective privacy concept by revisiting ex-

isting privacy definitions. We highlight that collective privacy is insufficiently addressed in current privacy landscapes.
2. We present real-world cases to illustrate the practical implications and necessity of addressing collective privacy, and its impact on collectives, delineating its dynamic scope and contextual relevance across domains.
3. We discuss the limitations of existing privacy-preserving techniques, such as k-anonymity and DP, given the static and dynamic nature of group identities.
4. We outline emerging challenges, identify gaps, and propose future research directions related to the governance and regulation of collective privacy.
5. We pose a set of guiding questions for future work examining how xAI can enhance transparency and accountability in algorithmic group profiling and support human-centric collective privacy research.

This chapter is organized as follows: Section 6.2 presents foundational definitions of individual and collective privacy in Section 6.2. We then discuss the dynamicity of collectives in Section 6.3, comparing it with traditional definitions. In Section 6.4, we analyze the social implications of collective privacy. Section 6.7 evaluates the limitations of existing privacy-preserving methods. In Section 6.5, we discuss regulatory challenges and open problems related to collective privacy. Finally, Section 6.8 summarizes our findings and outlines future research directions.

6.2 Definition of Individual, Group and Interdependent Privacy

Methodology We conducted a systematic search of scholarly literature published between 2010 and 2025 across multiple databases, including Scopus, Elsevier, IEEE Xplore, Frontiers, the ACM Digital Library, Taylor & Francis, and Google Scholar. The search strategy employed a comprehensive set of keywords encompassing diverse privacy concepts, including collective privacy, interdependent privacy, group privacy, mutual privacy, peer privacy, and relational privacy. To capture broader dimensions of the field, we additionally included terms related to data privacy, information privacy, privacy interdependence, collective privacy harm, algorithmic grouping, profiling, data aggregation, identity reconstruction, group anonymity, collective data rights, collective consent, and privacy governance. Inclusion criteria required that publications be peer-reviewed, demonstrate interdisciplinary relevance, and provide either conceptual or empirical contributions. After excluding papers that did not meet the inclusion criteria, the final corpus comprised 150 publications.

Table 6.1. Definitions of Privacy in Literature

Definition	**Source**	**Key Focus**
Right to be let alone	Warren & Brandeis (1890) [236]	Dignitary protection against unwanted publicity and press intrusion into private life
Tort-based privacy	Prosser (1960) [237]	Four actionable privacy wrongs as legal remedies (intrusion, disclosure, false light, appropriation of name)
Control over personal information	Westin (1967) [42]	Individual choice over disclosure, use, and circulation of personal data, especially amid surveillance
Informational self-determination	German Constitutional Court (1989) [238]	Constitutional limits on (state) data processing: purpose limits, proportionality, safeguards
Contextual integrity	Nissenbaum (2004) [239]	Context-specific informational norms for appropriate flow: actors, attributes, transmission principles
Taxonomy of privacy problems	Solove (2006) [240]	Map of privacy harms across the information lifecycle: collection, processing, dissemination, invasion
Normative control of access	Moore (2008) [241]	Moral right to regulate access to persons, spaces and personal information: access-control account
Personal information protection	Korea PIPA (in force since 2011, amended incl. 2023) [242]	Consent/rights and security duties for personal information with strong enforcement and cross-border transfer constraints
Seven types of privacy	Finn, Wright & Friedewald (2013) [243]	Multidimensional impacts beyond data privacy: bodily, spatial, communicational, decisional, etc.
Data Privacy	EU GDPR (2016) [15]	Rights and duties for processing personal data: lawful basis, transparency, access/erasure, accountability
Privacy as an Ethical Principle	Montréal Declaration for Responsible AI (2018) [228]	Privacy and intimacy as an ethical requirement to prevent intrusive data acquisition/archiving for AI systems
Privacy with AI governance	EU AI Act (2024) [16]	Safety/fundamental-rights governance for AI systems: risk tiers, bans/controls, transparency, conformity duties
Collective and group privacy	Contemporary debates (2010s-present)	Collective/Group/Interdependence/Mutual harms from inference and Data Collection

Definition of Privacy Privacy is commonly defined as "the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others" [244]. Similarly, privacy is described as "the ability of an individual or group to seclude themselves or information about themselves, and thereby express themselves selectively" [245]. Specifically, Individual privacy [246, 247, 248] centers on the right of individuals to maintain confidentiality of personal matters and to control the collection and processing of their personal information. Fundamentally, it enables individuals to shape their identity by safeguarding sensitive information, aligning with the definition of privacy as the individual's right to determine the boundaries of information communication about them.

Table 6.1 presents how the concept of privacy has evolved, reflecting shifts in technology, law, and social values. Early definitions, such as Warren and Brandeis *right to be let alone* [236], emphasized individual dignity and protection from press intrusion, while Prosser [237] framed privacy as a set of legal remedies against personal harms. With Westin [244], the focus shifted toward informational control, highlighting individual autonomy in managing personal data, a principle later reinforced by the German Constitutional Court's [238] principle of informational self-determination. As digital technologies expanded, Nissenbaum's contextual integrity [239] and Solove's taxonomy [240] mapped privacy harms across information flows, while Moore underscored moral rights to regulate access. Regulatory frameworks such as Korea's PIPA [242] and the EU GDPR [15] institutionalized privacy as rights and duties, complemented by multidimensional accounts, such as Finn et al.'s [243] seven types of privacy. More recently, ethical and governance perspectives emerged, with the Montréal Declaration [228] and the EU AI Act [16] addressing privacy in the age of AI. Crucially, contemporary debates recognize that privacy is no longer solely individual but collective and relational: data-driven inferences can harm groups, communities, and societies, making collective privacy protections essential to address interdependent risks in an era of pervasive surveillance and algorithmic decision-making.

Definition of Collective Privacy Table 6.2 presents the literature on collective and group-oriented privacy, illustrating a shift from individual-centric notions toward recognizing privacy as a relational, interdependent, and often co-owned phenomenon. Early work on *interdependent privacy (IDP)* highlighted how one person's data-sharing decisions can affect others, especially in networked or genetic contexts. Building on this, philosophers such as Floridi argued for *group privacy* as a moral and ethical necessity in the age of big data analytics, where algorithmically formed groups face profiling harms that extend beyond individual consent. Legal scholars extended this framing, emphasizing governance mechanisms to protect groups against collective inference. Conceptual refinements, such as Loi and Christen's distinction between within-group

confidentiality and inferential harms, further clarified the scope of group privacy. Parallel developments introduced *collective privacy*, emphasizing that harms are not merely the aggregation of individual losses but can target groups as collective entities, while *mutual privacy* framed privacy as a public good requiring coordinated protection. Other definitions, such as *networked privacy* and *peer privacy concerns*, emphasized the social negotiation of boundaries in online contexts. Technical and managerial approaches, including *multiparty privacy* and *collective privacy management*, addressed conflicts in shared data ownership and access control, while *collaborative privacy management* and *multiple-subject personal data* underscored the need for mechanisms to reconcile overlapping rights and preferences. Taken together, these perspectives converge on the idea that privacy is not only about individual autonomy but also about safeguarding collective interests and shared vulnerabilities.

To provide a formal, non-philosophical definition, we aim to unify prior definitions into a single definition: *"Collective privacy is the right of a group, community, or collectively constituted entity to control the collection, use, and dissemination of information pertaining to it as a whole or to its members in ways that generate collective harms. It extends beyond the aggregation of individual privacy claims to encompass both group-level information (e.g., stereotypes, profiling, reputational narratives) and interdependent privacy dimensions (in which one member's disclosure affects others). Collective privacy recognizes that harms can arise even when individuals are not de-identified, such as exposure of group existence, membership inference, or reinforcement of stigmatizing generalizations."*

We add to the definition that collective privacy is the principle that a group, rather than just an individual (that may be part of one or multiple groups), has the right to control how its data (i.e., information about the group) is collected and used, especially when the harm is on the collective and not on the individual. Collective privacy could include an explicit *representation or authority*: a group may act through a legitimate authority or designated representatives empowered to *permit, veto, or tailor* data uses on behalf of the group, going beyond the simple sum of individual consents. Such authority should be grounded in an appropriate basis of legitimacy for the group (e.g., recognized governance structures, delegation, or collective mandate) and should support transparency to address internal disagreements and enhance accountability.

This definition of collective privacy is important because it: (1) protects collective narratives and reputations by preventing data-driven profiling that can stigmatize or misrepresent groups, (2) preserves collective autonomy by recognizing that groups hold sensitive knowledge whose exposure cannot be governed through individual consent alone, (3) strengthens collective resilience and democratic participation by limiting surveillance and algorithmic inference that can enable manipulation or disruption of group activity, and (4) enables governance that address collective-level risks and externalities, allowing legitimate representation and coordinated decision-making where

Table 6.2. Collective, Mutual, Interdependent, Multiparty, and Group Privacy: Definitions in the Literature

Definition	Source	Key focus
Collective privacy management, early access-control	Squicciarini, Shehab & Paci (2009) [249]	Privacy-aware access control for shared content in OSNs (e.g., tagged), recognizes multiple affected users beyond the uploader
Multiparty privacy (MP)	Thomas et al. (2010) [250], Such & Criado (2018) [251]	Privacy boundaries are co-held when an item involves multiple people, focusing on resolving conflicts among co-owners preferences
Collective privacy management, conflict resolution	Squicciarini et al. (2010) [252]	Detects and resolves privacy policy conflicts for multi-party shared items, supports automated/assisted negotiation among individuals
Collaborative privacy management, collaborative control	Squicciarini, Xu & Zhang (2011) [253]	Personal information can be co-owned and should be co-managed by individuals. Focus on mechanisms for joint decision-making about sharing
Interdependent privacy (IDP)	Biczók & Chia (2013) [254], Humbert et al. (2015-2019) [255, 256], Kamleitner et al. [257]	An individual's privacy can be affected by others data sharing, externalities, and dependency in networks, genetics, location, etc.
Multiple-subject personal data	Gnesi et al. (2014) [258]	Data items may pertain to multiple subjects whose preferences/rights can conflict, representing and reconciling multiple subjects' privacy preferences.
Group privacy (philosophical/data ethics)	Floridi (2014) [259], Taylor, Floridi & Van der Sloot (2017) [260]	Privacy harms and protections should apply at the group level (including profiling), motivated by big-data analytics dealing with algorithmically formed groups rather than individuals
Networked privacy (social norms + context negotiation)	Marwick & boyd (2014) [261]	Privacy is negotiated in a networked audience with shifting contexts, with a focus on social norms, context collapse, and relational management of disclosure
Group privacy (legal/-data protection framing)	Taylor, Floridi & van der Sloot (eds.) (2017) [260]	Data protection law grants individual rights, but group privacy debates argue for governance against group profiling/inference beyond individual consent/rights
Collective privacy (collective interests/harms)	Mantelero (2017) as discussed in Puri (2023) [262]	Frames privacy as non-aggregative collective interests (not just summed individuals), focus on limiting harms to a group from invasive/discriminatory data processing
Relational Privacy	Reviglio et al. (2020) [263]	Privacy as a socially embedded condition that protects individuals within relationships emphasizing that autonomy is not isolated but shaped by interactions, trust, and shared contexts
Group privacy (Conceptualization)	Loi & Christen (2020) [230]	(i) shared confidential info inside a group vs outsiders, (ii) inferences about individuals sharing standard features
Peer privacy (peer-caused boundary loss)	Zhang et al. (2022) [264]	A user's privacy concern caused by peers' actions (tagging/sharing/disclosing about them)
Mutual privacy (as a group right / public good framing)	Puri (2023) [262]	Shared vulnerabilities (genetic/social/democratic/algorithmic grouping) create a collective interest, argue for a group right grounded in coordinated protection

individual choices are insufficient.

6.3 Definition of Collective: Bridging Social and Digital Collective Identities

We define key concepts relative to the definition of group, identity, and collective categories by drawing on existing research from social science and philosophy [230, 231, 232, 233, 265, 266]. This section also aims to differentiate between social privacy and digital privacy with regard to the collective.

A *group* is a collection of individuals bound by shared attributes, interests, or objectives, often underpinned by a common group identity [231]. These shared values, influenced by sociocultural norms and aspirations, shape group identity, behavior, and interaction. Following Floridi in [233], we posit that group formation is a function of analytical abstraction rather than a process of simple discovery or invention. This definition serves as a foundational element for our analysis of social and digital collective identities. Authors in [230] propose a binary typology of groups, distinguishing between those that exhibit interactions and those defined by shared attributes. The first groups are characterized by a shared interaction history and/or collective goals, demonstrating meaningful agency through intentional coordination and collective self-awareness. The latter category encompasses groups defined by shared features, excluding those with unified plans or a focus on the common good. For instance, city residents share a common geographical attribute but may lack a collective identity. Also, authors in [265] propose a group category typology, distinguishing among the categories collective, ascriptive, and ad hoc. Collective groups are formed through intentional association, driven by shared interests, backgrounds, or explicit common goals. Ascriptive groups are defined by membership based on inherited or incidentally developed characteristics, such as ethnicity or nationality. Ad-hoc groups are assembled by third parties, often for specific purposes like profiling, based on perceived connections between individuals. These different definitions illustrate how digital identities are represented in datasets released by data providers, whether from self-selected communities or from externally imposed sources. In the next paragraphs, we discuss social collective identity and link it to digital identity to safeguard digital collective identities that represent these social identities, and make a clear distinction between the nature of groups.

Social Collective Identity Social collective identity, also termed group identity, distinguishes a group from a mere collection of individuals by denoting a shared sense of belonging and unity among collective members. This concept, central to various social science disciplines, encompasses the collective principles, norms, and values that

Table 6.3. Privacy categories by scope of identity/collective

Category	**Description**
Individual Privacy	Concerns information that uniquely identifies a specific individual and their personal data (e.g., name, ID, biometrics, location traces, communications, health/financial records, device identifiers). Risks include de-identification, profiling, surveillance, targeted manipulation, and harms tied to a single person's exposure.
Family Privacy	Extends from one person to immediate relatives through shared attributes (e.g., genetics, hereditary conditions, household address, family relationships, shared routines, joint accounts/devices: bystander privacy). A disclosure about one family member can reveal sensitive facts about others.
Social Privacy	Concerns networks of friends, colleagues, and other close social ties, including interaction data (graphs, messaging, co-location, co-attendance, shared photos on OSNs) and the ability to infer membership, influence, workplace ties, or private associations. Risks include guilt by association and inference of hidden relationships.
Organizational Privacy	Collectives with shared objectives, such as companies, NGOs, or teams, including internal and strategic operations (employee roles, organizational structure, projects, customer lists, internal communications). Risks include industrial espionage, exposure of confidential partnerships, and attacks enabled by metadata.
Cultural Privacy	Collectives defined by cultural features (e.g., ethnicity, religion, language, regional identity). Even without identifying individuals, aggregated data can reveal sensitive patterns (such as segregation, stigmatization, or political pressure). Risks include stereotyping, collective harm, discrimination, and suppression of minorities.
Public Privacy	Broad collective identity and global trends at a large scale (population-level, public health, mobility, economic behavior, crowd analytics). Risks include harmful generalizations, the enabling of mass-surveillance infrastructure, and policy misuse that affects entire populations (e.g., biased resource allocation).

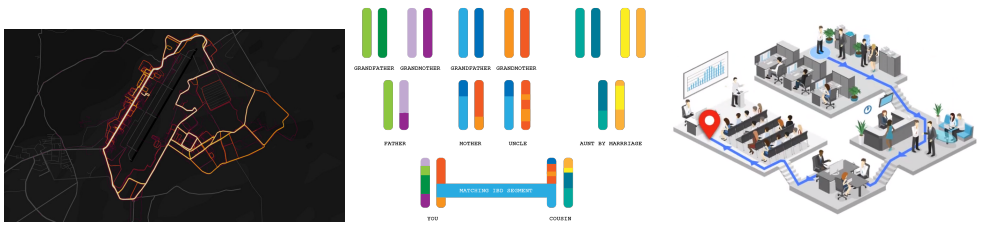
define a group's character and guide its members attributes, actions, and interactions [231, 267]. As articulated by [268], collective identity is "an interactive and shared definition produced by several individuals (or groups at a more complex level) and concerned with the orientation of action and the field of opportunities and constraints in which the action takes place".

Digital Collective Identity We propose a privacy-oriented operationalization of digital collective identity: a group identity that is materialized as group-referent data and becomes a target of inference/profiling/social sorting, which is exactly where collective privacy risks arise. Specifically, digital collective identity is an emergent "we-ness" and sense of collective agency that is formed, expressed, and negotiated through digital interaction, referring to a group's identity as embedded in data generated by or associated with its members. In contrast to traditional sociological definitions, which rely on qualitative assessments of shared norms and values, digital group identity is grounded in the recognition of a pre-existing social group within datasets. This representation can manifest in various forms, including shared sensitive attributes, relationships, communication and interaction patterns, preferences, and behaviors. For example, friends who frequently visit the same locations and use the same bike-sharing routes build a digital collective identity, as evidenced by location and transportation data. Digital group identity is not always static, particularly in algorithmic grouping applications, as it evolves as the group interacts in digital spaces, leaving a data trail that reflects its changing composition, interests, and activities. These data-driven definitions enable quantitative analysis of group dynamics and provide insights into the formation, maintenance, and dissolution of group identities in digital environments. Therefore, digital identity either reflects the group's real social identity or assigns a new social identity based on the grouping algorithm used to generate the groups. This definition makes the link explicit: digital collective identity is the object, collective privacy is the right to control the collection, use, and dissemination of that object, and to prevent group-level harms from inference and social sorting.

Collective Categories Building on the definition of digital collective identity, we acknowledge the multifaceted and dynamic nature of social and digital group identities and outline representation to define and complement these identities. We recognize that an individual may belong to several overlapping collectives that differ in their degree of closeness and influence. To capture this, Table. 6.3 is used to categorize collective identity, visualizing different levels of closeness and connection that lead to collective representation. Rather than relying on vague sociological labels for groups, it provides a concrete categorization for diagnosing, comparing, and designing targeted

interventions for future complex privacy challenges. This layered view clarifies why *one-size-fits-all* privacy rules cannot work, consent mechanisms that make sense for family and knit ties break down for data shared within an organization or inferred by an algorithm grouping. In Section 6.5, we discuss how collectives should determine the appropriate balance between individual consent and collective data-governance rights. Table 6.3 presents the categories from individual (1), familial (2), social (3), organizational (4), socio-cultural (5), and finally public-wide (6), each capturing a broader sphere of association. As one moves outward, the immediacy of the relationship loosens while the collective grows in size and diversity. In other words, every category marks a distinct layer of social inclusion, mapping how identity and the data that describe wider social collectives relate to one another. In digital terms, the table illustrates how data originating from close, personal relationships can aggregate into large-scale indicators of collective identity. When aligned with the theoretical distinctions proposed by [230] and [265], these categories complement their definitions by providing a clear view of the collective dynamicity.

Building on these representations, a promising direction for future regulation is to introduce a collectiveness factor or scale that assesses both the sensitivity of specific data types and the social acceptability of using them. For instance, analyzing broad, population-level data may be acceptable in many contexts, while examining smaller datasets, particularly those involving individuals, families, and indigenous or minoritized groups, could raise ethical concerns. Researchers can gain deeper insights into how privacy considerations shift across different levels and, in turn, develop more nuanced guidelines for ethically responsible collective data use. Finally, this representation highlights potential ethical concerns regarding individual privacy, particularly when individuals and groups are grouped (social identity presented through digital identity) without awareness or consent.



(a) Strava heatmap (b) DNA sharing (c) Indoor Localization

Figure 6.1. Examples of Collective Privacy use cases

6.4 Use Cases and Social Implications of Collective Privacy

To demonstrate the real-world vulnerabilities and social implications of collective privacy, we discuss multiple possible use cases, including human mobility, location-based services, social networks, genetic research, and algorithmic grouping.

Case 1: Genetic Research In genetic studies, inferences about disease predispositions or ancestral origins, though anonymized, can still be traced to specific populations, such as Indigenous tribes or familial communities. Genetic research involves investigating DNA, the molecule that carries human genetic information, to understand how genes influence human characteristics, how genetic variations contribute to differences between individuals and populations, and how traits pass from one generation to the next [269, 270, 271, 272]. Collective privacy risks are formalized from the interdependence of familial genomic and inherited data (Table 6.3 Family, Societal, and Cultural Privacy). Research has indeed demonstrated how different inferential methods are explored to infer information about family members and populations (Figure 6.1 b) [1]) [273, 274, 275, 276, 277]. These works build upon the genomic privacy concept, such as kin [278, 279, 280], probabilistic and statistical methods such as Bayesian networks [276, 281], game and graph theory [282], prove statistical connections within families and populations and extract inheritance genomic information. Similarly, other research investigates the evolution and the interdependence of genes over generations and inheritance [283]. Additionally, released models exhibit vulnerabilities to genetic privacy breaches, ML models used in genotype-based treatments are susceptible to model inversion attacks, as proven to infer sensitive genetic markers with regression models [284].

Genetic research presents unique collective privacy vulnerabilities beyond individual concerns [285] for both group and interdependent privacy [286]. Tools also exist to estimate the target of a family member, whose genomic privacy is affected [2]. For example, the Havasupai Tribe DNA case illustrates this (Table 6.3 Societal and Cultural Privacy), where research intended for diabetes studies inadvertently revealed predispositions to other diseases and theories of the tribe's geographical origins that contradict their traditional stories [287]. This unintended revelation caused significant distress within the group and prompted legal action, underscoring how group-based genetic research can have profound, unforeseen consequences for collective identity, cultural integrity, and shared well-being. These issues are fundamentally different from indi-

[1] https://www.swedishfinnhistoricalsociety.org/2018/06/26/the-lost-art-of-viewing-humanity/
[2] https://santeperso.unil.ch/privacy/tool/

vidual privacy violations. These factors raise critical concerns regarding group and interdependent privacy, wherein the compromise of one individual's genetic data can affect the privacy of their related population, family members, and social. In addition, direct-to-consumer genetic services such as 23andMe and GenomeLink, and platforms such as openSNP, in which individuals share their data, pose unique privacy risks, as one individual's decision to share their genetic information can affect family members, relatives, and the group to which they belong. Therefore, consent management in the collective remains an open issue in such cases.

Case 2: Human Mobility and Smart Environments Human mobility captures how and why people move, increasingly mediated by digital systems that collect and infer patterns across transportation modes, routines, and social–spatial interactions [288, 289]. Mobility and location traces create distinctive collective privacy risks: shared routes, hotspots, and synchronized behaviors can be inferred even under anonymization, exposing neighborhoods or communities to targeting or stigmatization [290]. Similar dynamics arise in Location-Based Services, and the Internet of Vehicles, where protecting an individual query does not prevent adversaries from learning group trajectories or coordinated movement patterns [291]. Internet of Things (IoT) environments further illustrate multi-subject exposure: sensors routinely capture the behaviors of households, families, and bystanders, making the affected entity a collective rather than a single user [292]. These settings reveal a core challenge for usable privacy: groups face collective risks but lack mechanisms to coordinate protections, articulate shared preferences, or manage interdependent privacy decisions, leaving defenses fragmented at the individual level [293].

The Strava fitness heatmap is another illustration of how aggregated mobility data can expose sensitive collective-level information: despite individual anonymity, the visualization revealed the location of a military installation [294] (Figure 6.1 a)). Similar issues arise in urban-planning datasets, where anonymous mobility flows derived from Google users who opted into location history are aggregated to compute city-level movement patterns [295, 296]. Although individual contributions are bounded within each partition, these techniques protect only user-level guarantees, they do not prevent inferences about neighborhoods, institutions, or other collectives. From a usable privacy perspective, such cases highlight a persistent gap between users expectations of anonymity and the collective vulnerabilities created by large-scale aggregation. Location-Based Services can infer an individual's unexposed location using data shared by their social contacts, as attackers can exploit publicly available geo-tags and interaction histories within the group [297, 298]. Social cues such as interaction strength and value similarity further amplify these risks, compromising both group and individual privacy [297] (Table 6.3 Social Privacy). Co-location on online social networks

similarly creates interdependent privacy harms: sharing one user's location can reveal information about others, with both direct and indirect effects documented in prior work [299]. Image sharing raises comparable issues. ConsenShare [300] demonstrates the need for co-consent when photos include multiple subjects. Repeated mobility traces can reveal home and work locations [301], while aggregated travel data expose group routines, such as common routes or organizational affiliations (Table 6.3 Organizational Privacy). Shared or overlapping trip patterns can also reveal social ties (Table 6.3 Social and Family Privacy). Finally, the ability to monitor attendance at public gatherings [302] poses risks, potentially chilling protest and democratic activities (Table 6.3 Social Privacy).

Moreover, modern localization technologies combine communication signals (e.g., WiFi, Bluetooth, cellular) with geographic positioning [303, 304, 305, 306, 307, 308, 309] (Figure 6.1 c) [3]). Entities with access to such data can reconstruct fine-grained location traces that enable behavioral analysis, profiling, and surveillance [310, 311, 312, 313] (Table 6.3 Organizational Privacy). Service providers can infer room-, floor-, or building-level positions by analyzing signal patterns from mobile devices. Crowdsourced Wi-Fi scanning further amplifies these risks: phones periodically collect nearby access-point fingerprints even when GPS is disabled, allowing distance estimation from MAC addresses and signal strength [314, 315]. Even with individual anonymity, highly accurate indoor localization enables the detection of gatherings and the monitoring of organizational or institutional activity. As a result, collective privacy concerns extend beyond social or demographic groups to include commercial, institutional, and organizational entities whose activities can be inferred from aggregated data.

Case 3: Online applications, Social Networks and Language Models Many modern systems extract value from group-level regularities, allowing privacy harms to persist even when individual identifiers are hidden. Social networks can surface population- or community-level signals that enable profiling, manipulation, or exclusion of groups [316]. Recommendation and trust-based models similarly support collective profiling and social sorting by learning sensitive group attributes from behavioral correlations [317, 318]. Language models introduce parallel risks: memorized or leaked content can reveal organizational or community practices, narratives, or operational details, producing collective confidentiality harms [319]. Online social networks further illustrate that data access and inference are inherently collective: one user's actions (e.g., installing apps, sharing, tagging) can expose non-consenting others and reveal group structure. Third-party apps amplify this collateral exposure, as permissions granted by a subset enable profiling of their broader networks and communities [320]. These

[3] https://github.com/HUbbm409/bbm406-project-wi-fi-based-indoor-positioning

dynamics motivate collective privacy framings and mitigations that treat exposures as externalities across friends, families, and communities rather than isolated individual leaks [256, 321, 322]. Even systems intended for social discovery can operationalize collective risk: peer-to-peer social search (e.g., Pythia) shows how routing and querying over social connections can reveal community membership and network structure, turning search itself into a channel for group-level exposure [323].

Moreover, a large share of social content is co-owned (e.g., group photos, conversations, tags), so privacy failures often manifest as collective boundary loss: one member's disclosure can reveal shared context, relationships, or group narratives even when no single individual is uniquely exposed. Empirical work shows that disclosure is socially negotiated, yet platform controls remain uploader-centric, enabling peer-driven privacy loss and mismatches between collective expectations and actual visibility [324, 325, 326]. The social media privacy model explains why such failures persist: platform affordances (e.g., context collapse, persistence, scalability) systematically complicate group boundary setting and make coordination harder than individual choice alone [327]. Recent work formalizes these dynamics as a collective calculus around co-owned disclosure and ownership perceptions, highlighting that the relevant unit of harm is often the group, friend circles, or organizations, rather than the individual record [328, 329].

The widespread adoption of LLMs across industries [330] and by everyday users [331, 332, 333] introduces new collective privacy risks. LLMs process large volumes of text [334], images, and documents containing diverse personal data [335, 336], enabling inferences about attributes such as location, income, and gender [337] from shared communications, social interactions, and aggregated datasets. They can reconstruct social ties and group narratives [338], revealing sensitive collective information that no single record discloses. Models trained on private chats (e.g., customer service [339]) may leak snippets, while membership inference and inversion attacks [340] further expose training data [341, 342, 343]. Real-world incidents underscore these risks: hundreds of thousands of Grok chatbot conversations became publicly searchable[4], and AI meeting tools have raised concerns about unauthorized access to recordings that reveal shared contexts and group dynamics[5]. LLMs also create interdependent risks: one person's data can reveal information about others who never interacted with them. Training on public datasets or family medical records [344, 345] can produce associations between surnames, regions, and rare diseases, or regenerate traits of specific ethnic or genetic subgroups [346, 347]. Retrieval-based LLMs trained on internal documents, Slack logs, or meeting transcripts increase the likelihood of

[4] https://www.bbc.com/news/articles/cdrkmk00jy0o

[5] https://www.zscaler.com/cxorevolutionaries/insights/privacy-security-concerns-ai-meeting-tools

leaking organizational data [348]. Models may reproduce phrases or ideologies that signal affiliation with vulnerable groups, and in domains such as law [349], they risk exposing strategic information that affects multiple clients or populations. Additional harms include identification of subsets within legal datasets [350], generation of demographic personas [351], group-level recommendation biases [352], and shifts in perceived trustworthiness toward stigmatized groups [353]. Finally, national or ethnic group profiling [354, 355] can yield stereotyped or discriminatory summaries [356], highlighting the collective nature of LLM-driven privacy risks.

Case 4: Algorithmic Grouping Algorithmic grouping, via clustering [357] or classification [358], sorts individuals into data-driven categories based on behavioral or attribute patterns [359], enabling prediction, decision-making, and targeted engagement [231, 360]. From a usable privacy and security perspective, these systems create risks that extend beyond individuals: users rarely understand how groupings are formed, how inferences propagate across members, or how to contest collective-level decisions. A key distinction arises between individual and collective rights. When algorithms form broad, ad-hoc interest groups (Table 6.3 Public and Collective Privacy), privacy concerns center on individual attribute inference, but grouping can also surface real social identities or reveal intentionally organized groups, directly implicating collective privacy. Grouping on sensitive attributes further risks discriminatory outcomes or the exposure of private information [361], which neither traditional privacy protections nor individual fairness metrics adequately address.

These challenges appear across domains. Healthcare clustering can inadvertently create identifiable subgroups based on rare genetic conditions [362]. Aggregated search data can reveal politically active clusters, and flawed aggregation may wrongly place individuals into sensitive groups, enabling unwarranted surveillance and democratic harms [363]. Collective harms also emerge when systems propagate discriminatory outputs, as in the Korean chatbot Luda incident targeting minority communities [364], echoing concerns about state surveillance of marginalized groups [365]. Federated Learning of Cohorts (FLoC) [366] illustrates similar issues: cohort-based advertising shifts tracking from individuals to groups, enabling inference about users in small or unique cohorts and raising privacy concerns [363, 367]. Its successor, the Topics API [368], faces similar criticism. Marketing analyses of Black Twitter further show how algorithmic grouping can misrepresent or profile communities, undermining collective identity and self-determination [369]. Across these cases, the core usable-privacy problem is governance: individuals cannot meaningfully manage or contest group-level inferences, and groups lack mechanisms to coordinate privacy expectations or assert collective boundaries. Algorithmic grouping therefore exemplifies why collective privacy requires protections beyond individual consent.

6.5 Challenges to Regulating Collective Privacy

Current regulatory frameworks focus on protecting individual privacy, as seen in the EU AI Act [16], GDPR [15], the Montreal Declaration [228], South Korean acts such as SKAIA and PIPA [227], Swiss Federal Act on Data Protection (nFADP) [17], and the US Privacy Act [226], but exhibit a gap in addressing collective privacy. The focus of these regulatory frameworks is on *natural persons*; for instance, in the privacy-related definitions of Article 4 of the GDPR, the data subject is an identified or identifiable natural person. While the emphasis on the individual is valuable and necessary, we argue that these definitions should also account for collective concerns. Future technical approaches may involve group-level protection, collective risk auditing of models, or differential privacy mechanisms tailored to groups rather than individuals. However, defining group sensitivity and consent at scale remains an open challenge, and implementing regulations will involve complex trade-offs. In this section, we discuss open problems in regulating collective privacy and distinguish between two key dimensions: *individual right to collective privacy* and *group right to collective privacy*.

The categories (Table 6.3) are valuable for addressing privacy issues in the legal sector by shifting the focus from broad sociological labels of groups to a practical method for diagnosing, comparing, and designing effective interventions for real-world privacy challenges. Specifically, it provides:

- *Clarity for policymakers:* Each category has a clear definition, cohesion, and data-sharing norms, so multidisciplinary teams can calibrate features, policies, and impact assessments against the same collective target.
- *Regulatory alignment and gap analysis:* Existing regulations align well with individualistic concerns. In contrast, proposals for collective governance map to the other categories, while algorithmic accountability laws typically fall into the broader categories.
- *Threat modeling:* Tell the difference between privacy harms from small groups and those from large-scale data collection. For example, a breach at the family level can cause personal and reputational damage, while a breach at the societal level can lead to discrimination or widespread profiling.

6.5.1 Individual Right to Collective Privacy

The individual right to collective privacy extends beyond protecting isolated individual privacy to safeguard collective-level information about the individual. This approach acknowledges that individuals are embedded within social groups and underscores the need to safeguard group affiliations and other social values to ensure their privacy and autonomy [231]. Specifically, we discuss the following rights:

Right to Interdependent Privacy Individuals in close groups, such as families, communities, or genetic relatives, have a right to privacy to protect information originating from shared relationships and interconnected data. This right acknowledges that one individual's data can reveal sensitive information about other individuals belonging to the group and emphasizes the need for safeguards when processing data with interdependent privacy implications [262]. *However, this right also raises regulation concerns: How can the boundaries of close groups be defined as legal entities? How social proximity, as shown in the Table 6.3, can be taken into consideration in this context? How can the individual's privacy rights be balanced with the group's collective privacy requirements? Moreover, what practical mechanisms can ensure that data processing respects interconnected privacy concerns without limiting data uses for scientific purposes?*

Right to membership transparency and self-determination Under current regulations, anti-discrimination laws do not apply to ad hoc groups. Individuals have the right to transparency regarding membership and to understand the basis of their group formation and classifications for fairness and accountability purposes (see Table 6.3 for Public and Collective Privacy). This right ensures access to information concerning the criteria and algorithms used in the grouping, whether in pre-existing social groups or through algorithmic processing [363] and the inferences drawn from their group membership. Such a right is accompanied by the right to opt out of algorithmically generated groups, to protect against unwanted or false categorization with potential collective and individual consequences. *This also raises critical questions: How to achieve such transparency in practice, especially with complex algorithmic methods? What are the limits of the right to opt-out, considering the similarity and interconnectedness of individual data points? How to enforce and understand the trade-off of collective machine unlearning in cases of deployed models? And how to communicate such membership transparency to individuals?*

Right to Transparency and Accountability The right of individuals in a defined group is to receive explanations of how algorithms profiled or grouped them and how this affects their understanding of how they perceive themselves within the group. This includes mechanisms to audit, challenge, and hold developers accountable for discriminatory or harmful individual or collective outcomes. *However, this also raises questions on how meaningful information can be translated into practical, accessible disclosures for diverse groups and how audit mechanisms can be designed to effectively and respect intellectual property?*

6.5.2 Group Right to Collective Privacy

The group's right to collective privacy emphasizes protecting collective information, such as digital and societal identity, and recognizes that groups, as distinct entities, have the will to protect sensitive group-specific information, along with other activities, from unauthorized disclosure. This right acknowledges that a group's privacy is not present when aggregating its individual members privacy and gives importance to the protection of shared knowledge, practices, and interactions that define the group's identity. Specifically, we discuss the following rights:

Right to Collective Representation The right of a collective, defined by a shared social identity, to protect how that identity is represented and understood in digital spaces, including protecting shared interests and characteristics. Furthermore, the right that such collectives may designate or establish representative bodies to act on their behalf, advocating for and negotiating matters affecting their collective privacy, presents a novel approach to collective privacy. *However, this raises critical questions: How can shared social identity be defined in a legally sound and inclusive manner, considering the dynamicity and intersectionality of identities in the societal and digital space? How can representative bodies be established and held accountable to ensure they reflect the interests of the diverse individuals within the collective?*

Right of Organized Groups Existence Organized groups with a shared purpose and awareness have the right to exist independently and not to be revealed through inferred digital identities, including the right to maintain the confidentiality of their existence [231]. *This right aims to protect the autonomy of organized groups but also opens up questions: How can organized groups be legally sound and consistent, considering the diverse forms of organization and the potential for abuse? And how can the right to maintain the confidentiality of existence be balanced with principles of transparency and accountability, particularly when group activities may have societal implications?*

Right to Collective Consent Management The right of a defined group to manage consent collectively for data processing that affects the group or its identity. This includes setting the scope of consent and having data practices with collective privacy implications. *How can a defined group be established for collective consent? How can collective consent mechanisms be designed to be both effective and accessible, especially in diverse and complex groups? And how can we ensure that collective consent does not inadvertently override individual rights or create new forms of exclusion?*

Right to Collective Erasure or Restricted Processing The right of a defined group to collectively request the erasure or restriction of data that impacts the collective, especially when it is inaccurate, discriminatory, no longer necessary, or poses risks to collective privacy. *However, this raises fundamental questions: How can this collective data management mechanism work practically with current data processing frameworks, including those with distributed data?*

Right to Collective Privacy Impact Assessment The right of a group to have a comprehensive impact assessment, involving meaningful consultation, to evaluate potential collective privacy risks and societal impacts before the implementation of new technologies, data practices, or policies that could significantly affect the group. *Who should be responsible for calling for such an assessment: the affected group, a regulatory body, or the entity proposing the new technology or policy? How can the results of these assessments be translated into actionable policy changes, and how can accountability be established for implementing those changes?*

Right to Collective Data Governance Defined groups have the right to establish and enforce data governance practices governing the collection, processing, storage, and sharing of data concerning them. This includes the right to govern collective data minimization and purpose limitation strategies about the collective. Specifically, such groups should be involved in defining data policies, conducting audits, and determining the permissible use of data that affects the collective privacy. *However, a question remains: how can this participatory data governance mechanism at the collective level be integrated into existing data protection frameworks to ensure consistent application across different sectors?*

6.6 Potential Role of Explainable AI in Supporting Collective Privacy

The role of xAI in supporting collective privacy remains largely exploratory and underexamined. Nevertheless, it offers a promising direction for analyzing how algorithmic systems encode, infer, and act upon group-level information. Many collective privacy risks emerge not from the disclosure of isolated individual records, but from latent inferences, cross-individual dependencies, and the aggregation of shared social, behavioral, or demographic attributes. In this context, xAI may provide analytical tools for making such mechanisms more visible. Methods such as feature attribution, counterfactual reasoning, influence functions, and representation-level analyses could help examine whether models rely on group-specific signals or propagate interdependent

privacy risks. For example, explanation methods may reveal that a model bases its predictions on features correlated with social groups, neighborhoods, family relations, or collective behavioral patterns. Similarly, global and concept-based interpretability techniques may offer insights into how models construct, represent, or use groupings. This raises the possibility that individuals and collectives could better understand, challenge, or contest the bases on which algorithmic classifications are made.

At the same time, xAI may provide a foundation for future governance mechanisms to evaluate group-level impacts. Explanation patterns could reveal disparities in explanation quality, stability, or fidelity across demographic or social groups, thereby exposing forms of collective harm that may not be captured by conventional individual-level privacy or fairness metrics. Influence-based explanations and data-attribution methods could also support investigations into how group-specific data contribute to model behavior, informing debates on collective consent, machine unlearning, restricted processing, and group-level contestation rights.

However, xAI output should not be treated as a definitive solution to collective privacy challenges. Its use in this context requires further methodological development, interdisciplinary evaluation, and human-centered validation. Beyond technicality, explanations must be evaluated for their relevance and usefulness to impacted stakeholders, including individuals, communities, domain experts, regulators, and civil society actors. This is particularly important in collective privacy settings, where harms may arise from group-level inferences, shared attributes, or relational dependencies, and where the affected collective may be implicit or difficult to identify. Human-centered validation is therefore needed to assess whether explanations enable stakeholders to recognize collective privacy risks, understand how algorithmic systems use group-level information, and meaningfully challenge, contest, or govern such uses. Rather than resolving collective privacy concerns on its own, xAI should be understood as a potential analytical lens for identifying, studying, and eventually governing group-level privacy risks.

6.7 Challenges to Preserving and Designing for Collective Privacy

Having discussed the relevance of collective privacy, we now turn to the challenges of preserving it and designing related systems and user interfaces.

Anonymity One of the known and employed properties for privacy-preserving data release is k-anonymity [370]. It aims to protect individual privacy by ensuring that each record in a dataset is indistinguishable from at least $k-1$ other records with re-

spect to quasi-identifiers. This property obfuscates individual identities within groups, thereby protecting against re-identification. However, k-anonymity primarily focuses on the individual level and neglects the inherent interconnectedness of individuals within groups [371]. For instance, k-anonymity groups individuals to obscure individual identity, it primarily focuses on ensuring each record is indistinguishable within that group, not necessarily on concealing the existence or characteristics of the group itself. Consequently, it does not adequately address collective privacy concerns. This issue is particularly amplified in tabular data, where relationships between individuals can reveal sensitive information [372]. For example, even if a dataset contains information about patients from the same workplace and each record satisfies k-anonymity, an attacker with auxiliary knowledge (e.g., knowledge of specific individuals who share a last name) might still be able to infer information about the group.

To mitigate the shortcomings of k-anonymity, extensions such as l-diversity and t-closeness have been proposed [373, 374]. Specifically, the aim is to mitigate homogeneity while remaining susceptible to background knowledge and skewness attacks, in which sensitive attribute distributions compromise privacy. For instance, homogeneity attacks exploit the lack of diversity within k-anonymous groups, enabling adversaries to infer sensitive attributes. In contrast, background knowledge attacks exploit external data to correlate and infer sensitive information about individuals. To address the issue of attribute homogeneity within equivalence classes, l-diversity principle states that an equivalence class is l-diverse if it contains at least l well-represented values for the sensitive attribute. This ensures that even if an individual is identified within a group, the sensitive attribute remains sufficiently diverse to prevent precise inference. Despite these advancements, l-diversity and t-closeness still struggle to address individual privacy issues fully [371]. The focus remains mainly on attribute diversity and distribution within equivalence classes, potentially overlooking other relationships and dependencies that can reveal sensitive group-level information. For example, even in a dataset satisfying t-closeness, if multiple members of the same family, workplace, or minority group appear in the same equivalence class, correlations in their data reveal collective traits. Moreover, the applicability of k-anonymity and its extensions is inherently limited to tabular data and does not handle other data types, such as graphs, text, sequences, or spatiotemporal data, in which relationships and interactions between entities can leak information and where group-level attributes that are not captured by quasi-identifiers alone.

Differential Privacy *Differential Privacy (DP)* [22] is widely adopted across various platforms to enhance user privacy while still allowing valuable data analysis, such as in Facebook [290, 375], Apple [376], Google [377, 378, 379], LinkedIn [380, 381] and also for governmental purposes [382]. The definition of DP (as explained in 2 means

that an attacker cannot readily infer whether a particular individual's data is included in the dataset from the output of a query, because the probability of any given output changes only slightly when that individual's data is present [19].

DP provides protection for individual privacy in data and model releases, but its suitability for safeguarding collective privacy, particularly group and interdependent privacy, remains questionable [255]. Specifically, it focuses on bounding the impact of a single individual's data on query output and is effective for individual-level protection, but it does not inherently extend to protecting arbitrarily defined groups or collections of individuals. In the context of DP, the term group privacy refers to protecting a collection of participants whose data may reveal sensitive information. However, this notion does not account for mutual or interdependent privacy concerns and is limited when applied to large groups, as privacy guarantees are expected to weaken with increasing group size [22, 255, 298, 383]. It is therefore crucial to consider the limitations of DP when collective privacy is a concern.

System Design and User Interaction From a Human-Computer Interaction perspective, addressing collective privacy requires consideration of the distinct roles and constraints of service providers, system designers, and end users. Service providers must balance data-driven functionality with safeguards that prevent collective-level inference and profiling, ensuring that systems do not inadvertently expose group referent information. System designers face the challenge of embedding usable, transparent mechanisms that communicate collective-level risks that individuals may neither anticipate nor directly control, while supporting legitimate group representation and consent. End users, who often interact with systems as individuals rather than as members of collectives, need interfaces that make interdependent and group-level privacy consequences visible without overwhelming them. Designing for collective privacy, therefore, requires moving beyond individualistic models of privacy toward interaction paradigms that reflect shared vulnerabilities, overlapping rights, and the dynamic formation of digital collective identities.

Open Questions

- *How can explainability contribute to the governance of collective data by providing interpretable summaries that inform collective consent?*
- *Could explainability support collective transparency, allowing groups to understand how they are represented by models trained on their data?*
- *How should systems communicate the collective implications of data collected at the individual level, especially when a user's seemingly personal disclosure can expose or reshape the privacy of an entire collective?*

- *How to effectively model and protect collective privacy (group and interdependent), especially for dynamic groups as represented in Table 6.3).*
- *Can we develop mechanisms that account for the inherent dependencies within these collective categories without compromising data utility and privacy?*
- *To what extent can structural information in non-tabular data (e.g., sequences, graphs) be exploited to infer sensitive group-level attributes, and how do traditional privacy models fall short in such settings?*
- *How can we design systems and related user interfaces that account for the tensions between individual and collective rights and privacy protections? How does that impact the users interactions with the system?*

6.8 Conclusion

In this chapter, we highlighted distinct collective privacy considerations and distinguished among collectives that vary in their inherent characteristics, ranging from families characterized by interpersonal relationships to broader social contexts. The aim of our work is to advance this discussion and call for interdisciplinary collaboration, bringing together expertise from law, computer science, social sciences, and ethics to ensure that the development and deployment of data practices at the collective level are guided by respect for collective rights and collective identities. Our contribution to the definition of collective and digital identity highlights the substantial information about a collective that can be inferred from publicly released data and models. Moreover, we discuss potential collective rights and their challenges, including consent, data governance, erasure, and related system design considerations. The rights discussed emphasize key findings regarding the collective dimensions of privacy as we move beyond traditional, individual-centric data protection frameworks. We acknowledge that significant work remains to be researched and call for future research on:

1. Explore legal ways to recognize collective privacy rights within existing data protection regulations and examine how current regulations can be interpreted or adapted to address the unique aspects of collective privacy.
2. Propose frameworks that consider collective privacy protections and investigate the implementation of technical solutions and related user interfaces that enable groups to express and manage their data preferences collectively.
3. Research on algorithmic impacts to collective privacy focuses on identifying and measuring group-level harms and developing privacy-enhancing technologies.
4. Pose questions for the role of xAI in supporting collective privacy, on how interpretable model outputs can help groups understand, contest, and influence how their data is used and how they are represented in algorithmic grouping.

Part II

Fairness and Explainability

Chapter 7

Impact of Fair Learning on Quality of Explanations

In this chapter, we investigate how fairness-enhancing learning techniques affect the quality of model explanations. We propose the Fair-FLIP approach for improving fairness in deepfake detection while preserving predictive performance and explanatory behavior. Through empirical evaluation, we compare Fair-FLIP with baseline fairness methods and analyze its impact on both fairness metrics and feature-attribution-based explanations. This chapter shows that fairness interventions can be designed to improve model behavior while limiting undesirable changes in explanations. This chapter is based on the content available in the following work:

1. Tomasz Szandala, **Fatima Ezzeddine**, Natalia Rusin, Silvia Giordano, and Omran Ayoub. Fair-FLIP: Fair Deepfake Detection with Fairness-Oriented Final Layer Input Prioritizing, Swiss Data Science Conference (SDS) 2025.

7.1 Introduction

Ensuring fairness and transparency in ML systems has become a central concern as models increasingly affect decisions with real social consequences. Many fairness-enhancing techniques, such as adversarial training, regularization, or weight pruning, require substantial modifications to a model's internal architecture or full retraining from scratch [384]. Although effective in reducing bias, these interventions often degrade predictive performance and compromise explainability [384, 385, 386]. Existing bias-mitigation strategies often rely on retraining with protected attributes, threshold adjustments, or auxiliary models, limiting their generalisability and complicating real-world deploy-

ment. Altering large portions of a model's weights can distort learned representations, making it difficult to trace how decisions are made and undermining the reliability of explanation tools such as saliency maps or attention heatmaps. This tension between fairness and interpretability is particularly problematic in sensitive, high-stakes applications where both properties are essential.

In this chapter, we focus on image datasets, particularly on the domain of deepfake detection, where fairness and explainability are not merely desirable but foundational to trustworthy deployment. Deepfake detection has emerged as a critical line of defence [387, 388, 389, 390, 391, 392, 393], given the rapid diffusion of AI-generated content, driven by advances in generative modeling across vision, audio, and language [387, 389, 394], that has amplified societal risks associated with synthetic media. Deepfakes, defined as manipulated or fabricated digital content that convincingly simulates real individuals or events [395], have been weaponized to spread misinformation, perpetrate harassment, and erode public trust [391, 396]. High-profile incidents, such as fabricated speeches by political leaders or privacy-violating hoaxes targeting private citizens [396, 397], illustrate the scale of potential harm.

However, despite substantial progress in accuracy, fairness, and explainability in deepfake detection, it remains underexplored. Unfair detectors risk systematically misclassifying individuals from certain ethnicities or demographic groups [398], thereby offering unequal protection and disproportionately exposing marginalized communities to reputational or psychological harm [399].

To address these limitations, we introduce Fair-FLIP (Fairness-Oriented Final Layer Input Prioritizing), a lightweight post-processing method that mitigates subgroup disparities without retraining or modifying the model's architecture and internal reasoning. Unlike regularization-based or pruning approaches that alter internal weights, Fair-FLIP intervenes only at the final layer by reweighting the contributions of the second-to-last-layer activations. The method downweights features that exhibit high variability across protected subgroups, which are hypothesized to encode protected-attribute-specific information, while promoting features with lower variability, which are more likely to capture general, task-relevant characteristics. Because deeper layers typically encode abstract semantic features [400], adjusting only the final-layer inputs allows us to reduce bias while preserving the integrity of earlier representations. This design enables Fair-FLIP to improve fairness across ethnicities with minimal intervention, no retraining, and broad applicability to architectures such as CNNs, ViTs, and even GAN discriminators, as well as minimal intervention for model explanations (addressing RQ4).

The remainder of this chapter is organized as follows: Sec. 7.2 reviews related work. Sec. 7.3 introduces the Fair-FLIP method. Sec. 7.4 presents the experimental setup and results; Sec. 7.5 discusses limitations and broader implications, and Sec. 7.6

concludes the chapter.

7.2 Related Work

Several studies have focused on deepfake detection across image, audio, and video modalities [387, 388, 389, 390, 391, 392, 393]. These approaches leverage statistical techniques [401, 402, 403, 404], ML-based methods [405, 406, 407], and deep learning-based methods [408, 409, 410, 411, 412, 413].

For example, Matern et al. examine subtle facial-region manipulations, such as changes in eye shading or the addition of accessories, which can be easily overlooked [414]. Sahla et al. propose a lightweight multi-layer perceptron to detect visual artifacts in the face region with minimal computational overhead [406]. A more advanced strategy uses a Generative Adversarial Network (GAN) simulator that synthesizes collective GAN-generated image artifacts as input features for a classifier, thereby enhancing the capacity to distinguish deepfakes [405]. Raza et al. propose a multi-model convolutional neural network (CNN) to further boost artifact detection accuracy [390]. Other notable studies include frequency-domain analysis to distinguish real from manipulated images [392], as well as tensor-based representations of compressed and resized images enriched with discrete cosine transform features [415]. More recent works explore refined model architectures, such as fine-tuned Vision Transformers (ViTs) integrated with support vector machines [393] or CNNs [416], to effectively detect deepfake images generated by Diffusion Models.

Fair deepfake detection, including the assessment of fairness and bias, has increasingly drawn attention. Pu et al. highlight a bias factor of approximately 4% in favour of female subjects when evaluating MesoInception-4 [417]. Similar biases are prevalent in other areas of computer vision and facial recognition [418], where specific subgroups, such as non-Caucasian individuals or individuals wearing glasses, experience disproportionately high error rates [419, 420]. Moreover, studies on XceptionNet reveal performance drops exceeding 30% for certain ethnicities and genders [421].

In terms of proposed de-biasing methods, Ezeakunne et al. propose a pre-processing technique that generates synthetic images by applying transformations such as scaling and rotation to balance under-represented demographics [422]. Ju et al. propose in-processing demographic-agnostic and demographic-aware methods that modify the training objective by incorporating fairness metrics, ensuring a more uniform treatment of subpopulations [423]. Lin et al. propose a framework that detaches demographic and forgery features, albeit at the cost of higher complexity [424]. Despite their effectiveness, such methods can significantly increase training overhead [425] and require detailed model modifications, which are not always feasible [89]. Liu et al. propose

a post-processing technique, namely Bias Pruning with Fair Activations (BPFA), which prunes network weights that exhibit demographic-related biases [426]. While this approach can yield fairer results, it often degrades performance, and the procedure can be time-consuming or unstable.

As in previous work, our work focuses on mitigating bias. However, we take a post-processing approach that requires minimal modifications to the trained model, thereby preserving its predictive performance while ensuring it remains lightweight. Our approach operates on the hypothesis that activations exhibiting high variance across protected attributes tend to encode ethnicity-specific features, while those with lower variance capture more general, ethnicity-independent characteristics. Leveraging this insight, we prioritize final-layer inputs accordingly. The most closely related work to ours is BPFA, which mitigates bias by removing certain weights entirely. In contrast, our proposed approach selectively modifies weights, ensuring that essential features are not lost. For a comprehensive evaluation, we compare the performance of our method with that of Fair-FLIP in Section 7.4.3.

7.3 Problem Formulation and Methodology

7.3.1 Problem Formulation

The problem can be formally stated as follows. Given a dataset containing real and fake images across various ethnicities, our goal is to develop a technique that improves the fairness metrics outlined in chapter 2, while preserving high predictive performance, measured by accuracy and F1-score. Ethnicity is treated as a protected attribute, and fairness metrics are computed accordingly. Note that the de-biasing process must not require the model to be aware of protected attributes during training or inference, nor should it demand substantial architectural alterations or computational resources to be applied. Therefore, our final objective is a lightweight, post-processing method that enhances fairness by countering demographic imbalances without compromising the usability and performance of the original deepfake detector.

7.3.2 Fair-FLIP

Fair-FLIP is a post-processing technique designed to mitigate subgroup disparities by intervening in a trained deepfake detection model. Its core principle is to down-weight activations of the second-to-last layer, often called *features* on which the classification is based, that exhibit high variability across protected subgroups while promoting those with lower variability. This is based on the hypothesis that activations with relatively high variance across protected attributes are biased toward ethnicity-specific features,

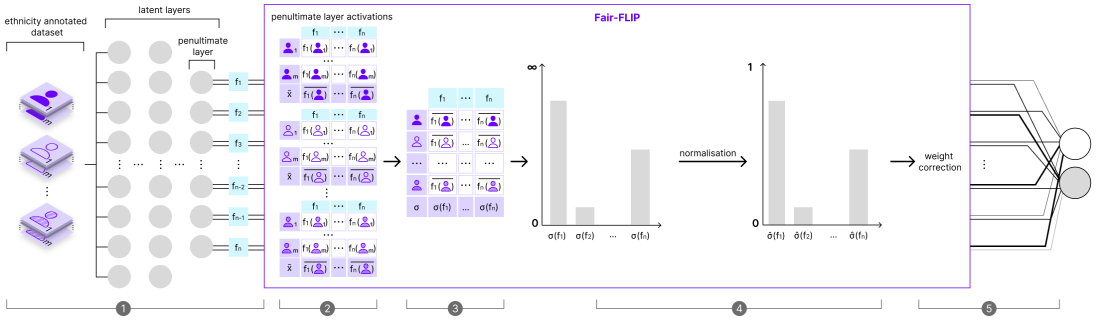


Figure 7.1. Schematic representation of Fair-FLIP's methodology.

whereas those with lower variance capture more general, non-ethnicity-dependent characteristics. We chose to apply de-biasing exclusively to the final layer because deeper layers typically encode more abstract features relevant to classification [400], making it possible to reduce bias without extensively altering multiple earlier layers effectively. By doing so, we aim to mitigate biases present across ethnicities. Overall, Fair-FLIP does not require retraining and is equally applicable to any neural network base architectures, such as ViTs, CNNs, or even a discriminator within a GAN—making it a highly versatile solution.

Figure 7.1 shows a schematic representation of the working of Fair-FLIP, which we describe in the following steps:

1. **Capture penultimate-layer activations**: For each image in the ethnicity-annotated dataset, record the activation vector from second-to-last layer of the deepfake detection model.

2. **Group activations by ethnicity**: Group the activation vectors according to their associated ethnic subgroup and calculate the mean value of activations for each subgroup.

3. **Measure between-group variability**: For each activation f_i, compute the standard deviation of its subgroup-specific mean activations, $\sigma_i(f_i) = std(\bar{f}_i)$).

4. **Normalise variability**: In order to provide comparable values for altering weights, we have to transform each standard deviation into a value within the range $[0, 1]$. We achieve this by normalisation: $\hat{\sigma}(f_i) = \frac{\sigma_i - min(\sigma)}{max(\sigma) - min(\sigma)}$.

5. **Adjust final-layer weights**: Use the normalised standard deviation to modify the original feature weights w_i according to Eq. 7.1, where α is a term controlling how strongly each feature is promoted or demoted.

$$w_i' = w_i \times \big(1 + \alpha - \hat{\sigma}(f_i)\big), \tag{7.1}$$

As a result, the weights of the final layer will be modified so that features heavily influenced by ethnicity are suppressed, while more general, reliable signals remain central to the detector's decision-making.

7.3.3 Benchmark Approaches

We compare the performance of Fair-FLIP to a baseline model and state-of-the-art bias mitigation techniques.

Baseline. The baseline model is a ViT-based classifier trained to label each image as *authentic* or *deepfake*, with no regard to fairness. It represents the model with the highest predictive performance.

Pre-processing. To reduce data-level bias, we perform pre-processing by undersampling overrepresented ethnic subgroups. This preserves the original model architecture and training procedure as it does not add or alter any hyper-parameters. However, since some examples from overrepresented subgroups are discarded, potentially informative data is lost, harming accuracy. We opt for simple undersampling over more sophisticated methods (e.g. Tomek links, SMOTE, ADASYN) to avoid artefacts and simplify the pipeline.

In-Processing. Building on Zafar *et al.* [89] in-processing strategy, we embed fairness penalties directly in the loss function. Specifically, we augment the training objective (eq. 7.2), encouraging balanced performance across subgroups. Increasing λ intensifies the fairness objective at the cost of diminishing accuracy.

$$\text{total\_loss} = \text{loss} + \lambda\,(1 - \text{TPP})^2\,(1 - \text{FPP})^2 \tag{7.2}$$

Threshold-based Post-Processing. In threshold-based post-processing, we adjust the decision threshold (eq. 7.3) to maximise a fairness objective [93].

$$\text{TPP}^2 \times \text{FPP}^2 \times \Big(1 - \max(\text{FPP}_a, \text{FPP}_b, \dots)\Big)^2 \times \text{PPV}^2 \times \text{NPV}^2 \tag{7.3}$$

This approach seeks to balance the True/False Positive Parities (TPP, FPP) and predictive values (PPV, NPV) without retraining. A global threshold from 0.0 to 1.0 is explored in increments of 0.001 to identify the configuration that maximises fairness. While this technique optimises fairness, it may come at the cost of a significant decline in accuracy.

Bias Pruning with Fair Activations. We include another post-processing strategy, namely, *Bias Pruning with Fair Activations (BPFA)* [426], which removes specific weights deemed responsible for demographic-related biases in a post-processing manner. This approach does not alter the data or training procedure but refines network parameters to mitigate biased representations.

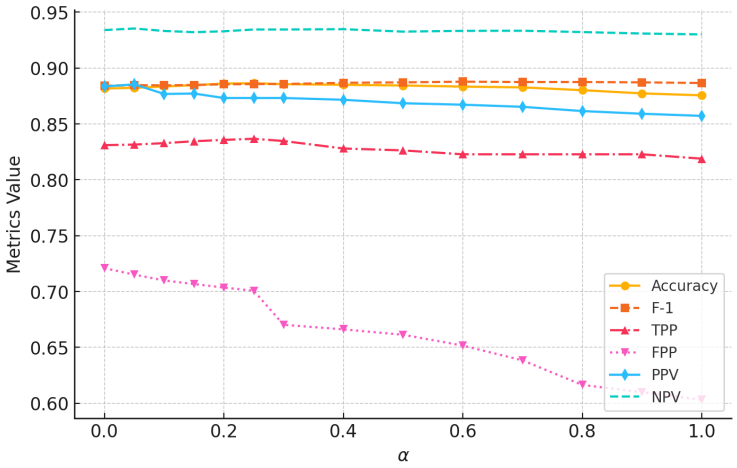


Figure 7.2. Sensitivity analysis of trade-offs between fairness metrics and accuracy as α changes.

7.4 Numerical Results

7.4.1 Experimental Settings

We used a Kaggle dataset of 190335 face images, evenly split between manipulated (95,134) and authentic (95,201) [427]. Each 256×256 JPEG was assigned an ethnicity label. Ethnicity annotations (White, Black, Asian, Indian, Latino, Middle Eastern) were obtained using a 97% accurate facial attribute model (HyperExtended LightFace [428]) and later verified by manual inspection.

Our baseline model for experiments is a Vision Transformer [416] (google/vit-base-patch16-224-in21k, pre-trained on `ImageNet-21k` trained for 15 epochs with standard cross-entropy loss. We used Adam (LR=0.001, $\beta = (0.9, 0.999)$) in `PyTorch 1.13.1+cu116`, halving the learning rate every five epochs, with a batch size of 64 and no data augmentation.

As evaluation metrics, we consider accuracy and F1-score as metrics for predictive performance and TPP, FPP, PPV, and NPV as fairness metrics.

7.4.2 Determining the value of α

To determine the value of α parameter of Fair-FLIP for the subsequent experiments, we performed a sensitivity analysis varying the value of α.

Figure 7.2 reports the accuracy and fairness metrics achieved for values of α ranging between 0.0 and 1.0. Results show that α may have a crucial impact on NPV while not heavily impacting other metrics. For our experiments, we select $\alpha = 0.25$, which

achieves the best balance between accuracy and fairness metrics, in our opinion.

7.4.3 Comparative Analysis

To validate Fair-FLIP, we conduct a five-fold cross-validation against a baseline model and state-of-the-art fairness interventions discussed in Section 7.2. Table 7.1 reports the performance of the various approaches in terms of both predictive performance and fairness metrics.

As expected, the *Baseline* approach delivers the highest predictive performance, achieving an average accuracy of 0.8731 and an average F1-score of 0.8769 across the five splits. *Fair-FLIP* demonstrates performance closely comparable to that of *Baseline*, with an average accuracy of 0.8708 (a marginal decrease of only 0.0023) and an average F1-score of 0.8696 (0.0073 lower than *Baseline*). Moreover, Fair-FLIP outperforms *BPFA*, which achieves an average accuracy of 0.8487 and an average F1-score of 0.8509, as well as other benchmark approaches, some of which exhibit significantly lower accuracy, such as the *Pre-Processing* technique, which reaches only 0.7648.

In terms of the fairness metrics, we first notice that *Baseline* demonstrates an acceptable TPP (0.8523), PPV (0.8184) and NPV (0.9064) but suffers largely in terms of FPP (0.5171). The *Pre-Processing* approach, which compensated heavily on accuracy, demonstrates an improved TPP and FPP with respect to *Baseline* but a slightly degraded PPV and NPV. This shows that while *Pre-Processing* approach can noticeably improve fairness metrics, it considerably reduces accuracy by discarding potentially informative samples. Similarly, *In-process*, in which the predictive performance is like-wise degraded (up to 4%), demonstrates minor improvement (TPP from 0.8523 to 0.8588) and FPP (from 0.5171 to 0.6622) but shows a slight degradation in terms of PPV and NPV. This shows that the *In-Processing* approach also improves parity measures but at the expense of diminishing accuracy and, as we will show later, additional complexity. *Post-processing: Threshold-based* approach, on the contrary, demonstrates an improvement in terms of all fairness metrics (TPP by 0.01, FPP by 0.17, PPV by 0.02 and NPV 0.02) at the cost of only 4% decline in predictive performance in comparison to the *Baseline*. Similarly, *BPFA* effectively address some subgroup bias (FPP increased by 0.17 and PPV by 0.01) but still incur moderate decreases in classification metrics (by 3%). In contrast, *Fair-FLIP*, which preserves to a large extent the classifier's performance (drop by 0.002), while demonstrating state-of-the-art improvement across all fairness metrics (TPP by 0.01, FPP by 0.16, PPV by 0.06 and NPV 0.02). These results show the effectiveness of Fair-FLIP in mitigating bias (over 10% in relative numbers) while incurring negligible impact on predictive performance(less than 0.25%).

Table 7.1. Table summarising different metrics for model's performance during five-fold cross-validation

	Fold	Baseline	Pre-Process	In-Process	Post-Process	Pruning (BPFA)	Fair-FLIP
Accuracy	1	0.8865	0.7561	0.8681	0.8772	0.8637	0.8856
	2	0.8880	0.7902	0.8532	0.8391	0.8602	0.8846
	3	0.8715	0.7537	0.8496	0.8226	0.8437	0.8691
	4	0.8520	0.7342	0.7921	0.8031	0.8332	0.8486
	5	0.8677	0.7899	0.8278	0.8188	0.8429	0.8663
	Avg	**0.8731**	**0.7648**	**0.8382**	**0.8322**	**0.8487**	**0.8708**
F1-Score	1	0.8876	0.7561	0.8782	0.8772	0.8637	0.8864
	2	0.8912	0.7902	0.8601	0.8421	0.8572	0.8836
	3	0.8755	0.7537	0.8343	0.8286	0.8477	0.8678
	4	0.8579	0.7542	0.8037	0.8078	0.8332	0.8430
	5	0.8723	0.7989	0.8181	0.8104	0.8529	0.8673
	Avg	**0.8769**	**0.7706**	**0.83888**	**0.8332**	**0.8509**	**0.8696**
TPP	1	0.8394	0.8905	0.8751	0.8333	0.8200	0.8467
	2	0.8662	0.8842	0.8385	0.8652	0.8410	0.8710
	3	0.8575	0.8835	0.8580	0.8725	0.8680	0.8605
	4	0.8575	0.8780	0.8403	0.8590	0.8601	0.8680
	5	0.8409	0.8804	0.8820	0.8698	0.8679	0.8567
	Avg	**0.8523**	**0.8833**	**0.8588**	**0.8600**	**0.8514**	**0.8606**
FPP	1	0.5749	0.7340	0.7164	0.6988	0.6821	0.7007
	2	0.4225	0.6524	0.6322	0.7002	0.6788	0.7007
	3	0.5661	0.6990	0.6509	0.6880	0.6988	0.6978
	4	0.5966	0.6845	0.6989	0.6782	0.6637	0.6702
	5	0.4256	0.5858	0.6124	0.6501	0.7070	0.5989
	Avg	**0.5171**	**0.6711**	**0.6622**	**0.6831**	**0.6861**	**0.6737**
PPV	1	0.8391	0.7852	0.8092	0.8794	0.8152	0.8854
	2	0.8530	0.7912	0.7796	0.8802	0.8413	0.8732
	3	0.8224	0.7968	0.8255	0.8699	0.8237	0.8779
	4	0.7935	0.8045	0.8103	0.8760	0.8381	0.8637
	5	0.7841	0.7860	0.7998	0.8809	0.8252	0.8802
	Avg	**0.8184**	**0.7927**	**0.8049**	**0.8773**	**0.8287**	**0.8761**
NPV	1	0.9096	0.8040	0.9062	0.9401	0.8843	0.9315
	2	0.9131	0.8838	0.8976	0.9327	0.8909	0.9345
	3	0.9038	0.8224	0.9124	0.9234	0.8790	0.9238
	4	0.8837	0.8108	0.8734	0.9033	0.8648	0.9036
	5	0.9219	0.8638	0.9001	0.9315	0.8814	0.9288
	Avg	**0.9064**	**0.8370**	**0.8979**	**0.9262**	**0.8801**	**0.9244**

7.4.4 Algorithms Complexity

We now discuss the computational complexity of the various approaches using *big-O* notation[1].

Table 7.2. Computational complexity of the considered fairness methods

Method	**Complexity (Big O)**
In-Process	$O(e \cdot x/b \cdot a)$
Post-Process: Threshold	$O(x + 1000 \cdot a) = O(x + a)$
Post-Process: BPFA	$O(x + a \cdot n \cdot w)$
Post-Process: Fair-FLIP	$O(x + a \cdot 2 \cdot w) = O(x + a \cdot w)$

Table 7.2 reports each algorithm's complexity, omitting pre-processing since it has no dedicated fairness step. The in-process approach adds a fairness penalty (*fairness_loss*) to each batch. With a protected groups, x total samples, batch size b, and e epochs, the complexity grows multiplicatively. By contrast, Post-Process (Threshold-based) performs one pass to collect statistics and then applies fairness adjustments. Threshold tuning explores around 1000 cut-offs between 0.0 and 1.0, adding a constant factor. Bias pruning computes variance scores for each of n neurons and their w weights across a groups, leading to a higher multiplicative term. Fair-FLIP focuses solely on the final layer's weights. We measure how each feature's variance differs across a protected groups and then re-scale the weights w of each neuron. For a binary classifier like a deepfake detector, we have two output neurons, thus the Fair-FLIP offers simplified complexity in comparison to BPFA.

7.5 Further Analysis and Limitations

Impact on Explainability. We extend our analysis to investigate the impact of our approach on the model's explainability. We apply *Activation Rollout* [429] on three variants of our deepfake detector (i.e., *Baseline*, *BPFA*, and our proposed approach, Fair-FLIP) and compare the resulting heatmaps. Fig. 7.3 shows an example of resulting heatmaps on one real and two manipulated images.

Our observations indicate that while both Fair-FLIP and BPFA attention heatmaps are very close to that of the Baseline, the BPFA's aggressive weight pruning may alter the attention patterns more drastically, potentially discarding certain intermediate

[1]Due to the different nature and principles of the fairness de-biasing methods, a comparison of execution time is not fair.

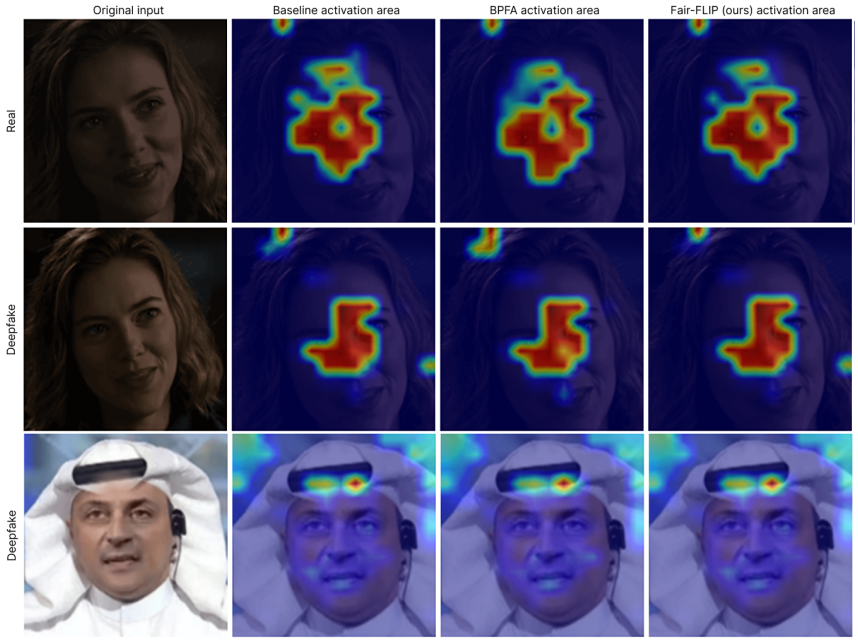


Figure 7.3. Attention heatmaps for the baseline, BPFA, and Fair-FLIP. In the first row (genuine image), baseline and Fair-FLIP match closely, whereas BPFA's activations are more blurred near the right eyebrow. In the second row, BPFA shows slightly stronger activation in the top□left and an extra highlight around the mouth, with some missing heat signature on the right; Fair-FLIP's region of interest remains only marginally narrower than the baseline's. In the final example, there are no visible differences between activations, implying no changes in how the model perceives the image.

features. On the contrary, Fair-FLIP's final-layer reweighting is comparatively less invasive, preserving most of the model's discriminative cues. While further analysis is still necessary to draw conclusions, these preliminary investigations show that Fair-FLIP demonstrates an attention similar to that of the Baseline approach.

Ethical Considerations. Since Fair-FLIP mitigates bias without exposing sensitive user data, i.e., it does not require access to protected attributes or deduce them during inference, it can be considered demographically unintrusive. This characteristic allows us to uphold key ethical principles by promoting fairness without compromising user privacy. This is crucial in real-world applications where demographic profiling raises legal and ethical concerns, such as re-identification risks or unintended discrimination. Moreover, since Fair-FLIP applies adjustments to the final layer rather than intervening in the training process, it maintains transparency and interpretability more than other approaches, such as in-processing techniques.

Limitations and Future Work. The limitations and future works can be sum-

marised as follows.

- We demonstrated improvements with respect to a single protected attribute; similarly, unfair discrepancies may be seen simultaneously for multiple attributes. It is valuable to research how Fair-FLIP can counter multi-attribute bias.
- The method was tested on a single dataset with annotated demographic attributes. Real-world scenarios often involve more diverse data, including non-human objects or broader environmental contexts.
- Further research involves generalizing Fair-FLIP to multi-modal settings, like video and audio manipulations, which involve sequential data.

7.6 Conclusion

In this chapter, we addressed the problem of fairness in deepfake detection, with particular attention to the tension between bias mitigation, predictive performance, and explainability. We proposed Fair-FLIP (*Fairness-Oriented Final Layer Input Prioritizing*), a lightweight post-processing method that improves subgroup fairness without retraining the model or altering its internal architecture, thereby preserving explanations and reasoning. By limiting the intervention to the final layer, Fair-FLIP avoids disrupting earlier learned representations and remains applicable to a wide range of neural architectures, including CNNs, Vision Transformers, and GAN-based discriminators.

The experimental results demonstrate that Fair-FLIP achieves a favorable balance between fairness and predictive performance. Compared with the baseline model, Fair-FLIP substantially improves fairness-related metrics while preserving almost the same classification accuracy. Furthermore, the analysis of attention heatmaps suggests that Fair-FLIP has a limited impact on the model's explanation patterns. While future work on extensive explainability studies is required, preliminary evidence indicates that Fair-FLIP preserves the model's main regions of attention more closely than other post-processing methods, such as BPFA. This is an important result because fairness interventions should not only improve numerical parity but should also preserve the interpretability and auditability of the underlying model.

Chapter 8

Generating Fair Counterfactual Explanations

In this chapter, we address the fairness of CFs by distinguishing between fair predictions and fair recourse, arguing that a model that satisfies fairness constraints does not necessarily provide equitable explanations or actionable paths to different protected groups. We propose an RL-based framework for generating CFs that satisfy individual, group, and hybrid fairness constraints. Through experiments across multiple datasets, we evaluate the proposed method's ability to balance CF validity, quality, and fairness. This chapter is based on the content available in the following works:

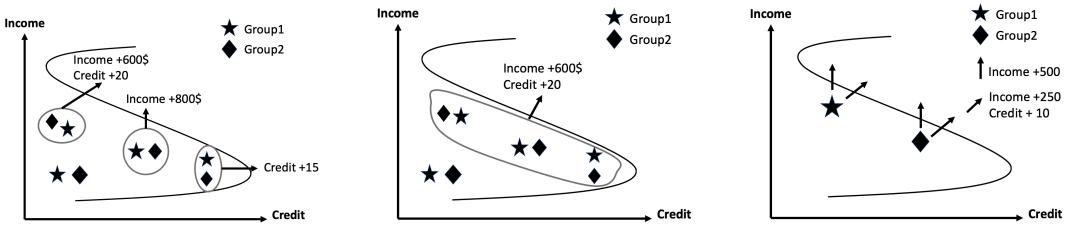
1. **Fatima Ezzeddine**, Omran Ayoub, Davide Andreoletti, and Silvia Giordano. SAC-FACT: Soft Actor-Critic Reinforcement Learning for Counterfactual Explanations. World Conference on eXplainable Artificial Intelligence (xAI 2023).

2. **Fatima Ezzeddine**, Obaida Ammar, Silvia Giordano, and Omran Ayoub. Fair Recourse for All: Ensuring Individual and Group Fairness in Counterfactual Explanations. IEEE Transactions on Artificial Intelligence Journal, 2026.

8.1 Introduction

Beyond technical transparency, fairness is indispensable for protecting individuals from discriminatory treatment and ensuring equitable outcomes in ML–driven decisions. Indeed, fairness is a concern that must be addressed at both the model level and the explanation level [430]. At the model level, fairness is violated when two individuals with comparable characteristics receive different outcomes due to attributes that

should not influence decisions, such as protected attributes like gender or race [431]. For instance, if two individuals have comparable financial profiles but different genders, the model may approve a loan for one and not for the other. This may occur due to biased training data that reflects inequalities, such as those based on certain demographic groups.



(a) Equal Effectiveness-Individual

(b) Equal Effectiveness-Group

(c) Equal Choice for Recourse-Group

Figure 8.1. Illustration of Equal Effectiveness (Individual and Group) and Equal Choice for Recourse.

Ensuring fairness in explanations goes beyond the fairness of the underlying ML model, as the methods used to generate explanations can themselves differ and potentially propagate or amplify unfairness [432, 433, 434]. Even minor feature perturbations can compromise the fairness of CF recourse by limiting its consistency and similarity across comparable individuals from different groups [430, 435], as authors in [436] show that minor data perturbations can result in up to 20 times higher recourse costs for marginalized groups. For instance, consider two individuals with comparable financial profiles who are both denied a loan by an ML model. For the first individual, the CF suggests increasing the salary by 10k. For the second individual, the CF advises adding a co-signer and improving their credit score, without mentioning salary. This discrepancy, in which individuals in similar circumstances receive different CFs, indicates bias. Such bias can be attributed to the fact that CF generation methods aim to find the smallest possible change needed to alter the prediction, however, if the smallest change for individuals within a particular group is to modify an immutable or disproportionately difficult feature, such as age or educational background, the resulting CFs become not only impractical but also unfair. Biased CFs thus have consequences beyond a single explanation, as they erode user confidence and undermine trust not only in the AI system but also in the organization deploying it, as the reliance on biased AI systems can result in reputational harm and legal liability.

In this context, ensuring fairness in CFs, a problem formally referred to as Fair CF, is equally critical to generating a traditional CF [437, 438]. A fair CF generation method should explicitly integrate notions of fairness into its optimization objectives. Fairness

in CFs (discussed formally in Sec. 8.2) focuses on ensuring that similar individuals (i.e., have similar attributes) and individuals pertaining to different demographic groups have comparable explanations.

Fair CFs can be considered at the individual level, group level, or, as we will show later, at a hybrid level, involving both individuals and the group they pertain to. More specifically, individual fairness [435, 439] refers to treating individuals with comparable attributes similarly. Group fairness [430, 440] aims to ensure fairness at a broader level, focusing on the equal treatment of different groups to avoid discrimination based on group-sensitive attributes, such as gender or race. Individual and group fairness in ML often exhibit a complex, contradictory, and orthogonal relationship as each has a specific focus and aims [439]. Recently, the notion of *hybrid fairness* emerged [439, 441], where, instead of prioritizing a single fairness consideration (individual or group), it acknowledges the diverse and often conflicting nature of fairness demands, and aims to achieve multiple fairness requirements simultaneously. On the model level, the research in [442] explores the conditions under which group fairness and individual fairness can be achieved simultaneously. It provides theoretical analyses and methods to balance the trade-offs between these levels of fairness for recommender systems. On the explanation level, and more specifically, in the context of CF explanations, hybrid fairness indicates that the explanations provided are not only individually fair, giving comparable explanations to similar individuals with a comparable set of attributes, but also collectively fair, avoiding disparate impacts across different demographic groups. To date, little to no work has specifically explored approaches that aim to balance group-level and individual-level fairness in generating CFs.

In this chapter, we address the problem of fair CF generation considering individual, group, and hybrid fairness by first defining metrics that capture fairness at these levels, and then formulating the problem as a model-agnostic optimization task solved using reinforcement learning (RL). We then customize reward functions that account for fairness, such as optimizing for the proportion of individuals and groups achieving recourse, diversity of choices for groups, and quality of CFs in terms of common metrics such as plausibility, actionability, and proximity. Specifically, to take fairness into account, we shift the definitions of common properties such as equal effectiveness (EE) and equal choice of recourse (ECR), which were originally introduced to audit fairness in ML models [2], and use them as guiding principles for the RL. We focus on *EE (Individual Fig. 8.1 a), Group Fig. 8.1 b) and 8.1 c))* that indicate that individuals across different demographic groups receive comparable CFs when seeking to achieve a favorable outcome, and ensures that individuals from different groups have equivalent diversity of recourse options (Detailed later in Sec. 8.4). To the best of our knowledge, this is the first work to introduce the notion of hybrid fairness for CF. We investigate to what extent considering hybrid fairness still allows individuals to achieve recourse,

examine the trade-offs with respect to both individual- and group-level fairness, and analyze its impact on the quality and accessibility of recourse. The contributions of our work can be summarized as follows:

- We model the problem of generating fair CFs as an optimization problem and propose a model-agnostic RL-based approach to optimize for fair CFs across different groups. We design customized reward functions to account for diverse CF fairness requirements, such as equal effectiveness and equal choice of recourse, as well as CF quality metrics, including validity and plausibility.
- We formulate individual, group, and hybrid fairness for CF generation and quantify the effectiveness of our proposed RL-based approach in guaranteeing fairness and achieving recourse for individuals and groups.
- We provide a comprehensive analysis that quantifies the impact of different fairness levels, including hybrid, individual, and group fairness, on CF quality. This analysis systematically compares our method against well-established baselines, highlighting the trade-offs between fairness enforcement and CF quality.

8.2 Preliminaries on Fair Counterfactual Explanations

We consider fair CF generation at three levels, namely, individual, group, and hybrid.

Individual-level fairness Drawing upon the principle of *treat similar individuals similarly*, individual-level fairness constrains that minor variations in input features should not result in significantly different CFs, as shown in Fig. 8.1 a). Fairness metrics employed for individual-level fair CFs often incorporate stability and robustness metrics [435], which aim to quantify the extent to which similar individuals (i.e., individuals with nearly identical input attributes) receive similar CF explanations (i.e., CF explanations that suggest changes in the same or similar attributes, and in comparable magnitude).

Group-level fairness Drawing upon the principle of *fairness through unawareness*, group-level fairness refers to ensuring that different protected groups, which are defined as groups of individuals sharing a protected group (e.g. race, gender, nationality), have equal recourse opportunities to achieve favorable outcomes (i.e., the ability to achieve a recourse through CFs should not significantly vary in difficulty across protected groups). Fig. 8.1 a) illustrates this concept by supposing the data subset in the shaded area, where the two protected groups are expected to receive similar CF adjustments (represented by the arrow) exemplifying how groups get a similar CF to cross the decision boundary. These adjustments aim to ensure a recourse by reducing

disparities in achieving favorable outcomes, thereby preventing one group from facing disproportionately greater difficulty in obtaining recourse. Fig. 8.1 c) additionally demonstrates the equal opportunity concept, where both protected groups have equal choices to achieve recourse (a similar number of possible and valid recourses).

Hybrid-level fairness In the context of CF generation, hybrid fairness aims to generate fair CFs that are actionable for individuals while addressing disparities across groups collectively. For example, in an educational setting, individual-level fairness ensures that two students with comparable academic performance and engagement receive similar CF suggestions (e.g., taking a specific preparatory course to qualify for a scholarship), while group-level fairness ensures that students from underrepresented or disadvantaged backgrounds are not consistently given harder or less attainable recommendations. Hybrid fairness ensures both personalized academic guidance and equitable opportunities across different demographic groups. Generating CFs under hybrid-level fairness is challenging, as it must balance individual and group fairness, which often have different focuses and conflicting objectives.

8.3 Related Work

8.3.1 Generating Counterfactual Explanations

Works aim to produce CFs aligned with standard optimization criteria (e.g., closeness to the original instance or actionability), do not explicitly incorporate fairness objectives (presented in previous chapters). More recently, research efforts began investigating fairness in relation to CFs [430, 443, 444, 445, 446]. The literature on fair CFs can be broadly distinguished into model-agnostic and model-specific approaches. While the former enforce fairness in the generation of CFs themselves, such as Artelt et al. [430] propose a model-agnostic method for generating group fair CFs while minimizing the complexity of CFs across protected groups. The latter embeds CF-based fairness criteria into the training process to promote or evaluate fair model predictions. For instance, Asher et al. [443] investigate and prove that CFs can hide model unfairness. Von et al. [444] defines fairness criteria for recourse at both the individual and group levels that account for causal relationships between features. Similarly, Russel et al. [447] explores fairness under multiple causal models, enabling fair decisions without requiring a precise causal specification. Rawal et al. [448] optimizes resource correctness, interpretability, and cost efficiency across populations and subpopulations to ensure fairness. Also, Han et al. [449] enhance model fairness by retraining the model to optimize for group fairness and counterfactual fairness, flipping the sensitive attribute

without optimizing fairness in the explainer.

Other research has focused on robustness, relating it to individual-fairness in CF generation [116, 450, 451, 452], as unstable CFs can undermine reliability and lead to unfair treatment of specific groups [435]. Artelt 2024 et al. [450] propose a methodology that clusters similar instances and generates a single CF for multiple instances, but does not explicitly account for fairness.

8.3.2 Counterfactual Explanations for Auditing Fairness

Beyond the works that incorporate fairness into CF generation, several studies have leveraged CFs to audit fairness in models by using the insights CFs provide to detect global discriminatory behavior, revealing whether similar individuals are treated differently in terms of explanations [1, 2, 74, 437, 445, 453, 454, 455, 456]. For example, in [453], the authors propose an approach to detect model bias by comparing sensitive and correlated attributes of CFs across protected groups while in [454], the authors propose a bias detection framework that utilizes CFs to identify minimal-cost positive outcomes to detect proxies for sensitive features. Kuraomi et el- [437], define two metrics to quantify model fairness, burden, and normalized accuracy-weighted burden, which integrate accuracy and counterfactual-based fairness measures using CFs. Also, Ma el al. [457] propose graph counterfactual fairness to evaluate fairness from a causal perspective, comparing each individual's original prediction to its CF. Coston et al. [458] investigates connections between standard fairness metrics and their CF counterparts. Ley et al. [1] introduce *GLOBE-CE*, a method to generate summaries for global CF-based explanations that are used to detect fairness. Also, Kavouras et al. [2] presents a framework, referred to as *FACTS*, to audit fairness through CF resources at both individual and subgroup levels, addressing individual and group-level fairness through detecting statically embedded unfairness by applying item set mining algorithms on the data and computing the proposed metrics. Similarly, Fragkathoulas et al. [3] introduce a graph-based framework for generating group CF with shared properties of explanations with the goal of audit model fairness. For a more in-depth analysis on the difference between our work and works that audit for fairness with CF, we highlight in Tab. 8.1 the main distinctions between our work and the most relevant approaches, namely GLOBE-CE [1], FACTS [2], and FGCE [3].

In this work, we focus on generating fair CFs, in contrast to prior studies that primarily use CFs as tools for auditing fairness. We propose a model-agnostic RL–based framework that explicitly incorporates novel fairness constraints into the CF generation process of a fair model, thereby enabling the systematic treatment of individual, group, and hybrid fairness on the explanation level. Moreover, we contribute to the formulation of hybrid fairness, as it unifies individual- and group-level considerations, which

Table 8.1. Comparison of GLOBE-CE [1], FACTS [2], FGCE [3], and our RL-based CF generation approach.

	GLOBE-CE	**FACTS**	**FGCE**	**RL-Based**
Objective	Summarize global recourse patterns	Audit subgroup recourse disparities	Audit feasible recourse for groups	Generate CFs under fairness constraints
Scope	Global	Individual, Group	Group	Individual, group, hybrid
Method	Translation directions that cover many instances while balancing coverage and cost	Itemset mining to identify common affected subgroups and shared recourse	Feasibility graphs over instances and CFs, then extracts groups	RL to explore and exploit mixed feature spaces under fairness-aware rewards
Evaluation	Coverage, cost, speed, and bias-detection user study	Rankings and cost–effectiveness trade-offs	Coverage–cost–CF trade-offs, saturation behavior	Fairness and CF quality metrics
Strength	Interpretable global view	Finds hidden subgroup gaps	Captures feasible group trade-offs	Optimizes fairness during generation
Limitation	No direct fairness optimization	Auditing only	No individualized fair CFs	High training cost

is, to the best of our knowledge, has never been addressed before. Additionally, our method, by design, generates a diverse collection of CFs from which individuals across different groups can equitably select, ensuring broad and fair access to viable recourse options while simultaneously optimizing well-established CF quality metrics such as actionability, sparsity, and proximity. We also extend the fairness definitions introduced by [2], shifting the focus from auditing model fairness to positioning fair CF generation as a primary objective, and design customizable RL reward functions accordingly. Finally, by embedding fairness constraints directly into the generation process, our approach moves beyond cost-centered objectives and provides a principled foundation for quantifying the trade-offs between fairness and other conventional evaluation metrics.

8.4 Problem Formulation and Methodology

8.4.1 Data, Model and Counterfactual Representation

We follow the definitions outlined in [2] regarding the data, model, and input space representation. We also illustrate how a CF is represented as an action in the feature space and formally define what constitutes a group. The feature space, denoted by $\mathscr{X}$, is defined as the Cartesian product of n individual feature spaces, that is, $\mathscr{X} = \mathscr{X}_1 \times \mathscr{X}_2 \times \cdots \times \mathscr{X}_n$. Among these, the feature space $\mathscr{X}_n$ represents a sensitive or protected attribute that takes binary values 0 or 1. Each individual is represented by a single instance $x \in \mathscr{D}$ where $\mathscr{D}$ is the training dataset. We consider a binary classifier $h : \mathscr{X} \to \{0, 1\}$, which maps each instance $x \in \mathscr{D}$ to a label in the set $\{0, 1\}$. The label 1 denotes a favorable outcome, while the label 0 signifies an unfavorable outcome. The classifier h operates over the feature space $\mathscr{X}$, providing a decision for each instance based on its attributes.

A *group* G is a subset of the dataset D and it consists of distinct individuals who each possess a sensitive or protected attribute $\mathscr{X}_n$. Individuals of the same group share similar attributes and are positioned similarly within the feature space. Considering a sensitive attribute, we define two sub-groups:

$$G_0 = \{ x \in D : \mathscr{X}_n = 0\} \quad \text{and} \quad G_1 = \{ x \in D : \mathscr{X}_n = 1\}. \tag{8.1}$$

where G_0 consists of individual instances with the sensitive attribute $\mathscr{X}_n = 0$, and G_1 consists of individuals with $\mathscr{X}_n = 1$. Our analysis focuses on the *affected dataset* $D_{affected} \subseteq \mathscr{X}$, which consists of instances from $\mathscr{X}$, including individuals who have been classified unfavorably, i.e., for whom $h(x) = 0$.

An *action* represents a CF as a transformation that modifies feature values and is denoted by $a \in \mathscr{A}$, where $\mathscr{A} = \{a_1, a_2, \ldots, a_i\}$ is the set of possible actions, where each action defines a way to alter the features of an instance $x \in D$. We call x' the *CF instance* of x under the action a ($x' = a(x)$) if a provides *recourse* to x and change the classifier's decision from unfavorable to favorable, that is, if $h(x) = 0$ and $h(x') = 1$. For example, consider a loan approval classifier h that predicts whether a loan application is approved ($h(x) = 1$) or denied ($h(x) = 0$). Let an instance x represent a loan application with a credit score of 580 and an annual income of 30k. Suppose the classifier outputs $h(x) = 0$, indicating the loan application is denied. An action $a \in \mathscr{A}$ can be to increase the credit score by 40 and the income by 5k. Applying this action transforms the instance x into a new instance $a(x)$, which has a credit score of 620 and an income of 35k. If the classifier now outputs $h(a(x)) = 1$, the action a is considered to provide *recourse* to the applicant, as it changes the decision from unfavorable to favorable.

8.4.2 Counterfactual Fairness Metrics

To achieve these fairness objectives, we need to optimize for two terms, namely, *equal effectiveness (EE)* and *equal choice of recourse (ECR)*. Following the work of [2], we aim to define *EE* and *ECR*, which were primarily used as metrics to audit the fairness of ML models.

We start with *effectiveness*, as a metric to evaluate the fairness of a single action a (one CF) on a subset of the data. The *effectiveness* (also referred to as *coverage*) of an action a quantifies its utility across a group of individuals. Specifically, it measures the proportion of individuals in an affected group $G \subseteq D$ who would receive a favorable outcome if the action a were employed for all of them. This metric reflects how broadly applicable or beneficial a given action is within the group. Formally, the effectiveness of action a for group G is:

$$\text{eff}(a, G) = \frac{1}{|G|} \left|\{x \in G \mid h(a(x)) = 1\}\right|. \tag{8.2}$$

To evaluate the individual and group fairness of a set of actions $\mathscr{A}$ for a group G. We consider individual-effectiveness (accounting for individual fairness) and group-effectiveness (accounting for group fairness), respectively.

Individual-Effectiveness: From a individual perspective, individuals are examined independently of the group they belong to. Similar individuals select an action from $\mathscr{A}$ that is most advantageous for them, as shown in Fig. 8.1 a). Individual-effectiveness measures the proportion of individuals who can achieve recourse by choosing one action a from $\mathscr{A}$ that, when employed, achieves their desired outcome at the least cost:

$$\text{aeff}_{\mu}(\mathscr{A}, G) = \frac{1}{|G|} \left|\{x \in G \mid \exists a \in \mathscr{A}, h(a(x)) = 1\}\right|. \tag{8.3}$$

Group-Effectiveness: From a group perspective, a group G is treated collectively with a single action applied to all individuals, meaning examining the overall outcome for the group as a whole, as shown in Fig. 8.1 b). The group-effectiveness of a set of actions $\mathscr{A}$ for a group G is defined as the maximum proportion of individuals in G who can achieve recourse through the same action:

$$\text{aeff}_{M}(\mathscr{A}, G) = \max_{a \in \mathscr{A}} \frac{1}{|G|} \left|\{x \in G \mid h(a(x)) = 1\}\right|. \tag{8.4}$$

Maximizing group-effectiveness across different CFs ensures that the framework can identify actions that are collectively beneficial across the group, minimizing disparities and ensuring fair outcomes for all members.

8.4.3 Problem Formulation and Objectives

Given a fair binary classifier $h : \mathcal{X} \rightarrow \{0, 1\}$, where 0 indicates a unfavorable outcome. An affected dataset $D_{affected}$ consists of individuals who received unfavorable outcomes ($h(a(x) = 0)$. $D_{affected}$ consists of protected groups $G = \{G_1, G_2\}$, which partition $D_{affected}$, based on sensitive attributes (e.g., ethnicity). We ensure fair training for the models, as our objectives aim to avoid fairwashing through the explanations we provide. A potential concern in any fairness-aware explanation framework is fairwashing: the risk that explanations are designed to give a false impression of fairness while the underlying model remains biased. To guard against this, we assume the classifier itself is fair and focus on ensuring that the explanations provided to users are fair as well, thereby reinforcing rather than masking model-level fairness. The objective is to generate a set of CF $\mathscr{A}$ that enhances fairness at both individual and group levels. Specifically $\mathscr{A}$ should:

1. Maximize the proportion of individuals that can find effective recourse by applying actions from $\mathscr{A}$.

2. Guarantee *Fair Opportunity*, so that all individuals and groups should have a fair chance to improve their outcomes through the provided CFs.

3. Identify valid, actionable, plausible CFs and similar to the original instance, and that offer practical steps that individuals across groups can realistically implement.

4. Ensure that every group has access to a diverse set of CFs, preventing any unfairness in terms of diversity, and is not limited to only one hard-to-implement action.

Equal Effectiveness (EE) requires that similar proportions of individuals from the two protected groups G0 and G1 achieve recourse through the provided set of actions. This ensures that no group is disproportionately disadvantaged in terms of the effectiveness of the CF. Given the individual- and group-effectiveness metrics $\text{aeff}_\mu(\mathscr{A}, G)$ and $\text{aeff}_M(\mathscr{A}, G)$ for groups G_0 and G_1, EE is achieved when:

$$|\text{aeff}_\mu(\mathscr{A}, G_0) - \text{aeff}_\mu(\mathscr{A}, G_1)| \leq \epsilon \tag{8.5}$$

$$\text{and} \quad |\text{aeff}_M(\mathscr{A}, G_0) - \text{aeff}_M(\mathscr{A}, G_1)| \leq \epsilon \tag{8.6}$$

where ϵ is a small, predefined threshold quantifying disparities.

Equal Choice for Recourse (ECR) ensures that individuals from different protected groups have equal opportunities (or number of actions) to achieve recourse from $\mathscr{A}$.

This metric addresses disparities in the diversity and feasibility of actionable explanations provided to different groups. Let $\mathscr{A}_G$ represent the subset of actions from $\mathscr{A}$ that are effective for group G, an action $a \in \mathscr{A}$ is considered effective if it achieves recourse for at least a proportion of individuals (ϕ) from group G. ECR is achieved when the two groups $G0$ and $G1$ have the same number of valid actions:

$$|\mathscr{A}_{G_0}| = |\mathscr{A}_{G_1}| \tag{8.7}$$

8.4.4 RL-based Approach For Generating Fair Counterfactuals

We design an RL agent with an SAC (detailed chapter 2) algorithm to serve as a CF generator. The choice of RL, specifically SAC, is due to its sequential and exploratory nature in feature space, which makes it suitable for the sequential nature of achieving a recourse. Specifically, the entropy regularization step ensures stochastic policies, making it well-suited for high-dimensional action spaces. Our objective is not to identify a single static CF or recourse path, but rather to generate diverse, realistic, and feasible CFs that provide recourse to data points to reach a desired target. An RL sequential environment supports our aim by enabling the model to condition on previous actions, such that each modification of an actionable feature (for example, education, lifestyle, or employment) alters the feasible space of subsequent actions. In contrast, static optimizers treat all changes as independent, which can lead to unrealistic one-shot interventions. Furthermore, RL frameworks balance short-term and long-term trade-offs, as sequential decision-making naturally incorporates cumulative costs or constraints, such as effort, time, or limited resources, which are challenging to encode in static solvers. By learning a policy rather than solving a fixed optimization problem for each instance, the SAC sequential agent can explore multiple feasible paths and generate diverse recourse sets, rather than a single deterministic solution.

This makes the RL promising as a robust value estimation and scalable optimization to support the process of CF generation. As shown in Sec. 8.3, RL is a suitable method for generating CF and offers a flexible and robust framework that addresses several limitations of traditional optimization approaches. Unlike gradient-based, manifold-based, or convex optimization methods, which often treat CF generation as a static optimization problem, RL frames it as a sequential decision-making process. RL also enables effective exploration of the CF space, which is especially important in non-differentiable or discrete domains. The SAC takes as input the classifier h and the affected dataset $D\_affected$ and provides as output a set of CFs that represent $\mathscr{A}$. We follow the Markov Decision Process (MDP) problem formulation, which involves defining state and action spaces and customized reward functions. Since SAC is a model-free RL method, it does not require defining a transition function. In the following subsection,

we detail the state and action representation along with the reward functions. Specifically, we design 3 reward functions that incorporate the EE and ECR objectives R_{EE} (at both, individual and group levels) and R_{ECR} (group) and R_{EE-ECR} (hybrid), respectively, to ensure that the CFs adhere to the fairness metrics in Eq. 8.5, 8.6 and 8.7, respectively.

State Representation

Suppose n is the number of desired actions a (CF) to be generated to form $\mathscr{A}$ ($n = |\mathscr{A}|$), $\mathscr{F}_{\text{act}}$ the set of actionable features (mutable features), and $l = |F_{\text{act}}|$ (ensuring diversity and actionability). The state s is represented as an N-dimensional vector where $N = n \times l$ that corresponds to a concatenation of features representing the amount of change to the value of each actionable feature. Fig. 8.2 shows an example of the state vector where $\mathscr{F}_{\text{act}} = \{Income, Credit\}$ and the number of actions to generate is equal to 3, then the state will have a dimension of 6.

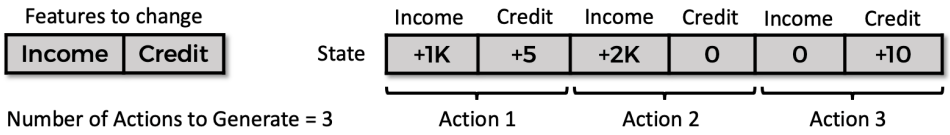


Figure 8.2. State representation, given the set of features to change and the maximum number of actions to generate.

Action Representation

An action is represented as a 2-dimensional vector $a = \{a_1, a_2\}$, where a_1 denotes the index of the feature to be modified and a_2 denotes the amount of change applied to that feature at the current time step. Specifically, a_1 selects one of the mutable features, while a_2 determines the magnitude and direction of the modification for the selected feature. By applying actions sequentially, one feature modification per step, the RL agent incrementally constructs a CF instance, enabling the policy to learn both the order and the extent of modifications required to achieve a valid and fair set of CFs.Thus, feature modifications are applied sequentially, allowing the agent to construct a multi-step transformation path rather than making all changes at once. This sequential structure is essential, as it enables the agent to evaluate intermediate states, adjust its strategy dynamically, and ultimately learn a sequence of modifications that leads to a feasible and effective CF. To improve the stability and efficiency of learning, we apply normalization and denormalization to the action feature space so that all features are handled on a consistent scale, and features with larger numeric ranges do not

dominate the optimization. In addition, the proposed feature values are bounded by the valid range of each feature to preserve plausibility and consistency with the similarity constraints. Fig. 8.3 illustrates an example in which the RL agent selects the feature at index 2 (feature *income*) and applies a change of +1k, resulting in an update from 2k to 3k.

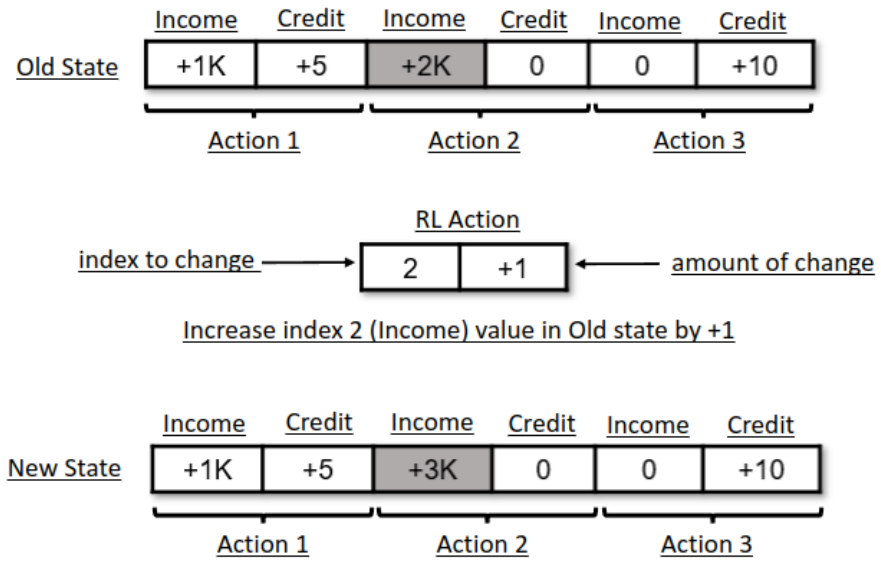


Figure 8.3. RL action representation, and the change in state after applying the action.

Reward Function:

The RL reward function incorporates the fairness metrics to be optimized, EE and ECR, and to the Gower distance, to ensure group fairness and similarity, respectively. We define R_{EE}, R_{ECR}, and R_{EE-ECR}. Specifically, R_{EE} is separately used when we aim to optimize only for EE, while R_{ECR} is separately used when we aim to optimize for ECR, while R_{EE-ECR} is used when we aim to optimize for EE and ECR simultaneously.

The reward for the Individual- and Group- EE case consists of the following terms (Eq. 8.5 and Eq. 8.6), presented in Eq. 8.8:

- *Average success rate (ASR)*: Average proportion of individuals who achieve recourse using any of the proposed actions for the two protected groups: $(SR_1 + SR_2)/2$, where SR_i represents the success rate (SR) of group i.

- *Proportions difference (PD)*: Absolute difference between the proportions of individuals across protected groups who can achieve recourse using any of the proposed actions: $|SR_1 - SR_2|$.

- *Similarity (Sim)*: The average Gower distance of points that were able to achieve recourse through an action a from $\mathscr{A}$: $Gower(x, h(a(x))$.

- *Active actions (Act):* number of active actions (an action that could be used to achieve recourse), it is used to ensure a diverse set of non-zero actions: $|a \in \mathscr{A}, \quad eff(a, G) \geq \alpha|$.

- $\mathbf{C_{0,1,2,3}}$: regularization terms allowing to balance trade-offs according to desired priorities.

$$\begin{aligned} R_{EE}(s_{t-1}, a, s_t) = ASR_{s_t} \times C_0 + Act_{s_t} \times C_1 \\ - PD_{s_t} \times C_2 - Sim_{s_t} \times C_3 \end{aligned} \tag{8.8}$$

To ensure ECR (Eq. 8.7), we customize the reward function to encourage diverse CF choices across protected groups by maximizing the number of actions available to each group. Additionally, we include terms that maintain fairness between the groups, ensuring equal potential for recourse, for example, by considering the number of active actions. Specifically, the reward function (Eq. 8.9) is:

$$\begin{aligned} R_{ECR}(s_{t-1}, a, s_t) = A_{0s_t} \times C_0 + A_{1s_t} \times C_1 - AD_{s_t} \times C_2 \\ - Act_{s_t} \times C_3 - Sim_{s_t} \times C_4 \end{aligned} \tag{8.9}$$

- A_0 *and* A_1*:* number of actions for the two protected subgroups G_0 and G_1, respectively (e.g., males and females that can be used to achieve recourse).

- *Action Difference (AD):* number of actions difference represents the absolute difference between the number of actions available for each protected subgroup to achieve recourse: $|A_0 - A_1|$.

- *Similarity (Sim)*: The average Gower distance of points that were able to achieve recourse through an action a from $\mathscr{A}$: $Gower(x, h(a(x))$.

- *Active actions (Act):* number of active actions (an action that could be used to achieve recourse), it is used to ensure a diverse set of non-zero actions: $|a \in \mathscr{A}, \quad eff(a, G) \geq \alpha|$.

- $\mathbf{C_{0,1,2,3,4}}$: regularization terms allowing to balance trade-offs according to desired priorities [1].

For the case of the hybrid approach, which optimizes for the hybrid fairness, we specify the reward function R_{EE-ECR} that represents a combination of R_{EE} and R_{ECR}:

$$R_{EE\text{-}ECR}(s_{t-1}, a, s_t) = R_{EE}(s_{t-1}, a, s_t) + R_{ECR}(s_{t-1}, a, s_t) \tag{8.10}$$

Stopping criteria

The episode-stopping condition that specifies the reach of the RL goal is met when the following criteria are satisfied:

- For EE: $ASR >= target\_success\_rate$ and $PD <= target\_proportion\_difference$.
- For ECR: $A_0 >= target\_group0\_nb\_actions$ and $A_1 >= target\_group1\_nb\_actions$ and $AD <= target\_nb\_actions\_difference$.

8.4.5 Counterfactual Evaluation Metrics

We design our approach to satisfy several key criteria shown in the rewards function. Actionability is ensured by allowing users to specify the features they are willing to change. We achieve validity through direct optimization, while plausibility is enforced by constraining the search space of our optimizer. Finally, we optimize for similarity using Gower distance. A CF a is evaluated based on several key metrics:

- *Validity.* A CF is considered *valid* if it successfully changes the classifier's prediction compared to the original instance. Formally, for a given instance $x \in D$ with $h(x) = 0$, a CF $a(x)$ is valid if $h(a(x)) = 1$. For a dataset D, the validity is the proportion of individuals for whom the generated CF achieves the desired outcome:

$$\text{Validity} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left[h(a(x_i)) = y_{\text{target}}\right] \tag{8.11}$$

- *Plausibility.* A CF is *plausible* if its feature values lie within the data distribution, i.e., they do not correspond to outliers. This is needed to ensure that the CF

[1] The regularization parameters can be adjusted to accommodate user constraints and preferences. In this work, equal weight is assigned.

represents a realistic instance. One of the robust methods to compute plausibility is based on the reconstruction error from a denoising autoencoder (following [76, 459]):

$$\text{Plausibility} = \|a(x_i) - a(\hat{x}_i)\|^2 \tag{8.12}$$

where $a(\hat{x}_i)$ is the autoencoder reconstruction of the CF $a(x_i)$. Lower the reconstruction error, meaning the CFs fit the training data distribution and are more realistic.

- *Similarity and Minimality.* A CF should be similar to the original instance. *Similarity (proximity)* refers to how close the CF is to the original instance in feature space, often measured using a distance function $d(x, a(x))$ such as Gower distance. *Minimality* (or sparsity) focuses on minimizing the number of features changed:

$$\text{Similarity} = \frac{1}{N}\sum_{i=1}^{N} \text{Gower}(x_i, a(x_i)) \tag{8.13}$$

$$\text{Minimality} = \frac{1}{N}\sum_{i=1}^{N} \|x_i - a(x_i)\|_0 \tag{8.14}$$

- *Actionability.* A CF is *actionable* if it only modifies features that the individual can feasibly change. Let $\mathscr{F}_{\text{act}} \subseteq \mathscr{F}$ denote the set of actionable features. Then, a CF $a(x)$ is actionable if and only if $\text{supp}(x - a(x)) \subseteq \mathscr{F}_{\text{act}}$, where $\text{supp}(\cdot)$ denotes the set of features with changes.

8.5 Evaluation Setup

This section outlines our experimental setup, including the datasets utilized for analysis and the RL configuration employed to evaluate the proposed methodology. Our framework's implementation is available online.

Experimental Scenarios and Baselines We consider four fairness scenarios, as follows:

Individual - Equal Effectiveness (Individual-EE) In this scenario, each individual, regardless of their protected attribute value, chooses the recourse action from the list of generated actions that best addresses their unique situation. Each individual chooses the action with the lowest Gower distance (i.e., the lowest recourse cost) relative to the original instance. We run the RL considering Eq. 8.3 and 8.5 and optimize for R_{EE} (Eq. 8.8).

Group - Equal Effectiveness (Group-EE) In this scenario, the aim is to identify a single widely applicable action that achieves recourse for a least effectiveness of $eff(a, G) = 75\%$) of individuals across both protected groups. This approach prioritizes broad effectiveness while aiming to provide recourse to all individuals within a group, where we run the RL on Eq. 8.4 and 8.6 and optimize for R_{EE}. (Eq. 8.8).

Group Equal Choice for Recourse (Group-ECR) This scenario represents a group ECR. It identifies a set of actions that offer ECR to both protected subgroups. An action is considered effective if it results in recourse for more than $\phi = 60\%$ [2]of individuals in that affected dataset, by ensuring Eq. 8.7 and optimizing for R_{ECR} (Eq. 8.9).

Hybrid Individual-Equal Effectiveness and Group-Equal Choice for Recourse (Hybrid-EE-ECR) This scenario represents the hybrid fairness, which balances between individual and group fairness. It explores feature space that allows for individual fairness while ensuring that a significant proportion of individuals from each protected group can achieve recourse through a set of collectively effective actions, based on Eq. 8.3, 8.5 and 8.7 simultaneously and optimize for R_{EE-ECR}. (Eq. 8.10).

Clustering In our experiments, we cluster the affected dataset into 3[3] clusters using the k-means algorithm [460, 461], creating groups of individuals with similar feature values close in space regardless of the protected attributes. This clustering step allows us to explore fair CF both globally across the entire dataset and within each chosen cluster, allows for the formation of clusters by grouping instances using Euclidean distance and minimizing within-cluster variance. We evaluate fairness metrics on the whole affected dataset (we refer to this by *Whole* in the rest of the chapter) as well as on each cluster separately (C1-C3). For each individual, we select the best valid CF with the lowest Gower distance and treat it as the final CF. Additionally, we report a clustering-based approach, in which each individual is assigned the CFs generated for the cluster to which they belong.

In our experiments, we aim to generate a maximum of 5 actions for EE, ECR and EE-ECR. We ensure that most individuals can successfully achieve recourse and that the PD across protected groups does not exceed 10% (the upper bound is 100%), to maintain fairness in the distribution of successful recourse outcomes. For ECR, we

[2]This represents the minimum boundary, however, the SR is optimized by the reward function, which aims to maximize the ϕ.

[3]We fix the number of clusters to three for simplicity and consistency across datasets, also to reflects a balance between granularity and interpretability: too few clusters would fail to capture heterogeneity, while too many would lead to overly fragmented groups, making analysis and comparison less meaningful.

impose constraints to ensure that the *number of actions* for each protected group is at least 1 to guarantee sufficient recourse options, and that the *number of actions difference* between the two protected groups is 0, to ensure protected groups receive ECR.

Baselines To evaluate the effectiveness of our RL-based CF generation methods, we compare them against 3 widely-adopted CF methods such: Nearest Instance Counterfactual Explanations (NiCE) [63], Diverse Counterfactual Explanations (DiCE) [80], and prototype-guided CFs [459]. Each method is configured with hyperparameters optimized to maximize the quality of CFs, balancing proximity, sparsity, and plausibility.

NiCE generates CFs by iteratively modifying the original instance by replacing feature values with those from the nearest unlike neighbor. We configure NiCE with the distance metric as Heterogeneous Euclidean-Overlap Metric to handle mixed feature types with min-max scaling for numerical features. The optimization objective is set to proximity and plausibility constraints.

DiCE Generates diverse CFs by optimizing validity, proximity, and diversity. We set the Generation method to random sampling to generate 5 CFs per instance, ensuring diversity. The feature constraints are set to only mutable features, which are chosen as the ones used in our RL method.

Prototype-guided Generate CFs by moving instances toward class prototypes. The search method is the KD-tree acceleration with Loss weighting. The prototype weight is set to 10 to ensure plausibility. The search process is optimized for 200 iterations and 15 perturbation steps. Along with Feature bounds derived from training data to avoid unrealistic values.

Datasets Description We evaluate our approach using 4 widely-adopted datasets in fairness evaluation: Adult, SSL, Alzheimer, and Law, each selected for their relevance and the presence of actionable features that enable meaningful interventions. These datasets collectively span diverse domains: economic, judicial, and medical, each with distinct fairness challenges and actionable recourses. Their inclusion ensures that our approach is robust, generalizable, and grounded in real-world impact.

Adult [462]: Its use is motivated by the societal importance of fair decision-making in employment and financial contexts. Consists of 48842 data points. The classifier predicts whether an individual's income exceeds 50k/year. The dataset consists of 15 features representing demographic and socioeconomic factors. The protected attribute is Race, and the actionable features are continuous: capital-gain, hours-per-week, and categorical: educational-num. Actionable features such as capital-gain, hours-per-week, and educational-num represent realistic features that might be adjusted to improve outcomes.

SSL Consists of 398,582 data points related to criminal justice. The classifier pre-

dicts the SSL score based on 10 features, including arrests, assaults, shootings, affiliation, and criminal activity. The protected attribute is Gender, and the actionable features are age at latest arrest (categorical) and trend in criminal activity (continuous). The actionable features can be influenced through policy or rehabilitation programs in justice-related applications.

Alzheimer's [4]*:* Consists of 2149 data points, designed to predict whether an individual will develop Alzheimer. It consists of 21 demographic, lifestyle, and health-related features. The protected attribute is Gender, and the actionable features are continuous: functional assessment, ADL, and MMSE. The actionable features represent clinical and lifestyle indicators that can be monitored or improved through intervention.

Law's [463]: Consists of 20,798 data points describing admission records for students, each described by 12 attributes. The task is to determine whether a student will pass the exam. The dataset has been widely used in fairness research due to well-documented disparities across demographic groups. The protected attribute is gender, and the actionable features are the school GPA (zgpa) and the LSAT score (lsat).

Models training We evaluate the performance of the classifier, focusing on standard metrics (accuracy, precision, recall, and F1-score) and fairness measures (Equalized Odds (EQ) and Demographic Parity differences (DP)) to assess model fairness across protected groups. We also employ in-processing fairness training using the Exponentiated Gradient to enhance the fairness of the underlying ML model prior to running our experiments. Each dataset is split into 80% for training and 20% for testing. We train different ML models and choose the one that performs best in terms of fairness metrics and predictive performance. More specifically, for *Alzheimer*, we train a Random Forest classifier, with max depth of 15, max features of 'sqrt', min samples leaf of 2, min samples split of 5, and 200 estimators. For *Adult*, we train a Catboost classifier with default configuration. For *SSL*, we train an XGBoost classifier with a subsample of 0.9, number estimators of 200, a max depth of 3, and a learning rate of 0.1. For *Law* we train a Gradient Boosting Classifier with default parameters.

The ML models trained on the various datasets achieved high predictive performance and acceptable fairness, as indicated by EQ and PD values below 10%. Given that these metrics have an upper bound of 100%, a level under 10% is considered satisfactory, as it supports that our generated CFs do not contribute to fairwashing. For *Alzheimer*, the model achieves an accuracy of 95.8%, a precision, recall, and F1-score of 96%. Fairness evaluation shows an EO difference of 2% and a DP of 8% between the two protected groups. For the *Adult* dataset, the model achieves an accuracy of 87.3% and precision, recall, and F1-score of 87%. Fairness evaluation revealed EO and

[4]https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset

DP differences of 4% and 6%, respectively. For the *SSL*, the model achieves 90% accuracy, 95% precision, 90% recall, and an F1-score of 92%. Fairness evaluation shows an EO difference of 2%, and a DP of 0.9%. For *Law* , the model achieves an accuracy of 90%, with precision, recall, and F1-score of 88%, 90%, and 88%, respectively. Fairness evaluation reveals EO of 5% and DP of 2%.

8.6 Experimental Results

In this section, we discuss experimental results in terms of the fairness of CFs generated across the various scenarios and the quality of the CFs. All results averaged over 5 runs. For each run, we randomly sampled a seed ranging from 1 to 64 and used it consistently throughout the experimental pipeline. All components, including the RL model initialization and the environment, were initialized using this same seed. [5]

Table 8.2. Validity in term of SR across Whole and clusters C1–C3. Values are shown as [G1 (std), G2 (std)] with the absolute (PD) in parentheses.

Dataset	Scenario	Whole	C1	C2	C3
Alzheimer	Individual-EE	[80 (± 7) , 81 (± 3)] (**1**)	[80 (± 5) , 81 (± 5)] (**1**)	[84 (± 5) , 91 (± 3)] (**7**)	[91 (± 8) , 86 (± 9)] (**5**)
	Group-EE	[83 (± 6) , 82 (± 3)] (**1**)	[87 (± 6) , 92 (± 6)] (**5**)	[86 (± 10) , 93 (± 7)] (**6**)	[91 (± 4) , 85 (± 6)] (**6**)
	Group-ECR	[74 (± 12) , 71 (± 12)] (**3**)	[66 (± 8) , 66 (± 5)] (**0**)	[72 (± 11) , 81 (± 12)] (**9**)	[85 (± 2) , 78 (± 6)] (**7**)
	Hybrid-EE-ECR	[80 (± 10) , 83 (± 9)] (**3**)	[84 (± 5) , 82 (± 10)] (**2**)	[93 (± 7) , 96 (± 5)] (**3**)	[87 (± 5) , 83 (± 6)] (**4**)
Adult	Individual-EE	[97 (± 2) , 97 (± 2)] (**0**)	[96 (± 2), 95 (± 3)] (**1**)	[98 (± 0) , 98 (± 1)] (**0**)	[97 (± 2) , 98 (± 2)] (**1**)
	Group-EE	[97 (± 2) , 97 (± 2)] (**0**)	[96 (± 0) , 96 (± 2)] (**0**)	[97 (± 1) , 97 (± 2)] (**0**)	[97 (± 3) , 99 (± 2)] (**2**)
	Group-ECR	[96 (± 1) , 97 (± 2)] (**1**)	[96 (± 0) , 96 (± 2)] (**0**)	[97 (± 4) , 96 (± 6)] (**1**)	[96 (± 1) , 98 (± 2)] (**2**)
	Hybrid-EE-ECR	[96 (± 2) , 98 (± 5)] (**2**)	[96 (± 1) , 95 (± 2)] (**1**)	[97 (± 1) , 98 (± 1)] (**1**)	[95 (± 7) , 98 (± 5)] (**3**)
SSL	Individual-EE	[99 (± 1) , 96 (± 3)] (**3**)	[96 (± 2) , 92 (± 7)] (**4**)	[93 (± 5) , 91 (± 6)] (**2**)	[100 (± 0) , 94 (± 4)] (**6**)
	Group-EE	[98 (± 1) , 97 (± 2)] (**1**)	[94 (± 1) , 90 (± 4)] (**4**)	[91 (± 4) , 94 (± 5)] (**3**)	[100 (± 0) , 92 (± 3)] (**8**)
	Group-ECR	[99 (± 1) , 97 (± 4)] (**2**)	[88 (± 7) , 79 (± 10)] (**3**)	[89 (± 4) , 90 (± 4)] (**1**)	[97 (± 4) , 88 (± 10)] (**9**)
	Hybrid-EE-ECR	[99 (± 1) , 96 (± 3)] (**3**)	[93 (± 3) , 90 (± 5)] (**7**)	[91 (± 5) , 93 (± 6)] (**2**)	[100 (± 0), 96 (± 3)] (**4**)
Law	Individual-EE	[76 (± 2) , 79 (± 2)] (**3**)	[75 (± 1) , 79 (± 3)] (**4**)	[80 (± 3) , 80 (± 3)] (**0**)	[75 (± 2) , 80 (± 3)] (**5**)
	Group-EE	[78 (± 3) , 82 (± 3)] (**4**)	[80 (± 5) , 83 (± 3)] (**3**)	[86 (± 3) , 84 (± 4)] (**2**)	[81 (± 5) , 81 (± 5)] (**0**)
	Group-ECR	[74 (± 6) , 76 (± 4)] (**2**)	[72 (± 8) , 74 (± 10)] (**2**)	[83 (± 12) , 75 (± 12)] (**8**)	[66 (± 4) , 71 (± 3)] (**6**)
	Hybrid-EE-ECR	[81 (± 10) , 82 (± 9)] (**1**)	[75 (± 3) , 79 (± 1)] (**4**)	[77 (± 3) , 78 (± 3)] (**1**)	[81 (± 10), 81 (± 7)] (**0**)

8.6.1 Success Rates and Proportion Difference

Tab. 8.2 reports the validity in term of SR achieved by our approach across *Whole*, the clusters C1, C2, and C3, and across the protected G1 and G2, considering the various fairness scenarios on 4 datasets. For each cell, we report the average SR for both groups

[5] Experiments were conducted on a workstation equipped with an Intel Core i7 CPU, an NVIDIA RTX 3070 GPU, and 32 GB of RAM.

in the format *[G1 (±), G2 (±)]*, followed by the *PD* highlighted in bold. The ± is the standard deviation of SR over the 5 runs.

Success Rate Across Fairness Scenarios Results show that the validity achieved by our approach across diverse scenarios and datasets are generally high, indicating its ability to generate valid CFs (i.e., CFs that enable individuals to perform recourse) for *Whole* and across a variety of clusters (subgroups). Starting with *Adult*, our approach consistently achieves high SR, ranging between 95 and 98 across groups and clusters. Similarly, high SR is observed in *SSL*, with *Whole* values as high as [99 (± 1), 96 (± 3)], [98 (± 1), 97 (± 2)], and [99 (± 1), 97 (± 4)] across the different fairness scenarios. The SR observed in *Alzheimer* is likewise high but relatively lower than in other datasets. For instance, for the *Hybrid-EE-ECR* scenario, our approach attains an SR of [80 (± 10), 83 (± 9)] on *Whole*, [84 (± 5), 82 (± 10)] in C1, [93 (± 7), 96 (± 5)] in C2, and [87 (± 5), 83 (± 6)] in C3. Particularly, and only for the *Group-ECR* scenario, a relatively low SR is observed in *Alzheimer*, with [74 (± 12), 71 (± 12)] on *Whole* and [66 (± 8), 66 (± 5)] in C1, which indicates difficulties in generating CFs. A similar intermediate pattern is observed for *Law*, where SR on *Whole* ranges from [74 (± 6), 76 (± 4)] to [81 (± 10), 82 (± 9)] across the considered scenarios. Despite limitations in the *Group-ECR* scenario, results suggest that our approach is generally effective in generating valid CFs that guarantee recourse.

Success Rate Across Clusters We now turn to comparing the SR achieved across *Whole* and the diverse clusters, in each of the scenarios. A noteworthy, though expected, observation is that the SR values across clusters are consistently higher than those obtained across the *Whole*. For instance, in *Alzheimer*, under *Group-EE*, SR is [83 (± 6), 82 (± 3)] across *Whole*, and [87 (± 6), 92 (± 6)], [86 (± 10), 93 (± 7)], and [91 (± 4), 85 (± 6)] in C1, C2, and C3, respectively. Similarly, in the *Hybrid-EE-ECR* scenario, SR is [80 (± 10), 83 (± 9)] for *Whole*, while it is [84 (± 5), 82 (± 10)], [93 (± 7), 96 (± 5)], and [87 (± 5), 83 (± 6)] for C1, C2, and C3, respectively. In *Adult*, the results show that SR on *Whole* is generally very close to that of the clusters. For example, in *Group-EE*, SR achieved for *Whole* is [97 (± 2), 97 (± 2)], while for the three clusters it is [96 (± 0), 96 (± 2)], [97 (± 1), 97 (± 2)], and [97 (± 3), 99 (± 2)]. A clearer understanding of these results will emerge in the subsequent analysis of CF similarity and the overall quality of the generated CFs.

Takeaway 1: Our approach demonstrates that recourse can be reliably achieved across diverse datasets while sustaining high SR across different demographic groups. This indicates that integrating fairness objectives does not compromise the validity of the generated CFs. Furthermore, our proposed CF generation strategy benefits from operating on more

homogeneous subgroups represented by clusters, where the underlying structure and decision boundaries are easier to capture, as opposed to working on the entire dataset, where heterogeneity increases the complexity of generating valid CFs.

Proportion Difference We now turn to the PD values obtained across groups G1 and G2, reported in Tab. 8.2 (shown in parentheses) for all scenarios. Recall that PD serves as a fairness metric by quantifying the disparity in SR between the two groups and theoretically ranges from 0 (no disparity) to 100 (complete disparity). Overall, the results show that PD is relatively low, ranging between 0 and, in worst case, 12, which indicates the effectiveness of our approach in ensuring fairness across the two protected groups. More specifically, in *SSL*, our approach keeps PD values within a narrow range of 0–8 across all scenarios and clusters except for *Group-ECR*, demonstrating that the CF generation process introduces minimal group disparity. Similarly, in the *Adult*, PD remains between 0–3, while in *Alzheimer* it ranges from 0–7, and *Law* between 0 and 8. Results show that the reward function in our proposed approach consistently captures PD, keeping it relatively low across datasets, fairness scenarios, and clusters. It is also noteworthy to highlight that our proposed approach shows comparable performance across all scenarios.

Takeaway 2: Our approach not only achieves high validity in terms of SR but also upholds fairness by minimizing disparities between groups. This is reflected in consistently low PD values, which indicate that validity differences across groups are effectively reduced. Consequently, our method guarantees actionable recourse for both individuals and groups, without introducing systematic bias.

Impact of Fairness Constraints on SR and PD We now turn to comparing the diverse fairness scenarios in each of the datasets in terms of SR and PD. The key question we seek to address is: *To what extent does the introduction of additional fairness constraints (i.e., group and hybrid) influence SR and PD?* In *Alzheimer*, our approach attains an SR ranging from 80 to 96, and a PD ranging between 0 and 7, across all scenarios, except for *Group-ECR*. Specifically, in *Individual-EE* and *Group-EE*, SR is substantially high (80–91) and PD is consistently low (1–7). In *Hybrid-EE-ECR* we see a more favorable balance with a relatively high SR (ranging between 80 and 96) and a PD between 2 and 4. In contrast, *Group-ECR* shows the lowest SR (61.6 in *Whole* and 62–79 across clusters) and the highest PD (11.1 for C2 and 9.4 for C3). A similar pattern is observed in *Adult*, *Law* and *SSL*, where *Group-ECR* consistently emerges as the least favorable setting, yielding lower SR and higher PD compared to the other fairness scenarios, yet still within a comparable range.

Takeaway 3: Group-only fairness constraints (Group-ECR) are associated with reduc-

tions in SR and increases in PD across datasets. By contrast, the Hybrid-EE-ECR approach combines individual and group fairness objectives, yielding comparatively high SR while keeping PD at lower levels. These results suggest that multi-level fairness integration offers a more balanced trade-off between performance and equity than group-only constraints.

8.6.2 Fairness in terms of CF Similarity

So far, we have discussed the SR and PD metrics. The results indicate that the SR values obtained across the different scenarios are highly comparable, while the PD remains relatively low between the two groups in each case. We now turn our attention to analyzing the similarity (measured using Gower distance) between the generated CFs and their corresponding original instances, measured using the Gower distance. This analysis provides a deeper understanding of fairness, as comparable SR values across groups may still conceal disparities, as one group's CFs might be less similar to the original instances, thereby implying more challenging or less feasible recourse options. This analysis allows us to address the question: *to what extent do CFs preserve fairness not only in terms of their SR but also in the quality and feasibility of the recourse they provide across different groups?*

Fig. 8.4 shows the average Gower distance between the CFs generated using our approach and the original instances in the four dataset, across the diverse scenarios, clusters, and across protected groups. We first compare the similarity across the diverse scenarios and then we focus on the group granularity.

CF Similarity Across Scenarios Overall, the values of Gower distance range approximately between 0 and 0.2 (with 0 being similar and 1 dissimilar). This indicates that the generated CFs are similar to their respective original instances in the feature space. For instance, in the *Adult*, in *Individual-EE*, Gower distances as low as 0.025 are observed. Similarly, in *SSL*, Gower distances are mostly below 0.25. Additionally, we note that the variance across scenarios is generally low. For example, in the *Alzheimer* and *Group-ECR* scenario, *the observed Gower distance is consistently around 0.177 with very low standard deviation (e.g.,* ± 0.05 *and* ± 0.033 *for G1 and G2, respectively)*. Importantly, we observe that *Hybrid-EE-ECR* maintains a low Gower distance and generates CFs similar to original instances as much as those generated in *Individual-EE* and *Group-ECR* when considered separately.

Impact of Clustering on CF Similarity Comparing CFs Gower distance values between *Whole* and the three clusters, we observe that the CFs in *Whole* generally exhibit equal or slightly higher distances than those in the clusters. This trend is shared across all scenarios and datasets. For instance, in *Adult*, for *Group-EE*, the average Gower distance

for *Whole* is 0.17 ± 0.05, while that of the diverse clusters ranges from 0.10 ± 0.04 to 0.16 ± 0.01. Also in *SSL*, for *Hybrid-EE-ECR*, the average Gower distance observed for *Whole* is around 0.25 ± 0.01 higher than those of C1, C2, and C3, which range from 0.12 to 0.17 ± 0.04. These findings suggest that clustering generally yields CFs that are more tightly grouped in the feature space, leading to greater internal consistency compared to those derived from the *Whole*.

Group Fairness in terms of CF Similarity We now examine the extent to which our proposed approach promotes fairness in the feasibility of recourse by comparing the Gower distance achieved for G1 and G2. Results show that across all scenarios, the difference between the Gower distances of G1 and G2 is relatively small and, in most cases, even negligible. This pattern is consistent also across *Whole* and the diverse clusters. Specifically, in *Alzheimer*, the differences between CF Gower of groups are negligible, around 0.03, with the largest difference appearing in cluster C3 in *Individual-EE*, where it is around 0.05, which is still considered relatively low. A similar trend is observed for *Adult* and *SSL*, for instance, *Group-ECR* scenario of *Adult*, G1 and G2 exhibit nearly identical distances of 0.033 and *SSL* for *Hybrid-EE-ECR* scenario, the Gower distances for G1 and G2 in cluster C3 are 0.127 and 0.139, respectively, showing a minimal difference.

These results lead us to the following two key takeaways: *Takeaway 4: The generated CFs at the cluster level produce more precise and similar outcomes compared to generating them for the entire dataset. This improvement stems from the ability to customize CFs based on subgroup-specific patterns, resulting in more targeted solutions with reduced distance from original instances and fairness across groups, as reflected by a comparable CF similarity across groups.*

Takeaway 5: The Hybrid-EE-ECR *approach maintains CF similarity while balancing individual- and group-level objectives, reinforcing its role as a principled formulation of hybrid fairness. The generated CFs remain close to the original data points, as reflected by consistently low Gower distances, making them both effective and actionable solutions. This closeness is preserved across groups, ensuring fairness due to minimal differences in CF similarity.*

8.6.3 Number of Actions

We now focus our analysis on the number of generated actions for *Group-ECR* and *Hybrid-EE-ECR*, as these scenarios are directly optimizing for the diversity of CFs. This analysis allows to determine if the ECR constraint is consistently met by ensuring equal number of choices between the protected groups and, consequently, if our approach is capable of ensuring diversity in the CFs by providing many recourse options. Fig. 8.5

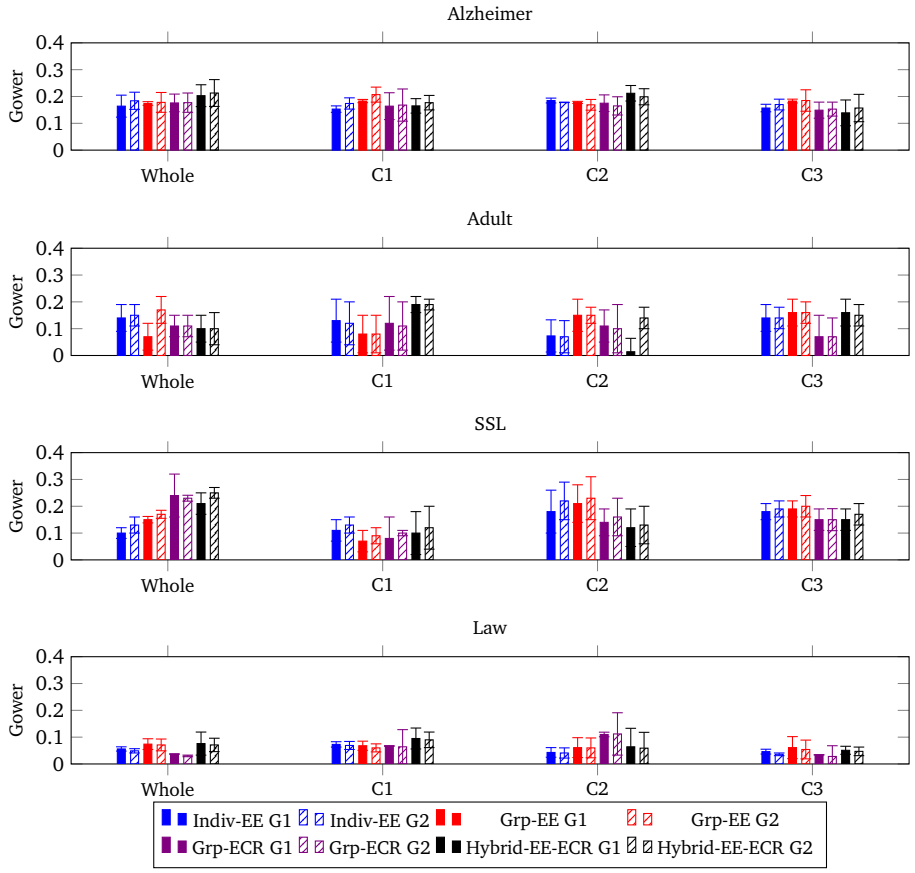


Figure 8.4. Similarity in terms of Gower distance between generated CFs and original instances across the four datasets. Each subplot corresponds to a dataset and shows the similarity for each CF explainer trained on the whole dataset or clusters (C1–C3), separately for protected groups G1 and G2.

shows the number of actions achieved for *Whole* and the diverse clusters in the *Group-ECR* and *Hybrid-EE-ECR*, across the four datasets. Note that we show only one bar per case (scenario-cluster), as the number of actions for G1 and G2 resulted the same in each of the cases. First, we highlight that the fact that the number of actions resulted the same for both groups across all cases indicates that our approach was capable of identifying an equal choice of recourse for individuals to choose from, which proves the effectiveness of our approach in providing and ensuring fair and equitable recourse choices, regardless of which group the individual belongs to.

Impact of Clustering on the Number of Actions We now move to discussing how the clustering impacts the number of actions (available recourse options). This is rel-

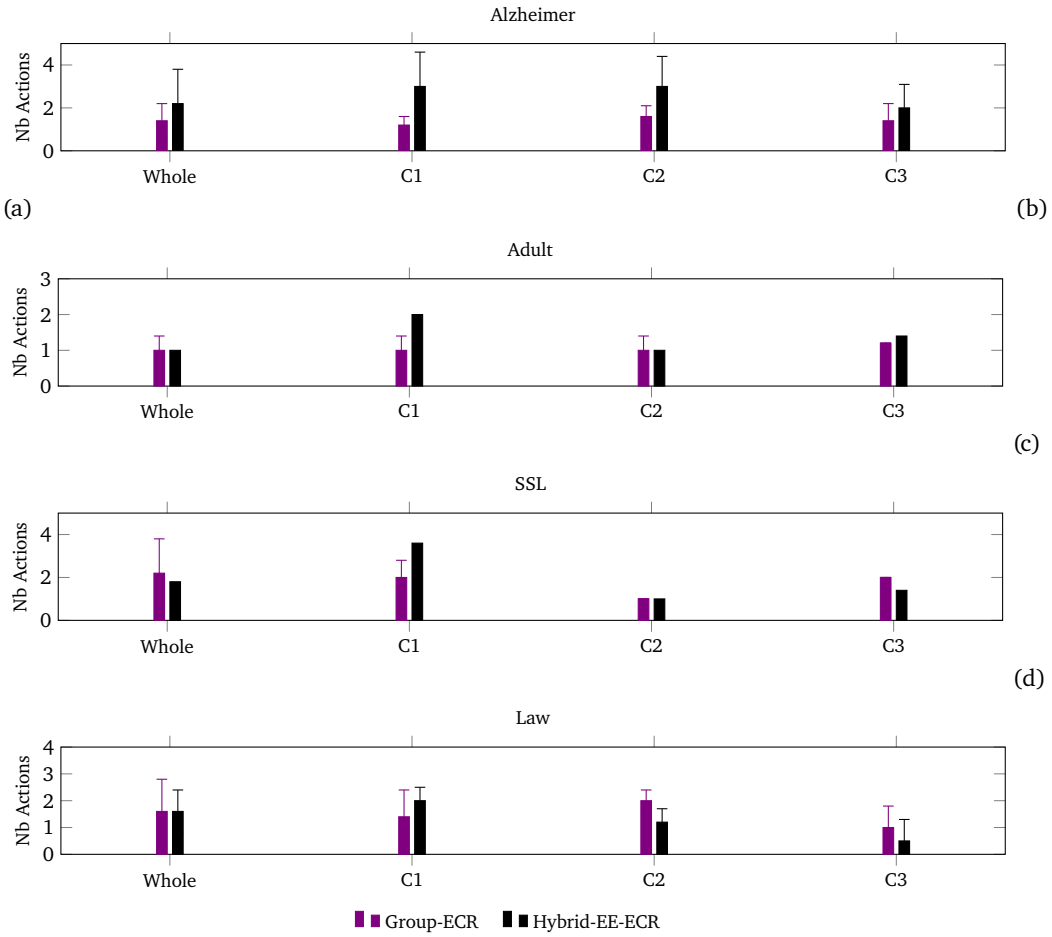


Figure 8.5. Minimality in terms of the number of actions required to reach a recourse for protected groups across the four datasets. G1 and G2 yield identical values; therefore, only one bar per method is displayed. Results are shown for Group-ECR and Hybrid-EE-ECR across the whole dataset and clusters C1–C3.

evant as it reveals how clustering, as expected, can increase the total availability of resources across the datasets. A clear and consistent trend emerges: the clustered approach significantly increases the total number of unique actions compared to when considering the dataset as a *Whole*, across all datasets. This is because the clustering identifies distinct local recourse options, allowing for the generation of tailored CF options for different subgroups that the *Whole* scenario would often miss. For example, in *Alzheimer*, our approach identified on average 2 actions for *Whole* while for C1, C2 and C3, it identified 3, 3, and 2, respectively, resulting in a total of 8 recourse options.

Takeaway 6: Clustering enables the exploitation of local data patterns to identify flexible and varied recourse options. The approach also maintains fairness by guaranteeing that protected groups receive equivalent numbers of actionable CFs, balancing recourse diversity with fair access.

Comparing Group-ECR to Hybrid-EE-ECR We now analyze the impact of introducing fairness constraints through *Hybrid-EE-ECR* on the number of recourse options, compared to that of *Group-ECR*. This will allow us to answer: *Does optimizing for* Hybrid-EE-ECR *impact the number of available recourse options, with respect to only optimizing for* Group-ECR*?* The results reveal an interesting and non-uniform trend. Contrary to the expectation that the *Hybrid-EE-ECR* approach would generate fewer actions, our findings show a variable relationship across different clusters and datasets. This can be attributed to the nature of the optimization problem: The *Hybrid-EE-ECR* scenario imposes additional fairness constraints such as those of individual-level fairness, hence, it effectively narrows the search space for valid CFs. In some cases, *Hybrid-EE-ECR* leads to a simpler, more precise solution that satisfies both individual and group fairness criteria simultaneously. In contrast, the *Group-ECR*, with its broader objective, may find a more complex CF that works for an entire group but is not necessarily the minimal for any single individual. These findings are consistent with the results presented in Tab. 8.2 and Fig. 8.4, where the *Group-ECR* exhibited a lower SR and a higher Gower distance compared to the *Hybrid-EE-ECR*.

Takeaway 7: Hybrid-EE-ECR *introduces no additional complexity in generating CFs in term of the number of actions, when compared to* Group-ECR*. However,* Group-ECR *tends to yield more complex CFs that deviate further from the original instances and are associated with lower SR and higher PD.*

8.6.4 Comparative Evaluation Against Baseline Approaches

We now compare the quality of CFs generated by our diverse scenarios of our approach to that of baseline approaches, namely, NICE, DICE, and prototype-guided, in terms of validity, plausibility, similarity, and minimality, as reported in Tab. 8.3. This analysis

Table 8.3. CF quality metrics across datasets and scenarios. Values are shown for [G1, G2] with absolute difference in parentheses. For reference, training set plausibility is 0.822 (Alzheimer), 0.382 (Adult), 0.15 (SSL), and 0.292 (Law).

Dataset	Scenario	Validity	Plausibility	Similarity	Minimality
Alzheimer	**Individual-EE**	[86% - 86.4%] (0.4)	[0.849 - 0.798] (0.051)	[0.174 - 0.182] (0.008)	[2.897 - 2.914] (0.017)
	Group-EE	[90% - 91.5%] (1.5)	[0.852 - 0.797] (0.055)	[0.192 - 0.199] (0.007)	[2.804 - 2.825] (0.021)
	Group-ECR	[76.1% - 73.25%] (2.85)	[0.853 - 0.795] (0.058)	[0.186 - 0.187] (0.001)	[2.663 - 2.666] (0.003)
	Hybrid-EE-ECR	[88.6% - 86.3%] (2.3)	[0.85 - 0.799] (0.051)	[0.185 - 0.186] (0.001)	[2.821 - 2.852] (0.031)
	NiCE	[100% - 100%] (0.0)	[0.941 - 0.987] (0.046)	[0.446 - 0.463] (0.017)	[1.658 - 1.361] (0.297)
	Prototype	[82.6% - 77.8%] (4.8)	[0.917 - 0.97] (0.053)	[0.147 - 0.146] (0.001)	[15.254 - 15.25] (0.004)
	DiCE	[100% - 100%] (0.0)	[0.889 - 0.836] (0.053)	[0.349 - 0.346] (0.003)	[1.251 - 1.313] (0.062)
Adult	**Individual-EE**	[96.5% - 96.6%] (0.1)	[0.454 - 0.383] (0.071)	[0.13 - 0.131] (0.001)	[2.257 - 2.283] (0.026)
	Group-EE	[96.5% - 96.9%] (0.4)	[0.454 - 0.382] (0.072)	[0.159 - 0.157] (0.002)	[2.155 - 2.187] (0.032)
	Group-ECR	[93.3% - 93.7%] (0.4)	[0.454 - 0.383] (0.071)	[0.142 - 0.135] (0.007)	[2.128 - 2.138] (0.010)
	Hybrid-EE-ECR	[96.0% - 96.6%] (0.6)	[0.441 - 0.371] (0.070)	[0.195 - 0.192] (0.003)	[2.498 - 2.516] (0.018)
	NiCE	[100% - 100%] (0.0)	[0.488 - 0.548] (0.060)	[0.201 - 0.21] (0.009)	[1.398 - 1.601] (0.203)
	Prototype	[44.88 % - 39.1%] (5.78)	[0.493 - 0.565] (0.072)	[0.37 - 0.39] (0.02)	[5.9 - 6.2] (0.3)
	DiCE	[100% - 100%] (0.0)	[0.541 - 0.571] (0.030)	[0.246 - 0.248] (0.002)	[1.669 - 1.71] (0.041)
SSL	**Individual-EE**	[91.7% - 88.5%] (3.2)	[0.01 - 0.021] (0.011)	[0.149 - 0.178] (0.029)	[1.636 - 1.733] (0.097)
	Group-EE	[91.5% - 89.3%] (2.2)	[0.005 - 0.019] (0.014)	[0.123 - 0.176] (0.053)	[1.257 - 1.485] (0.228)
	Group-ECR	[88.3% - 81.7%] (6.6)	[0.012 - 0.021] (0.009)	[0.126 - 0.167] (0.041)	[1.148 - 1.378] (0.230)
	Hybrid-EE-ECR	[92.8% - 90.9%] (1.9)	[0.017 - 0.026] (0.009)	[0.187 - 0.213] (0.026)	[1.287 - 1.482] (0.195)
	NiCE	[100% - 100%] (0.0)	[0.067 - 0.233] (0.166)	[0.05 - 0.1] (0.050)	[1.147 - 1.317] (0.170)
	Prototype	[73% - 37.14%] (35.86)	[0.06 - 0.186] (0.126)	[0.034 - 0.039 (0.005)]	[1 - 1] (0)
	DiCE	[100% - 100%] (0.0)	[0.176 - 0.274] (0.098)	[0.29 - 0.265] (0.025)	[1.497 - 1.638] (0.141)
Law	**Individual-EE**	[77.4% - 79.6%] (2.2)	[0.277 - 0.268] (0.009)	[0.059 - 0.051] (0.008)	[1.886 - 1.832] (0.054)
	Group-EE	[82% - 83%] (1)	[0.277 - 0.268] (0.009)	[0.072 - 0.066] (0.006)	[1.784 - 1.694] (0.090)
	Group-ECR	[73.6% - 73.4%] (0.2)	[0.282 - 0.276] (0.006)	[0.069 - 0.066] (0.003)	[1.894 - 1.845] (0.049)
	Hybrid-EE-ECR	[78.8% - 80%] (0.2)	[0.278 - 0.27] (0.008)	[0.079 - 0.07] (0.009)	[1.823 - 1.768] (0.055)
	NiCE	[100% - 100%] (0.0)	[0.639 - 0.733] (0.094)	[0.174 - 0.156] (0.018)	[1.679 - 1.5] (0.179)
	Prototype	[35.5% - 37.2%] (1.7)	[0.671 - 0.658] (0.013)	[0.096 - 0.07] (0.026)	[2.051 - 2.1] (0.049)
	DiCE	[100% - 100%] (0.0)	[0.712 - 0.801] (0.089)	[0.197 - 0.181] (0.016)	[1.161 - 1.148] (0.013)

is relevant as it clarifies the implications of enforcing fairness constraints compared to traditional CFs.

Validity Across all datasets and scenarios, our approach consistently achieves high CF validity, typically ranging from 86% to 91% for *Alzheimer*, from 88% to 96% on *Adult and SSL*, and from 77% to 83% on *Law*, with the exception of under *Group-ECR*. This performance is comparable to the validity of CFs generated with DiCE and NiCE for the *SSL* and *Adult* datasets, where it is more stable than prototype-based CFs, with substantial drops in certain datasets like *Alzheimer* and *Law*. For instance, our approach under *Hybrid-EE-ECR* scenario attains validity rates of 88.6–86.3% on *Alzheimer*, 96.0–96.6% on *Adult*, 92.8-90.9% on *SSL*, and around 78.8-80% on *Law* for both groups. DiCE reaches full validity (100%) on all datasets. By contrast, prototype-based CFs exhibit highly inconsistent behavior and achieve 82.6-77.8% on *Alzheimer*, while their validity is low on *Adult* (as low as 39.1%) and *SSL* (37.14%) for certain groups. As observed, some values are lower than baselines, but this comes at the cost of other metrics, as explained later in details.

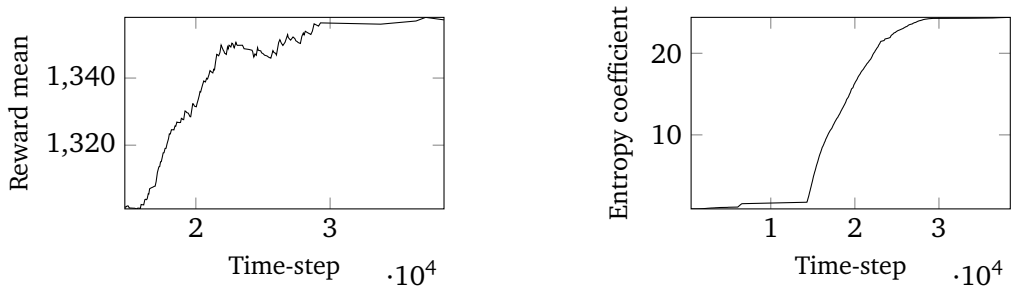
We now compare the results achieved across the diverse fairness scenarios. For *Individual-EE* and *Group-EE*, validity remains very high generally in the upper 90s across *SSL* and *Adult*, above 86% for *Alzheimer*, and above 77% for *Law*. In contrast, the *Group-ECR* scenario yields slightly lower validity in some cases (e.g., 73.25% on *Alzheimer*, 81.7% on *SSL*), reflecting the stricter constraints imposed by enforcing ECR. Importantly, our approach under the *Hybrid-EE-ECR* scenario balance validity at consistently high levels, ranging from 88.6–86.3% on *Alzheimer*, 96–96.6% on *Adult*, and around 92% on *SSL*. These results demonstrate that our approach maintains consistently high CF validity across datasets and fairness scenarios, with *Hybrid-EE-ECR* balancing quality metrics relative to baselines and group constraints.

Plausibility In terms of plausibility, measured as reconstruction error (Sec. 8.4), where lower values indicate more plausible CFs, our approach across fairness scenarios achieves plausibility scores that are consistently close to the training reconstruction error, indicating that the generated CFs remain realistic and aligned with the data distribution. For example, in the *Alzheimer*, plausibility ranges between 0.797 and 0.85 compared to the training error of 0.822, with similar patterns observed on *Adult*, *SSL* and *Law*. Compared to the baselines, a significant difference emerges: NiCE and DiCE achieve very high validity at the expense of plausibility, yielding reconstruction errors that are noticeably higher than those of our approach (e.g. > 0.9 on NiCE on Alzheimer, and > 0.545 for *Adult*). Prototype-based CFs also show less plausibility in many cases, making them less reliable overall.

Similarity and Minimality Similarity and minimality assess how closely generated CFs resemble the original instances. NiCE and DiCE produce CFs with relatively high Gower distances compared to the fairness scenarios, for example, 0.446–0.463 on *Alzheimer* of NiCE, and 0.246–0.248 of DiCE on *Adult*, indicating less similarity to the original inputs. Prototype CFs perform differently on different datasets, with as high as 0.37–0.39 on *Adult*, they alter up to 15 features on *Alzheimer* and > 6 on *Adult*. In contrast, our approach consistently produces CFs that are both closer to the original instances and require only a small number of modifications across all scenarios. For example, in *Individual-EE* in *Adult*, the approach generates CFs with Gower distances of 0.13 while modifying an average of 2.28 features. In *Group-EE*, our proposed approach results in a similar number of feature changes (2.1 features on *Adult*) but still maintains low Gower values (0.159–0.157). In *Hybrid-EE-ECR*, the approach provides a balanced trade-off, producing CFs with moderate edits (e.g., two features and a similarity of 0.195-0.192 on *Adult*, and 2.5 features, with similarity values of 0.203–0.210 on *Alzheimer*). For *Law*, the emphasis on similarity and minimality helps explain its lower validity score compared to baselines, as our approach achieves notably higher similarity than the baselines, for instance, 0.197 on DiCE compared to 0.079 on *Hybrid-EE-ECR*, highlighting its stronger alignment performance. Overall, across datasets, our approach preserves proximity to the original data while avoiding excessive edits.

Takeaway 8: Our approach demonstrates that fairness constraints do not necessarily impose significant drawbacks on CF quality, but balance different metrics. While baseline approaches often exhibit trade-offs between fairness and performance metrics, such as achieving high validity at the expense of similarity. The integration of fairness objectives preserves validity, plausibility, and similarity at levels comparable to existing baselines. In addition, hybrid fairness balances individual and group fairness while preserving the quality of CFs and maintaining fairness at both levels.

8.6.5 Proof of Convergence of RL



(a) Reward Mean (b) Entropy

Figure 8.6. Visualization of RL performance metrics for a sample experiment on the Alzheimer dataset.

Fig. 8.6 illustrates the convergence plots of the RL-based approach with Hybrid-EE-ECR during training, showing the evolution of two key SAC metrics: mean reward and the entropy coefficient. The mean reward curve shows that the RL agent's performance improves over time, indicating effective learning that generates CFs. The entropy coefficient plot reflects the degree of exploration during training. The rising mean reward indicates that the agent is learning to maximize cumulative rewards. The upward trend in the entropy coefficient confirms that the agent balances reward maximization with exploration, a desirable property for the search for CF solutions.

8.7 Conclusion

In this chapter, we address the problem of generating fair counterfactual explanations (CFs) with the aim of ensuring that individuals who share comparable attributes or belong to different protected groups (e.g., demographic groups), are provided with comparable explanations. We define fairness in CFs at the individual, group, and hybrid levels and propose a model-agnostic, reinforcement-learning-based approach. Our approach optimizes CF fairness by employing customized reward functions that enforce equal effectiveness and equal choice of recourse, thereby balancing fairness with the quality of generated CFs. Experiments across four datasets demonstrate the effectiveness of our approach for generating CFs that adhere to fairness constraints while preserving their quality. Additionally, we demonstrate the trade-offs of applying hybrid fairness constraints and show that this approach does not inherently compromise CF quality. In terms of limitations, our proposed approach introduces computational

overhead, which may limit scalability in higher-dimensional settings, and provides a direction for future work to investigate efficient optimization strategies.

Chapter 9

Conclusion

This thesis investigated the interplay between *privacy, explainability, and fairness* in Responsible Artificial Intelligence (AI). Its central hypothesis is that these dimensions should not be treated as isolated desiderata, but as interdependent requirements whose interactions shape the behavior, risks, and societal impact of Machine Learning (ML) models. While each principle has been widely studied independently, this thesis showed that interventions targeting one dimension can affect the others in significant and sometimes counterintuitive ways. Privacy-preserving mechanisms may alter explanations, explanations may expose sensitive information, and fairness interventions may reshape model explanations. Understanding these interactions is therefore essential for developing AI systems that are not only accurate but also transparent, privacy-preserving, and equitable. In particular, this thesis focused on two pairwise interplays: *privacy and explainability*, and *fairness and explainability*.

The first part of the thesis addressed the interplay between privacy and explainability. With respect to RQ1, the results showed that explanations, particularly counterfactual explanations (CFs), can increase the privacy attack surface of deployed ML models. In realistic Machine Learning as a Service (MLaaS) settings, CFs provide useful information to adversaries performing model extraction and membership inference attacks. By revealing local properties of the decision boundary and the actionable changes required to alter predictions, CFs can help attackers approximate the target model and infer information about its training data. These findings show that access to explanations should be treated as part of the model's information disclosure channel and explicitly considered in privacy threat models.

With respect to RQ2, the thesis analyzed how privacy-preserving mechanisms affect both explanation-based attacks and the quality of explanations. Differential privacy (DP) and active learning were investigated as mitigation strategies. The results showed that incorporating these mechanisms at the model or explanation-generation level can

reduce privacy leakage but introduce trade-offs that impact predictive performance and the realism, plausibility, stability, and usefulness of explanations. In particular, privacy-preserving CF generation reduced leakage risks, as measured by the success of attacks, while the study of DP in anomaly detection showed that the impact of privacy noise on SHAP explanations is model- and data-dependent. These findings highlight that privacy defenses should be evaluated not only in terms of attack reduction or model performance but also in terms of their impact on the explanations provided to users and auditors.

With respect to RQ3, the thesis broadened the discussion of privacy beyond the individual level by looking into collective privacy and its relationship to explainable AI. This work argued that existing privacy frameworks remain limited when harms emerge at the level of groups, communities, relationships, or interdependent identities. Explainability may support collective privacy by making collective risks more visible and by helping users, regulators, and system designers understand how group-level inferences are generated. At the same time, explanations must be carefully designed, since they may themselves expose collective patterns. This frames collective privacy as both a technical and socio-legal challenge that requires interdisciplinary approaches to AI governance.

The second part of the thesis investigated the interplay between fairness and explainability. With respect to RQ4, the thesis examined how fairness-enhancing techniques affect the quality of explanations. The results showed that improving fairness does not necessarily require sacrificing predictive performance or explanation consistency. In particular, Fair-FLIP substantially improved fairness metrics while incurring only a marginal reduction in model performance. The preliminary explainability analysis further indicated that its attention patterns remained close to those of the baseline model, suggesting that fairness-oriented interventions can be designed to preserve explanatory behavior.

With respect to RQ5, the thesis addressed the fairness of explanations themselves using a reinforcement learning-based framework to generate fair CFs. This work distinguished between fair models and fair explanations, showing that satisfying fairness constraints at the prediction level does not automatically ensure fair or equitable recourse. The proposed approach defined fairness in CFs at the individual, group, and hybrid levels, and optimized for equal effectiveness and equal choice of recourse. Across multiple datasets, the results showed that CFs can be generated in a way that balances explanation quality with fairness constraints.

In conclusion, this thesis shows that privacy, explainability, and fairness are deeply interconnected dimensions of Responsible AI. Treating them independently risks overlooking how progress in one dimension can create vulnerabilities or inequities in another. By analyzing these interactions, proposing mitigation strategies, and developing

fairness-aware explanation methods, this thesis contributes to the design of Responsible AI systems that better balance transparency, privacy protection, and equitable treatment.

Limitations and Future Perspectives

This thesis advances the understanding of the interactions among privacy, explainability, and fairness, but several limitations remain.

Pairwise Rather Than Triadic Analysis. The empirical and theoretical investigations primarily examine pairwise relationships between privacy and explainability, and between fairness and explainability. This design enables focused analysis of specific tensions such as privacy leakage induced by CFs, the impact of differential privacy on explanation fidelity, and fairness-driven shifts in explanation structure. However, real-world deployments might not exhibit such separability, and future work should therefore develop unified evaluation frameworks that jointly quantify privacy risk, fairness disparities, explanation quality, and predictive utility, enabling multi-objective optimization and trade-off analysis.

Model and Explanation Families. The empirical studies rely on selected model classes for classification, datasets, and explanation methods, with emphasis on counterfactuals and feature attributions. Although these methods are widely adopted in the real world, they represent a subset of explainability techniques. Extending the analysis to gradient-based, example-based, concept-based, and intrinsically interpretable models, as well as explanation methods for large language models and multimodal systems, would provide a more comprehensive characterization of the triadic interplay.

Computational Metrics and Human Interpretability. Explanation quality is assessed primarily through computational metrics (e.g., validity, plausibility, proximity, sparsity, stability). While these metrics enable reproducible comparison, they only approximate human interpretability. Explanations that satisfy numerical criteria may still be perceived by users as unrealistic, unactionable, or inequitable. Future work should therefore incorporate human-centered evaluation through user studies involving affected individuals, domain experts, auditors, and system designers, to assess whether private and fair explanations are understandable, trustworthy, actionable, and equitable in practice.

Fairness Definitions and Explanation-Level Fairness. The fairness analyses focus on selected notions of fairness, including group fairness, individual fairness, and metrics such as equal effectiveness and equal recourse. These notions are central, but fairness remains inherently contextual. Future research should develop richer metrics for explanation-level fairness, including intersectional, temporal, causal, and user-perceived fairness, and examine fairness-aware explanations when protected attributes are unavailable, noisy, or legally restricted.

Formalization of Collective Privacy. The discussion of collective privacy is primarily conceptual and socio-technical. While it highlights the limitations of individual-centric privacy frameworks and the potential role of explainability in supporting collective governance, collective privacy requires practical formalization. Future work should investigate how collective privacy risks can be measured, how group-level harms can be represented, and how privacy guarantees can be extended to dynamic and overlapping collectives. Designing explanation interfaces that communicate collective risks without overwhelming users or revealing sensitive information will require interdisciplinary collaboration across machine learning, privacy engineering, human-computer interaction, law, ethics, and the social sciences.

Toward Deployment-Aware Responsible AI. Finally, the findings of this thesis underscore the need for deployment-oriented Responsible AI systems in which privacy, fairness, and explainability are monitored throughout the model lifecycle. These dimensions should be continuously assessed during data collection, model training, explanation generation, deployment, and post-deployment monitoring. Developing practical tools capable of detecting instability in explanations, increases in privacy risk, or emergent fairness disparities would translate the conceptual contributions of this thesis into operational governance mechanisms for real-world AI systems.

Ultimately, these limitations underscore the need for a fundamental shift in how Responsible AI systems are designed and evaluated. Privacy, explainability, and fairness must be treated as interdependent, not orthogonal, objectives. This thesis pushes toward integrated, human-centered, and deployment-aware frameworks that can reason about their trade-offs and synergies as a unified whole.

Bibliography

[1] Dan Ley, Saumitra Mishra, and Daniele Magazzeni. Globe-ce: A translation based approach for global counterfactual explanations. In *International conference on machine learning*, pages 19315–19342. PMLR, 2023.

[2] Loukas Kavouras, Konstantinos Tsopelas, Giorgos Giannopoulos, Dimitris Sacharidis, Eleni Psaroudaki, Nikolaos Theologitis, Dimitrios Rontogiannis, Dimitris Fotakis, and Ioannis Emiris. Fairness aware counterfactuals for subgroups. *Advances in Neural Information Processing Systems*, 36, 2024.

[3] Christos Fragkathoulas, Vasiliki Papanikou, Evaggelia Pitoura, and Evimaria Terzi. Fgce: Feasible group counterfactual explanations for auditing fairness. In *Greeks in AI Symposium 2025*.

[4] Debidutta Pattnaik, Sougata Ray, and Raghu Raman. Applications of artificial intelligence and machine learning in the financial services industry: A bibliometric review. *Heliyon*, 10(1), 2024.

[5] Lukas Buess, Matthias Keicher, Nassir Navab, Andreas Maier, and Soroosh Tayebi Arasteh. From large language models to multimodal ai: A scoping review on the potential of generative ai in medicine. *Biomedical Engineering Letters*, pages 1–19, 2025.

[6] Debesh Jha, Gorkem Durak, Vanshali Sharma, Elif Keles, Vedat Cicek, Zheyuan Zhang, Abhishek Srivastava, Ashish Rauniyar, Desta Haileselassie Hagos, Nikhil Kumar Tomar, et al. A conceptual framework for applying ethical principles of ai to medical practice. *Bioengineering*, 12(2):180, 2025.

[7] Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, et al. Managing extreme ai risks amid rapid progress. *Science*, 384(6698):842–845, 2024.

[8] Patrick Mikalef, Kieran Conboy, Jenny Eriksson Lundström, and Aleš Popovič. Thinking responsibly about responsible ai and 'the dark side'of ai, 2022.

[9] Petar Radanliev, Omar Santos, Alistair Brandon-Jones, and Adam Joinson. Ethics and responsible ai deployment. *Frontiers in Artificial Intelligence*, 7: 1377011, 2024.

[10] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM computing surveys*, 55(9): 1–33, 2023.

[11] Julien Ferry, Ulrich Aïvodji, Sébastien Gambs, Marie-José Huguet, and Mohamed Siala. Taming the triangle: On the interplays between fairness, interpretability, and privacy in machine learning. *Computational Intelligence*, 41(4): e70113, 2025.

[12] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14, 2020.

[13] Dhananjay Kulkarni. Data privacy and regulatory concerns. *AI, Machine Learning, and Image Processing in Market Research and Branding*, page 247, 2026.

[14] Cynthia Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer, 2008.

[15] European Parliament and Council of the European Union. General data protection regulation (GDPR)). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC*, 2016.

[16] Regulation (EU) 2024/1689 of the european parliament and of the council of 12 july 2024 establishing harmonized rules on artificial intelligence, 2024. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206. EU Artificial Intelligence Act. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.

[17] Swiss Confederation. New federal act on data protection (nfadp). *https://www.kmu.admin.ch/kmu/en/home/facts-and-trends/digitization/data-protection/new-federal-act-on-data-protection-nfadp.html*, 2023.

[18] Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1310–1321, 2015.

[19] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[20] Aleksei Triastcyn. *Data-Aware Privacy-Preserving Machine Learning*. PhD thesis, EPFL, Switzerland, 2020.

[21] Maria Rigaki and Sebastian Garcia. A survey of privacy attacks in machine learning. *ACM Computing Surveys*, 56(4):1–34, 2023.

[22] Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006.

[23] Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In *28th USENIX security symposium (USENIX security 19)*, pages 1895–1912, 2019.

[24] Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, et al. Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106:102301, 2024.

[25] Vincent Damen, Menno Wiersma, Gokce Aydin, and Rens van Haasteren. Explainable ai for eu ai act compliance audits. *Maandblad voor Accountancy en Bedrijfseconomie*, 99(4):231–242, 2025.

[26] Giulia Vilone and Luca Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106, 2021.

[27] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Neural Information Processing Systems*, 2017. URL https://api.semanticscholar.org/CorpusID:21889700.

[28] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[29] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

[30] Matthew G Hanna, Liron Pantanowitz, Brian Jackson, Octavia Palmer, Shyam Visweswaran, Joshua Pantanowitz, Mustafa Deebajah, and Hooman H Rashidi. Ethical and bias considerations in artificial intelligence/machine learning. *Modern Pathology*, 38(3):100686, 2025.

[31] Tiago P Pagano, Rafael B Loureiro, Fernanda VN Lisboa, Rodrigo M Peixoto, Guilherme AS Guimarães, Gustavo OR Cruz, Maira M Araujo, Lucas L Santos, Marco AS Cruz, Ewerton LS Oliveira, et al. Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big data and cognitive computing*, 7(1): 15, 2023.

[32] Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, pages 1–7, 2018.

[33] Aimee Kendall Roundtree. Ai explainability, interpretability, fairness, and privacy: an integrative review of reviews. In *International Conference on Human-Computer Interaction*, pages 305–317. Springer, 2023.

[34] Yongjie Wang, Hangwei Qian, and Chunyan Miao. DualCF: Efficient model extraction attack from counterfactual explanations. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1318–1329, 2022.

[35] Reza Shokri, Martin Strobel, and Yair Zick. On the privacy risks of model explanations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 231–241, 2021.

[36] Saifullah Saifullah, Dominique Mercier, Adriano Lucieri, Andreas Dengel, and Sheraz Ahmed. The privacy-explainability trade-off: unraveling the impacts of differential privacy and federated learning on attribution methods. *Frontiers in Artificial Intelligence*, 7:1236947, 2024.

[37] David Pujol and Ashwin Machanavajjhala. Equity and privacy: More than just a tradeoff. *IEEE Security & Privacy*, 19(6):93–97, 2021.

[38] Sushant Agarwal. Trade-offs between fairness and privacy in machine learning. In *IJCAI 2021 Workshop on AI for Social Good*, 2021.

[39] David Pujol, Ryan McKenna, Satya Kuppam, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. Fair decision making using privacy-protected data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 189–199, 2020.

[40] Apple Card algorithm sparks gender bias allegations against goldman sachs. https://www.washingtonpost.com/business/2019/11/11/apple-card-algorithm-sparks-gender-bias-allegations-against-goldman-sachs/. Accessed: 2025-10-20.

[41] Hongyan Chang and Reza Shokri. On the privacy risks of algorithmic fairness. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 292–303. IEEE, 2021.

[42] Alan F Westin. Privacy and freedom. *Washington and Lee Law Review*, 25(1): 166, 1968.

[43] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction {APIs}. In *25th USENIX Security Symposium (USENIX Security 16)*, pages 601–618, 2016.

[44] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.

[45] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37, 2022.

[46] Jun Niu, Peng Liu, Xiaoyan Zhu, Kuo Shen, Yuecong Wang, Haotian Chi, Yulong Shen, Xiaohong Jiang, Jianfeng Ma, and Yuqing Zhang. A survey on membership inference attacks and defenses in machine learning. *Journal of Information and Intelligence*, 2024.

[47] Martin Pawelczyk, Himabindu Lakkaraju, and Seth Neel. On the privacy risks of algorithmic recourse. In *International Conference on Artificial Intelligence and Statistics*, pages 9680–9696. PMLR, 2023.

[48] Ahmed Salem, Yang Zhang, Mathias Humbert, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In *Network and Distributed Systems Security Symposium 2019*. Internet Society, 2019.

[49] David Gunning. Explainable artificial intelligence (xai). *Defense advanced research projects agency (DARPA), nd Web*, 2(2):1, 2017.

[50] Timo Speith. A review of taxonomies of explainable artificial intelligence (xai) methods. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2239–2250, 2022.

[51] Jenny Bunn. Working in contexts for which transparency is important: A record-keeping view of explainable artificial intelligence (xai). *Records Management Journal*, 30(2):143–153, 2020.

[52] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7): e0130140, 2015.

[53] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[54] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017.

[55] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[56] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, 20(177):1–81, 2019.

[57] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 0210–0215. IEEE, 2018.

[58] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[59] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. Face: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350, 2020.

[60] Raluca-Ioana Cobzaru. *From Theory to Practice: Improving Causal Conclusions from Healthcare Data*. PhD thesis, Massachusetts Institute of Technology, 2025.

[61] Satyam Kumar, Yelleti Vivek, Vadlamani Ravi, and Indranil Bose. A comprehensive review of causal inference in banking, finance, and insurance. *ACM Computing Surveys*, 2025.

[62] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, 2020.

[63] Dieter Brughmans, Pieter Leyman, and David Martens. Nice: an algorithm for nearest instance counterfactual explanations. *Data mining and knowledge discovery*, 38(5):2665–2703, 2024.

[64] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[65] Pei Wang and Nuno Vasconcelos. Scout: Self-aware discriminant counterfactual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8981–8990, 2020.

[66] Ilia Stepin, Jose M Alonso, Alejandro Catala, and Martin Pereira-Fariña. Generation and evaluation of factual and counterfactual explanations for decision trees and fuzzy rule-based classifiers. In *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8. IEEE, 2020.

[67] Ana Lucic, Harrie Oosterhuis, Hinda Haned, and Maarten de Rijke. Focus: Flexible optimizable counterfactual explanations for tree ensembles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5313–5322, 2022.

[68] Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. In *International Conference on Artificial Intelligence and Statistics*, pages 895–905. PMLR, 2020.

[69] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 10–19, 2019.

[70] Gabriele Tolomei, Fabrizio Silvestri, Andrew Haines, and Mounia Lalmas. Interpretable predictions of tree-based ensembles via actionable feature tweaking. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 465–474, 2017.

[71] David Martens and Foster Provost. Explaining data-driven document classifications. *MIS quarterly*, 38(1):73–100, 2014.

[72] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. *Advances in neural information processing systems*, 31, 2018.

[73] Jonathan Moore, Nils Hammerla, and Chris Watkins. Explaining deep learning models with constrained adversarial examples. In *PRICAI 2019: Trends in Artificial Intelligence: 16th Pacific Rim International Conference on Artificial Intelligence, Cuvu, Yanuca Island, Fiji, August 26–30, 2019, Proceedings, Part I 16*, pages 43–56. Springer, 2019.

[74] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models. *arXiv preprint arXiv:1905.07857*, 2019.

[75] Masoud Hashemi and Ali Fathi. Permuteattack: Counterfactual explanation of machine learning credit scorecards. *arXiv preprint arXiv:2008.10138*, 2020.

[76] Daniel Nemirovsky, Nicolas Thiebaut, Ye Xu, and Abhishek Gupta. Countergan: Generating counterfactuals for real-time recourse and interpretability using residual gans. In *Uncertainty in Artificial Intelligence*, pages 1488–1497. PMLR, 2022.

[77] Danilo Numeroso and Davide Bacciu. Meg: Generating molecular counterfactual explanations for deep graph networks. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

[78] Tri Minh Nguyen, Thomas P Quinn, Thin Nguyen, and Truyen Tran. Counterfactual explanation with multi-agent reinforcement learning for drug target prediction. *arXiv preprint arXiv:2103.12983*, 2021.

[79] Daniel Nemirovsky, Nicolas Thiebaut, Ye Xu, and Abhishek Gupta. Countergan: Generating counterfactuals for real-time recourse and interpretability using residual gans. In *Uncertainty in Artificial Intelligence*, pages 1488–1497. PMLR, 2022.

[80] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 607–617, 2020.

[81] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and machine learning. fairmlbook. org, 2019.

[82] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[83] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[84] Faisal Kamiran, Asim Karim, Sicco Verwer, and Heike Goudriaan. Classifying socially sensitive data without discrimination: An analysis of a crime suspect dataset. In *2012 IEEE 12th International Conference on Data Mining Workshops*, pages 370–377. IEEE, 2012.

[85] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

[86] Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. Fairtest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 401–416. IEEE, 2017.

[87] Tomasz Szandała and Henryk Maciejewski. Discriminating feature ratio: Introducing metric for uncovering vulnerabilities in deep convolutional neural networks. *Knowledge-Based Systems*, 302:112306, 2024.

[88] Saeid Tizpaz-Niari, Ashish Kumar, Gang Tan, and Ashutosh Trivedi. Fairness-aware configuration of machine learning libraries. In *Proceedings of the 44th International Conference on Software Engineering*, pages 909–920, 2022.

[89] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.

[90] Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. In *International Conference on Machine Learning*, pages 1397–1405. PMLR, 2019.

[91] Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015.

[92] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[93] Shengjia Zhao, Michael Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. Calibrating predictions to decisions: A novel approach to multi-class calibration. *Advances in Neural Information Processing Systems*, 34:22313–22324, 2021.

[94] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

[95] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.

[96] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008.

[97] Gopinath Muruti, Fiza Abdul Rahim, and Zul-Azri bin Ibrahim. A survey on anomalies detection techniques and measurement methods. In *2018 IEEE conference on application, information and network security (AINS)*, pages 81–86. IEEE, 2018.

[98] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4794–4802, 2019.

[99] Richard S Sutton, Andrew G Barto, et al. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998.

[100] Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

[101] Vijay R Konda and John N Tsitsiklis. Onactor-critic algorithms. *SIAM journal on Control and Optimization*, 42(4):1143–1166, 2003.

[102] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[103] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

[104] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.

[105] Pramila P Shinde and Seema Shah. A review of machine learning and deep learning applications. In *2018 Fourth international conference on computing communication control and automation (ICCUBEA)*, pages 1–6. IEEE, 2018.

[106] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521 (7553):436–444, 2015.

[107] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 739–753. IEEE, 2019.

[108] Akram Aslani and Mostafa Ghobaei-Arani. Machine learning inference serving models in serverless computing: a survey. *Computing*, 107(1):47, 2025.

[109] Google auto ml. https://cloud.google.com/automl?hl=en, . Accessed: 2024-01-04.

[110] Microsoft explainable ai. https://learn.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-interpretability?view=azureml-api-2&WT.mc_id=docs-article-lazzeri. Accessed: 2024-01-04.

[111] Ibm explainable ai. https://aix360.mybluemix.net/. Accessed: 2024-01-04.

[112] Google explainable ai. https://cloud.google.com/explainable-ai, . Accessed: 2024-01-04.

[113] Aws explainable ai. https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-model-explainability.html. Accessed: 2024-01-04.

[114] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006.

[115] Subhasish Ghosh, Amit Kr Mandal, and Agostino Cortesi. Enhancing deep learning model privacy against membership inference attacks using privacy-preserving oversampling. *SN Computer Science*, 6(4):1–21, 2025.

[116] Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Thanh Toan Nguyen, Phi Le Nguyen, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of privacy-preserving model explanations: Privacy risks, attacks, and countermeasures. *arXiv preprint arXiv:2404.00673*, 2024.

[117] Bjørn Aslak Juliussen. The right to an explanation under the gdpr and the ai act. In *International Conference on Multimedia Modeling*, pages 184–197. Springer, 2025.

[118] Cecilia Panigutti, Ronan Hamon, Isabelle Hupont, David Fernandez Llorca, Delia Fano Yela, Henrik Junklewitz, Salvatore Scalzo, Gabriele Mazzini, Ignacio Sanchez, Josep Soler Garrido, et al. The role of explainable ai in the context of the ai act. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pages 1139–1150, 2023.

[119] Duy Nguyen, Ngoc Bui, and Viet Anh Nguyen. Feasible recourse plan via diverse interpolation. In *International Conference on Artificial Intelligence and Statistics*, pages 4679–4698. PMLR, 2023.

[120] Duong Chi Thang, Nguyen Thanh Tam, Nguyen Quoc Viet Hung, and Karl Aberer. An evaluation of diversification techniques. In *Database and Expert Systems Applications: 26th International Conference, DEXA 2015, Valencia, Spain, September 1-4, 2015, Proceedings, Part II 8*, pages 215–231. Springer, 2015.

[121] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1312, 2019.

[122] Catherine Huang, Martin Pawelczyk, and Himabindu Lakkaraju. Explaining the model, protecting your data: Revealing and mitigating the data privacy risks of post-hoc model explanations via membership inference. In *ICML 2024 Next Generation of AI Safety Workshop*, .

[123] Catherine Huang, Chelse Swoopes, Christina Xiao, Jiaqi Ma, and Himabindu Lakkaraju. Accurate, explainable, and private models: Providing recourse while minimizing training data leakage. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*, .

[124] Soham Pal, Yash Gupta, Aditya Shukla, Aditya Kanade, Shirish Shevade, and Vinod Ganapathy. Activethief: Model extraction using active learning and unannotated public data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 865–872, 2020.

[125] Kalpesh Krishna, Gaurav Singh Tomar, Ankur P. Parikh, Nicolas Papernot, and Mohit Iyyer. Thieves on sesame street! model extraction of bert-based apis. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Byl5NREFDr.

[126] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Knockoff nets: Stealing functionality of black-box models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4954–4963, 2019.

[127] Mika Juuti, Sebastian Szyller, Samuel Marchal, and N Asokan. Prada: protecting against dnn model stealing attacks. In *2019 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 512–527. IEEE, 2019.

[128] Seong Joon Oh, Bernt Schiele, and Mario Fritz. Towards reverse-engineering black-box neural networks. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 121–144, 2019.

[129] Kaiyi Pang, Tao Qi, Chuhan Wu, Minhao Bai, Minghu Jiang, and Yongfeng Huang. Modelshield: Adaptive and robust watermark against model extraction attack. *IEEE Transactions on Information Forensics and Security*, 2025.

[130] Xinwei Zhang, Haibo Hu, Qingqing Ye, Li Bai, and Huadi Zheng. Mer-inspector: Assessing model extraction risks from an attack-agnostic perspective. In *Proceedings of the ACM on Web Conference 2025*, pages 4300–4315, 2025.

[131] Yulian Sun, Vedant Bonde, Li Duan, and Yong Li. Obfuscation for deep neural networks against model extraction: Attack taxonomy and defense optimization. In *International Conference on Applied Cryptography and Network Security*, pages 391–414. Springer, 2025.

[132] Matthew Jagielski, Nicholas Carlini, David Berthelot, Alex Kurakin, and Nicolas Papernot. High accuracy and high fidelity extraction of neural networks. In *29th USENIX security symposium (USENIX Security 20)*, pages 1345–1362, 2020.

[133] Ulrich Aïvodji, Alexandre Bolot, and Sébastien Gambs. Model extraction from counterfactual explanations. *arXiv preprint arXiv:2009.01884*, 2020.

[134] Abdullah Caglar Oksuz, Anisa Halimi, and Erman Ayday. Autolycus: Exploiting explainable ai (xai) for model extraction attacks against decision tree models. *arXiv preprint arXiv:2302.02162*, 2023.

[135] Anli Yan, Ruitao Hou, Xiaozhang Liu, Hongyang Yan, Teng Huang, and Xianmin Wang. Towards explainable model extraction attacks. *International Journal of Intelligent Systems*, 37(11):9936–9956, 2022.

[136] Kacper Sokol and Peter Flach. Counterfactual explanations of machine learning predictions: Opportunities and challenges for ai safety. In *2019 AAAI Workshop on Artificial Intelligence Safety, SafeAI 2019*. CEUR Workshop Proceedings, 2019.

[137] Smitha Milli, Ludwig Schmidt, Anca D Dragan, and Moritz Hardt. Model reconstruction from model explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 1–9, 2019.

[138] Takayuki Miura, Toshiki Shibahara, and Naoto Yanai. Megex: Data-free model extraction attack against gradient-based explainable ai. In *Proceedings of the 2nd*

ACM Workshop on Secure and Trustworthy Deep Learning Systems, pages 56–66, 2024.

[139] Neel Patel, Reza Shokri, and Yair Zick. Model explanations with differential privacy. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1895–1904, 2022.

[140] Fan Yang, Qizhang Feng, Kaixiong Zhou, Jiahao Chen, and Xia Hu. Differentially private counterfactuals via functional mechanism. *arXiv preprint arXiv:2208.02878*, 2022.

[141] Sikha Pentyala, Shubham Sharma, Sanjay Kariyappa, Freddy Lecue, and Daniele Magazzeni. Privacy-preserving algorithmic recourse. *arXiv preprint arXiv:2311.14137*, 2023.

[142] Give me some credit dataset. https://www.kaggle.com/c/GiveMeSomeCredit. Accessed: 2024-01-04.

[143] Credit card fraud dataset. https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud. Accessed: 2024-01-04.

[144] California housing dataset. https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_california_housing.html. Accessed: 2024-01-04.

[145] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*, 2019.

[146] Wisam Abbasi, Paolo Mori, and Andrea Saracino. Further insights: Balancing privacy, explainability, and utility in machine learning-based tabular data analysis. In *Proceedings of the 19th International Conference on Availability, Reliability and Security*, pages 1–10, 2024.

[147] O. Roesler. Eeg eye state. UCI Machine Learning Repository, 2013. URL https://doi.org/10.24432/C57G7J. Accessed: 2024-01-04.

[148] Abdullah Caglar Oksuz, Anisa Halimi, and Erman Ayday. Autolycus: Exploiting explainable artificial intelligence (xai) for model extraction attacks against interpretable models. *Proceedings on Privacy Enhancing Technologies*, 2024.

[149] Michael Veale, Reuben Binns, and Lilian Edwards. Algorithms that remember: model inversion attacks and data protection law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133): 20180083, 2018.

[150] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*, pages 5558–5567. PMLR, 2019.

[151] Rob Munro. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Manning, 2021.

[152] Noveen Sachdeva and Julian McAuley. Data distillation: A survey. *arXiv preprint arXiv:2301.04272*, 2023.

[153] Shiye Lei and Dacheng Tao. A comprehensive survey of dataset distillation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):17–32, 2023.

[154] Tian Dong, Bo Zhao, and Lingjuan Lyu. Privacy for free: How does dataset condensation help privacy? In *International Conference on Machine Learning*, pages 5378–5396. PMLR, 2022.

[155] Tianhang Zheng and Baochun Li. Differentially private dataset condensation. 2023.

[156] Han Liu, Yuhao Wu, Zhiyuan Yu, and Ning Zhang. Please tell me more: Privacy impact of explainability through the lens of membership inference attack. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 4791–4809. IEEE, 2024.

[157] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE, 2018.

[158] Milad Nasr, Reza Shokri, and Amir Houmansadr. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, pages 634–646, 2018.

[159] Jiacheng Li, Ninghui Li, and Bruno Ribeiro. Membership inference attacks and defenses in classification models. In *Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy*, pages 5–16, 2021.

[160] Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong. Memguard: Defending against black-box membership inference attacks via adversarial examples. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, pages 259–274, 2019.

[161] Md Atiqur Rahman, Tanzila Rahman, Robert Laganière, Noman Mohammed, and Yang Wang. Membership inference attack against differentially private deep learning model. *Trans. Data Priv.*, 11(1):61–79, 2018.

[162] Dayong Ye, Sheng Shen, Tianqing Zhu, Bo Liu, and Wanlei Zhou. One parameter defense—defending against data inference attacks via differential privacy. *IEEE Transactions on Information Forensics and Security*, 17:1466–1480, 2022.

[163] Xiaoyong Yuan and Lan Zhang. Membership inference attacks and defenses in neural network pruning. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 4561–4578, 2022.

[164] Rakshit Naidu, Aman Priyanshu, Aadith Kumar, Sasikanth Kotti, Haofan Wang, and Fatemehsadat Mireshghallah. When differential privacy meets interpretability: A case study. *arXiv preprint arXiv:2106.13203*, 2021.

[165] Harsha Nori, Rich Caruana, Zhiqi Bu, Judy Hanwen Shen, and Janardhan Kulkarni. Accuracy, interpretability, and differential privacy via explainable boosting. In *International conference on machine learning*, pages 8227–8237. PMLR, 2021.

[166] O. Roesler. Eeg eye state [dataset]. *UCI Machine Learning Repository*, 2013.

[167] Joaquín Torres-Sospedra, Raúl Montoliu, Adolfo Martínez-Usó, Joan P Avariento, Tomás J Arnau, Mauri Benedito-Bordonau, and Joaquín Huerta. Ujiindoorloc: A new multi-building and multi-floor database for wlan fingerprint-based indoor localization problems. In *2014 international conference on indoor positioning and indoor navigation (IPIN)*, pages 261–270. IEEE, 2014.

[168] D. Martens D. Brughmans, P. Leyman. Nice: an algorithm for nearest instance counterfactual explanations. *Data Mining and Knowledge Discovery*, 2023.

[169] Mohiuddin Ahmed, Abdun Naser Mahmood, and Jiankun Hu. A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, 60:19–31, 2016.

[170] Honglu Jiang, Jian Pei, Dongxiao Yu, Jiguo Yu, Bei Gong, and Xiuzhen Cheng. Applications of differential privacy in social network analysis: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):108–127, 2021.

[171] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[172] Basma Alharbi, Zhenwen Liang, Jana M Aljindan, Ammar K Agnia, and Xiangliang Zhang. Explainable and interpretable anomaly detection models for production data. *SPE Journal*, 27(01):349–363, 2022.

[173] Khushnaseeb Roshan and Aasim Zafar. Using kernel shap xai method to optimize the network anomaly detection model. In *2022 9th International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 74–80. IEEE, 2022.

[174] Mengwei Yang, Linqi Song, Jie Xu, Congduan Li, and Guozhen Tan. The tradeoff between privacy and accuracy in anomaly detection using federated xgboost. *arXiv preprint arXiv:1907.07157*, 2019.

[175] Rudolf Mayer, Markus Hittmeir, and Andreas Ekelhart. Privacy-preserving anomaly detection using synthetic data. In *Data and Applications Security and Privacy XXXIV: 34th Annual IFIP WG 11.3 Conference, DBSec 2020, Regensburg, Germany, June 25–26, 2020, Proceedings 34*, pages 195–207. Springer, 2020.

[176] Kwassi H Degue, Karthik Gopalakrishnan, Max Z Li, and Hamsa Balakrishnan. Differentially private outlier detection in correlated data. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 2735–2742. IEEE, 2021.

[177] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1): 18, 2020.

[178] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[179] Mohiuddin Ahmed, Abdun Naser Mahmood, and Md Rafiqul Islam. A survey of anomaly detection techniques in financial domain. *Future Generation Computer Systems*, 55:278–288, 2016.

[180] Muneeb Ul Hassan, Mubashir Husain Rehmani, and Jinjun Chen. Differential privacy in blockchain technology: A futuristic approach. *Journal of Parallel and Distributed Computing*, 145:50–74, 2020.

[181] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International conference on learning representations*, 2018.

[182] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019.

[183] Kingsly Leung and Christopher Leckie. Unsupervised anomaly detection in network intrusion detection using clusters. In *Proceedings of the Twenty-eighth Australasian conference on Computer Science-Volume 38*, pages 333–342, 2005.

[184] Mohsin Munir, Shoaib Ahmed Siddiqui, Andreas Dengel, and Sheraz Ahmed. Deepant: A deep learning approach for unsupervised anomaly detection in time series. *Ieee Access*, 7:1991–2005, 2018.

[185] Watson Jia, Raj Mani Shukla, and Shamik Sengupta. Anomaly detection using supervised learning and multiple statistical methods. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 1291–1297. IEEE, 2019.

[186] Zhaomin Chen, Chai Kiat Yeo, Bu Sung Lee, and Chiew Tong Lau. Autoencoder-based network anomaly detection. In *2018 Wireless telecommunications symposium (WTS)*, pages 1–5. IEEE, 2018.

[187] Shiyao Ma, Jiangtian Nie, Jiawen Kang, Lingjuan Lyu, Ryan Wen Liu, Ruihui Zhao, Ziyao Liu, and Dusit Niyato. Privacy-preserving anomaly detection in cloud manufacturing via federated transformer. *IEEE Transactions on Industrial Informatics*, 18(12):8977–8987, 2022.

[188] Min Du, Ruoxi Jia, and Dawn Song. Robust anomaly detection and backdoor attack detection via differential privacy. *arXiv preprint arXiv:1911.07116*, 2019.

[189] Jairo Giraldo, Alvaro Cardenas, Murat Kantarcioglu, and Jonathan Katz. Adversarial classification under differential privacy. In *Network and Distributed Systems Security (NDSS) Symposium 2020*, 2020.

[190] Federico Angelini, Jiawei Yan, and Syed Mohsen Naqvi. Privacy-preserving online human behaviour anomaly detection based on body movements and objects positions. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8444–8448. IEEE, 2019.

[191] Shagufta Mehnaz and Elisa Bertino. Privacy-preserving real-time anomaly detection using edge computing. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 469–480. IEEE, 2020.

[192] Abdulatif Alabdulatif, Ibrahim Khalil, Heshan Kumarage, Albert Y Zomaya, and Xun Yi. Privacy-preserving anomaly detection in the cloud for quality assured decision-making in smart cities. *Journal of Parallel and Distributed Computing*, 127:209–223, 2019.

[193] Ragav Sridharan, Rajib Ranjan Maiti, and Nils Ole Tippenhauer. Wadac: Privacy-preserving anomaly detection and attack classification on wireless traffic. In *Proceedings of the 11th ACM conference on security & privacy in wireless and mobile networks*, pages 51–62, 2018.

[194] Pei Zhang, Xiaohong Huang, Xin Sun, Hao Wang, and Yan Ma. Privacy-preserving anomaly detection across multi-domain networks. In *2012 9th International Conference on Fuzzy Systems and Knowledge Discovery*, pages 1066–1070. IEEE, 2012.

[195] Lingjuan Lyu, Yee Wei Law, Sarah M Erfani, Christopher Leckie, and Marimuthu Palaniswami. An improved scheme for privacy-preserving collaborative anomaly detection. In *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, pages 1–6. IEEE, 2016.

[196] Marwa Keshk, Elena Sitnikova, Nour Moustafa, Jiankun Hu, and Ibrahim Khalil. An integrated framework for privacy-preserving based anomaly detection for cyber-physical systems. *IEEE Transactions on Sustainable Computing*, 6(1):66–79, 2019.

[197] Rina Okada, Kazuto Fukuchi, and Jun Sakuma. Differentially private analysis of outliers. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part II 15*, pages 458–473. Springer, 2015.

[198] Sai Sree Laya Chukkapalli, Priyanka Ranade, Sudip Mittal, and Anupam Joshi. A privacy preserving anomaly detection framework for cooperative smart farming ecosystem. In *2021 Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, pages 340–347. IEEE, 2021.

[199] Jianting Guo, Peijia Zheng, and Jiwu Huang. Efficient privacy-preserving anomaly detection and localization in bitstream video. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(9):3268–3281, 2019.

[200] Shuhan Yuan and Xintao Wu. Trustworthy anomaly detection: A survey. *arXiv preprint arXiv:2202.07787*, 2022.

[201] Egawati Panjei, Le Gruenwald, Eleazar Leal, Christopher Nguyen, and Shejuti Silvia. A survey on outlier explanations. *The VLDB Journal*, 31(5):977–1008, 2022.

[202] Khushnaseeb Roshan and Aasim Zafar. Utilizing xai technique to improve autoencoder based model for computer network anomaly detection with shapley additive explanation (shap). *arXiv preprint arXiv:2112.08442*, 2021.

[203] Ambareesh Ravi, Xiaozhuo Yu, Iara Santelices, Fakhri Karray, and Baris Fidan. General frameworks for anomaly detection explainability: comparative study. In *2021 IEEE International Conference on Autonomous Systems (ICAS)*, pages 1–5. IEEE, 2021.

[204] Julian Tritscher, Anna Krause, and Andreas Hotho. Feature relevance xai in anomaly detection: Reviewing approaches and challenges. *Frontiers in Artificial Intelligence*, 6:1099521, 2023.

[205] Aso Bozorgpanah, Vicenç Torra, and Laya Aliahmadipour. Privacy and explainability: The effects of data protection on shapley values. *Technologies*, 10(6): 125, 2022.

[206] Fatima Ezzeddine, Omran Ayoub, Davide Andreoletti, Massimo Tornatore, and Silvia Giordano. Vertical split learning-based identification and explainable deep learning-based localization of failures in multi-domain nfv systems. In *2023 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*, pages 46–52. IEEE, 2023.

[207] Helena Montenegro, Wilson Silva, and Jaime S Cardoso. Privacy-preserving generative adversarial network for case-based explainability in medical image analysis. *IEEE Access*, 9:148037–148047, 2021.

[208] Thijs Veugen, Bart Kamphorst, and Michiel Marcus. Privacy-preserving contrastive explanations with local foil trees. *Cryptography*, 6(4):54, 2022.

[209] Dimitar Jetchev and Marius Vuille. Xorshap: Privacy-preserving explainable ai for decision tree models. *Cryptology ePrint Archive*, 2023.

[210] Fábio Manuel Neves de Araújo. *XAIPrivacy-XAI with Differential Privacy*. PhD thesis, Universidade do Porto (Portugal), 2023.

[211] Frederik Harder, Matthias Bauer, and Mijung Park. Interpretable and differentially private predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4083–4090, 2020.

[212] Kevin S Woods, Christopher C Doss, Kevin W Bowyer, Jeffrey L Solka, Carey E Priebe, and W Philip Kegelmeyer Jr. Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(06):1417–1436, 1993.

[213] Guansong Pang, Chunhua Shen, and Anton van den Hengel. Deep anomaly detection with deviation networks. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 353–362, 2019.

[214] Ettore Mariotti, Adarsa Sivaprasad, and Jose Maria Alonso Moral. Beyond prediction similarity: Shapgap for evaluating faithful surrogate models in xai. In *World Conference on Explainable Artificial Intelligence*, pages 160–173. Springer, 2023.

[215] Ettore Mariotti, Jose M Alonso-Moral, and Albert Gatt. Measuring model understandability by means of shapley additive explanations. In *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8. IEEE, 2022.

[216] Patrick Dunleavy and Helen Margetts. Data science, artificial intelligence and the third wave of digital era governance. *Public Policy and Administration*, 40 (2):185–214, 2025.

[217] Nicolaus Henke and London Jacques Bughin. The age of analytics: Competing in a data-driven world. 2016.

[218] Omar Y Al-Jarrah, Paul D Yoo, Sami Muhaidat, George K Karagiannidis, and Kamal Taha. Efficient machine learning for big data: A review. *Big Data Research*, 2(3):87–93, 2015.

[219] Eswar Prasad Galla, Chandrababu Kuraku, Hemanth Kumar Gollangi, Janardhana Rao Sunkara, and Chandrakanth Rao Madhavaram. *AI-DRIVEN DATA ENGINEERING TRANSFORMING BIG DATA INTO ACTIONABLE INSIGHT*. JEC PUBLICATION.

[220] Senthil Kumar Jagatheesaperumal, Mohamed Rahouti, Kashif Ahmad, Ala Al-Fuqaha, and Mohsen Guizani. The duo of artificial intelligence and big data for industry 4.0: Applications, techniques, challenges, and future research directions. *IEEE Internet of Things Journal*, 9(15):12861–12885, 2021.

[221] Marc Langheinrich. Privacy by design—principles of privacy-aware ubiquitous systems. In *International conference on ubiquitous computing*, pages 273–291. Springer, 2001.

[222] Eugenia Politou, Efthimios Alepis, and Constantinos Patsakis. Forgetting personal data and revoking consent under the gdpr: Challenges and proposed solutions. *Journal of cybersecurity*, 4(1):tyy001, 2018.

[223] Yuanyuan Feng, Yaxing Yao, and Norman Sadeh. A design space for privacy choices: Towards meaningful privacy control in the internet of things. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.

[224] Jay Pil Choi, Doh-Shin Jeon, and Byung-Cheol Kim. Privacy and personal data collection with information externalities. *Journal of Public Economics*, 173:113–124, 2019.

[225] Priyank Jain, Manasi Gyanchandani, and Nilay Khare. Big data privacy: a technological perspective and review. *Journal of Big Data*, 3:1–25, 2016.

[226] Office of Privacy and Civil Liberties, U.S. Department of Justice. Overview of the privacy act: 2020 edition. *https://www.justice.gov/opcl/overview-privacy-act-1974-2020-edition/introduction*, 2020.

[227] Republic of Korea. Basic act on artificial intelligence, 2023. URL https://likms.assembly.go.kr/bill/billDetail.do?billId=PRC_R2V4H1W1T2K5M1O6E4Q9T0V7Q9S0U0. Enacted by the National Assembly of South Korea.

[228] Montreal declaration for a responsible development of artificial intelligence, 2017. URL https://declarationmontreal-iaresponsable.com/. A declaration outlining ethical principles for AI development and deployment. Available at https://declarationmontreal-iaresponsable.com/.

[229] Tunde Toyese Oyedokun. Balancing innovation and privacy in the age of artificial intelligence. In *Digital Citizenship and the Future of AI Engagement, Ethics, and Privacy*, pages 343–376. IGI Global Scientific Publishing, 2025.

[230] Michele Loi and Markus Christen. Two concepts of group privacy. *Philosophy & Technology*, 33(2):207–224, 2020.

[231] Anuj Puri. A theory of group privacy. *Cornell JL & Pub. Pol'y*, 30:477, 2020.

[232] Linnet Taylor, Luciano Floridi, and Bart Van der Sloot. *Group privacy: New challenges of data technologies*, volume 126. Springer, 2016.

[233] Luciano Floridi. Group privacy: A defence and an interpretation. *Group privacy: New challenges of data technologies*, pages 83–100, 2017.

[234] Hadas Kotek, Rikker Dockum, and David Sun. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pages 12–24, 2023.

[235] Messi Lee, Jacob Montgomery, and Calvin Lai. The effect of group status on the variability of group representations in llm-generated text. In *Socially Responsible Language Modelling Research*, 2024.

[236] Samuel D Warren and Louis D Brandeis. The right to privacy. *Harv. L. Rev.*, 4: 193, 1890.

[237] William L. Prosser. Privacy. *California Law Review*, 48(3):383–423, 1960. doi: 10.2307/3478805.

[238] Paul Schwartz. The computer in german and american constitutional law: Towards an american right of informational self-determination. *The American Journal of Comparative Law*, 37(4):675–701, 1989.

[239] Helen Nissenbaum. Privacy as contextual integrity. *Wash. L. Rev.*, 79:119, 2004.

[240] Daniel J Solove and Chris Jay Hoofnagle. A model regime of privacy protection. *U. Ill. L. Rev.*, page 357, 2006.

[241] Adam D Moore. Defining privacy. *Journal of Social Philosophy*, 39(3):411–428, 2008.

[242] Republic of Korea. Personal information protection act. Act No. 10465 (promulgated Mar. 29, 2011) (S. Korea), as amended. URL https://elaw.klri.re.kr/eng_mobile/viewer.do?hseq=62389&key=4&type=part. English translation (Statutes of the Republic of Korea, KLRI e-Law). Addenda: entered into force six months after promulgation.

[243] Rachel L Finn, David Wright, and Michael Friedewald. Seven types of privacy. In *European data protection: coming of age*, pages 3–32. Springer, 2012.

[244] Alan F Westin. Special report: legal safeguards to insure privacy in a computer society. *Communications of the ACM*, 10(9):533–537, 1967.

[245] Péter Vörös and Attila Kiss. Openwebcrypt—securing our data in public cloud. *Modern Approaches for Intelligent Information and Database Systems*, pages 479–489, 2018.

[246] Edward J Bloustein and Nathaniel J Pallone. *Individual and group privacy*. Routledge, 2018.

[247] Xavier Caron, Rachelle Bosua, Sean B Maynard, and Atif Ahmad. The internet of things (iot) and its impact on individual privacy: An australian perspective. *Computer law & security review*, 32(1):4–15, 2016.

[248] Jordi Soria-Comas, Josep Domingo-Ferrer, David Sánchez, and David Megías. Individual differential privacy: A utility-preserving formulation of differential privacy guarantees. *IEEE Transactions on Information Forensics and Security*, 12 (6):1418–1429, 2017.

[249] Anna Cinzia Squicciarini, Mohamed Shehab, and Federica Paci. Collective privacy management in social networks. In *Proceedings of the 18th International Conference on World Wide Web (WWW 2009), Madrid, Spain, April 20–24, 2009*, pages 521–530. ACM, 2009. doi: 10.1145/1526709.1526780. URL https://doi.org/10.1145/1526709.1526780.

[250] Kurt Thomas, Chris Grier, and David M. Nicol. unfriendly: Multi-party privacy risks in social networks. In *Privacy Enhancing Technologies (PETS 2010)*, volume 6205 of *LNCS*, pages 236–252, 2010. doi: 10.1007/978-3-642-14527-8_14.

[251] Jose M. Such and Natalia Criado. Multiparty privacy in social media. *Communications of the ACM*, 2018. doi: 10.1145/3208039.

[252] Anna C. Squicciarini, Mohamed Shehab, and Joshua Wede. Privacy policies for shared content in social network sites. *The VLDB Journal*, 19:777–796, 2010. doi: 10.1007/s00778-010-0193-7. URL https://doi.org/10.1007/s00778-010-0193-7.

[253] Anna C. Squicciarini, Heng Xu, and Xiaolong (Luke) Zhang. Cope: Enabling collaborative privacy management in online social networks. *Journal of the Association for Information Science and Technology*, 62(3):521–534, 2011. doi: 10.1002/asi.21473.

[254] Gergely Biczók and Pern Hui Chia. Interdependent privacy: Let me share your data. In *Financial Cryptography and Data Security: 17th International Conference, FC 2013, Okinawa, Japan, April 1-5, 2013, Revised Selected Papers 17*, pages 338–353. Springer, 2013.

[255] Mathias Humbert, Benjamin Trubert, and Kévin Huguenin. A survey on interdependent privacy. *ACM Computing Surveys (CSUR)*, 52(6):1–40, 2019.

[256] Mathias Humbert. *When Others Impinge upon Your Privacy-Interdependent Risks and Protection in a Connected World.* PhD thesis, EPFL, Switzerland, 2015.

[257] Bernadette Kamleitner and Vince Mitchell. Your data is my data: a framework for addressing interdependent privacy infringements. *Journal of Public Policy & Marketing*, 38(4):433–450, 2019.

[258] Stefania Gnesi, Ilaria Matteucci, Corrado Moiso, Paolo Mori, Marinella Petrocchi, and Michele Vescovi. My data, your data, our data: managing privacy preferences in multiple subjects personal data. In *Annual Privacy Forum*, pages 154–171. Springer, 2014.

[259] Luciano Floridi. Open data, data protection, and group privacy. *Philosophy & Technology*, 27:1–3, 2014. doi: 10.1007/s13347-014-0157-8.

[260] Linnet Taylor, Luciano Floridi, and Bart van der Sloot, editors. *Group Privacy: New Challenges of Data Technologies*. Springer, 2017.

[261] Alice E. Marwick and danah boyd. Networked privacy: How teenagers negotiate context in social media. *New Media & Society*, 2014. doi: 10.1177/1461444814543995.

[262] Anuj Puri. The group right to mutual privacy. *Digital Society*, 2(2):22, 2023.

[263] Urbano Reviglio and Rogers Alunge. "i am datafied because we are datafied": An ubuntu perspective on (relational) privacy. *Philosophy & Technology*, 33(4): 595–612, 2020.

[264] Nan Zhang, Chong Wang, Elena Karahanna, and Yan Xu. Peer privacy concern: Conceptualization and measurement. *MIS Quarterly*, 46(1):491–530, 2022.

[265] Brent Mittelstadt. From individual to group privacy in big data analytics. *Philosophy & Technology*, 30(4):475–494, 2017.

[266] Linnet Taylor, Luciano Floridi, and Bart van der Sloot. Introduction: A new perspective on privacy. *Group privacy: New challenges of data technologies*, pages 1–12, 2017.

[267] Shalom H Schwartz. An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1):11, 2012.

[268] Alberto Melucci. The process of collective identity. In *Social movements and culture*, pages 41–63. Routledge, 2013.

[269] Saiful Singar, Ravinder Nagpal, Bahram H Arjmandi, and Neda S Akhavan. Personalized nutrition: tailoring dietary recommendations through genetic insights. *Nutrients*, 16(16):2673, 2024.

[270] Yusriani Yusriani, Dinda Sari, and Puput Mustika. Assessing the benefits and ethical risks of genomic medicine in personalized public health care. *Journal of Asian-african Focus in Health*, 2(1):10–17, 2024.

[271] Francisca Chibugo Udegbe, Ogochukwu Roseline Ebulue, Charles Chukwudalu Ebulue, and Chukwunonso Sylvester Ekesiobi. Precision medicine and genomics: A comprehensive review of it-enabled approaches. *International Medical Science Research Journal*, 4(4):509–520, 2024.

[272] Shaik Aminabee. The future of healthcare and patient-centric care: Digital innovations, trends, and predictions. In *Emerging Technologies for Health Literacy and Medical Practice*, pages 240–262. IGI Global Scientific Publishing, 2024.

[273] WeiKang Liu, Yanchun Zhang, Hong Yang, and Qinxue Meng. A survey on differential privacy for medical data analysis. *Annals of Data Science*, 11(2): 733–747, 2024.

[274] Mathias Humbert, Erman Ayday, Jean-Pierre Hubaux, and Amalio Telenti. Addressing the concerns of the lacks family: quantification of kin genomic privacy. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 1141–1152, 2013.

[275] Rosa Moreno, Alberto Gutierrez-Ramirez, Patelle Jivalagian, Stephanie Lopez, Matthew Deardorff, Xiaowu Gai, and Sabrina Derrington. P564: Parental perspectives on pediatric genomic testing, research, and data management in a multicultural population. *Genetics in Medicine Open*, 3, 2025.

[276] Mathias Humbert, Erman Ayday, Jean-Pierre Hubaux, and Amalio Telenti. Quantifying interdependent risks in genomic privacy. *ACM Transactions on Privacy and Security (TOPS)*, 20(1):1–31, 2017.

[277] Iman Deznabi, Mohammad Mobayen, Nazanin Jafari, Oznur Tastan, and Erman Ayday. An inference attack on genomic data using kinship, complex correlations, and phenotype information. *IEEE/ACM transactions on computational biology and bioinformatics*, 15(4):1333–1343, 2017.

[278] Erman Ayday and Mathias Humbert. Inference attacks against kin genomic privacy. *IEEE Security & Privacy*, 15(5):29–37, 2017.

[279] Amalio Telenti, Erman Ayday, and Jean Pierre Hubaux. On genomics, kin, and privacy. *F1000Research*, 3:80, 2014.

[280] Fatma Balci, Handan Kulan, Can Alkan, and Erman Ayday. A new inference attack against kin genomic privacy. *Proc. Privacy-Aware Computational Genomics*, 2015.

[281] Zaobo He and Junxiu Zhou. Inference attacks on genomic data based on probabilistic graphical models. *Big Data Mining and Analytics*, 3(3):225–233, 2020.

[282] Zhiyu Wan, Yevgeniy Vorobeychik, Weiyi Xia, Ellen Wright Clayton, Murat Kantarcioglu, and Bradley Malin. Expanding access to large-scale genomic data while promoting privacy: a game theoretic approach. *The American Journal of Human Genetics*, 100(2):316–322, 2017.

[283] Michael Backes, Pascal Berrang, Mathias Humbert, Xiaoyu Shen, and Verena Wolf. Simulating the large-scale erosion of genomic privacy over time. *IEEE/ACM transactions on computational biology and bioinformatics*, 15(5): 1405–1412, 2018.

[284] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An {End-to-End} case study of personalized warfarin dosing. In *23rd USENIX security symposium (USENIX Security 14)*, pages 17–32, 2014.

[285] Lawrence O Gostin. Genetic privacy. *Journal of Law, Medicine & Ethics*, 23(4): 320–330, 1995.

[286] Mathias Humbert, Didier Dupertuis, Mauro Cherubini, and Kévin Huguenin. Kgp meter: Communicating kin genomic privacy to the masses. In *2022 IEEE*

7th European Symposium on Security and Privacy (EuroS&P), pages 410–429. IEEE, 2022.

[287] Jim Wilson/The New York Times. Indian tribe wins fight to limit research of its dna, Apr 2010. URL https://www.nytimes.com/2010/04/22/us/22dna.html.

[288] Luca Pappalardo, Ed Manley, Vedran Sekara, and Laura Alessandretti. Future directions in human mobility science. *Nature computational science*, 3(7):588–600, 2023.

[289] John RB Palmer, Thomas J Espenshade, Frederic Bartumeus, Chang Y Chung, Necati Ercan Ozgencil, and Kathleen Li. New approaches to human mobility: Using mobile phones for demographic research. *Demography*, 50(3):1105–1128, 2013.

[290] Amaç Herdağdelen, Alex Dow, Bogdan State, Payman Mohassel, and Alex Pompe. Protecting privacy in facebook mobility data during the COVID-19 response, 2020. Available at: https://research.facebook.com/blog/2020/06/protecting-privacy-in-facebook-mobility-data-during-the-covid-19-response/.

[291] Shushu Liu, An Liu, Zheng Yan, and Wei Feng. Efficient lbs queries with mutual privacy preservation in iov. *Vehicular Communications*, 16:62–71, 2019.

[292] Ahmed Alshehri, Joseph Spielman, Amiya Prasad, and Chuan Yue. Exploring the privacy concerns of bystanders in smart homes from the perspectives of both owners and bystanders. *Proceedings on Privacy Enhancing Technologies*, 2022.

[293] Yuxi Wu, W Keith Edwards, and Sauvik Das. "a reasonable thing to ask for": Towards a unified voice in privacy collective action. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2022.

[294] The New York Times. Strava fitness app can reveal military sites, analysts say, January 2018. URL https://www.nytimes.com/2018/01/29/world/middleeast/strava-heat-map.html.

[295] Aleix Bassolas, Hugo Barbosa-Filho, Brian Dickinson, Xerxes Dotiwalla, Paul Eastham, Riccardo Gallotti, Gourab Ghoshal, Bryant Gipson, Surendra A Hazarie, Henry Kautz, et al. Hierarchical organization of urban mobility and its connection with city livability. *Nature communications*, 10(1):4817, 2019.

[296] Aleix Bassolas, Hugo Barbosa-Filho, Brian Dickinson, Xerxes Dotiwalla, Paul Eastham, Riccardo Gallotti, Gourab Ghoshal, Bryant Gipson, Surendra A Hazarie, Henry Kautz, et al. Reply to: On the difficulty of achieving differential privacy in practice: user-level guarantees in aggregate location data. *Nature Communications*, 13(1):30, 2022.

[297] Luca Luceri, Davide Andreoletti, Massimo Tornatore, Torsten Braun, and Silvia Giordano. Measurement and control of geo-location privacy on twitter. *Online social networks and media*, 17:100078, 2020.

[298] Alexandra-Mihaela Olteanu, Kévin Huguenin, Reza Shokri, Mathias Humbert, and Jean-Pierre Hubaux. Quantifying interdependent privacy risks with location data. *IEEE Transactions on Mobile Computing*, 16(3):829–842, 2016.

[299] Alexandra-Mihaela Olteanu, Mathias Humbert, Kévin Huguenin, and Jean-Pierre Hubaux. The (co-) location sharing game. *Proceedings on Privacy Enhancing Technologies*, 2019(2):5–25, 2019.

[300] Alexandra-Mihaela Olteanu, Kévin Huguenin, Italo Dacosta, and J-P Hubaux. Consensual and privacy-preserving sharing of multi-subject and interdependent data. In *Proceedings of the 25th network and distributed system security symposium (NDSS)*, pages 1–16. Internet Society, 2018.

[301] Michela Papandrea, Karim Keramat Jahromi, Matteo Zignani, Sabrina Gaito, Silvia Giordano, and Gian Paolo Rossi. On the properties of human mobility. *Computer Communications*, 87:19–36, 2016.

[302] Xun Zhou, Amin Vahedian Khezerlou, Alex Liu, Zubair Shafiq, and Fan Zhang. A traffic flow approach to early detection of gathering events. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 1–10, 2016.

[303] Andrey Sesyuk, Stelios Ioannou, and Marios Raspopoulos. A survey of 3d indoor localization systems and technologies. *Sensors*, 22(23):9380, 2022.

[304] Zdenek Chaloupka. Technology and standardization gaps for high accuracy positioning in 5g. *IEEE Communications Standards Magazine*, 1(1):59–65, 2017.

[305] José A del Peral-Rosado, Ronald Raulefs, José A López-Salcedo, and Gonzalo Seco-Granados. Survey of cellular mobile radio localization methods: From 1g to 5g. *IEEE Communications Surveys & Tutorials*, 20(2):1124–1148, 2017.

[306] Jean-Pierre Chanet, D Boffety, Géraldine André, T Humbert, P Rameau, A Amamra, Gil de Sousa, E Piron, KM Hou, F Vigier, et al. Technologies sans fil pour lacquisition de données de terrain. 2005.

[307] Bashima Islam, Md Tamzeed Islam, Jasleen Kaur, and Shahriar Nirjon. Lorain: Making a case for lora in indoor localization. In *2019 IEEE international conference on pervasive computing and communications workshops (PerCom workshops)*, pages 423–426. IEEE, 2019.

[308] Yizheng Wang, Zhimin Cao, Wei Gong, and Mingwu Chen. Survey of deep learning applications in wifi localization. In *2024 IEEE 7th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, volume 7, pages 517–522. IEEE, 2024.

[309] Chamara Sandeepa, Bartlomiej Siniarski, Nicolas Kourtellis, Shen Wang, and Madhusanka Liyanage. A survey on privacy for b5g/6g: New privacy challenges, and research directions. *Journal of Industrial Information Integration*, 30:100405, 2022. ISSN 2452-414X. doi: https://doi.org/10.1016/j.jii.2022.100405. URL https://www.sciencedirect.com/science/article/pii/S2452414X22000723.

[310] Faheem Zafari, Athanasios Gkelias, and Kin K Leung. A survey of indoor localization systems and technologies. *IEEE Communications Surveys & Tutorials*, 21 (3):2568–2599, 2019.

[311] Angelo Coluccia and Alessio Fascista. A review of advanced localization techniques for crowdsensing wireless sensor networks. *Sensors*, 19(5):988, 2019.

[312] Da Zhang, Feng Xia, Zhuo Yang, Lin Yao, and Wenhong Zhao. Localization technologies for indoor human tracking. In *2010 5th international conference on future information technology*, pages 1–6. IEEE, 2010.

[313] Zheng Yan, Xinren Qian, Shushu Liu, and Robert Deng. Privacy protection in 5g positioning and location-based services based on sgx. *ACM Transactions on Sensor Networks (TOSN)*, 18(3):1–19, 2022.

[314] Alberto Berenguer, David Fernández Ros, Andrea Gómez-Oliva, Josep A Ivars-Baidal, Antonio J Jara, Jaime Laborda, Jose-Norberto Mazón, and Angel Perles. Crowd monitoring in smart destinations based on gdpr-ready opportunistic rf scanning and classification of wifi devices to identify and classify visitors' origins. *Electronics*, 11(6):835, 2022.

[315] Mark Lindquist and Paul Galpern. Crowdsourcing (in) voluntary citizen geospatial data from google android smartphones. *J Digit Landsc Archit*, 1:263–272, 2016.

[316] Simone Fischer-Hübner and Leonardo A Martucci. Privacy in social collective intelligence systems. In *Social Collective Intelligence: Combining the Powers of Humans and Machines to Build a Smarter Society*, pages 105–124. Springer, 2014.

[317] Somayeh Moghaddam Zadeh Kashani and Javad Hamidzadeh. Feature selection by using privacy-preserving of recommendation systems based on collaborative filtering and mutual trust in social networks. *Soft Computing-A Fusion of Foundations, Methodologies & Applications*, 24(15), 2020.

[318] Aylin Caliskan Islam, Jonathan Walsh, and Rachel Greenstadt. Privacy detective: Detecting private information and collective privacy behavior in a large social network. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society*, pages 35–46, 2014.

[319] Xudong Pan, Mi Zhang, Shouling Ji, and Min Yang. Privacy risks of general-purpose language models. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 1314–1331. IEEE, 2020.

[320] Iraklis Symeonidis, Gergely Biczók, Fatemeh Shirazi, Cristina Pérez-Solà, Jessica Schroers, and Bart Preneel. Collateral damage of facebook third-party applications: a comprehensive study. *Computers & Security*, 77:179–208, 2018.

[321] Chutikulrungsee Tharntip Tawnie and Burmeister Oliver Kisalay. Interdependent privacy. *The ORBIT Journal*, 1(2):1–14, 2017.

[322] Shuaishuai Liu and Gergely Biczók. Idpfilter: Mitigating interdependent privacy issues in third-party apps. *Computers & Security*, 151:104321, 2025.

[323] Shirin Nilizadeh, Naveed Alam, Nathaniel Husted, and Apu Kapadia. Pythia: a privacy aware, peer-to-peer network for social search. In *Proceedings of the 10th annual ACM workshop on Privacy in the electronic society*, pages 43–48, 2011.

[324] Airi Lampinen, Vilma Lehtinen, Asko Lehmuskallio, and Sakari Tamminen. We're in it together: interpersonal management of disclosure in social network services. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 3217–3226, 2011.

[325] Jin Chen, Jerry Wenjie Ping, Yunjie Xu, and Bernard CY Tan. Information privacy concern about peer disclosure in online social networks. *IEEE Transactions on Engineering Management*, 62(3):311–324, 2015.

[326] Ricarda Moll, Stephanie Pieschl, and Rainer Bromme. Whoever will read it–the overload heuristic in collective privacy expectations. *Computers in Human Behavior*, 75:484–493, 2017.

[327] Sabine Trepte. The social media privacy model: Privacy and communication in the light of social media affordances. *Communication Theory*, 31(4):549–570, 2021.

[328] Yafei Feng, Yongqiang Sun, Nan Wang, and Xiao-Liang Shen. Co-owned information disclosure and collective privacy calculus on social network platforms: the moderating role of information ownership. *Internet Research*, 35(3):1154–1180, 2025.

[329] Patricia Sánchez Abril, Avner Levin, and Alissa Del Riego. Blurred boundaries: Social media privacy and the twenty-first-century employee. *Am. Bus. LJ*, 49: 63, 2012.

[330] Zina Chkirbene, Ridha Hamila, Ala Gouissem, and Unal Devrim. Large language models (llm) in industry: A survey of applications, challenges, and trends. In *2024 IEEE 21st International Conference on Smart Communities: Improving Quality of Life using AI, Robotics and IoT (HONET)*, pages 229–234. IEEE, 2024.

[331] Yuxin Hou and Junming Huang. Natural language processing for social science research: A comprehensive review. *Chinese Journal of Sociology*, 11(1):121–157, 2025.

[332] François Théberge. Survey of generative methods for social media analysis. 2022.

[333] Sarah Jenner, Dimitris Raidos, Emma Anderson, Stella Fleetwood, Ben Ainsworth, Kerry Fox, Jana Kreppner, and Mary Barker. Using large language models for narrative analysis: a novel application of generative ai. *Methods in Psychology*, 12:100183, 2025.

[334] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *CoRR*, 2024.

[335] Noé Zufferey, Sarah Abdelwahab Gaballah, Karola Marky, and Verena Zimmermann. "ai is from the devil." behaviors and concerns toward personal data sharing with llm-based conversational agents. *Proceedings on Privacy Enhancing Technologies*, 2025(3):5–28, 2025.

[336] Lisa Mekioussa Malki et al. "hoovered up as a data point": Exploring privacy behaviours, awareness, and concerns among uk users of llm-based conversational agents. In *Proceedings on Privacy Enhancing Technologies*. ACM, 2025.

[337] Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models. In *International Conference on Learning Representations 2024*, 2024.

[338] Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Surabhi Adhikari, Hariram Veeramani, Mehwish Nasim, and Usman Naseem. Large language models (llm) in computational social science: prospects, current state, and challenges. *Social Network Analysis and Mining*, 15(1):1–30, 2025.

[339] Som Sekhar Bhattacharyya. Study of adoption of artificial intelligence technology-driven natural large language model-based chatbots by firms for customer service interaction. *Journal of Science and Technology Policy Management*, 2024.

[340] Wenjie Qu, Yuguang Zhou, Yongji Wu, Tingsong Xiao, Binhang Yuan, Yiming Li, and Jiaheng Zhang. Prompt inversion attack against collaborative inference of large language models. In *2025 IEEE Symposium on Security and Privacy (SP)*, pages 1695–1712. IEEE, 2025.

[341] Qizhang Feng, Siva Rajesh Kasa, SANTHOSH KUMAR KASA, Hyokun Yun, Choon Hui Teo, and Sravan Babu Bodapati. Exposing privacy gaps: Membership inference attack on preference data for llm alignment. In *International Conference on Artificial Intelligence and Statistics*, pages 5221–5229. PMLR, 2025.

[342] Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. Membership inference attacks against fine-tuned large language models via self-prompt calibration. *Advances in Neural Information Processing Systems*, 37: 134981–135010, 2024.

[343] Rui Wen, Zheng Li, Michael Backes, and Yang Zhang. Membership inference attacks against in-context learning. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 3481–3495, 2024.

[344] Syed Arsalan Ahmed Naqvi, Umair Ayub, Muhammad Ali Khan, Peter R Khoury, Deepak B Ravichandar, Owen H Witte, Daniel S Childs, Jacob Orme, Yousef Zakharia, Parminder Singh, et al. Large language models (llms) for inferring genomic characteristics and facilitating genomic literacy in prostate cancer (pca) patients., 2025.

[345] Ning Liao, Cheukfai Li, William J Gradishar, V Suzanne Klimberg, Joshua A Roshal, Taize Yuan, Sanjiv S Agarwala, Vincente K Valero, Sandra M Swain, Julie A Margenthaler, et al. Accuracy and reproducibility of chatgpt responses to breast cancer tumor board patients. *JCO clinical cancer informatics*, 9:e2500001, 2025.

[346] Jordan N Keels, Joanne Thomas, Kathleen A Calzone, Laurie Badzek, Sarah Dewell, Vinaya Murthy, Rosie O'Shea, Emma T Tonkin, and Andrew A Dwyer. Consumer-oriented (patient and family) outcomes from nursing in genomics: a scoping review of the literature (2012–2022). *Frontiers in Genetics*, 15:1481948, 2024.

[347] Jessica Mozersky and Galen Joseph. Case studies in the co-production of populations and genetics: The making of 'at risk populations' in brca genetics. *BioSocieties*, 5(4):415–439, 2010.

[348] Yangsibo Huang, Samyak Gupta, Zexuan Zhong, Kai Li, and Danqi Chen. Privacy implications of retrieval-based language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14887–14902, 2023.

[349] Marco Siino, Mariana Falco, Daniele Croce, and Paolo Rosso. Exploring llms applications in law: A literature review on current legal nlp approaches. *IEEE Access*, 2025.

[350] Ahmed Izzidien, Holli Sargeant, and Felix Steffek. Llm vs. lawyers: Identifying a subset of summary judgments in a large uk case law dataset. *CoRR*, 2024.

[351] Angelina Wang, Jamie Morgenstern, and John P Dickerson. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, pages 1–12, 2025.

[352] Cedric Waterschoot, Nava Tintarev, and Francesco Barile. The pitfalls of growing group complexity: Llms and social choice-based aggregation for group recommendations. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, pages 322–330, 2025.

[353] Zachary W Petzel and Leanne Sowerby. Prejudiced interactions with large language models (llms) reduce trustworthiness and behavioral intentions among members of stigmatized groups. *Computers in Human Behavior*, 165:108563, 2025.

[354] Mahammed Kamruzzaman and Gene Louis Kim. Exploring changes in nation perception with nationality-assigned personas in llms. *CoRR*, 2024.

[355] Shera Potka. *Three Ethical Dimensions of AI: Fairness in Social Recommenders, Bias Detection in LLMs, and Privacy in NLP*. PhD thesis, University of Victoria, 2025.

[356] Jayanta Sadhu, Maneesha Rani Saha, and Rifat Shahriyar. Social bias in large language models for bangla: An empirical study on gender and religious bias. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 204–218, 2025.

[357] Lior Rokach and Oded Maimon. Clustering methods. *Data mining and knowledge discovery handbook*, pages 321–352, 2005.

[358] Sotiris B Kotsiantis, Ioannis Zaharakis, P Pintelas, et al. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1):3–24, 2007.

[359] Roger Brownsword. Knowing me, knowing you–profiling, privacy and the public interest. In *Profiling the European Citizen: Cross-Disciplinary Perspectives*, pages 345–363. Springer, 2008.

[360] Paola Mavriki and Maria Karyda. Automated data-driven profiling: threats for group privacy. *Information & Computer Security*, 28(2):183–197, 2019.

[361] Bart Custers. Data mining with discrimination sensitive and privacy sensitive attributes. *Custers BHM (2010), Data Mining with Discrimination Sensitive and Privacy Sensitive Attributes. In: Proceedings of ISP*, pages 31–38, 2010.

[362] Louis-Philippe Lemieux Perreault, Marc-André Legault, Amina Barhdadi, Sylvie Provost, Valérie Normand, Jean-Claude Tardif, and Marie-Pierre Dubé. Comparison of genotype clustering tools with rare variants. *BMC bioinformatics*, 15: 1–12, 2014.

[363] Anuj Puri. *The group right to privacy*. PhD thesis, University of St Andrews, 2022.

[364] AI INCIDENT DATABASE. Korean chatbot luda made offensive remarks towards minority groups, Apr 2020. URL https://incidentdatabase.ai/cite/106/.

[365] AI INCIDENT DATABASE. Government deployed extreme surveillance technologies to monitor and target muslim minorities in xinjiang, Apr 2020. URL https://incidentdatabase.ai/cite/249/.

[366] Google. Federated learning of cohorts (floc). https://privacysandbox.com/intl/en_us/proposals/floc/, 2021. Accessed: 2025-04-07.

[367] The Economist. Why is FLoC, google's new ad technology, taking flak? https://www.economist.com/the-economist-explains/2021/05/17/why-is-floc-googles-new-ad-technology-taking-flak, 2021. Accessed: 7 April 2025.

[368] Google. Topics API. https://developers.google.com/privacy-sandbox/private-advertising/topics, 2023. Accessed: 7 April 2025. Published by The Privacy Sandbox.

[369] Luciano Floridi, Linnet Taylor, and Bart van der Sloot. Group privacy: new challenges of data technologies. 2017.

[370] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05):557–570, 2002.

[371] Abdul Majeed, Safiullah Khan, and Seong Oun Hwang. Group privacy: An underrated but worth studying research problem in the era of artificial intelligence and big data. *Electronics*, 11(9):1449, 2022.

[372] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE, 2008.

[373] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *Acm transactions on knowledge discovery from data (tkdd)*, 1(1):3–es, 2007.

[374] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd international conference on data engineering*, pages 106–115. IEEE, 2006.

[375] Solomon Messing, Christina DeGregorio, Bennett Hillenbrand, Gary King, Saurav Mahanti, Zagreb Mukerjee, Chaya Nayak, Nate Persily, Bogdan State, and Arjun Wilkins. Facebook privacy-protected full URLs data set. *Version DRAFT VERSION*, 2020.

[376] Apple Differential Privacy Team. Learning with privacy at scale. https://docs-assets.developer.apple.com/ml-research/papers/learning-with-privacy-at-scale.pdf, 2017. Accessed: 7 April 2025.

[377] Adam Boulanger, Akim Kumok, Arti Patankar, Badih Ghazi, Benjamin Miller, Chaitanya Kamath, Charlotte Stanton, Chris Scott, Damien Desfontaines, Dennis Kraft, et al. Google covid-19 vaccination search insights: Anonymization process description. *research.google*, 2021.

[378] Shailesh Bavadekar, Andrew Dai, John Davis, Damien Desfontaines, Ilya Eckstein, Katie Everett, Alex Fabrikant, Gerardo Flores, Evgeniy Gabrilovich, Krishna Gadepalli, et al. Google COVID-19 search trends symptoms dataset: Anonymization process description (version 1.0). *arXiv preprint arXiv:2009.01265*, 2020.

[379] Bing Zhang, Vadym Doroshenko, Peter Kairouz, Thomas Steinke, Abhradeep Thakurta, Ziyin Ma, Eidan Cohen, Himani Apte, and Jodi Spacek. Differentially private stream processing at scale. *Proceedings of the VLDB Endowment*, 17(12): 4145–4158, 2024.

[380] Ryan Rogers, Subbu Subramaniam, Sean Peng, David Durfee, Seunghyun Lee, Santosh Kumar Kancha, Shraddha Sahay, and Parvez Ahammad. LinkedIn's Audience Engagements API: A privacy preserving data analytics system at scale. *arXiv preprint arXiv:2002.05839*, 2020.

[381] Ryan Rogers, Adrian Rivera Cardoso, Koray Mancuhan, Akash Kaura, Nikhil Gahlawat, Neha Jain, Paul Ko, and Parvez Ahammad. A members first approach to enabling LinkedIn's labor market insights at scale. *arXiv preprint arXiv:2010.13981*, 2020.

[382] Andrew David Foote, Ashwin Machanavajjhala, and Kevin McKinney. Releasing earnings distributions using differential privacy: Disclosure avoidance system for post-secondary employment outcomes (pseo). *Journal of Privacy and Confidentiality*, 9(2), 2019.

[383] Nour Almadhoun, Erman Ayday, and Özgür Ulusoy. Differential privacy under dependent tuples—the case of genomic privacy. *Bioinformatics*, 36(6):1696–1703, 2020.

[384] Decheng Liu, Zongqi Wang, Chunlei Peng, Nannan Wang, Ruimin Hu, and Xinbo Gao. Thinking racial bias in fair forgery detection: Models, datasets and evaluations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5379–5387, 2025.

[385] Alessa Angerschmid, Jianlong Zhou, Kevin Theuermann, Fang Chen, and Andreas Holzinger. Fairness and explanation in ai-informed decision making. *Machine Learning and Knowledge Extraction*, 4(2):556–579, 2022.

[386] Emma AM Stanley, Matthias Wilms, Pauline Mouches, and Nils D Forkert. Fairness-related performance and explainability effects in deep learning models for brain image analysis. *Journal of Medical Imaging*, 9(6):061102–061102, 2022.

[387] Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, and Andrew H Sung. Deepfake detection: A systematic literature review. *IEEE access*, 10:25494–25513, 2022.

[388] Asad Malik, Minoru Kuribayashi, Sani M Abdullahi, and Ahmad Neyaz Khan. Deepfake detection for human face images and videos: A survey. *Ieee Access*, 10: 18757–18775, 2022.

[389] Shichao Dong, Jin Wang, Jiajun Liang, Haoqiang Fan, and Renhe Ji. Explaining deepfake detection by analysing image matching. In *European conference on computer vision*, pages 18–35. Springer, 2022.

[390] Ali Raza, Kashif Munir, and Mubarak Almutairi. A novel deep learning approach for deepfake image detection. *Applied Sciences*, 12:9820, 2022.

[391] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Deepfake detection by analyzing convolutional traces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020.

[392] Harsh Agarwal, Ankur Singh, and D Rajeswari. Deepfake detection using svm. In *2021 Second Int. Conference on Electronics and Sustainable Communication Systems (ICESC)*. IEEE, 2021.

[393] Agil Aghasanli, Dmitry Kangin, and Plamen Angelov. Interpretable-through-prototypes deepfake detection for diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2023.

[394] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip Yu, and Lichao Sun. A survey of ai-generated content (aigc). *ACM Computing Surveys*, 57(5): 1–38, 2025.

[395] Bobby Chesney and Danielle Citron. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.*, 107:1753, 2019.

[396] Amanda Margaret Narvali, Joshua August (Gus) Skorburg, and Maya J. Goldenberg. Cyberbullying girls with pornographic deepfakes is a form of misogyny. https://theconversation.com/cyberbullying-girls-with-pornographic-deepfakes-is-a-form-of-misogyny-217182. Accessed: 2024-11-28.

[397] Amber Davisson. Eros and ethos in celebrity deepfake pornography. In *Ethos, Technology, and AI in Contemporary Society*, pages 186–204. Routledge, 2024.

[398] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[399] Walter Matli. Extending the theory of information poverty to deepfake technology. *International Journal of Information Management Data Insights*, 4(2): 100286, 2024.

[400] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.

[401] Marissa Koopman, Andrea Macarulla Rodriguez, and Zeno Geradts. Detection of deepfake video manipulation. In *The 20th Irish machine vision and image processing conference (IMVIP)*, pages 133–136, 2018.

[402] Teun Baar, Wiger van Houten, and Zeno Geradts. Camera identification by grouping images from database, based on shared noise patterns. *arXiv preprint arXiv:1207.2641*, 2012.

[403] Bernard L Welch. The generalization of 'student's'problem when several different population varlances are involved. *Biometrika*, 34(1-2):28–35, 1947.

[404] Sakshi Agarwal and Lav R Varshney. Limits of deepfake detection: A robust estimation viewpoint. *arXiv preprint arXiv:1905.03493*, 2019.

[405] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)*, pages 1–6. IEEE, 2019.

[406] MA Sahla Habeeba, A Lijiya, and Anu Mary Chacko. Detection of deepfakes using visual artifacts and neural network classifier. In *Innovations in Electrical and Electronic Engineering: Proceedings of ICEEE 2020*, pages 411–422. Springer, 2021.

[407] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

[408] Disheng Feng, Xuequan Lu, and Xufeng Lin. Deep detection for face manipulation. In *Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 18–22, 2020, Proceedings, Part V 27*, pages 316–323. Springer, 2020.

[409] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*. IEEE, 2018.

[410] Yuval Nirkin, Lior Wolf, Yosi Keller, and Tal Hassner. Deepfake detection based on the discrepancy between the face and its context. *arXiv preprint arXiv:2008.12262*, 2020.

[411] Hua Qi, Qing Guo, Felix Juefei-Xu, Xiaofei Xie, Lei Ma, Wei Feng, Yang Liu, and Jianjun Zhao. Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms. In *Proceedings of the 28th ACM international conference on multimedia*, pages 4318–4327, 2020.

[412] Javier Hernandez-Ortega, Ruben Tolosana, Julian Fierrez, and Aythami Morales. Deepfakeson-phys: Deepfakes detection based on heart rate estimation. *arXiv preprint arXiv:2010.00400*, 2020.

[413] Mengnan Du, Shiva Pentyala, Yuening Li, and Xia Hu. Towards generalizable deepfake detection with locality-aware autoencoder. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 325–334, 2020.

[414] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 83–92. IEEE, 2019.

[415] Sara Concas, Gianpaolo Perelli, Gian Luca Marcialis, and Giovanni Puglisi. Tensor-based deepfake detection in scaled and compressed images. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3121–3125. IEEE, 2022.

[416] Muhammad Asad Arshed, Ayed Alwadain, Rao Faizan Ali, Shahzad Mumtaz, Muhammad Ibrahim, and Amgad Muneer. Unmasking deception: Empowering deepfake detection with vision transformer network. *Mathematics*, 11(17): 3710, 2023.

[417] Muxin Pu, Meng Yi Kuan, Nyee Thoang Lim, Chun Yong Chong, and Mei Kuan Lim. Fairness evaluation in deepfake detection models using metamorphic testing. In *Proceedings of the 7th international workshop on metamorphic testing*, pages 7–14, 2022.

[418] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[419] Loc Trinh and Yan Liu. An examination of fairness of ai models for deepfake detection. *arXiv preprint arXiv:2105.00558*, 2021.

[420] Laura Stroebel, Mark Llewellyn, Tricia Hartley, Tsui Shan Ip, and Mohiuddin Ahmed. A systematic literature review on the effectiveness of deepfake detection techniques. *Journal of Cyber Security Technology*, 7(2):83–113, 2023.

[421] Akshay Agarwal and Nalini Ratha. Deepfake: Classifiers, fairness, and demographically robust algorithm. In *2024 IEEE 18th Int. Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–9. IEEE, 2024.

[422] Uzoamaka Ezeakunne, Chrisantus Eze, and Xiuwen Liu. Deepfake detection, image manipulation detection, fairness, generalization. *arXiv preprint arXiv:2412.16428*, 2024.

[423] Yan Ju, Shu Hu, Shan Jia, George H Chen, and Siwei Lyu. Improving fairness in deepfake detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4655–4665, 2024.

[424] Li Lin, Xinan He, Yan Ju, Xin Wang, Feng Ding, and Shu Hu. Preserving fairness generalization in deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16815–16825, 2024.

[425] Xiaotian Han, Jianfeng Chi, Yu Chen, Qifan Wang, Han Zhao, Na Zou, and Xia Hu. Ffb: A fair fairness benchmark for in-processing group fairness methods. *arXiv preprint arXiv:2306.09468*, 2023.

[426] Decheng Liu, Zongqi Wang, Chunlei Peng, Nannan Wang, Ruimin Hu, and Xinbo Gao. Thinking racial bias in fair forgery detection: Models, datasets and evaluations. *arXiv preprint arXiv:2407.14367*, 2024.

[427] Manjil Karki. URL https://www.kaggle.com/datasets/manjilkarki/deepfake-and-real-images.

[428] Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *2021 Int. Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4. IEEE, 2021.

[429] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.

[430] André Artelt and Barbara Hammer. "explain it in the same way!"–model-agnostic group fairness of counterfactual explanations. *arXiv preprint arXiv:2211.14858*, 2022.

[431] Sina Shaham, Arash Hajisafi, Minh K Quan, Dinh C Nguyen, Bhaskar Krishnamachari, Charith Peris, Gabriel Ghinita, Cyrus Shahabi, and Pubudu N Pathirana. Privacy and fairness in machine learning: A survey. *IEEE Transactions on Artificial Intelligence*, 2025.

[432] Yuying Zhao, Yu Wang, and Tyler Derr. Fairness and explainability: Bridging the gap towards fair model explanations, 2022. URL https://arxiv.org/abs/2212.03840.

[433] James R. Foulds, Rashidul Islam, Kamrun Keya, and Shimei Pan. Differential fairness. 2019. URL https://api.semanticscholar.org/CorpusID:208538368.

[434] Sarah Bird, Miroslav Dudík, R Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna M. Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. 2020. URL https://api.semanticscholar.org/CorpusID:219772478.

[435] André Artelt, Valerie Vaquet, Riza Velioglu, Fabian Hinder, Johannes Brinkrolf, Malte Schilling, and Barbara Hammer. Evaluating robustness of counterfactual explanations. In *2021 IEEE symposium series on computational intelligence (SSCI)*, pages 01–09. IEEE, 2021.

[436] Dylan Slack, Anna Hilgard, Himabindu Lakkaraju, and Sameer Singh. Counterfactual explanations can be manipulated. *Advances in neural information processing systems*, 34:62–75, 2021.

[437] Alejandro Kuratomi, Evaggelia Pitoura, Panagiotis Papapetrou, Tony Lindgren, and Panayiotis Tsaparas. Measuring the burden of (un) fairness using counterfactuals. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 402–417. Springer, 2022.

[438] Vivek Gupta, Pegah Nokhiz, Chitradeep Dutta Roy, and Suresh Venkatasubramanian. Equalizing recourse across groups. *ArXiv*, abs/1909.03166, 2019. URL https://api.semanticscholar.org/CorpusID:202537227.

[439] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, Juntao Tan, Shuchang Liu, and Yongfeng Zhang. Fairness in recommendation: Foundations, methods, and applications. *ACM Transactions on Intelligent Systems and Technology*, 14 (5):1–48, 2023.

[440] Lucius E.J. Bynum, Joshua R. Loftus, and Julia Stoyanovich. A new paradigm for counterfactual reasoning in fairness and recourse. In *International Joint Conference on Artificial Intelligence*, 2024. URL https://api.semanticscholar.org/CorpusID:267212116.

[441] Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. Fairness in recommender systems: research landscape and future directions. *User Modeling and User-Adapted Interaction*, 34(1):59–108, 2024.

[442] Shizhou Xu and Thomas Strohmer. On the (in) compatibility between group fairness and individual fairness. *arXiv preprint arXiv:2401.07174*, 2024.

[443] Nicholas Asher, Soumya Paul, and Chris Russell. Fair and adequate explanations. In *Machine Learning and Knowledge Extraction: 5th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2021, Virtual Event, August 17–20, 2021, Proceedings 5*, pages 79–97. Springer, 2021.

[444] Julius Von Kügelgen, Amir-Hossein Karimi, Umang Bhatt, Isabel Valera, Adrian Weller, and Bernhard Schölkopf. On the fairness of causal algorithmic recourse. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 9584–9594, 2022.

[445] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

[446] Nicholas Asher, Lucas De Lara, Soumya Paul, and Chris Russell. Counterfactual models for fair and adequate explanations. *Machine Learning and Knowledge Extraction*, 4(2):316–349, 2022.

[447] Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. When worlds collide: integrating different counterfactual assumptions in fairness. *Advances in neural information processing systems*, 30, 2017.

[448] Kaivalya Rawal and Himabindu Lakkaraju. Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses. *Advances in Neural Information Processing Systems*, 33:12187–12198, 2020.

[449] Sungwon Han, Seungeon Lee, Fangzhao Wu, Sundong Kim, Chuhan Wu, Xiting Wang, Xing Xie, and Meeyoung Cha. Dualfair: fair representation learning at both group and individual levels via contrastive self-supervision. In *Proceedings of the ACM Web Conference 2023*, pages 3766–3774, 2023.

[450] André Artelt and Andreas Gregoriades. A two-stage algorithm for cost-efficient multi-instance counterfactual explanations. In *xAI (Late-breaking Work, Demos, Doctoral Consortium)*, 2024.

[451] Martin Pawelczyk, Chirag Agarwal, Shalmali Joshi, Sohini Upadhyay, and Himabindu Lakkaraju. Exploring counterfactual explanations through the lens of adversarial examples: A theoretical and empirical analysis. In *International Conference on Artificial Intelligence and Statistics*, pages 4574–4594. PMLR, 2022.

[452] Seung Hyun Cheon, Anneke Wernerfelt, Sorelle Friedler, and Berk Ustun. Feature responsiveness scores: Model-agnostic explanations for recourse. In *The Thirteenth International Conference on Learning Representations*.

[453] Sofie Goethals, David Martens, and Toon Calders. Precof: counterfactual explanations for fairness. *Machine Learning*, 113(5):3111–3142, 2024.

[454] Giandomenico Cornacchia, Vito Walter Anelli, Giovanni Maria Biancofiore, Fedelucio Narducci, Claudio Pomo, Azzurra Ragone, and Eugenio Di Sciascio. Auditing fairness under unawareness through counterfactual reasoning. *Information Processing & Management*, 60(2):103224, 2023.

[455] Zichong Wang, Zhibo Chu, Ronald Blanco, Zhong Chen, Shu-Ching Chen, and Wenbin Zhang. Advancing graph counterfactual fairness through fair representation learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 40–58. Springer, 2024.

[456] Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226, 2019.

[457] Jing Ma, Ruocheng Guo, Mengting Wan, Longqi Yang, Aidong Zhang, and Jundong Li. Learning fair node representations with graph counterfactual fairness. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 695–703, 2022.

[458] Amanda Coston, Alan Mishler, Edward H Kennedy, and Alexandra Chouldechova. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 582–593, 2020.

[459] Arnaud Van Looveren and Janis Klaise. Interpretable counterfactual explanations guided by prototypes. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 650–665. Springer, 2021.

[460] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[461] J MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press*, 1967.

[462] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

[463] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(3):e1452, 2022.