SCIENCE CHINA
Information Sciences

• REVIEW •

January 2025, Vol. 68, Iss. 1, 111101:1–111101:34
https://doi.org/10.1007/s11432-024-4123-4

Privacy-preserving explainable AI: a survey

Thanh Tam NGUYEN[1], Thanh Trung HUYNH[2], Zhao REN[3], Thanh Toan NGUYEN[4*], Phi Le NGUYEN[5], Hongzhi YIN[6*] & Quoc Viet Hung NGUYEN[1*]

[1]*School of Information and Communication Technology, Griffith University, Gold Coast QLD 4215, Australia;*
[2]*School of Computer and Communication Sciences, Ecole Polytechnique Federale de Lausanne, Lausanne 1015, Switzerland;*
[3]*Faculty of Mathematics and Computer Science, University of Bremen, Bremen 28359, Germany;*
[4]*Faculty of Information Technology, HUTECH University, Ho Chi Minh City 70000, Vietnam;*
[5]*Department of Computer Science, Hanoi University of Science and Technology, Hanoi 10000, Vietnam;*
[6]*School of Electrical Engineering and Computer Science, The University of Queensland, Brisbane QLD 4072, Australia*

Received 4 April 2024/Revised 26 June 2024/Accepted 7 August 2024/Published online 7 November 2024

Abstract As the adoption of explainable AI (XAI) continues to expand, the urgency to address its privacy implications intensifies. Despite a growing corpus of research in AI privacy and explainability, there is little attention on privacy-preserving model explanations. This article presents the first thorough survey about privacy attacks on model explanations and their countermeasures. Our contribution to this field comprises a thorough analysis of research papers with a connected taxonomy that facilitates the categorization of privacy attacks and countermeasures based on the targeted explanations. This work also includes an initial investigation into the causes of privacy leaks. Finally, we discuss unresolved issues and prospective research directions uncovered in our analysis. This survey aims to be a valuable resource for the research community and offers clear insights for those new to this domain. To support ongoing research, we have established an online resource repository, which will be continuously updated with new and relevant findings.

Keywords privacy-preserving explainable AI, privacy attacks, privacy defences, PrivEx, PPXAI

Citation Nguyen T T, Huynh T T, Ren Z, et al. Privacy-preserving explainable AI: a survey. Sci China Inf Sci, 2025, 68(1): 111101, https://doi.org/10.1007/s11432-024-4123-4

1 Introduction

In recent years, the push for automated model explanations has gained significant momentum, with key guidelines like the GDPR highlighting their importance [1], and tech giants such as Google, Microsoft, and IBM pioneering this initiative by integrating explanation toolkits into their machine learning solutions [2]. This movement toward transparency encompasses a variety of explanation types, from global and local explanations that offer broad overviews and specific decision rationales, respectively, to feature importance analysis that pinpoints the impact of individual data inputs [3]. Techniques like SHAP and LIME provide nuanced insights into feature contributions [4, 5], while counterfactual explanations explore how changes in input could lead to different outcomes [6]. Additionally, interactive visualization tools are becoming increasingly popular, making the interpretation of complex models more accessible to users [7–9].

However, this pursuit of transparency is not without its risks, especially privacy. The very act of providing explanations involves the disclosure of information that, while intended to illuminate, also carries the risk of inadvertently revealing sensitive details embedded in the models' training data. The balance between transparency and privacy becomes even more precarious when considering the granularity of explanations. Detailed explanations, although more informative, might offer direct inferences about individual data points used in training, thereby increasing the risk of privacy breaches. This paradox underscores a significant challenge within the field, as highlighted by recent research [2, 10, 11], which delves into the privacy implications of model explanations.

The degree to which model explanations reveal specifics about users' data is not fully understood. The unintended disclosure of sensitive details, such as a person's location, health records, or identity, through these explanations could pose serious concerns if such information were to be deciphered by a malicious

* Corresponding author (email: nt.toan@hutech.edu.vn, h.yin1@ug.edu.au, henry.nguyen@griffith.edu.au)

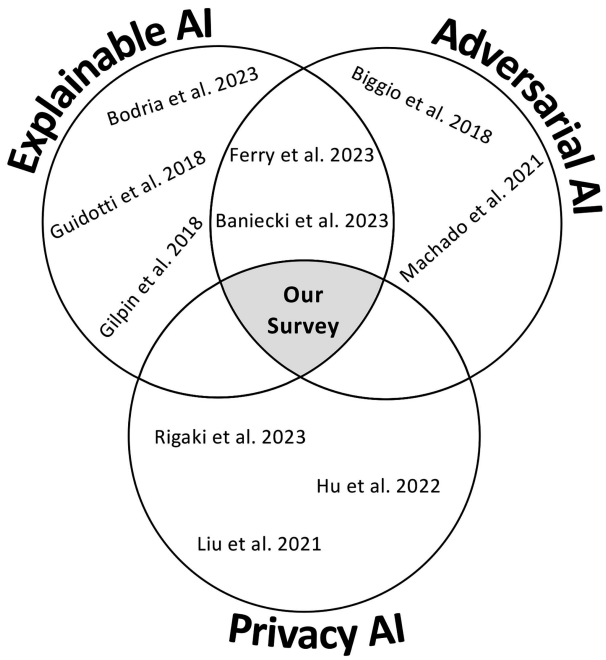

Figure 1 Existing surveys. Explainable AI involves model explanations (e.g., [7–9]). Adversarial AI includes adversarial attacks on ML models (e.g., [16, 17]). Privacy AI involves privacy issues in ML (e.g., [18–20]). Others [11, 21] discuss exploits on model explanations. Ours offers the first complete picture on privacy risks, attacks, and defenses on model explanations.

entity [12]. On the flip side, if private data are used without the rightful owner's permission, the same techniques aimed at exposing information could also detect unauthorized data utilization, thus potentially safeguarding user privacy [13]. Furthermore, there is a growing interest not just in the attacks themselves but in understanding the underlying causes of privacy violations and what makes a model explanation susceptible to privacy-related attacks [14]. The leakage of information via model explanations can be attributed to a range of factors. Some of these factors are intrinsic to how explanations are crafted and the methodologies behind them, while others relate to the data's sensitivity and the granularity of the information the explanations provide [15].

Given the paramount importance of protecting data privacy while simultaneously enhancing the transparency of machine learning (ML) models across domains, both the academic community and industry stakeholders are keenly focused on the privacy aspects of model explanations. To our knowledge, this article represents the inaugural comprehensive review of privacy-preserving mechanisms within model explanations. Through this work, we present an initial investigation that encapsulates both privacy breaches and their countermeasures in the context of model explanations, alongside explainable ML methodologies that inherently prioritize privacy. Furthermore, we develop taxonomies grounded in diverse criteria to serve as a reference for related research fields.

Comparisons with existing surveys. Figure 1 [7–9,11,16–21] compares our survey against existing ones. Many surveys have summarized different privacy issues on ML models [16, 17, 20, 22, 23], while others reviewed explanation methods for ML models [7, 9, 24], but not both. For example, Rigaki et al. [18] presented a thorough analysis of over 45 publications on privacy attacks in machine learning, spanning the last seven years. Hu et al. [19] surveyed a special type of privacy attacks, called membership inference. On the other hand, others [8,24,25] offered a comprehensive classification of model explanations to enhance interpretability and provided guidance for selecting suitable methods for specific ML models and desired explanations.

Some existing surveys summarized adversarial attacks but presented partial coverage of privacy attacks on model explanations with basic introductions and limited discussions of the methods. Ferry et al. [11] examined the interplay between interpretability, fairness, and privacy, which are critical for responsible AI, particularly in high-stakes decision-making like college admissions and credit scoring. Baniecki et al. [21] surveyed adversarial attacks on model explanations and fairness metrics, offered a unified taxonomy for clarity across related research areas, and discussed defensive strategies against such attacks. However, these studies are either too high level or too specialized in non-privacy attacks.

Our survey presents an in-depth examination of privacy attacks on model explanations, diverging from previous work by its comprehensive nature. Rather than addressing the full spectrum of adversarial attacks, our study is specifically tailored to privacy attacks. This focus is due to the recent surge in these attacks and their significant potential to compromise the right to explanation [1] and the right to privacy [26]. The threat posed by such privacy attacks could, in essence, challenge the very existence and usefulness of model explanations. Unlike the existing reviews that selected a very limited number of publications related to privacy attacks on model explanations (e.g., only two references are included in [21]), we conduct a comprehensive search and include magnitudes of related studies in this survey. We delve into the underlying principles, theoretical frameworks, methodologies, and taxonomies, while also mapping out potential trajectories for future research. Especially, our work encompasses the emerging field of privacy-preserving explainable AI (PrivEx or PPXAI), highlighting model explanations that inherently protect user privacy [27–29].

Paper collection methodology. Finding relevant research on this subject proved to be complex due to its incorporation of various topics such as data privacy, privacy attacks, explanations of models, explainable AI (XAI), and the development of privacy-preserving explanations. To navigate this breadth of concepts, we employed diverse keyword combinations about "privacy", "explanation", and specific attack types including "membership inference", "data reconstruction", "attribute inference", "model extraction", "model stealing", "property inference", and "model inversion". Our initial search utilized platforms like Google Scholar, Semantic Scholar, and Scite.ai — an AI-enhanced search tool — to assemble a preliminary collection of studies. This selection was expanded through backward searches, analyzing the references of initially chosen papers, and forward searches, identifying papers that cited the initial ones. Additionally, we manually verified the relevance and focus of these articles across various sources due to discrepancies, such as some studies addressing privacy in the context of safeguarding against manipulation attacks instead of privacy intrusions. Ultimately, this process culminated in magnitudes of pivotal research papers on the topic.

Contributions of the article. The main contributions of this article are as follows:

• **Comprehensive review.** To the best of our knowledge, this study represents the inaugural effort to thoroughly examine privacy-preserving model explanations. We have collated and summarized a substantial body of literature, including papers published or in pre-print up to April 2024.

• **Connected taxonomies.** We have organized all existing literature on PrivEx according to various criteria, including the types of explanations targeted and the methodologies employed in attacks and defences. Figure 2 showcases the taxonomy we have developed to structure these studies.

• **Challenges and future directions.** Designing privacy-preserving explanations for machine learning models is an emerging field of research. From the surveyed literature, we highlight unresolved issues and suggest several potential research directions into both the offensive and defensive aspects of privacy in model explanations.

• **Online updating resource.** To facilitate research in privacy-preserving model explanations, we have established an open-source repository, which aggregates a collection of pertinent studies, including links to papers and available code.

Organization of the article. The rest of the article is organized as follows. Section 2 revisits model explanations, acting as foundations for privacy attacks. Section 3 presents the taxonomy of privacy attacks on model explanations and provides in-depth descriptions, including threat model and attack scenarios. Section 4 discusses the causes of privacy leaks in model explanations. Section 5 explores countermeasures and a new class of privacy-preserving model explanations by design. Finally, Section 6 contains a discussion on ongoing and upcoming research directions and Section 7 concludes the survey.

2 Model explanations

Model explanations serve to clarify the decisions a model renders concerning a specific querying sample denoted by x represented as an n-dimensional feature vector ($x \in \mathbb{R}^n$). The explanation function ϕ ingests the dataset D, along with its labels — either the ground truth labels $\ell : D \to [C]$ or those inferred by a trained model f — and the query $x \in \mathbb{R}^n$. Such methods for explanation may require access to supplementary data [2], including the ability to query the model actively, a predefined notion of the data distribution, or familiarity with the class of the model [30].

Table 1 summarizes important notations in this paper.

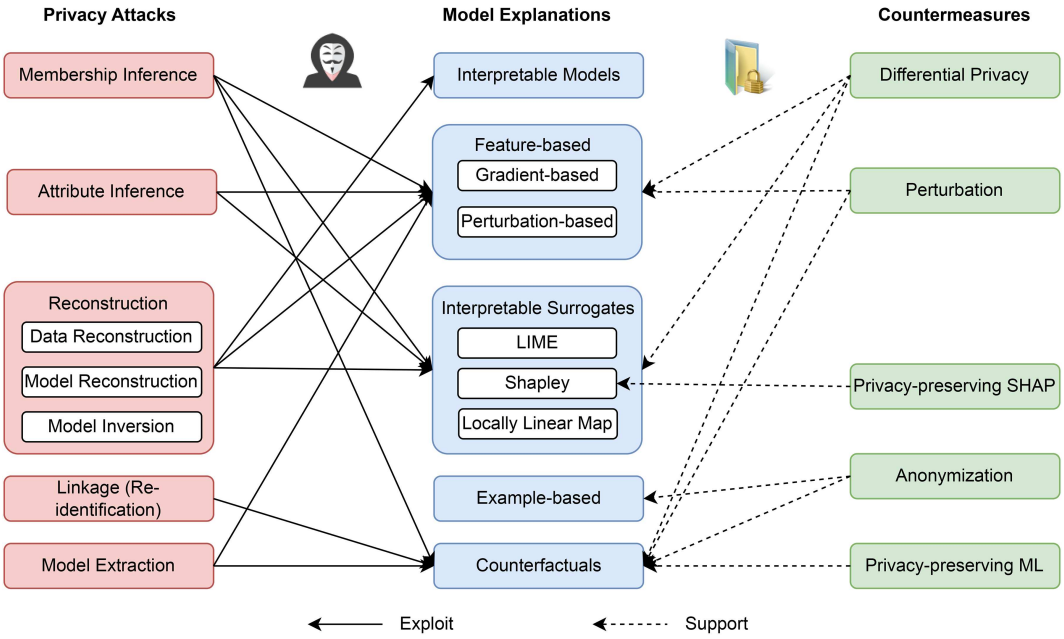

Figure 2 (Color online) Our taxonomy of privacy attacks and countermeasures on model explanations. "Exploit" arrows indicate existing studies about privacy attacks on targeted explanations. "Support" arrows indicate existing studies about privacy countermeasures for corresponding explanations. Some countermeasures target privacy attacks directly and their arrows are omitted for simplicity.

Table 1 Summary of important notations.

Notation	Description
$f : X \to Y$	A machine learning model
f_t	Target model of a privacy attack
f_a	Adversarial model by a privacy attack
D	Training data
$\phi(x) = \phi(D, f, x)$	Explanation on the input data x
$\phi^{\text{GRAD}}(x)$	Gradient-based explanation on input x
$\phi^{\text{INTG}}(x)$	Integrated gradient-based explanation on input x
$\phi^{\text{SMOOTH}}(x)$	Perturbation-based explanation on input x
$\phi^{\text{LIME}}(x)$	LIME explanation on input x
$\phi^{\text{SHAP}}(x)$	Shapley explanation on input x
$\phi^{\text{LLM}}(x)$	Locally linear map-based explanation on input x
$\phi^{\text{CF}}(x)$	Counterfactual explanation on input x
$\text{cf}(x)$	Counterfactual explanations/instances of the input data x
$\text{MI}_{\text{Distance}}(x)$	Distance-based membership inference attack on x
$\nabla_x f(x)$	Gradient of the model f on x
$\hat{f}(\cdot)$	Surrogate model produced by model extraction attack
ϵ-DP	Different privacy with ϵ degree or privacy budget

2.1 Feature-based explanations

The explanation function $\phi(D, f, x; \cdot)$ is predicated on identifying influential attributes (with the $\cdot$ symbol representing any potential additional inputs), and the explanation for the query x is frequently referred to simply as $\phi(x)$ [2]. The value at the i-th index of a feature-based explanation, $\phi_i(x)$, quantifies the extent of influence the i-th feature exerts on the label ascribed to x. Ancona et al. [3] have curated a comprehensive exposition of these attribution-focused explanation modalities, also termed attribution methods or numerical influential measures [31].

Backpropagation-based (also known as gradient-based). This type of explanation explains the decisions of neural network models through the lens of back propagation [30] (see Figure 3). It allows for the allocation of the model's predictive reasoning back to the individual input features [32–37].

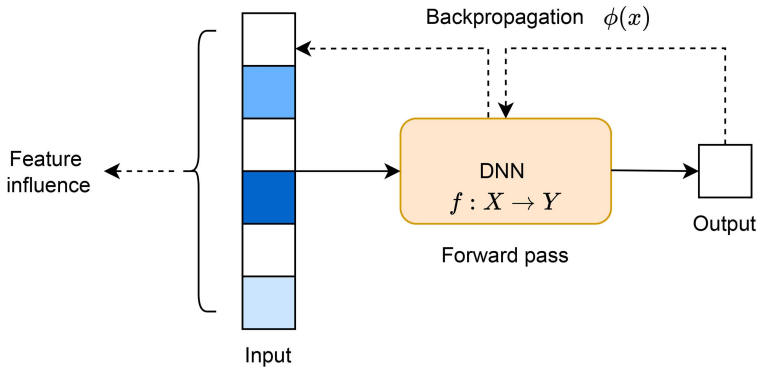

Figure 3 (Color online) Feature-based explanations via backpropagation.

• (Vanilla) Gradients. Simonyan et al. [32] introduced gradient-based explanations, originally for image classification models, to emphasize important image pixels that affect the predictive outcomes. The explanation vector is defined as $\phi^{\mathrm{GRAD}}(x) = \nabla_x f(x)$ or $\phi_i(x) = \frac{\partial f}{\partial x_i}(x)$ for each feature i. A high partial differential value indicates that a pixel significantly affects the prediction, and analyzing the map of these values (so-called gradient map) can explain a model's decision-making [38]. Shrikumar et al. [34] suggested enhancing numerical explanations using the input feature value multiplied by the gradient, $\phi_i(x) = x_i \times \frac{\partial f}{\partial x_i}(x)$.

• Integrated gradients. Sundararajan et al. [37] advocated for an alternative to standard gradient computation by averaging gradients along a straight path from a baseline input x^{BL} (often $x^{\mathrm{BL}} = \mathbf{0}$) to the actual input. This method follows critical axioms like sensitivity and completeness. Sensitivity ensures that if there is a prediction change due to x_i not equaling $x_{\mathrm{BL},i}$, then $\phi_i(x)$ should not be zero. Completeness dictates that the sum of all attributions equals the change in prediction from the baseline to the input.

$$\phi^{\mathrm{INTG}}(x_i) = (x_i - x_{\mathrm{BL},i}) \cdot \int_{\alpha=0}^{1} \frac{\partial c(x^\alpha)}{\partial x_i^\alpha}\bigg|_{x^\alpha = x + \alpha(x - x^{\mathrm{BL}})}. \tag{1}$$

• Guided backpropagation. Designed for networks with ReLU activations (others as well), guided backpropagation [39] modifies the gradient to only reflect paths with positive weights and activations, thereby considering only the positive evidence for a specific prediction.

• Layer-wise relevance propagation (LRP). LRP is proposed by Klauschen et al. [33] to assign relevance from the output layer back to the input features. The relevance in each layer is proportionally distributed according to the contribution from neurons in the previous layer. The final attributions for the input are referred to as $\phi^{\mathrm{LRP}}(x)$.

Perturbation-based. Perturbation-based explanations involve querying a model that needs to be explained with a series of altered inputs [30]. SmoothGrad [36] is a popular perturbation-based explanation method that produces several samples by injecting Gaussian noise into the input data and then computes the mean of the gradients from these samples. Formally, for a certain k samples, the explanation function is defined as

$$\phi^{\mathrm{SMOOTH}}(x) = \frac{1}{k} \sum_{k} \nabla_f (x + \mathcal{N}(0, \sigma)), \tag{2}$$

where $\mathcal{N}$ represents the normal distribution and σ stands for a hyperparameter that controls the level of perturbation.

2.2 Interpretable surrogates

This method explains a black-box ML model or complex deep neural networks by computing a surrogate model that is interpretable by design [8, 30, 40] that can emulate the overall predictive patterns of the original model [14].

LIME. Local interpretable model-agnostic explanations [4] generate a local interpretative approximation of a given model through sampling on the optimisation problem:

$$\phi^{\mathrm{LIME}}(\bar{x}) = \arg\min_{g \in G} \mathcal{L}(g, f, \pi_x) + \Omega(g), \tag{3}$$

where G is a collection of interpretable functions employed for explanatory purposes, $\mathcal{L}$ quantifies how well g approximates f in the neighborhood π_x of x, and Ω imposes a regularization on g to avoid overfitting. Usually, G involves one or multiple linear models and Ω is a Ridge regularization [30]. The loss function is typically computed as the expected squared difference between the outputs of f and g weighted by the probability distribution π_x [41]:

$$L(f, g, \pi_x) = \sum_{x' \in X'} [f(x') - g(x')]^2 \pi_x(x'), \tag{4}$$

where X' is the neighbourhood of x and $\pi_x(x')$ is a promixity measure between x and x'.

SHAP (local). The main distinction between LIME and SHAP is in the selection of the functions Ω and π_x. LIME takes a heuristic approach: $\Omega(g)$ represents the count of non-zero weights within the linear model, while $\pi_x(x')$ utilizes either cosine or l2 distance [41]. SHAP values provide a way to quantify the contribution of each feature in a model prediction [5, 42–45]. Specifically, for a given model f and a data point $x = [x_1, \ldots, x_M]$, the SHAP value for feature i is calculated as a weighted average of differences between the model prediction with and without feature i:

$$\phi_i^{\text{SHAP}}(x) = \sum_{S \subseteq \{1,\ldots,M\} \setminus \{i\}} \frac{1}{M} \frac{f_{S \cup \{i\}}(x) - f_S(x)}{\binom{M-1}{|S|}}, \tag{5}$$

where $|S|$ is the size of the subset S and M is the total number of features. For instance, let $x^0 = [x_i^0]_{i=1}^M$ be a reference sample of M features. Suppose $M = 4$, $x = [5, 2, 7, 3]$, $x^0 = [0, 0, 0, 0]$, and we want to compute the marginal contribution s_i of feature $i = 1$ to the feature set $S = \{2, 3\}$. Then $s_i = \frac{1}{4} \frac{f(x_{[1,2,3]}) - f(x_{[2,3]})}{3} = \frac{f([5,2,7,0]) - f([0,2,7,0])}{12}$.

Global Shapley values. The above Shapley values are local because the explanations are based on a singular reference sample x^0 and a single input sample x [41]. Begley et al. [46] proposed a global Shapley value by averaging local Shapley values over both foreground and background distributions, as given by

$$\Phi_i^{\text{SHAP}}(f, F, B) = \mathbb{E}[\phi_i(f, x, x^0)] \tag{6}$$

for each feature index $i = 1, 2, \ldots, M$. In other words, to conduct a global analysis of model behavior, it is necessary to consider predictions at multiple inputs $x \sim \mathcal{F}$ from a distribution $\mathcal{F}$ called the foreground. Since the choice of baseline x^0 is ambiguous, baselines $x^0 \sim \mathcal{B}$ are sampled from a distribution $\mathcal{B}$ called the background [47].

Locally linear maps. Harder et al. [29] introduced locally linear maps (LLM), a method aimed at providing both local and global explanations for models, which is more expressive than standard linear models and offers an efficient way to manage the number of parameters for a good privacy-accuracy trade-off.

$$\phi_k^{\text{LLM}}(x) = \sum_{m=1}^{M} \sigma(x)_m^k g_m^k(x), \text{ where } g_m^k(x) = w_m^k \cdot x + b_m^k, \tag{7}$$

and the weighting coefficients are computed via softmax:

$$\sigma_m^k(x) = \frac{\exp[\beta \cdot g_m^k(x)]}{\sum_{m=1}^{M} \exp[\beta \cdot g_m^k(x)]}. \tag{8}$$

The method optimizes a cross-entropy loss $\mathcal{L}(W, \mathcal{D})$ for the parameters of LLM collectively denoted by W, with the predictive class label $y_{n,k}(W)$ defined through a softmax function applied to the output of $\phi_k(x_n)$.

2.3 Example-based explanations

Example-based explanation (aka case-based interpretability or record-based explanation [31]) uses comparable examples to create transparent explanations for machine learning decisions, offering an accessible way to understand model predictions by contrasting similar cases from the model's database or generated data [48]. Case-based interpretability techniques can create a range of explanatory examples, including the following.

• Similar examples are the closest matches from the training data with corresponding predictions to the case being analyzed, identified through a defined measure of similarity.

• Typical examples represent the epitome of a particular prediction, frequently utilized in models that focus on prototype learning.

• Counterfactual examples are similar examples but with differing predictions, highlighting the minimal changes needed for a different outcome. We dedicate a separate discussion on counterfactuals in Subsection 2.4.

• Semi-factual examples are similar to the original case with the same prediction but positioned near the decision boundary, demonstrating the robustness of the prediction against variations typical of a different classification.

• Influential examples are key data points within a training set that have a significant impact on a model's prediction for a given query instance [49]. For explanatory purposes, we can provide the top k influential points [31].

These explanations can be sourced from existing datasets (i.e., $\phi(D, f, x; \cdot) \in D$) [49] or crafted based on the original data [50, 51].

Intrinsic methods for traditional ML. Case-based explanations in machine learning are derived from either distance-based or prototype-based interpretable methods. Distance-based methods utilize a measure of proximity to retrieve the most similar data points as explanations, while prototype-based methods classify and explain instances based on representative prototypes of clustered data. The k-nearest neighbors (KNN) algorithm exemplifies the former, offering explanations as similar or counterfactual examples based on label correspondence. The Bayesian case model (BCM) is a prototype-based method that explains decisions through typical examples representative of data clusters [52]. Both methods aim to make model decisions understandable by referencing specific, characteristic data points or clusters [48].

Post hoc methods for traditional ML. Post hoc interpretability techniques leverage traditional machine learning models as metrics for finding similar examples, with decision trees and rule-based models used to determine similarity between data samples [48]. Counterfactual examples, on the other hand, come from nodes with differing outcomes. Moreover, models like explanation oriented retrieval (EOR), built on the KNN algorithm, reorder neighbors to highlight those with the highest explanatory utility, thus providing semi-factual examples that maintain the same classification but are closer to the decision boundary [53].

Intrinsic methods for deep learning. In deep learning, intrinsic interpretability can be provided by prototype-based or distance-based methods [48]. For instance, the explainable deep neural network (xDNN) [54] and deep machine reasoning (DMR) [55] define prototypes as dense data points and classify observations based on the closest prototype. The prototype classifier method learns representative prototypes from training data, using an autoencoder for feature extraction and classification based on latent representations [56]. The prototypical part network (ProtoPNet) represents image parts in clusters in a latent space, which are used to predict and explain classifications [57]. Additionally, the deep k-nearest neighbors (DkNN) calculates neighbors at each model layer to ensure consistent predictions, offering explanations based on similar examples across the model's entirety [58].

Post hoc methods for deep learning. Post hoc interpretability methods in deep learning either utilize interpretable surrogate models to extract explanations from a primary model or directly analyze a "black-box" model to identify and retrieve the most similar data instances for explanation purposes [48]. Concept whitening, for example, organizes the latent space of a classification network around predefined concepts, enabling the measurement of distance between instances for similar example retrieval [59]. Interpretability-guided content-based image retrieval (IG-CBIR) enhances image retrieval using saliency maps to focus on relevant image regions [60]. Unsupervised clustering and the KNN algorithm within the twin systems framework are other surrogate models that categorize or find similar examples based on feature extraction techniques like perturbation and sensitivity analysis [61, 62].

2.4 Counterfactual explanations

Counterfactual explanations (aka algorithmic recourse) provide insights into how slight changes to input features could lead to different model outcomes, aiding in tasks like model debugging and ensuring regulatory compliance [10, 63]. Figure 4 gives an illustration of counterfactual and other four sample categories (i.e., adversarial examples, local robustness, invariant samples, and uncertainty samples) through the boundaries between human analysis and a learned model. The application of counterfactual explanations

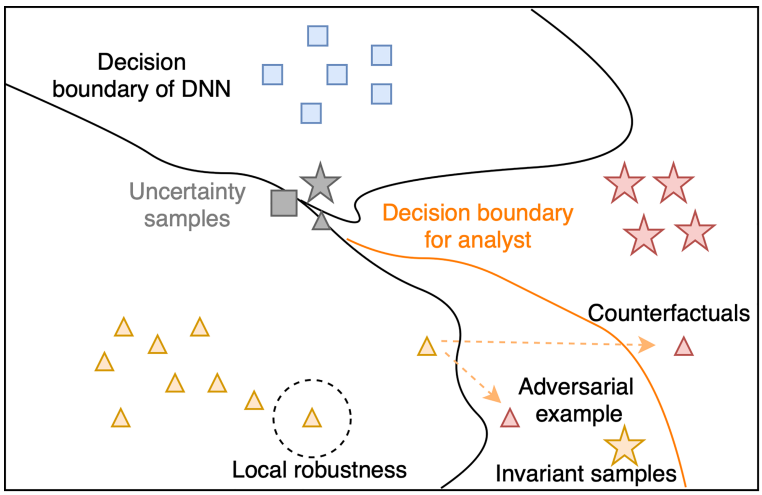

Figure 4 (Color online) Decision boundaries between human analyst and a learned model.

varies with the model's complexity and includes considerations such as model transparency, type compatibility, and adherence to constraints like feasibility and causality [64–66]. The concept overlaps with other areas of research such as algorithmic recourse, inverse classification, and contrastive explanations [67–70].

Single counterfactual. Formally, counterfactual explanation is the process of finding changes δ to an instance x that reverses a negative predictive outcome from a model $f_\theta(x) = 0$ to a positive one $f_\theta(x+\delta) = 1$, where θ are model parameters. The problem involves identifying a counterfactual $x' = x+\delta$ where the predictive model outputs a positive outcome and does so with minimal cost $c(x, x')$, which is easily implementable, often using ℓ_1 or ℓ_2 distance as cost functions. The optimization problem is defined as

$$\phi^{\mathrm{CF}}(x) = \arg\min_{x' \in A^P} L(f_\theta(x'), 1) + \lambda \cdot c(x, x'), \tag{9}$$

where A^P is the set of plausible or actionable counterfactuals and $L(\cdot, \cdot)$ is a differential loss such as binary cross entropy [71].

Example 1. Possible counterfactual explanations derived from the FICO explainable machine learning challenge dataset [12].

• If the number of satisfactory trade lines had been 10 or fewer, rather than the actual 20, the prediction would have been positive.

• If there had been no trade lines that were ever 60 days overdue and marked as derogatory in the public record, rather than the actual count of 2, the prediction would have shifted to positive.

Diverse counterfactuals. Recent articles study the generation of multiple alternative counterfactuals per input, offering a spectrum of potential changes rather than just one nearest option [72]. This approach empowers users by offering them various ways they could potentially modify their data to achieve a preferred result.

Kuppa et al. [63] noted that methods for creating counterfactual explanations (CF) bear resemblance to those for generating adversarial examples (AE) in the way they both employ gradient-based optimization and surrogate models to find CF/AE for a given model. Some privacy attacks on adversarial examples can be used on counterfactual explanations [63].

3 Privacy attacks

According to a classification system mentioned in [17, 21], explainable AI systems can fall prey to three main categories of attacks: (i) integrity attacks, such as evasion and backdoor poisoning, leading to incorrect categorization of certain data points [73–76]; (ii) availability attacks, characterized by poisoning efforts aimed at inflating the error rate in classification tasks [77]; and (iii) privacy and confidentiality attacks, aimed at extracting sensitive information from user data and the models themselves. Although all forms of interference in machine learning can be considered adversarial, "adversarial attacks" specifically denote those targeting the security aspect, particularly through malicious samples [41, 47, 78, 79].

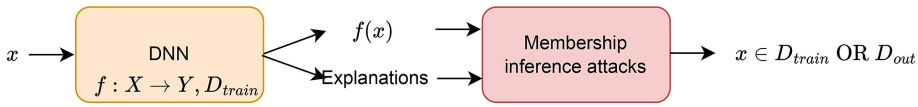

Figure 5 (Color online) Membership inference attacks.

This work is primarily concerned with breaches of privacy and confidentiality, including membership inference attacks, linkage attacks, reconstruction attacks, attribute/feature inference attacks, and model extraction attacks. The rationale behind including model extraction attacks is their frequent association with privacy violations in related literature [18], and the notion hijacking a model's functions could also infringe on privacy. Veale et al. [80] contended that privacy violations like membership inference attacks elevate the likelihood of machine learning models being deemed personal data under the European Union's General Data Protection Regulation (GDPR), as they could make individuals identifiable.

3.1 Membership inference attacks (MIA)

MIA aims to detect if data are part of a model's training set [30, 81]. Before model explanations, popular attacks are loss thresholding and likelihood ratio attack (LRT) [71]. Loss threshold identifies if a data point was in the training set by checking the model's error rate against a threshold, requiring access to labels and model details [82, 83]. LRT, in contrast, uses shadow models to compare confidence levels of data being in or out of the training set, calculating a likelihood ratio to predict membership without needing direct model access [84]. Pawelczyk et al. [71] designs a recourse-based attack (using counterfactual explanation) without access to the true labels and knowledge of the correct loss functions.

Threat model. The adversary is able to submit x to the black-box model [84–87] to receive the prediction $f(x)$ and any corresponding explanations, despite not having direct access to the model's internals [88] (see Figure 5). However, they are assumed to know the model's architecture and possess an auxiliary dataset similar to the model's training data, reflected in much of the current research on the topic [89].

• Threat model on gradient-based explanations. Most threat models are based on threshold-based attacks [30]. There are two key scenarios for this: the optimal threshold scenario, where the threshold is deduced from known data point memberships to gauge the maximum privacy risk; and the reference/shadow model scenario, which is more practical and assumes the attacker has some labeled data from the same distribution as the target model, as well as knowledge of the model's architecture and hyperparameters in line with Kerckhoffs's principle [90]. The attacker then trains a number of shadow models on this data to approximate the threshold, an approach that becomes more resource-intensive as the number of shadow models increases [30].

• Threat model on interpretable surrogates. Naretto et al. [14] investigated how global explanation methods can potentially compromise the privacy. Specifically, the authors focus on TREPAN [91], an algorithm that explains neural network decisions by creating a surrogate decision tree model.

• Threat model on counterfactuals. Pawelczyk et al. [71] formulated a membership inference game for attacking counterfactual explanations. The game features two participants: a model owner ($\mathcal{O}$) and an opponent ($\mathcal{A}$). Their actions are as follows. $\mathcal{O}$ selects a dataset for training from a population D^N, applying a training algorithm T with a loss function ℓ. Subsequently, $\mathcal{O}$ assigns a binary label $f_\theta(z)$ to each datapoint z in D_t. Let D_t^0 be the segment of training data for which $f_\theta(x) = 0$, and $D_{\theta,0}$ represent the conditional distribution $p(z)|f_\theta(z) = 0$. $\mathcal{O}$ tosses a coin, and based on the outcome, selects a sample x from either $D_{\theta,0}$ or D_t^+. Then, using the recourse algorithm ϕ, $\mathcal{O}$ generates an alternate instance x' from $\phi(f_\theta, x, D_t)$ and sends the pair (x', x) to $\mathcal{A}$. In addition to the sample pair, $\mathcal{A}$ has the capability to make queries to D. It is presumed that $\mathcal{A}$ is fully aware of $\mathcal{O}$'s implementation specifics, including the training algorithm T and the recourse algorithm ϕ. $\mathcal{A}$ concludes the game by providing a binary guess G signifying if x belongs to D_t (member) or does not ($x \notin D_t$, non-member).

General attacks. In the training set, data points are generally positioned away from the decision boundary, leading to lower loss scores that can be leveraged to detect membership in the training data [82, 83, 88]. This principle is utilized in the OPT-var method [30], in which the variance in the explanation $e = \phi(f, x)$ based on the logit score $f(x)$ could signal whether a point was in the training set. However, Quan et al. [88] argued that logit scores alone may not fully represent the prediction confidence of the victim model because they do not take into account the scores of other classes. Instead, Quan et

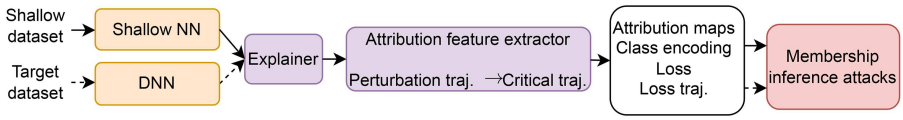

Figure 6 (Color online) Model-based membership inference attacks proposed in [89].

al. [88] suggested using the softmax function $\sigma(f(x))$, which reflects class interactions, to provide a more comprehensive membership indicator.

Liu et al. [89] proposed a model-based attack that involves four main stages: training a shadow model, extracting attribution features, training an attack model, and inferring membership (see Figure 6). The adversary starts by training a shadow model using an auxiliary dataset that is similar to the training data of the target model. Then, attribution maps are generated for a given sample, and perturbations are applied based on these maps to observe changes in predictions. Next, the adversary trains an attack model, typically a multi-layer perceptron (MLP), using the attribution features combined with other data such as loss values and one-hot encoded class information to construct features indicative of membership.

• Attacks on gradient-based explanations. Shokri et al. [30] used a threshold-based attack that infers membership based on the model's confidence or its explanation output. A data point is classified as a member if the variance of the confidence scores $\mathrm{Var}(f_\theta(x))$ or the variance of the explanation $\mathrm{Var}(\phi(x))$ is below or equal to a certain threshold τ. Attacks using explanation variance exploit the model's certainty: when a model is sure about a prediction, explanation variance is low. However, near the decision boundary, even small changes can increase explanation variance. Models with certain activation functions like tanh, sigmoid, or softmax have steeper gradients, affecting how training data points are positioned relative to these boundaries [30].

• Attacks on interpretable surrogates. Naretto et al. [14] developed an attacking procedure to assess the potential privacy risks of an interpretable surrogate (global explainer) that attempts to replicate the behavior of a black-box model. First, an MIA model, denoted as A_b, is trained to determine whether a specific data record x was included in the training dataset D^b_{train} of the black-box model b. This attack model leverages the black-box b itself to classify the training data for the attack, making it specifically aimed at b. The attack training dataset D^a_{train} is the same as D^B_{Attack}. Similarly, another MIA model A_c is developed to target the global explainer c, which serves as an interpretable stand-in for the black-box model b. This model is trained using D^a_{train}, but this time the labeling is done by c, not b.

• Attacks on counterfactual explanations. The adversary has access to both the original instance x and a counterfactual instance x'. Models often overfit to training points, resulting in lower losses for these points compared to those on the test set [30]. Pawelczyk et al. [71] designed a distance-based attack where if the loss is below a certain threshold τ, the point is considered a member of the training set. The counterfactual distance $c(x, x')$ is effectively the distance to the model boundary, and even though algorithms that produce realistic recourses may not optimize for this distance, it can still be viewed as an approximation to the distance to the model boundary [67, 92]. The counterfactual distance-based attack is defined by $\mathrm{MI}_{\text{Distance}}(x)$ as follows:

$$\mathrm{MI}_{\text{Distance}}(x) = \begin{cases} \text{Member}, & \text{if } c(x, x') \geqslant \tau_D, \\ \text{Non-member}, & \text{if } c(x, x') < \tau_D. \end{cases} \tag{10}$$

Another attack is using a likelihood ratio test on top of the counterfactual distance (CFD) [71]. The process involves calculating a baseline statistic t_0 using $c(x, x')$ from the recourse output. If the initial statistic t_0 surpasses the critical threshold $z_{1-\alpha}$, which is the $1-\alpha$ quantile of the normal distribution Z, the algorithm designates the data point as a 'Non-member'; and 'Member' otherwise. The key benefit is that it estimates the parameters $\mu_{\text{out}}, \sigma_{\text{out}}$ only once for the non-membership scenario, reducing the computational load when assessing multiple data points x' [83].

Huang et al. [93] proposed a CFD-based likelihood ratio test (LRT) for linear classifiers built on the above Pawelczyk method [71]. But the attack is simplified and one-sided as it only estimates parameters for data outside the training set, thus reducing computational complexity.

Kuppa et al. [63] developed an attack that leverages an auxiliary dataset D_{aux} to train a shadow model A_{MemInf}. This is done by generating counterfactual examples x_{cfi} for input samples x_i and training a 1-nearest neighbor (1-NN) classifier to predict class membership based on proximity to these counterfactuals. If the prediction probability difference between the shadow model A_{MemInf} and the

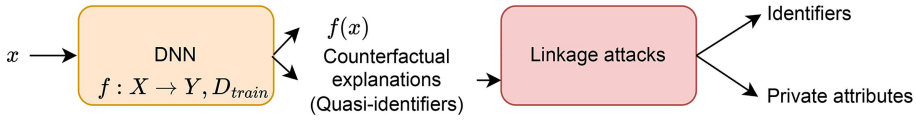

Figure 7 (Color online) Linkage attacks.

target model T is below a threshold t, the sample is deemed part of the training set. This inference is made under the assumption that if both models predict similarly for a sample, it implies the sample was significant in its prediction. The method is advantageous as it requires no direct access to the training set and iteratively uses counterfactuals to extract new data.

3.2 Linkage attacks

Threat model. Goethals et al. [10] introduced a privacy concern with counterfactual explanations when they are based on training instances. The data usually consist of identifiers (like name and social security number), quasi-identifiers (like age, zip code, gender), and private attributes. It has been shown that a significant portion of US citizens could be uniquely identified by combining their zip code, gender, and date of birth [94]. The attack setup assumes the adversary has access to identifiers and quasi-identifiers. There are two re-identification scenarios discussed: one where a specific individual is targeted to uncover their private attributes, and another where the adversary aims to prove that re-identification is possible, regardless of who the individual is. Counterfactual explanations, which do not include identifiers but may contain unique combinations of quasi-identifiers, could be exploited by an attacker to infer private attributes in what is termed an "explanation linkage attack" or "re-identification attack" [10] (see Figure 7).

Attacks on counterfactual explanations. Goethals et al. [10] presented a scenario where Lisa is denied credit and requests a counterfactual explanation, which inadvertently reveals Fionas' private information because Fiona is the nearest neighbor in the data. Native counterfactuals, which are real instances from the dataset, are more plausible but increase the risk of re-identification [95]. Perturbation-based counterfactuals, which synthetically generate explanations, pose less privacy risk but can still be vulnerable to sophisticated attacks if the perturbations are minor [15,96,97]. Aïvodji et al. [98] identified that diverse counterfactual explanations can inadvertently expose decision boundaries more, risking the leak of sensitive data like health or financial information. Linkage attacks exploit this by matching anonymized records with external datasets, combining various attributes to re-identify individuals.

3.3 Reconstruction attacks

Based on model predictions and explanations, reconstruction attacks involve dataset reconstruction attacks, model reconstruction attacks, and model inversion attacks (see Figure 8).

Dataset reconstruction attacks. It is important to preserve privacy in datasets due to several threats posed by inference attacks that seek to deduce sensitive information from model outputs [18, 99]. Ferry et al. [100, 101] reviewed the evolution of reconstruction attacks from databases to machine learning, where adversaries attempt to recover training data. Techniques range from linear programming to exploiting data memorization, even within frameworks meant to promote fairness [102,103]. The goal of data reconstruction attacks is to make models trained for fairness inadvertently reveal sensitive attributes, including leveraging auxiliary datasets and queries to an auditor for enhancing attacks [104, 105].

• Threat model. A machine learning model that is interpretable, like a decision tree, contains implicit information about its training dataset [100]. This information can be formalized into a probabilistic dataset $\mathcal{D}$ consisting of n examples, each with d attributes. Every attribute a_k has a domain V_k covering all possible attribute values. The knowledge about an attribute a_k for a given example x_i is represented by a probability distribution across all possible values for that attribute, using the random variable $\mathcal{D}_{i,k}$. If a value $\mathcal{D}_{i,k}$ within V_k has all the probability mass ($P(\mathcal{D}_{i,k} = v_{i,k}) = 1$), it is deterministic. Otherwise, the dataset encompasses some uncertainty about attribute values.

• Probabilistic reconstruction attacks. Earlier research [106] proposes a method for constructing a probabilistic dataset $\mathcal{D}^{\mathrm{DT}}$ from the structure of a trained decision tree DT. This probabilistic dataset reflects the decision tree's implicit knowledge about its training dataset $\mathcal{D}^{\mathrm{Orig}}$. The construction of this dataset is termed a probabilistic reconstruction attack, and by design, $\mathcal{D}^{\mathrm{DT}}$ is compatible with $\mathcal{D}^{\mathrm{Orig}}$,

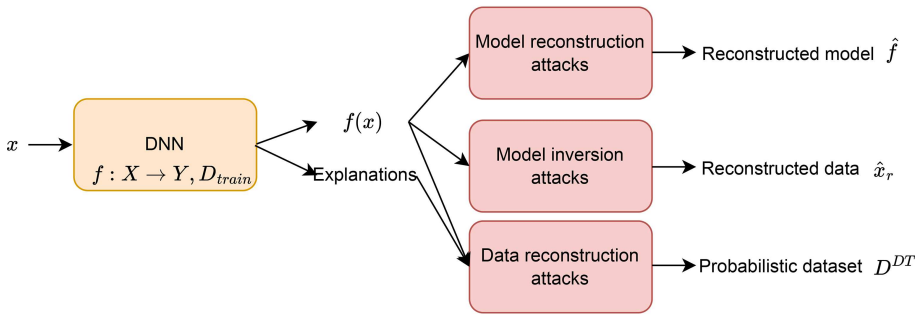

Figure 8 (Color online) Reconstruction attacks.

meaning the actual value $v_{i,k}^{\text{Orig}}$ of any attribute a_k for any example x_i is always among the set of possible values in the probabilistic reconstruction ($P(\mathcal{D}_{i,k}^{\text{DT}} = v_{i,k}^{\text{Orig}}) > 0$).

• Attacks on interpretable models. Ferry et al. [100] discussed the possibility of a probabilistic reconstruction attack on interpretable models. In the general case, the success of the attack is calculated using the joint entropy of the dataset's cells, which can be simplified if the variables of the model are statistically independent. For interpretable models like decision trees and rule lists, this assumption allows further decomposition of the computation [100].

Model reconstruction attacks. Model reconstruction is the process of replicating a classifier $\hat{f}$ when provided with membership and gradient queries to an oracle that, for any input x, reveals both the classifier's output $\hat{f}(x)$ and the gradient $\nabla_x \hat{f}(x)$. Milli et al. [107] examined a specific scenario involving a one hidden-layer neural network function $f : \mathbb{R}^d \to \mathbb{R}$ that uses ReLU activations, formulated as $f(x) = \sum_{i=1}^{h} w_i \max(A_i^{\mathrm{T}} x, 0)$.

• Threat model. For a DNN with parameters $A \in \mathbb{R}^{h \times d}$ and $w \in \mathbb{R}^h$, where A_i represents the ith row of A, three assumptions are posited: (1) each row $A_1, \ldots, A_h$ is a unit vector; (2) no pair of rows A_i and A_j are collinear for $i \neq j$, satisfying $\langle A_i, A_j \rangle \leqslant 1 - c$ for some $c > 0$; (3) the rows $A_1, \ldots, A_h$ are linearly independent. These assumptions are stated to be without loss of generality since they can be achieved by simple reparameterization of the network, such as scaling w or A, or by reducing the hidden layer dimension.

• General attacks. Under these assumptions, it is possible to learn the function with a sample complexity independent of the input dimension d [107]. Specifically, with a probability of $1 - \delta$, an algorithm can find a function $\hat{f}$ such that $\hat{f} = f$. If the algorithm cannot find such a function, it will report the failure. Regardless of the outcome, the algorithm requires only $O(h \log \frac{h}{\delta})$ queries to learn the function.

• Attacks on gradient-based explanations. The attack involves recovering a matrix Z and a sign vector s [107]. Z is composed of either $w_i A_i$ or $-w_i A_i$, with the signs encapsulated in s. The function f can then be reconstructed from Z and s, utilizing the recovered structure to make predictions. The approach exploits the gradient structure of f to identify the hyperplanes that partition the input space and uses binary search to recover the necessary components of Z and s.

Model inversion attacks. Model inversion attacks aim to deduce original data from predictions, such as recreating a person's face based on their predicted emotional state [108–110]. Initially, model inversion attacks showed limited success [108], but advancements in deep learning, especially through the use of transposed convolutional neural networks (CNNs), have significantly enhanced their effectiveness [109,111, 112]. Additional enhancements have been achieved by utilizing auxiliary information, including access to the model's internal workings and feature embeddings, or understanding the joint probability distribution between features and labels [82, 110, 112]. Especially the increasing demand for model explanations is likely to make these attacks more common [113].

• Threat model. We consider a machine learning model f_t that processes confidential data x from a set X_p (for instance, facial images). It employs these private inputs to generate a prediction $\hat{y}_t$ (such as identifying emotions). An issue arises when an attacker gains access to the target prediction $\hat{y}_t$ and the explanation ϕ_t (due to reasons like a data breach, interception during transmission, or sharing on social media). One scenario is to assume that the attacker only has the compromised data, an independent dataset $x \in X_a$, and the ability to interact with the target model via black-box [113]. The attacker does not require additional privileged information, such as blurred versions of the images. The objective of the attacker is to develop their own inversion model f_a to reconstruct the original image x from the

model's outputs $(\hat{y}_t, \phi_t)$. Such a reconstruction would allow them to predict sensitive information from the reconstructed image $\hat{x}_r$, including the possibility of re-identifying the individual from the facial emotion recognition system.

• Attack on a single gradient-based explanation. To invert the target model M_t, a transposed convolutional neural network (TCNN) [114] is devised to reconstruct a two-dimensional image x_r from the one-dimensional prediction vector y_t provided by M_t. The TCNN minimizes the mean squared error (MSE) loss to approximate the original image. This TCNN incorporates various input forms, such as saliency maps and 2D explanations [32,115], enhancing the reconstruction of x_r. Inputs can be processed by flattening the 2D explanations into a 1D vector and concatenating with the prediction vector, or using a CNN to convert 2D patterns into a 1D feature embedding, following the approach used in CNN encoder-decoder networks and super-resolution techniques [110, 116]. A U-Net architecture is employed to improve the reconstruction fidelity [117]. A hybrid model that combines flattened explanations with the U-Net structure is introduced in [113]. The training objective for these models is defined by the image reconstruction loss function:

$$L_r = \sum_x (M_i^a(M_t(x)) - x)^2, \tag{11}$$

where x represents the original image, $M_t(x) = y_t$ denotes the prediction from the target model, and $M_i^a(M_t(x)) = x_r$ is the reconstructed image output. Zhao et al. [113] conducted experiments on how different explanation methods, including gradients [32], CAM [118], LRP [33], and blurred versions of the input images, affect the inversion model's ability to capture information.

• Attack on multiple gradient-based explanations. While many explanations clarify the reasons that a model predicts a certain class within a set C, it is equally crucial to elucidate why it did not predict a different class $c' \neq c$, offering contrastive insights [119]. To facilitate this, certain techniques like Grad-CAM can generate explanations that are specific to a class based on the user's query [115]. Nevertheless, this approach increases the risk to privacy as it provides additional information. Zhao et al. [113] made use of these alternative CAMs (Σ-CAM) by merging explanations across all classes in $|C|$ into a three-dimensional tensor, and they train their inversion models on this tensor rather than on a two-dimensional matrix representing a single explanation.

• Attack on surrogate explanations. Interpretable surrogates could be harnessed for inversion attacks, even for models that do not provide target explanations. Zhao et al. [113] proposed an attack that predicts the target explanation and exploits that explanation to invert the original target data. Initially, an explainable surrogate target model f_a is trained using the attacker's dataset to generate a surrogate explanation $\widetilde{\phi}$. However, $\widetilde{\phi}_t$ is only accessible during the training phase and not during prediction. Consequently, an explanation inversion model f_e is trained to reconstruct $\widetilde{\phi}_t$ as $\widehat{\phi_r}$ based on the target prediction $\widehat{y_t}$. The loss function for minimizing the surrogate explanation error is

$$L_\phi = \sum_x \left(f_e(f_t(x)) - \phi(f_t(x))\right)^2, \tag{12}$$

where $\phi(f)$ denotes the explanation of the model f, $f_t(x) = y_t$ represents the surrogate target prediction, $\phi(f_t(x)) = \widetilde{\phi}_t$ is the surrogate explanation, and $f_e(f_t(x)) = \widehat{\phi_r}$ is the reconstructed surrogate explanation. This reconstructed explanation is available at prediction time. Finally, $\widehat{\phi_r}$ is fed into the image inversion model ϕ_i to finalize the model inversion attack. Given that $\widehat{\phi_r}$ is formatted similarly to $\widetilde{\phi}_t$, any explanation methods can be applied.

• Attacks on confidence scores. Fredrikson et al. [108] developed a model inversion attack using a maximum a posteriori (MAP) estimator to compute $f(x_1, \ldots, x_d)$ for all possible values of the sensitive feature x_1, while exploiting confidence information from model predictions. Fredrikson et al. [108] addressed the challenge of inverting high-dimensional features like facial recognition, where the inversion task becomes an optimization problem solved by gradient descent.

3.4 Attribute/feature inference attacks

Attribute inference attacks, aka feature inference attacks, are designed to deduce specific attributes, such as gender, from individual data records using accessible data like model predictions or explanations [82, 120] (see Figure 9). These types of attacks are distinct from property inference attacks, which seek to ascertain broader dataset characteristics, like the training data's gender ratio [121–123].

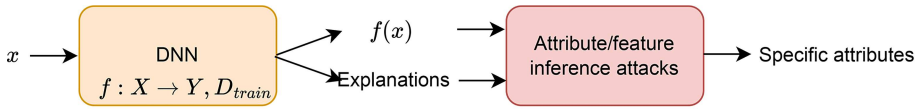

Figure 9 (Color online) Attribute/feature inference attacks.

Duddu et al. [124] investigated a scenario where a machine learning model, f_{target}, is cloud-deployed within an MLaaS framework (e.g., Google Cloud, Microsoft Azure), capable of providing predictions and required explanations for any given input. Users can submit a private sample $x = \{x_i\}_{i=1}^{n}$ to the service provider and receive a prediction vector $\hat{y} = \{\hat{y}_i\}_{i=1}^{c}$, along with an explanation vector $\phi = \{\phi_i\}_{i=1}^{n}$ that pertains to a specific class. Although the service provider has the capacity to return multiple explanation vectors corresponding to different classes [125], for practicality and without loss of generality, most studies focus on the use of one explanation vector for a specific class [13].

Threat models on feature-based explanations. Duddu et al. [124] considered two threat models (TM). (1) TM1 (with s in D): Here, the sensitive feature s is included in both the training dataset D and the input. $\mathcal{A}dv$ has access to the predictions $f_{\text{target}}(x \cup s)$ and explanations $\phi(x \cup s)$, but not the ability to pass inputs to the model. The adversary's goal is to train an attack model f_{adv} that maps the explanations $\phi(x)$ to s on D_{aux}, an auxiliary dataset known to $\mathcal{A}dv$. (2) TM2 (without s in D): In this scenario, s is not included in the dataset D or the input x. Unlike TM1, $\mathcal{A}dv$ can pass inputs x to the model and has black-box access to f_{target} and $\phi(x)$, making this a more practical threat where s is censored for privacy. $\mathcal{A}dv$'s goal remains the same, to infer s by training f_{adv} on D_{aux}. For both models, the adversary has an additional auxiliary dataset D_{aux} that contains data records with non-sensitive and sensitive attributes along with their corresponding labels.

Threat models on Shapley values. Unlike previous assumptions [30,126] that adversaries have an auxiliary dataset with a distribution similar to the target sample, Luo et al. [13] explored two relaxed scenarios. The first adversary has access to an explanation vector, an auxiliary dataset, and a black-box prediction model, aiming to reconstruct the target sample. The second adversary operates under more practical constraints with only black-box access to the machine learning services and the explanation vector, without any background knowledge of the target sample.

Attacks on feature-based explanations. Duddu et al. [124] developed an attribute inference attack based on thresholding. The attack model f_{adv} uses model explanations to infer sensitive attributes and chooses the threshold t^* that maximizes the F1-Score. This calibration step deviates from using the typical default threshold of 0.5 to increase the precision and recall of the attack, particularly when there is a moderate to large class imbalance of the sensitive attribute s. Duddu et al. [124] also showed low Pearson correlation coefficients between the sensitive attribute s and other entities like y, x, and $\phi(x)$ across different datasets and explanation methods, suggesting little to no direct correlation between the sensitive attribute and the model's predictions or explanations, challenging the notion that the attack is merely exploiting these correlations.

Attacks on Shapley values. Luo et al. [13] proposed an attack where an adversary, with access to a black-box model f, attempts to infer private input features from Shapley value explanations. To simplify the computation of Shapley values, the adversary uses a reference sample x^0 and a linear transformation function h. They aim to reduce mutual information between the input x_i and the Shapley value s_i to zero, meaning the adversary cannot gain any information about x_i from s_i. Luo et al. [13] assumed that the Shapley values follow a Gaussian distribution, and thus the probability $P(s_i)$ is modeled as a Gaussian function. To ensure that the mapping from the auxiliary input data X_{aux} to the Shapley values S_{aux} is bijective, Luo et al. [13] presented a theorem requiring X_{aux} to be finite. The adversary can then use a hypothesis ψ to map Shapley values back to the auxiliary input data. To execute the attack, the adversary collects the Shapley values for all $x_{\text{aux}} \in X_{\text{aux}}$, sends prediction queries to the MLaaS platform, and obtains explanations S_{aux}. They then train a regression model on X_{aux} to learn the mapping ψ from Shapley values S_{aux} to X_{aux}.

Another scenario is where an adversary lacks an auxiliary dataset to carry out a feature inference attack [13]. Without knowledge of the target's data distribution, it becomes challenging to learn an attack model by observing Shapley values. To mitigate these challenges, the adversary can use the linear correlation between feature values and Shapley values for important features. By drawing samples independently and using a generalized additive model (GAM) for approximation, the adversary can restore features from Shapley values. Luo et al. [13] noted that while their attacks work well with Shapley values,

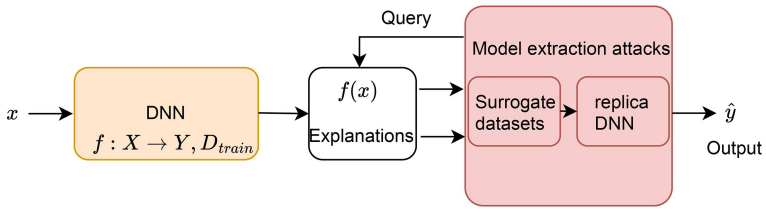

Figure 10 (Color online) Model extraction attacks.

other explanation methods like LIME and DeepLIFT may not be suitable due to their heuristic-based, unstable mappings between features and explanations.

3.5 Model extraction attacks

There is an increasing concern of model extraction attacks in machine learning as a service (MLaaS) [127], where attackers steal ML models using surrogate datasets to make queries through the MLaaS API, and then train replica models with the obtained predictions. The goal is to create a functionally equivalent version with identical predictions (see Figure 10). The difference between a model extraction attack and a model reconstruction attack is that the former does not need to know the model architecture.

Research on model extraction attacks targeting explainable AI systems is emerging. Milli et al. [107] developed a method that leverages the discrepancy in gradient-based explanations between an original AI model and its clone, demonstrating enhanced attack efficiency. Additionally, Aïvodji et al. [98] designed an attack utilizing counterfactual explanations to train a cloned model with greater effectiveness. Miura et al. [38] designed a data-free attack that does not require surrogate datasets in advance.

Threat models. An adversary duplicates a trained model, referred to as the victim model $f : X \to Y$, by utilizing its predictions to create a similar clone model $\hat{f} : X \to Y$. The adversary's goal is to replicate the victim model's accuracy using only the output predictions. On the one hand, typical model extraction attacks [107] involve the adversary collecting input data $x \in X$, querying the victim model to obtain predictions, and using the pairs $(x_i, f(x_i))$ to compile a dataset for training the clone model. In some scenarios, an adversary requires query access to the victim model but does not necessarily need the training data's ground-truth labels [88]. The attack relies on knowing the architecture of the victim model but not its parameter values. The attacker aims to produce a model that performs identically on the same test dataset, although the adversary's extracted model may not have been trained on the same data or in the same manner as the victim model.

On the other hand, data-free model extraction attacks [38] eliminates the need for input data collection, in which an adversary employs a generative DNN $G : \mathbb{R}^r \to X$ to convert Gaussian distribution noise into synthetic input data. The adversary then uses this data to query the victim model and gather training pairs $(x, f(x))$, which are used to train the clone model to emulate the victim model f. The generative model is designed to create data that, when predicted by the clone model, is different from the victim model's output, intending to maximize the clone model's loss function and improve parameter updates. Although the generative model G does not learn the actual distribution of the input data space X, it is optimized to produce data that facilitates the clone model's training process.

In the case of counterfactual explanations, the explanation API provides for each data point x_i, a corresponding counterfactual explanation $c(x_i)$, accompanied by the predicted outcome $\hat{y}_i$. When seeking a collection of diverse counterfactuals, the API will yield a collection $C(x_i)$ comprising multiple counterfactual instances, rather than just a single example.

Attacks on gradient-based explanations. In the data-free model extraction [38], an attacker crafts a surrogate model, denoted as $\hat{f} : X \to Y$, alongside a generative model $G : \mathbb{R}^r \to X$, responsible for creating synthetic data inputs. An iterative process is repeated between two steps. The first step generates N_G input samples and queries the target model to refine the generative model based on both predictions and explanations, utilizing these explanations to compute the gradient $\nabla_{\theta_G}\mathcal{L}$. The second routine produces N_C input samples for querying the target model and uses the resulting predictions to train the surrogate model. The process stops when the number of queries $(N_G + N_C)$ aligns with the allocated query budget Q. This strategy enables the attacker to leverage the gradient $\nabla G(x) = \nabla_x f(x)$ for the training of this generative model.

Adversarial attacks for model extraction without data rely on alternately calculating the gradients of

an objective function with the parameters of both a cloned model and a generative model. Training the clone requires calculating the gradient $\nabla_{\theta_f}\mathcal{L}$, achievable via back-propagation by the adversary. However, current methods do not provide the adversary with access to $\nabla_{\theta_G}\mathcal{L}$ for training the generative model. According to [38], it suffices to find $\nabla_x\mathcal{L}(x)$ as it leads to $\nabla_{\theta_G}\mathcal{L} = -\nabla_{\theta_G}G(z)\cdot\nabla_x\mathcal{L}(x)$. Unlike previous methods that only provided terms other than $\nabla_x f(x)$, the adversary now gains explanations through the standard Gradient $G(x) = \nabla_x f(x)$, enabling the computation of $\nabla_x\mathcal{L}(x)$ precisely. The adversary can employ almost any differentiable loss function for training the generative model.

Quan et al. [88] proposed another explanation-matching attack [107], focusing on replicating both the predictions and explanations of the original or victim model. The adversary's model minimizes two losses: the prediction loss (the difference in predictions between the two models) and the explanation matching loss (the difference in their explanations). The overall loss being minimized is a weighted combination of these two losses. Additionally, the method includes the use of LIME to ensure the interpretability of predictions matches that of the victim model.

Attacks on counterfactual explanations. Kuppa et al. [63] considered two main factors. (a) The auxiliary dataset D_{aux} should approximate the training set of f. This can be challenging if D_{aux} does not naturally follow the training distribution, but counterfactual explanations can provide samples from various classes that may bridge this gap. An attacker can iteratively query and obtain diverse class samples to better reflect the training set distributions. (b) Knowing the architecture of f can significantly enhance the fidelity of the extracted model. However, in realistic scenarios, attackers often lack this information, complicating the attack. To circumvent this obstacle, once data samples that mirror the training set are collected, knowledge distillation techniques are employed. This involves transferring insights from f to a surrogate model g. The knowledge transfer is quantified using a distillation loss, given by $L_{\text{Distill}}(f,g) = L_{\text{KL}}(P_f(x), P_g(x))$, where L_{KL} represents the Kullback-Leibler divergence loss. In this setup, the attacker leverages publicly available data and queries f, then applies the distillation loss to train g, thereby extracting the functionality of f.

Aïvodji et al. [98] proposed a model extraction attack [128] by compiling an attack set and training a surrogate model on the collected data from counterfactual samples. Counterfactual explanations typically change features with larger importance values to achieve the desired prediction, thus revealing the model's sensitive areas. However, this approach has limitations, such as the decision boundary shift issue caused using distant queries from the decision boundary as training samples [98]. This leads to an unstable substitute model and requires more queries to resolve, thus increasing the attack cost. Wang et al. [129] proposed a method called DualCF to mitigate this issue using pairs of CF and their corresponding counterfactual explanations (CCF) from the opposite class as training data. This helps to balance the substitute model's decision boundary and improve extraction efficiency. DualCF for a linear model is also discussed, illustrating that for binary linear models, it is possible to extract a substitute model with 100% agreement using CF and CCF pairs. While promising for linear models, extending this approach to nonlinear and complex models remains a challenge, and the effectiveness of DualCF in those scenarios is yet to be thoroughly evaluated [127].

4 Causes of privacy leaks

Research into the causes that lead to privacy leakage through model explanations has started to emerge in the past few years [2,14,15,30,71,88]. Certain types of explanations are prone to divulging data, often due to their inherent structure. For instance, case-based explanations, which utilize actual data points from the training set, can inadvertently reveal sensitive information [31,48]. Other explanations, such as surrogate models (e.g., support vector machine (SVM), linear classifiers), are relatively easy to leak their parameters by querying enough input/output data pairs [14,88,100].

4.1 Privacy leaks in counterfactual explanations

While counterfactual explanations aim to clarify AI decisions, they may inadvertently compromise privacy [12]. These explanations can give adversaries clues to manipulate the system, as seen in instances where the absence of a feature (like a savings account) leads to a better outcome than a suboptimal presence [12]. They provide insights into decision boundaries, potentially revealing model specifics and training data, such as feature splits in logical models, training points in k-nearest neighbors, or support vectors in SVMs. Moreover, the existence of multiple and varying-length counterfactuals for a single

data point could increase the ease of model theft, with longer, more complex counterfactuals potentially disclosing substantial model information with just one explanation.

Vo et al. [27] outlined essential privacy concepts relevant to public datasets. Identifiers are personal attributes capable of uniquely distinguishing an individual, such as names or government-issued numbers. Quasi-identifiers, while not individually unique, can collectively re-identify individuals; a mix of gender, birthdate, and ZIP code, for instance, can pinpoint 87% of American residents [94]. Sensitive attributes cover confidential information like salaries or medical records that need safeguarding to prevent personal or emotional harm. To protect against re-identification risks, public datasets need to undergo anonymization by removing direct identifiers, though vulnerability remains due to quasi-identifiers.

Example 2. In the given scenario from the FICO explainable ML dataset [12], the outcome of the credit evaluation could have shifted from negative to positive if one of the following conditions were met:

- # installment trades is less than 3 instead of 3.
- # revolving trades is less than 3 instead of 5.
- # trades with 60 days overdue and marked as derogatory in the public record is 0 instead of 2.
- # loans within 1 year is less or equal to 2 instead of 5.

Here, user privacy is violated as the exact values of the above sensitive attributes are revealed [12].

Diverse counterfactuals equip users with a range of actionable insights to potentially alter their outcomes favorably [72,130]. However, this also increases privacy risks as it may give away additional details that could be exploited [98]. Artelt et al. [15] identified a key problem with counterfactual explanations: their instability to minor input variations can lead to significantly different outcomes for similar cases. Addressing this, Ref. [131] proposed studying the robustness of counterfactual explanations and suggested using plausible rather than closest counterfactuals to enhance stability.

4.2 Causes of membership inference attacks

Membership inference attacks (MIAs) aim to predict whether a data point is in the training set or not [31]. The trade-off between explainability and privacy has been investigated and evaluated using membership inference attacks in [2, 14, 30, 71].

Global explainers. Naretto et al. [14] demonstrated that interpretable tree-based global explainers can increase the risk of privacy leakage. To explain f, an interpretable global surrogate classifier g is required to be trained to imitate the behavior of f, i.e., $g(X) = f(X)$. To compare the privacy exposure risk caused by f and g, two attack models are trained: one is learned by querying f, and the other queries g. It was found that the global explainer is more vulnerable to the membership inference attack model than the classifier [14], resulting in more privacy exposure.

Feature-based explanations. MIAs were also evaluated on feature-based explanations, including back-propagation and perturbation [30]. Backpropagation-based explanations were found to result in privacy leakage, which may be caused by high variances of explanations. A high variance of an explanation indicates that the point is close to the decision boundary and has an uncertain prediction, which is helpful for an adversary. Compared to backpropagation-based explanations, perturbation-based explanations are more robust to membership inference attacks. This might be because the query points are not used to train the model [30].

Repeated interaction. Kumari et al. [132] focused on repeated interactions. The author introduces attacks using explanation variance to infer data membership, modeled through a continuous-time stochastic signaling game. The study proves an optimal attack threshold exists, analyses equilibrium conditions, and uses simulations to assess attack effectiveness in dynamic settings.

Fairness. Pursuing fairness in explanations can also increase risks of privacy exposure [2]. When processing imbalanced data, fairness constraints require the model to memorize the training data in the smaller groups rather than learning a general pattern [2]. Such a way makes it easier for membership inference attacks to attack the model. Especially, when membership inference attacks are designed specifically for each group, they showed higher attack accuracy than that of a common membership inference attack for all groups [2]. Another study [31] also reports small groups in record-based explanations are more vulnerable to membership inference attacks than majority groups.

Influence of input dimension. Shokri et al. [30] evaluated how the input dimension influences the privacy risks of gradient-based explanations. Their experiments revealed that as the number of features grows (between 10^3 and 10^4), a correlation between gradient norms and training membership appears, indicating vulnerability to membership inference attacks. However, this effect is moderated by

the number of classes and is also dependent on model behavior, as overfitting can occur with too many features. While increasing the number of classes generally increases learning problem complexity, the actual impact on the correlation between gradient norms and membership depends on the specific range of features and the interval and amount of correlation vary.

Influence of overfitting. Yeom et al. [82] demonstrated that overfitting has a notable impact on the success of membership inference attacks. Shokri et al. [30] conducted tests varying the number of training iterations to achieve different levels of accuracy, in order to assess the effects of overfitting. Consistent with prior research on loss-based attacks, they found that their threshold-based attacks, which leverage explanations, are more effective when targeting overfitted models.

4.3 Causes of reconstruction attacks

Reconstruction attacks target on reconstructing the partial or complete training data. Ferry et al. [100] showed that post hoc explanations can disproportionately impact individual privacy, exacerbating risks for minority groups. This trend toward reduced privacy for minorities is also reflected in interpretability, as identified by Shokri et al. [30, 31, 81]. They discovered that the likelihood of discerning whether an individual's data was used in a model's training set from post hoc explanations is higher for outliers and certain minority groups that the model finds difficult to generalize. This increased risk is attributed to these groups being more frequently included in the generated explanations. Consequently, interpretability tools could inadvertently lead to greater information leakage about these already vulnerable groups.

Interpretable models enhance transparency but can inadvertently disclose information about their training data. Gambs et al. [106] used such data leakage to probabilistically reconstruct a decision tree's training set. The uncertainty within this reconstruction can be measured to determine how much information the model leaks. Ferry et al. [100,101] examined how optimal and heuristic decision trees and rule lists reveal information about their training data. The study finds that optimal models tend to leak less information than greedily-built ones for a given level of accuracy. It also notes significant variance in how much information individual training examples contribute to the overall entropy reduction, with some examples inherently leaking more information based on their position within the model's structure.

4.4 Causes of property inference attacks

Regularization techniques like dropout and ensemble learning can prevent models from memorizing private inputs, potentially reducing the risk of information leakage [122, 133]. Despite previous findings, Luo et al. [13] revealed that incorporating dropout in neural networks at varying rates (0.2, 0.5, 0.8) actually enhances the accuracy of certain attacks. This counterintuitive result is attributed to dropout preventing overfitting by smoothing the decision boundaries, which inadvertently benefits the attack. Nevertheless, a very high dropout rate (0.8) does decrease the success rates of one attack due to underfitting and increased randomness in the model, which disrupts the linearity between inputs and outputs.

Case-based explanation methods, often used in sensitive fields like medical diagnosis, risk privacy breaches when they share detailed visual data with unauthorized viewers [48]. To mitigate this, anonymization techniques must be applied to the images before they are shared, ensuring that the identity of individuals is not disclosed while still preserving the explanatory power and realism of the images. The anonymization process involves altering identity features in the latent vector to produce a privatized image, but there is no guarantee that other latent features do not inadvertently reveal the identity, especially if facial embeddings capture significant identifiable information.

4.5 Causes of model extraction attacks

Quan et al. [88] explored how model extraction attacks can benefit from explanations with fewer queries. A particular finding is that while certain explanation methods, such as Gradient, Integrated Gradient, and SmoothGrad, can be exploited to enhance attack efficiency, others like Guided Backprop and GradCam may result in poorer performance due to biases in gradient estimation.

While counterfactual explanations do not reveal the entirety of a cloud model's workings, their impact on security and privacy has been underestimated [12, 134, 135]. Some research argues that CFs only unveil a minimal amount of information, showing a limited set of dependencies for an individual instance which might seem insufficient for model extraction [64,136]. However, accumulating enough data through multiple queries can significantly facilitate the extraction process [129]. Aïvodji et al. [98] pioneers the use

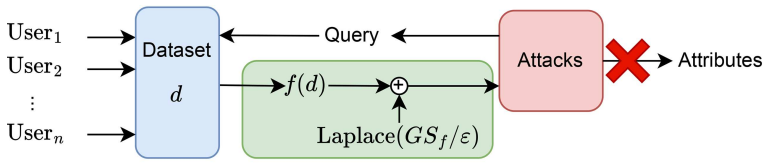

Figure 11 (Color online) Differential privacy.

of model extraction attacks on counterfactual explanations by treating these explanations near decision boundaries as supplementary training data. Wang et al. [129] also showed that adversaries can exploit CF explanations to extract a high-fidelity model by learning about the decision boundaries.

4.6 Causes of explanation linkage attacks

Vo et al. [27] reviewed key concepts relevant to data privacy, specifically in the context of public datasets. Identifiers are attributes that can uniquely identify an individual, like names or government numbers. Quasi-identifiers, while not unique on their own, can combine to uniquely identify a person. Sensitive attributes are confidential data that, if disclosed, could harm an individual. Public datasets are at risk of explanation linkage attacks, aka re-identification attacks, even after anonymization if quasi-identifiers are present [27]. Their experiments acknowledge that k-anonymity lowers the risks but it may still allow private information to be inferred through homogeneity and background knowledge attacks.

5 Privacy-preserving model explanations

5.1 Defences with differential privacy

Differential privacy (DP) is a solid, mathematical-based privacy standard that defines privacy loss using a quantifiable metric. It does so through mechanisms that guarantee the aggregated data output will obscure the involvement of any individual record in the dataset, as established by Dwork et al. [137]. Differential privacy is usually formalized as follows [93]. A randomized mechanism M with domain D and range R achieves ε-differential privacy (ε-DP) if, for all adjacent datasets d, d' differing by one row, and for any output set $S \subseteq R$, the following inequality holds:

$$\Pr[Q(d) \in S] \leqslant \mathrm{e}^{\varepsilon} \cdot \Pr[Q(d') \in S], \tag{13}$$

where ε is the privacy loss parameter, in which smaller values correspond to stronger privacy.

The Laplace mechanism of differential privacy is useful for queries on numerical data [93]. As shown in Figure 11, the mechanism adds noise to the sensitive query's output according to the Laplace distribution. Specifically, for a sensitive query function $Q(d)$, the ε-DP Laplace mechanism Q_{Lap} is given by $Q_{\mathrm{Lap}}(d) = Q(d) + \mathrm{Laplace}(\mathrm{GS}_Q/\varepsilon)$, where $\mathrm{Laplace}(\mathrm{GS}_Q/\varepsilon)$ represents a random variable from the Laplace distribution with a scale dependent on the global sensitivity GS_Q divided by ε. Global sensitivity GS_Q is the maximum norm-1 difference of Q across all pairs of adjacent datasets d, d'. Lastly, Dwork et al. [137] have demonstrated a post-processing property of differential privacy. If Q is ε-DP and G is any arbitrary deterministic mapping, then the composite function $G \circ Q$ is also ε-DP [93].

5.1.1 *Differentially private feature-based explanations*

An explanation $\phi(\cdot)$ is (ϵ, δ)-differentially private if the probability of any sequence of explanations does not change significantly with the addition or removal of a single data point in the training set [138]. For a sequence of queries $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k$, any two neighboring training sets $\mathcal{D}$ and $\mathcal{D}'$, and subsets $S_1, \ldots, S_k \subseteq \mathbb{R}^n$,

$$\Pr[\phi^1 \in S_1, \ldots, \phi^k \in S_k] \leqslant \mathrm{e}^{\epsilon} \cdot \Pr[\phi'^1 \in S_1, \ldots, \phi'^k \in S_k] + \delta, \tag{14}$$

where $\phi^i = \phi(\boldsymbol{z}_i, f_{\mathcal{X}}(\boldsymbol{x}))$ and $\phi'^i = \phi(\boldsymbol{z}_i, f_{\mathcal{D}'}(\boldsymbol{x}))$ for all i. The privacy for the explanation dataset $\mathcal{X}$ can follow a similar guarantee. Despite these measures, post hoc explanation algorithms, which are applied after the model has been trained, cannot fully prevent membership inference attacks, since they do not control the training process or parameters [138].

Single explanation algorithm. Patel et al. [138] focused on creating differentially private feature-based model explanations, where $\phi(\boldsymbol{z})$ is a vector in $\mathbb{R}^n$ that quantifies the impact of each feature on

the model's predicted label $f_{\mathcal{D}}(\boldsymbol{z})$. The aim is to find a local explanation function ϕ, centered at a point of interest $\boldsymbol{z}$, that minimizes the local empirical model error over an explanation dataset $\mathcal{X}$. The local empirical loss of ϕ over $\mathcal{X}$ is given by

$$\mathcal{L}(\phi, \boldsymbol{z}, f_{\mathcal{X}}) = \frac{1}{|\mathcal{X}|} \sum_{\boldsymbol{x} \in \mathcal{X}} \alpha(\|\boldsymbol{x} - \boldsymbol{z}\|)(\boldsymbol{x} - \boldsymbol{z})^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{z}) - f_{\mathcal{X}}(\boldsymbol{x})^2, \tag{15}$$

where α is a weight function that decreases with distance from $\boldsymbol{z}$. The optimal model explanation is the one that minimizes this loss:

$$\phi^*(\boldsymbol{z}, f_{\mathcal{X}}) = \arg\min_{\phi \in \mathcal{C}} \mathcal{L}(\phi, \boldsymbol{z}, f_{\mathcal{X}}). \tag{16}$$

To ensure differential privacy, Patel et al. [138] introduced a differentially private gradient descent (DPGD) algorithm, which utilizes the Gaussian mechanism to protect the explanation dataset $\mathcal{X}$. The privacy of the explanation dataset is protected by computing a private version of the gradient descent updates. The DPGD-Explain procedure iteratively updates ϕ using the gradient of the loss function perturbed by Gaussian noise, aiming to find the minimum of ϕ within a certain bound:

$$\phi^{(t+1)} \leftarrow \arg\min_{\phi \in \mathcal{C}_2, 1} \|\phi - \zeta^{(t)}\|, \tag{17}$$

where $\zeta^{(t)}$ is the perturbed gradient at iteration t. Patel et al. [138] provided conditions for bounded sensitivity for the gradient $\nabla\mathcal{L}(\cdot)$, which is crucial for the differential privacy guarantee. The authors specify a family of weight functions $\alpha(\cdot)$ that ensure the gradient sensitivity is bounded, which is a requisite for the differential privacy mechanisms employed. The authors also define a family of desirable weight functions $\mathcal{F}(\mathcal{C}, \boldsymbol{z})$ as those that are non-increasing and satisfy

$$\forall \boldsymbol{x} \in \mathbb{R}^n, \alpha(\|\boldsymbol{x} - \boldsymbol{z}\|) \leqslant \frac{c}{2\|\boldsymbol{x} - \boldsymbol{z}\|_2(\|\boldsymbol{x} - \boldsymbol{z}\|_2 + 1)}. \tag{18}$$

Adaptive algorithm for streaming explanation queries. Patel et al. [138] described an adaptive differentially private algorithm that involves sequentially explaining queries with the aid of differential privacy, using information from previously explained queries to optimize future explanations and manage the privacy budget. Key insights for this approach include reusing past explanations for similar new queries and ensuring that the initialization of the DPGD is as close as possible to the new query to achieve faster convergence and reduce privacy spending. The authors present a weight function $\alpha(\|\boldsymbol{x}-\boldsymbol{z}\|)$, defined as

$$\alpha(\|\boldsymbol{x} - \boldsymbol{z}\|) = \begin{cases} 1, & \text{if } \|\boldsymbol{x} - \boldsymbol{z}\| \leqslant r, \\ \frac{c}{2\|\boldsymbol{x}-\boldsymbol{z}\|_2(\|\boldsymbol{x}-\boldsymbol{z}\|_2+1)}, & \text{else.} \end{cases} \tag{19}$$

This weight function is used to identify points similar to $\boldsymbol{z}$ and is employed to ensure stable and consistent local explanations. Patel et al. [138] also introduced the idea of a non-interactive differential privacy mechanism to generate new explanations without additional privacy spending by constructing a proxy dataset from previous explanations.

5.1.2 *Differentially private counterfactual explanations*

Mochaourab et al. [28] developed a differentially private support vector machine (SVM) and introduced methods for generating robust counterfactual explanations. Yang et al. [139] created a differentially private autoencoder to produce privacy-preserving prototypes for each class label, optimizing perturbations to the input data that minimizes the distance to the counterfactual while favoring a specific class outcome. Hamer et al. [140] suggested data-driven recourse directions could be privatized, but does not elaborate on providing private multi-step recourse paths. Huang et al. [93] proposed generating privacy-preserving recourse using a differentially private logistic regression model but did not detail the provision of a multi-step path for recourse. Pentyala et al. [141] was a pioneer to offer a complete privacy-preserving pipeline that provides counterfactual explanations with differential privacy guarantees. Huang et al. [93] outlined a methodology for incorporating DP into logistic regression classifiers to offer recourse against membership inference (MI) attacks. Logistic regression is described with weights w that output a probability score $f(x) = w^{\mathrm{T}}x = \log \frac{P(y=1|x)}{1-P(y=1|x)}$. The counterfactual distance for instance x from the target score s

in logistic regression space is given by $c(x, x') = \frac{s-f(x)}{\|w\|_2^2}$. The decision boundary is set at $s = 0$, meaning that $P(y = 1|x)$ is 0.5 at the threshold. In particular, Huang et al. [93] introduced two DP methods for recourse generation.

• Differentially private model (DPM). It involves training the logistic regression classifier with DP. An ϵ-DP logistic regression model leads to ϵ-DP counterfactual recourse, using IBM's diffprivlib library [142] based on Chaudhuri et al.'s mechanism for DP empirical risk minimization [143, 144].

• Differentially private laplace recourse (LR). A new method is proposed for DP post hoc computation of counterfactual recourse that does not touch the underlying logistic regression model training process. It involves: (1) applying Laplace noise to the predicted probability score $\Pr'(y = 1|x) = \Pr(y = 1|x) + \text{Laplace}(1/\varepsilon)$, (2) clamping $\Pr'(y = 1|x)$ to $[0, 1]$, (3) computing the noisy logistic regression score $f'(x)$ based on $\Pr'(y = 1|x)$, and (4) calculating the noisy CFD as $c'(x, x') = \frac{s-f'(x)}{\|w\|_2^2}$.

Huang et al. [93] claimed that these methods are ϵ-DP. This is explained by starting with applying Laplace noise to the predicted probability and noting that the global sensitivity $\text{GS}_{p(y=1|x)}$ is 1. The process from calculating $\Pr(y = 1|x)$ to $M_{\text{CFD,Lap}}(x)$ is argued to be a post-processing step that retains ϵ-DP, according to the post-processing invariance property of DP [137]. Pawelczyk et al. [71] proposed that applying DP to a recourse generation algorithm can limit an adversary's balanced accuracy, with a bound expressed as $BA_A \leqslant \frac{1}{2} + \frac{1}{2} \cdot \mathrm{e}^{-\epsilon}$, where ϵ is the privacy loss parameter. However, the authors also acknowledge that while DP offers robust privacy assurances, it is not a fail-safe measure and can significantly reduce accuracy, posing a challenge in maintaining the utility of the explanation. Pentyala et al. [141] proposed "PrivRecourse", a framework for generating privacy-preserving counterfactual explanations. The method relies on a two-phase approach: a training phase and an inference phase. The training phase involves training a differentially private ML model f, clustering the dataset into K subsets with (ϵ_k, δ_k)-DP guarantees, and constructing a graph G with clusters as nodes [145, 146]. Nodes are connected by edges based on distance and density without violating actionable constraints, and the entire graph is published ensuring (ϵ, δ)-differential privacy [137, 147]. During the inference phase, for any query instance Z, a recourse path P and a counterfactual instance Z^* that would flip the model's decision to a favorable outcome are computed. This is done by first identifying the nearest node Z_1 to Z in G, and then using Dijkstra's algorithm to find the shortest path to the favorable counterfactuals in Z_{CF} [148].

Hamer et al. [140] proposed another framework to generate counterfactuals, called the stepwise explainable paths (StEP). The framework begins by partitioning the dataset X into k clusters $\{X_1, \ldots, X_k\}$. For a point of interest $\tilde{x}$, if the model prediction $f(\tilde{x}) = -1$ indicating an unfavorable outcome, StEP generates a direction $\tilde{d}_c$ for each cluster using the formula:

$$\tilde{d}_c = \sum_{x' \in X_c} (x' - \tilde{x})(\alpha(||x' - \tilde{x}||)f(x') = 1), \tag{20}$$

where α is a non-negative function, and $||\cdot||$ is a rotation invariant distance metric [35]. This process repeats iteratively, with the user updating their point of interest $\tilde{x}$, until a favorable outcome is achieved. StEP can be adapted to satisfy (ϵ, δ)-differential privacy by adding Gaussian noise to the directions computed. When the distance metric is the ℓ_2 norm, the sensitivity of StEP is upper-bounded by a constant C, and therefore, Gaussian noise with a mean of 0 and standard deviation $\sigma \geqslant \frac{C^2\beta}{\epsilon}$ where $\beta \geqslant 2\log(1.25/\delta)$ can be added to each feature to achieve differential privacy. When multiple directions are provided to a user, and each is (ϵ, δ)-differentially private, the overall mechanism is $(k\epsilon, k^\delta)$-differentially private [137].

Yang et al. [139] proposed another DP-based method through the use of a functional mechanism. The functional mechanism does not add noise directly to the optimal parameter set w^*, but to the loss function $\tilde{L}_D(w)$ by injecting Laplace noises into the coefficients of its polynomial representation. The process involves constructing class prototypes in the latent space using a well-trained autoencoder and the functional mechanism through a perturbed training loss. Counterfactual samples are then searched for in the latent space based on these prototypes. Yang et al. [139] provided that if the prototype construction process is ϵ-differentially private, then the counterfactual explanation process also satisfies DP under the same privacy budget ϵ. This relies on the post-processing immunity of DP [137], which allows for certain noises to be added in the prototype construction process without further affecting subsequent computations.

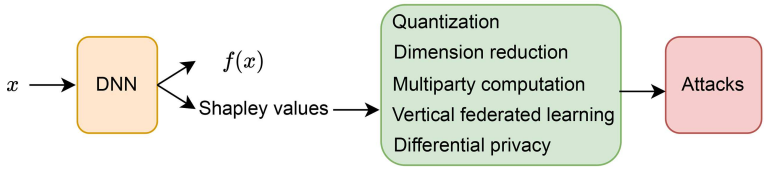

Figure 12 (Color online) Defences with privacy-preserving Shapley values.

5.1.3 *DP-locally linear maps*

To create differentially private LLM, Harder et al. [29] employed the moments accountant technique combined with differentially private stochastic gradient descent (DP-SGD) [147]. The perturbation process involves two main steps per iteration for each minibatch of size L. (1) Clipping the norm of the datapoint-wise gradient $h_t(x_n)$ using a threshold C and adding Gaussian noise to it, resulting in $\hat{h}_t$: $\hat{h}_t \leftarrow \frac{1}{L}\sum_{n=1}^{L} h_t(x_n) + \mathcal{N}(0, \sigma^2 C^2 I)$. (2) Updating the LLM parameters in the descending direction: $W_{t+1} \leftarrow W_t - \eta \hat{h}_t$. This process ensures that the final LLM is (ϵ, δ)-differentially private. To improve the privacy-accuracy trade-off, especially for high-dimensional inputs like images, Ref. [29] suggested reducing the dimensionality of the parameters by first projecting them onto a lower-dimensional space using a shared matrix R_m, and then perturbing the gradients of the projected parameters.

5.2 Defences with privacy-preserving SHAP

Several studies have focused on preserving the privacy of users from explanation using Shapley values, including quantization, dimension reduction, multi-party computation, federated learning, and differential privacy (see Figure 12).

Quantized Shapley values. Luo et al. [13] proposed quantization of Shapley values to protect privacy by reducing mutual information between input features and their corresponding Shapley values. By restricting the Shapley values to a set number of discrete levels (e.g., 5, 10, or 20 distinct values), the entropy of the Shapley values $H(s_i)$ and hence the mutual information $I(x_i; s_i)$ can be reduced. While quantization has minimal effects on the effectiveness of one attack strategy, it does compromise the accuracy and success rate of another due to the increased range of candidate estimations for a feature, leading to larger estimation errors as per the bounds established earlier. Quantization might also result in two different input samples yielding the same explanation, which disturbs the privacy-utility balance.

Low-dimensional Shapley values. Luo et al. [13] discussed a defensive strategy by suggesting a reduction in the dimensionality of Shapley values. Since the number of Shapley values for a class corresponds to the number of input features, the defence involves only releasing the Shapley values of the top k features based on their variance, rather than their magnitude.

Multi-party Shapley values. Jetchev et al. [42] introduced secure multiparty computation (MPC), which allows multiple parties to jointly evaluate a public function on their private data without revealing anything other than the function's output. The authors developed a privacy-preserving algorithm, XorSHAP, which operates on top of the Manticore MPC framework. This algorithm is a variant of the TreeSHAP method and retains agnosticism toward the underlying MPC framework. Jetchev et al. [42] discussed the secret sharing of binary decision trees within an MPC setting, where decision trees can be shared secretly and then used in the computation of privacy-preserving algorithms like XorBoost. Jetchev et al. [42] proved that all subsequent operations and variables in the algorithm are secret and data-independent.

Federated Shapley values. Wang [149] discussed interpreting models in the context of vertical federated learning (VFL) where different parties possess different slices of the feature space. Traditional model interpretation methods like Shapley values can reveal sensitive data across parties, making it unsuitable for VFL. To address this, a variant called SHAP Federated is proposed for VFL, particularly for dual-party scenarios involving a host and guest. The host and guest collaboratively develop a machine learning model, with the host owning the label data and part of the feature space, and the guest owning another part. The algorithm involves setting values in the instance x to their original or reference values based on whether a feature is hosted or federated and encrypting IDs when necessary to maintain privacy. Then, predictions are made for each combination of features, and feature importance is calculated from the aggregated prediction results using Shapley values. Features that cannot handle missing values are set to either NA or the median [5].

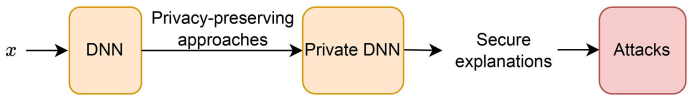

Figure 13 (Color online) Defences with privacy-preserving ML models.

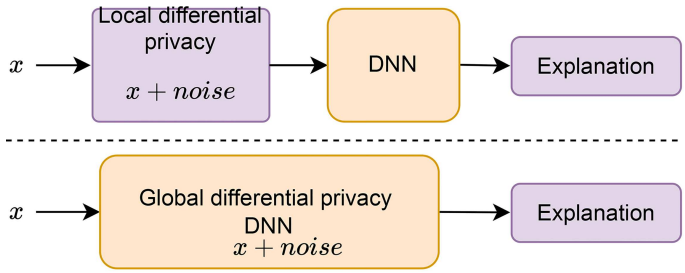

Figure 14 (Color online) Local and global differential privacy schemes proposed in [151].

Differentially private Shapley values. Luo et al. [13] pointed out that DP is not suitable for local interpretability methods. For DP to be effective, the explanations for any two different private samples must be indistinguishable, which would reduce the utility of Shapley values as they would become too similar across different samples. As a result, DP cannot be applied to the current problem of maintaining interpretability while defending against attacks that leverage Shapley values.

Watson et al. [150] discussed the computational challenges of calculating Shapley values due to their expensive nature and the privacy concerns in using large portions of datasets for each query. Watson et al. [150] introduced an estimation algorithm that utilizes only a small fraction of data, taking advantage of the property that larger datasets reduce the marginal contributions of individual data points, which are proportionally smaller. The algorithm is shown to satisfy ϵ-differential privacy with a coalition sample complexity of $O(\ln(n))$ [150]. Watson et al. [150] emphasized the cost advantages of the layered Shapley approach, which uses fewer data points and has lower computational and data access costs, offering privacy benefits.

5.3 Defences with privacy-preserving ML models

To protect user privacy, privacy-preserving ML models have been trained to resist against attacks (see Figure 13). Naidu et al. [151] discussed two primary models of implementing differential privacy: Local DP, where noise is added directly to user data before it is shared, ensuring data privacy against untrusted parties; and Global DP, where a trusted central entity applies differentially private algorithms like DP-SGD [147] to the collected data to produce models or analyzes with limited information leakage (see Figure 14). Interpreting models trained with differential privacy is challenging due to the noise added during training, which obfuscates the model's decision-making process [138]. Naidu et al. [151] investigated the interpretability of differentially private models by establishing the first benchmark for interpretability in deep neural networks (DNNs) trained with differential privacy.

Liu et al. [89] developed a model-level defense by employing differentially-private stochastic gradient descent (DP-SGD) [152], to build inherently private models. The process involves the automatic configuration of gradient clipping and the selection of 'MixOpt' as the clipping model, uniformly applied across all model layers. While DP-SGD can reduce the effectiveness of membership inference attacks, it also significantly decreases classification accuracy, even with a large epsilon ϵ. Findings indicate that attribution maps become less informative than even methods not considering model parameters [153]. This underscores the challenge of balancing between defense capability and performance utility, as effective defense mechanisms like DP-SGD can significantly impact model accuracy and the quality of explanations provided.

Mochaourab et al. [28] outlined a method for providing differential privacy to SVM classifiers by perturbing the optimal weight vector w^* with additive Laplace noise. The perturbed weight vector $\tilde{w}$ is given by $\tilde{w} := w^* + \mu$, where μ consists of i.i.d. Laplace random variables $\mu_i \sim \text{Lap}(0, \lambda)$. This perturbation ensures β-differential privacy for $\lambda \geqslant 4C_k\sqrt{F}/(\beta n)$, with certain conditions on the kernel function ϕ. Mochaourab et al. [28] introduced robust counterfactual explanations for SVM classifiers, providing explanations for classification results that account for the uncertainty introduced by the differential privacy

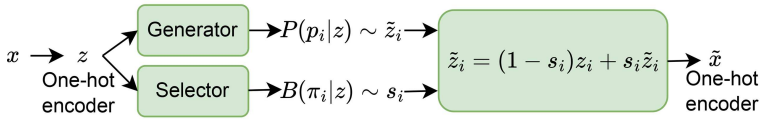

Figure 15 (Color online) Approach of generating diverse counterfactuals to reduce the privacy risk in re-identification [27].

mechanism. For the optimization problem, a root of the function g, defined as

$$y' f_\phi(x, \tilde{w}) - \lambda\sqrt{2\ln(2/(1-p))}\|\phi(x)\| \leqslant 0, \tag{21}$$

is considered a robust counterfactual explanation. Efficient solutions to this optimization problem are proposed using convex optimization solvers like CVXPY for linear SVM or a bisection method for non-linear SVM. The solution implies that a domain expert's input is required to determine prototypes representing each class when direct access to test data is not available due to privacy considerations. A bisection method used for finding robust counterfactual explanations in non-linear SVMs is also developed [28]. In a similar setting, Veugen et al. [154] used local foil trees to explain the decisions of a black-box model without accessing its training data. By generating synthetic data points that are close to the user's data point, classifying them through the model, and then training a decision tree in a secure manner, the method constructs explanations in terms of feature thresholds [155]. This process utilizes secret-shared data and secure multi-party computation [156] to ensure that no sensitive information from the model or its training data is disclosed, except for the minimal necessary details required to provide the user with an explanation for the classification outcome.

5.4 Defences with perturbations

Jia et al. [157] introduced a defence technique called MemGuard, differing from other strategies that modify the training process. MemGuard cleverly injects perturbations into the confidence scores produced by the model for each input, transforming these altered scores into adversarial examples aimed at misleading attack models. However, the primary limitation of MemGuard is its focus on distorting the model's output by adding noise, which does not protect the attribution maps, thus failing to completely deter the attacks [89]. Vo et al. [27] described a methodology for addressing the trade-off between diversity and sparsity in the features modified to form a counterfactual. As shown in Figure 15, it introduces a local feature-based perturbation distribution $P(\tilde{z}_i|z)$ for each mutable feature z_i, along with a selection distribution Bernoulli$(\pi_i|z)$ to control sparsity. To form a counterfactual example $\tilde{z}$, the method samples from these distributions and updates mutable features, maintaining validity by maximizing the likelihood of the counterfactuals to alter the original outcome.

Olatunji et al. [158] discussed a defence mechanism for feature-based explanations. It involves perturbing each explanation bit, where an explanation is represented as a bit mask, using a randomized response mechanism. The perturbation probability for flipping each bit $\mathcal{E}_{xi}$ is determined by a privacy budget ϵ:

$$\Pr(\mathcal{E}'_{xi} = 1) = \begin{cases} \frac{\mathrm{e}^\epsilon}{\mathrm{e}^\epsilon+1}, & \text{if } \mathcal{E}_{xi} = 1, \\ \frac{1}{\mathrm{e}^\epsilon+1}, & \text{if } \mathcal{E}_{xi} = 0, \end{cases}$$

where $\mathcal{E}_{xi}$ and $\mathcal{E}'_{xi}$ are the true and perturbed ith bits of explanation, respectively. This method ensures $d\epsilon$-local differential privacy for an explanation with d dimensions.

5.5 Defences with anonymisation

k-anonymity. Goethals et al. [10] presented a unique application of k-anonymity aimed at ensuring anonymity within counterfactual explanations, as opposed to anonymizing an entire dataset. This approach is particularly relevant when the dataset is not intended to be fully public. Goethals et al. [10] defined a counterfactual instance as k-anonymous if its quasi-identifiers, the partially identifying attributes, could apply to at least k individuals within the training set. In turn, a counterfactual explanation inherits this k-anonymity if it is derived from such a k-anonymous instance. However, while counterfactual explanations usually aim to change the outcome of a model's prediction, k-anonymous counterfactuals can include a range of instances beyond those used to generate the explanation, leading to uncertainty about whether all values in this range would lead to a change in the prediction.

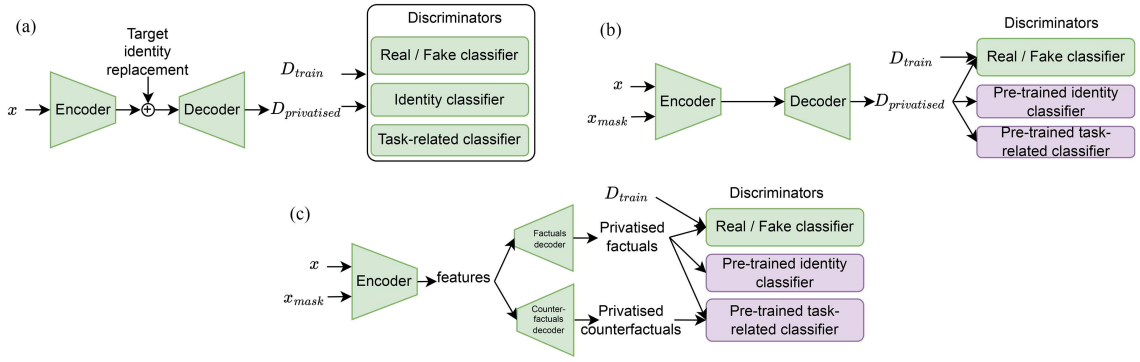

Figure 16 (Color online) Three approaches to defend with anonymization. (a) The PPRL-VGAN model proposed in [160]. (b) The WGAN-GP framework using the pre-trained identifier and task-related classifier [159]. (c) Privatised counterfactual samples are generated by a counterfactual decoder [159]. x_{mask} is the mask data, and $D_{\text{privatised}}$ is the generated privatised data.

Privatised factual samples. Montenegro et al. [48] argued that an explanation should not reveal sensitive personal identity information while remaining realistic and informative regarding the decision-making process. The author outlines an optimization objective which involves minimizing three loss functions, one for privacy, one for realism, and one for explanatory evidence, each weighted by a non-negative parameter. The distance between a privatized image and the source image is minimized, ensuring that the privatized image is sufficiently different from any identity in the training data to preserve anonymity. In another setting, Montenegro et al. [159] developed a privacy-preserving network with multi-class identity recognition designed for case-based explanations. The network seeks to preserve privacy by promoting a uniform distribution across identities, making identity recognition akin to random guessing. The PPRL-VGAN model [160] (see Figure 16(a)), which intentionally collapses to the replacement identity and task-related class, is replaced with a WGAN-GP framework that uses a Wasserstein loss with a gradient penalty to stabilize the discriminator (see Figure 16(b)). This change, alongside using interpretability saliency maps for reconstruction of relevant task-related features, aims to retain the explanatory value in the privatized images [161]. Montenegro et al. [159] also introduced another privacy-preserving network that utilizes a Siamese identity recognition framework to enhance privacy in domains with scarce images per subject. They employ a contrastive loss function for training, defined as $\text{ContrastiveLoss} = \frac{1}{2} \times Y \times \text{ED}^2 + \frac{1}{2} \times (1-Y) \times [\max(0, m - \text{ED})]^2$, where Y is the label indicating if the image pair is of the same identity, ED is the Euclidean distance between embeddings, and m is a margin. The Siamese network ensures the privatized image is distinct in identity from the original and others in the dataset.

Privatised counterfactual samples. Montenegro et al. [159] also generated counterfactual explanations from the privatized samples. As shown in Figure 16(c), a counterfactual generation module, in the form of a decoder, is added to the above privacy-preserving network to map an image's latent representation to its counterfactual. This decoder is designed to make minimal alterations to the privatized factual explanations to change their predicted class, thereby minimizing the pixel-wise distance between the factual and counterfactual explanations while altering the image's task-related prediction. Saliency masks and explanatory features are used to guide changes to image regions that are relevant to the explanation. The loss function for the counterfactual decoder training is represented as $L_C = E_{I,M\sim p_{\text{data}}}[\lambda_x[F(I) \times (1-M) - C(I) \times (1-M)]^2 + \lambda_D \text{Exp}(D_{\text{exp}}(I) \times \log(1 - D_{\text{exp}}(C(I))))]$, where $F(I)$ and $C(I)$ denote the privatized factual and counterfactual explanations, respectively, and λ_x and λ_D are weights controlling the importance of each term in the loss function.

6 Future research directions

6.1 Ethical implications

The push for explainable AI has led to the development of tools and startups like MS InterpretML, Fiddler Explainable AI Engine, IBM Explainability 360, Facebook Captum AI, and H2O Driverless AI [162]. Our survey explores the privacy risks of making ML models explainable, highlighting the potential for malicious exploitation of these explanations, especially for high-risk data such as medical

records and financial transactions. This raises concerns about the conflict between the right to explain and user privacy [1], necessitating discussions involving legal experts and policymakers [26]. Additionally, the tension between explainability and privacy may disproportionately impact minority groups by either exposing their data or providing lower-quality explanations [30].

This survey contributes to the broader research on AI transparency and privacy, igniting discussions on AI governance. The trade-off between privacy and explainability is a well-known legal issue [163], but we aim to develop explanation methods that protect user privacy, even if they impact explanation quality. Though explanation quality is subjective, explanations that do not provide useful model insights while protecting user data are likely less beneficial to end-users [30].

Looking into the future, the ethical implications of privacy-preserving techniques include balancing privacy protection with transparency and fairness [164]. Techniques like differential privacy and federated learning secure data by adding noise or decentralizing processing, but they can reduce model accuracy and transparency, complicating trust and understanding [165, 166]. These methods can also introduce biases, affecting certain groups disproportionately and amplifying discrimination [167]. Ensuring informed consent and user autonomy is crucial, necessitating clear communication about how these techniques impact data use and model performance [168].

6.2 Regulatory compliance

Privacy attacks on model explanations pose significant challenges under regulatory frameworks like the GDPR, which emphasizes the protection of personal data and transparency in automated decision-making. Such attacks can lead to unauthorized data disclosure, complicating compliance with GDPR's requirements for data subject rights, including access and erasure [169, 170]. Additionally, privacy-preserving techniques that obscure model explanations may hinder transparency, making it difficult for organizations to demonstrate compliance and for individuals to understand AI decisions, thereby affecting accountability [171]. Moreover, these techniques must balance privacy and utility, as overly restrictive measures can impact the effectiveness and fairness of AI systems, posing further challenges for legal and ethical standards [168].

6.3 Privacy tradeoffs

Li et al. [172] discussed the impact of differential privacy on the interpretability of deep neural networks. It examines how injected noise into the model parameters affects the gradient-based interpretability method. The analysis reveals that while noise in the fully connected layer directly affects the feature map used for interpretability, noise in the convolutional layer alters the output of the activation function, thus impacting the feature map indirectly. Chang et al. [2] examined the relationship between algorithmic fairness and privacy. It points out that while fair machine learning models strive to reduce discrimination by equalizing behavior across different groups, this process can alter the influence of training data points on the model, leading to uneven changes in information leakage. Fair algorithms may inadvertently memorize and leak more information about under-represented subgroups in an attempt to equalize errors across different groups based on protected attributes. The findings indicate a trade-off where achieving fairness for protected or unprivileged groups amplifies their privacy risks. Moreover, the greater the initial bias in the training data, the higher the privacy cost when making the model fair for these groups. These findings are relevant to model explanations, which also impact fairness [65, 173].

6.4 Underexplored privacy attacks

Aïvodji et al. [47] presented techniques for manipulating and detecting manipulation of SHAP values. To manipulate SHAP values, a brute-force sub-sampling method is used to minimize the differences in SHAP values, with a clever re-weighting strategy to make the sampling appear legitimate. Detection of such manipulation employs statistical tests to compare model outputs from manipulated and unmanipulated samples [174]. Slack et al. [41] outlined a framework for constructing adversarial classifiers that deceive post hoc explanation techniques, such as LIME and SHAP. The framework produces an adversarial classifier that mimics the biased classifier on real distribution data but reverts to unbiased predictions on out-of-distribution (OOD) data [175]. Regarding data reconstruction attacks, an interesting direction is to utilize the inner workings of learning algorithms in some interpretable models (e.g., decision trees) to reduce the entropy of probabilistically reconstructed datasets. For example, since greedy algorithms

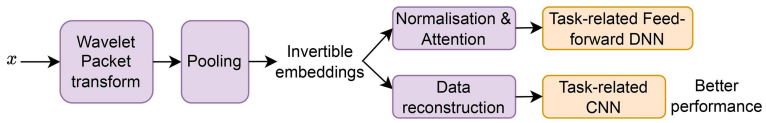

Figure 17 (Color online) "DeepFixCx" model uses compression techniques (i.e., wavelet packet transform and a pooling function) for preserving privacy and explicability [188].

for constructing decision trees select features based on Gini impurity, we can identify and discard certain attribute combinations that do not contribute to an optimal decision tree [100].

6.5 Underexplored model explanations

Gillenwater et al. [176] introduced a novel method for computing multiple quantiles in sensitive data with differential privacy. Traditional methods compromise on accuracy by either splitting the privacy budget across quantiles or inefficiently summarizing the entire distribution. The proposed approach uses an exponential mechanism to estimate multiple quantiles efficiently, achieving better accuracy and efficiency compared to existing methods. This is particularly relevant because there are emerging explainability measures based on quantiles [177–179]. Alvarez et al. [180] proposed the concept of self-explaining models that incorporate interpretability from the onset of learning. The authors design self-explaining models in a stepwise manner, starting from simple linear classifiers and advancing to more complex structures with built-in interpretability [181]. They introduce specialized regularization techniques to maintain faithfulness and stability. Olatunji et al. [158] pioneered the examination of privacy risks tied to feature explanations in graph neural networks (GNNs), presenting scenarios where adversaries attempt to unveil hidden relationships within the data, despite having limited access to the network's structure [182]. The paper delves into various explanation methods for GNNs such as gradient-based, perturbation-based, and surrogate methods. Furthermore, it outlines potential adversarial attacks aimed at exploiting these explanations to compromise privacy and introduces a novel defense mechanism based on perturbing explanation bits to adhere to differential privacy standards. Other studies [183, 184] examine the role of knowledge graphs as model explanations, positing that integrating structured, domain-specific knowledge can lead to more understandable, insightful, and trustworthy AI systems. However, knowledge graphs can be used to fuel privacy attacks such as de-anonymization and membership inference [185, 186].

6.6 Underexplored countermeasures

Domingo-Ferrer et al. [187] presented methods for collaborative rule-based model approximation without the direct use of a model simulator. It suggests that users can employ simulators to interact with a concealed model to obtain responses for certain feature sets, which although limited and controlled, can help deduce how the model makes decisions. While simulators prevent full transparency of the model and often limit the number of queries to prevent misuse, users can collaborate by querying the model for various feature sets and publishing the predictions. This collective data can then be mined for decision rules to approximate the model's logic.

Gaudio et al. [188] proposed the "DeepFixCx" model, an approach that utilizes wavelet packet transforms and spatial pooling for image compression that preserves privacy and explicability (see Figure 17). The method relies on analyzing images with multi-scale wavelet-based methods, allowing local regions of pixels to be summarized at multiple scales. The wavelet packet transform offers several benefits, such as facilitating image processing with deep learning libraries, ensuring that all coefficient values represent equally-sized pixel regions, and maintaining consistency with boundary effects. "DeepFixCx" provides a trade-off between compressing images for efficiency while still retaining enough detail for reconstruction and privacy preservation. Gaudio et al. [188] also outlined methods for inverse wavelet packet transform for image reconstruction, which can restore images from compressed representations to their original size. This model offers a privacy-conscious method to process images for various applications, including medical imaging, by removing local spatial information, allowing for the preservation of privacy without the need for additional learning.

6.7 Underexplored data modalities

Graph data. The rapid development in the area of GNNs highlights a special treatment for GNN explainability [189, 190]. Yuan et al. [191] discussed explainability methods specifically designed for graph

neural networks (GNNs) such as gradients/features-based, perturbation-based, surrogate, and decomposition methods. Prado-Romero et al. [192] provided a comprehensive overview of graph counterfactual explanations for GNNs. Privacy attacks on GNNs are also an emerging direction [193].

Audio data. Audio signals consist of speech signals and other non-speech audio signals. Speech processing involves tasks like automatic speech recognition, speaker identification, and paralinguistic information recognition, while non-speech audio signal processing contains many more applications, such as human heart sound analysis, bird sound analysis, and environmental sound classification. Current research has separately focused on data/model privacy and explanation approaches [194–197]. While explainable models are essential for audio-based healthcare applications [198–200], there is still a large gap to further explore the privacy risks of audio-based model explanations.

6.8 Privacy-preserving models

Exploring how privacy-preserving models, such as differentially private decision trees, reduce the success of privacy attacks represents a valuable research direction [100,171,201]. Li et al. [172] presented an adaptive differential privacy (ADP) mechanism aimed at improving the interpretability of machine learning models without compromising privacy. This mechanism selectively injects noise into the less critical weights of a model's parameters, thereby preserving the interpretability of important features which conventional differential privacy methods may obscure [202, 203].

6.9 Privacy-protecting explanations

Using model explanations to counter adversarial attacks is a novel direction. Belhadj-Cheikh et al. [204] outlined a framework (called FOX) to safeguard social media users' privacy using adversarial reactions to trick classifiers. It constructs a dataset of social media interactions, employs an explainability tool to extract influential adversarial features, and filters them to create a robust list. These features are then used to generate adversarial reactions, aiming to mislead the classifier away from the correct classification and toward a predetermined label, thus preserving the user's privacy.

6.10 Time complexity

Time complexity is crucial in privacy attacks on model explanations. Fast run-time methods pose higher risks by enabling rapid exploitation, while more complex iterative attacks are less practical due to longer execution times. The feasibility of these attacks depends on computational resources and scalability. Effective countermeasures must balance protection and performance to mitigate risks from fast, real-time attacks. Unfortunately, only a few studies thoroughly discuss time complexity such as Shapley approximation [205] and DP-quantiles [176].

7 Conclusion

Summary. As model explanations become more prevalent, interest in their impacts, such as fidelity, fairness, stability, and privacy, is growing. This survey thoroughly investigates recent privacy-centric attacks on model explanations, classifying them based on their characteristics. It also explores advanced research on defensive strategies and privacy-focused model explanations, identifying common privacy design approaches and their variations.

Our survey highlights several unresolved issues needing further investigation. First, current research is limited, focusing mainly on membership inference attacks, counterfactual explanations, and differential privacy, while many widely used algorithms and models require more detailed scrutiny. Second, there is a lack of deep theoretical insight into the origins of privacy breaches, affecting the development of protective measures and understanding of privacy attack limitations. Experimental research has provided valuable knowledge, but studies evaluating attacks under realistic conditions are scarce. This survey aims to be a crucial resource for readers interested in exploring the privacy implications of model explanations.

Challenges. The challenges for new work in this field, as highlighted in the survey, include the following.

• Balancing transparency and privacy. Providing detailed explanations improves transparency but increases the risk of privacy breaches by revealing sensitive information embedded in the training data.

• Granularity of explanations. Detailed explanations can lead to direct inferences about data points, making it challenging to protect privacy without losing interpretability.

• Understanding privacy leaks. Identifying the causes of privacy leaks through model explanations is complex and requires a thorough investigation of different explanation methods and their vulnerabilities.

• Diverse attack models. Developing comprehensive defenses against a wide range of privacy attacks, including membership inference, model inversion, and reconstruction attacks, is necessary but challenging due to the evolving nature of these attacks.

• Countermeasure effectiveness. Evaluating and improving the effectiveness of countermeasures, such as differential privacy and perturbation techniques, to ensure they do not compromise the utility of model explanations.

• Dynamic interaction scenarios. Assessing the impact of repeated interactions between adversaries and the model in dynamic settings adds complexity to designing robust privacy-preserving methods.

• Interpretable surrogates. Surrogate models used for providing explanations can themselves become targets for privacy attacks, necessitating additional safeguards.

• Scalability and practicality. Implementing privacy techniques in the real-world must balance scalability and practicality without significantly affecting model performance.

Acknowledgements This work was supported by ARC Discovery Early Career Researcher Award (Grant No. DE200101465) and ARC DP Project (Grant No. DP240101108).

References

1 Goodman B, Flaxman S. European Union regulations on algorithmic decision making and a "right to explanation". AI Mag, 2017, 38: 50–57
2 Chang H, Shokri R. On the privacy risks of algorithmic fairness. In: Proceedings of IEEE European Symposium on Security and Privacy, 2021. 292–303
3 Ancona M, Ceolini E, Oztireli C, et al. Towards better understanding of gradient-based attribution methods for deep neural networks. In: Proceedings of International Conference on Learning Representations, 2018
4 Ribeiro M T, Singh S, Guestrin C. "Why should I trust you?" explaining the predictions of any classifier. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2016. 1135–1144
5 Lundberg S M, Lee S-I. A unified approach to interpreting model predictions. In: Proceedings of Conference on Neural Information Processing Systems, 2017
6 Guidotti R. Counterfactual explanations and how to find them: literature review and benchmarking. Data Min Knowl Disc, 2024, 38: 2770–2824
7 Bodria F, Giannotti F, Guidotti R, et al. Benchmarking and survey of explanation methods for black box models. Data Min Knowl Disc, 2023, 37: 1719–1778
8 Guidotti R, Monreale A, Ruggieri S, et al. A survey of methods for explaining black box models. ACM Comput Surv, 2018, 51: 1–42
9 Gilpin L H, Bau D, Yuan B Z, et al. Explaining explanations: an overview of interpretability of machine learning. In: Proceedings of IEEE International Conference on Data Science and Advanced Analytics, 2018. 80–89
10 Goethals S, Sörensen K, Martens D. The privacy issue of counterfactual explanations: explanation linkage attacks. ACM Trans Intell Syst Technol, 2023, 14: 1–24
11 Ferry J, Aïvodji U, Gambs S, et al. SoK: taming the triangle–on the interplays between fairness, interpretability and privacy in machine learning. 2023. ArXiv:2312.16191
12 Sokol K, Flach P. Counterfactual explanations of machine learning predictions: opportunities and challenges for AI safety. In: Proceedings of the AAAI Workshop on Artificial Intelligence Safety, 2019
13 Luo X, Jiang Y, Xiao X. Feature inference attack on Shapley values. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, 2022. 2233–2247
14 Naretto F, Monreale A, Giannotti F. Evaluating the privacy exposure of interpretable global explainers. In: Proceedings of IEEE International Conference on Cognitive Machine Intelligence, 2022. 13–19
15 Artelt A, Vaquet V, Velioglu R, et al. Evaluating robustness of counterfactual explanations. In: Proceedings of IEEE Symposium Series on Computational Intelligence, 2021. 1–9
16 Machado G R, Silva E, Goldschmidt R R. Adversarial machine learning in image classification: a survey toward the defender's perspective. ACM Comput Surv, 2021, 55: 1–38
17 Biggio B, Roli F. Wild patterns: ten years after the rise of adversarial machine learning. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, 2018. 2154–2156
18 Rigaki M, Garcia S. A survey of privacy attacks in machine learning. ACM Comput Surv, 2023, 56: 1–34
19 Hu H, Salcic Z, Sun L, et al. Membership inference attacks on machine learning: a survey. ACM Comput Surv, 2022, 54: 1–37
20 Liu B, Ding M, Shaham S, et al. When machine learning meets privacy: a survey and outlook. ACM Comput Surv, 2022, 54: 1–36

21 Baniecki H, Biecek P. Adversarial attacks and defenses in explainable artificial intelligence: a survey. Inf Fusion, 2024, 107: 102303
22 Papernot N, McDaniel P, Goodfellow I, et al. Practical black-box attacks against machine learning. In: Proceedings of the ACM on Asia Conference on Computer and Communications Security, 2017. 506–519
23 Liu Z, Guo J, Yang W, et al. Privacy-preserving aggregation in federated learning: a survey. IEEE Trans Big Data, 2024. doi: 10.1109/TBDATA.2022.3190835
24 Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access, 2018, 6: 52138–52160
25 Došilović F K, Brčić M, Hlupić N. Explainable artificial intelligence: a survey. In: Proceedings of International Convention on Information and Communication Technology, Electronics and Microelectronics, 2018. 210–215
26 Banisar D. The right to information and privacy: balancing rights and managing conflicts. World Bank Institute Governance Working Paper, 2011. https://documents.worldbank.org/en/publication/documents-reports/documentdetail/847541468188048435/the-right-to-information-and-privacy-balancing-rights-and-managing-conflicts-access-to-information-program
27 Vo V, Le T, Nguyen V, et al. Feature-based learning for diverse and privacy-preserving counterfactual explanations. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2023. 2211–2222
28 Mochaourab R, Sinha S, Greenstein S, et al. Robust counterfactual explanations for privacy-preserving SVM. In: Proceedings of ICML Workshops, 2021
29 Harder F, Bauer M, Park M. Interpretable and differentially private predictions. In: Proceedings of AAAI Conference on Artificial Intelligence, 2020. 34: 4083–4090
30 Shokri R, Strobel M, Zick Y. On the privacy risks of model explanations. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2021. 231–241
31 Shokri R, Strobel M, Zick Y. Exploiting transparency measures for membership inference: a cautionary tale. In: Proceedings of the AAAI Workshop on Privacy-Preserving Artificial Intelligence, 2020
32 Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. 2013. ArXiv:1312.6034
33 Bach S, Binder A, Montavon G, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. Plos One, 2015, 10: e0130140
34 Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: Proceedings of International Conference on Machine Learning, 2017. 3145–3153
35 Sliwinski J, Strobel M, Zick Y. Axiomatic characterization of data-driven influence measures for classification. In: Proceedings of AAAI Conference on Artificial Intelligence, 2019. 33: 718–725
36 Smilkov D, Thorat N, Kim B, et al. SmoothGrad: removing noise by adding noise. 2017. ArXiv:1706.03825
37 Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: Proceedings of International Conference on Machine Learning, 2017. 3319–3328
38 Miura T, Hasegawa S, Shibahara T. MEGEX: data-free model extraction attack against gradient-based explainable AI. 2021. ArXiv:2107.08909
39 Springenberg J T, Dosovitskiy A, Brox T, et al. Striving for simplicity: the all convolutional net. 2014. ArXiv:1412.6806
40 Deng H. Interpreting tree ensembles with intrees. J Dialogue Studies, 2019, 7: 277–287
41 Slack D, Hilgard S, Jia E, et al. Fooling lime and shap: adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020. 180–186
42 Jetchev D, Vuille M. XorSHAP: privacy-preserving explainable AI for decision tree models. Cryptology ePrint Archive, 2023. https://eprint.iacr.org/2023/1859
43 Datta A, Sen S, Zick Y. Algorithmic transparency via quantitative input influence: theory and experiments with learning systems. In: Proceedings of IEEE Symposium on Security and Privacy, 2016. 598–617
44 Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. Knowl Inf Syst, 2014, 41: 647–665
45 Maleki S, Tran-Thanh L, Hines G, et al. Bounding the estimation error of sampling-based Shapley value approximation. 2013. ArXiv:1306.4265
46 Begley T, Schwedes T, Frye C, et al. Explainability for fair machine learning. 2020. ArXiv:2010.07389
47 Aïvodji U, Hara S, Marchand M, et al. Fooling shap with stealthily biased sampling. In: Proceedings of International Conference on Learning Representations, 2022
48 Montenegro H, Silva W, Gaudio A, et al. Privacy-preserving case-based explanations: enabling visual interpretability by protecting privacy. IEEE Access, 2022, 10: 28333–28347
49 Koh P W, Liang P. Understanding black-box predictions via influence functions. In: Proceedings of International Conference on Machine Learning, 2017. 1885–1894
50 Kenny E M, Ford C, Quinn M, et al. Explaining black-box classifiers using post-hoc explanations-by-example: the effect of explanations and error-rates in XAI user studies. Artif Intell, 2021, 294: 103459
51 Lipton Z C. The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. Queue, 2018, 16: 31–57
52 Kim B, Rudin C, Shah J A. The Bayesian case model: a generative approach for case-based reasoning and prototype classification. In: Proceedings of Conference on Neural Information Processing Systems, 2014
53 Nugent C, Doyle D, Cunningham P. Gaining insight through case-based explanation. J Intell Inf Syst, 2009, 32: 267–295
54 Angelov P, Soares E. Towards explainable deep neural networks (xDNN). Neural Netws, 2020, 130: 185–194
55 Angelov P, Soares E. Towards deep machine reasoning: a prototype-based deep neural network with decision tree inference. In: Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, 2020. 2092–2099
56 Li O, Liu H, Chen C, et al. Deep learning for case-based reasoning through prototypes: a neural network that explains its predictions. In: Proceedings of AAAI Conference on Artificial Intelligence, 2018
57 Chen C, Li O, Tao D, et al. This looks like that: deep learning for interpretable image recognition. In: Proceedings of Conference on Neural Information Processing Systems, 2019
58 Papernot N, McDaniel P. Deep k-nearest neighbors: towards confident, interpretable and robust deep learning. 2018. ArXiv:1803.04765
59 Chen Z, Bei Y, Rudin C. Concept whitening for interpretable image recognition. Nat Mach Intell, 2020, 2: 772–782
60 Silva W, Poellinger A, Cardoso J S, et al. Interpretability-guided content-based medical image retrieval. In: Proceedings of

International Conference on Medical Image Computing and Computer Assisted Intervention, 2020. 305–314
61 Kim S, Chae D K. What does a model really look at?: extracting model-oriented concepts for explaining deep neural networks. IEEE Trans Pattern Anal Mach Intell, 2024, 46: 4612–4624
62 Kenny E M, Keane M T. Twin-systems to explain artificial neural networks using case-based reasoning: comparative tests of feature-weighting methods in ANN-CBR twins for XAI. In: Proceedings of International Joint Conferences on Artificial Intelligence, 2019. 2708–2715
63 Kuppa A, Le-Khac N A. Adversarial XAI methods in cybersecurity. IEEE Trans Inform Forensic Secur, 2021, 16: 4924–4938
64 Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harv JL & Tech, 2017, 31: 841
65 Dodge J, Liao Q V, Zhang Y, et al. Explaining models: an empirical study of how explanations impact fairness judgment. In: Proceedings of the International Conference on Intelligent User Interfaces, 2019. 275–285
66 Binns R, van Kleek M, Veale M, et al. 'It's reducing a human being to a percentage' perceptions of justice in algorithmic decisions. In: Proceedings of CHI Conference on Human Factors in Computing Systems, 2018. 1–14
67 Karimi A-H, Schölkopf B, Valera I. Algorithmic recourse: from counterfactual explanations to interventions. In: Proceedings of ACM Conference on Fairness, Accountability, and Transparency, 2021. 353–362
68 Ustun B, Spangher A, Liu Y. Actionable recourse in linear classification. In: Proceedings of ACM Conference on Fairness, Accountability, and Transparency, 2019. 10–19
69 Laugel T, Lesot M-J, Marsala C, et al. Inverse classification for comparison-based interpretability in machine learning. 2017. ArXiv:1712.08443
70 Dhurandhar A, Chen P-Y, Luss R, et al. Explanations based on the missing: towards contrastive explanations with pertinent negatives. In: Proceedings of Conference on Neural Information Processing Systems, 2018
71 Pawelczyk M, Lakkaraju H, Neel S. On the privacy risks of algorithmic recourse. In: Proceedings of International Conference on Artificial Intelligence and Statistics, 2023. 9680–9696
72 Mothilal R K, Sharma A, Tan C. Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of ACM Conference on Fairness, Accountability, and Transparency, 2020. 607–617
73 Severi G, Meyer J, Coull S, et al. Explanation-guided backdoor poisoning attacks against malware classifiers. In: Proceedings of USENIX, 2021. 1487–1504
74 Kuppa A, Le-Khac N-A. Black box attacks on explainable artificial intelligence (XAI) methods in cyber security. In: Proceedings of International Joint Conference on Neural Networks, 2020. 1–8
75 Liu M, Liu X, Yan A, et al. Explanation-guided minimum adversarial attack. In: Proceedings of the International Conference on Machine Learning for Cyber Security, 2022. 257–270
76 Nguyen T, Lai P, Phan H, et al. XRand: differentially private defense against explanation-guided attacks. In: Proceedings of AAAI Conference on Artificial Intelligence, 2023. 873–881
77 Abdukhamidov E, Abuhamad M, Woo S S, et al. Hardening interpretable deep learning systems: investigating adversarial threats and defenses. IEEE Trans Dependable Secure Comput, 2024, 21: 3963–3976
78 Garcia W, Choi J I, Adari S K, et al. Explainable black-box attacks against model-based authentication. 2018. ArXiv:1810.00024
79 Zhang X, Wang N, Shen H, et al. Interpretable deep learning under fire. In: Proceedings of USENIX, 2020
80 Veale M, Binns R, Edwards L. Algorithms that remember: model inversion attacks and data protection law. Philos Trans R Soc A, 2018, 376: 20180083
81 Shokri R, Strobel M, Zick Y. Privacy risks of explaining machine learning models. 2019. ArXiv:1907.00164
82 Yeom S, Giacomelli I, Fredrikson M, et al. Privacy risk in machine learning: analyzing the connection to overfitting. In: Proceedings of IEEE Computer Security Foundations Symposium, 2018. 268–282
83 Sablayrolles A, Douze M, Schmid C, et al. White-box vs black-box: Bayes optimal strategies for membership inference. In: Proceedings of International Conference on Machine Learning, 2019. 5558–5567
84 Carlini N, Chien S, Nasr M, et al. Membership inference attacks from first principles. In: Proceedings of IEEE Symposium on Security and Privacy, 2022. 1897–1914
85 Liu Y, Zhao Z, Backes M, et al. Membership inference attacks by exploiting loss trajectory. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, 2022. 2085–2098
86 Li Z, Liu Y, He X, et al. Auditing membership leakages of multi-exit networks. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, 2022. 1917–1931
87 Ye J, Maddi A, Murakonda S K, et al. Enhanced membership inference attacks against machine learning models. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, 2022. 3093–3106
88 Quan P, Chakraborty S, Jeyakumar J V, et al. On the amplification of security and privacy risks by post-hoc explanations in machine learning models. 2022. ArXiv:2206.14004
89 Liu H, Wu Y, Yu Z, et al. Please tell me more: privacy impact of explainability through the lens of membership inference attack. In: Proceedings of IEEE Symposium on Security and Privacy, 2024
90 Petitcolas F A. Kerckhoffs' principle. In: Proceedings of Encyclopedia of Cryptography, Security and Privacy, 2023. 1–2
91 Craven M W, Shavlik J W. Using sampling and queries to extract rules from trained neural networks. In: Proceedings of Machine Learning Proceedings, 1994. 37–45
92 Pawelczyk M, Broelemann K, Kasneci G. Learning model-agnostic counterfactual explanations for tabular data. In: Proceedings of the Web Conference, 2020. 3126–3132
93 Huang C, Swoopes C, Xiao C, et al. Accurate, explainable, and private models: providing recourse while minimizing training data leakage. 2023. ArXiv:2308.04341
94 Sweeney L. Simple demographics often identify people uniquely. Health, 2000, 671: 1–34
95 Brughmans D, Leyman P, Martens D. NICE: an algorithm for nearest instance counterfactual explanations. Data Min Knowl Disc, 2024, 38: 2665–2703
96 Keane M T, Smyth B. Good counterfactuals and where to find them: a case-based technique for generating counterfactuals for explainable AI (XAI). In: Proceedings of Case-Based Reasoning Research and Development. Cham: Springer, 2020. 163–178
97 Pawelczyk M, Broelemann K, Kasneci G. On counterfactual explanations under predictive multiplicity. In: Proceedings of the Conference on Uncertainty in Artificial Intelligence, 2020. 809–818
98 Aïvodji U, Bolot A, Gambs S. Model extraction from counterfactual explanations. 2020. ArXiv:2009.01884
99 Dwork C, Smith A, Steinke T, et al. Exposed! A survey of attacks on private data. Annu Rev Stat Appl, 2017, 4: 61–84

100 Ferry J, Aïvodji U, Gambs S, et al. Probabilistic dataset reconstruction from interpretable models. 2023. ArXiv:2308.15099
101 Ferry J. Addresing interpretability fairness & privacy in machine learning through combinatorial optimization methods. Dissertation for Ph.D. Degree. Toulouse: Université Paul Sabatier-Toulouse III, 2023
102 Garfinkel S, Abowd J M, Martindale C. Understanding database reconstruction attacks on public data. Commun ACM, 2019, 62: 46–53
103 Song C, Ristenpart T, Shmatikov V. Machine learning models that remember too much. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, 2017. 587–601
104 Carlini N, Liu C, Erlingsson Ú, et al. The secret sharer: evaluating and testing unintended memorization in neural networks. In: Proceedings of USENIX, 2019. 267–284
105 Salem A, Bhattacharya A, Backes M, et al. Updates-leak: data set inference and reconstruction attacks in online learning. In: Proceedings of USENIX, 2020. 1291–1308
106 Gambs S, Gmati A, Hurfin M. Reconstruction attack through classifier analysis. In: Proceedings of Data and Applications Security and Privacy XXVI, 2012. 274–281
107 Milli S, Schmidt L, Dragan A D, et al. Model reconstruction from model explanations. In: Proceedings of ACM Conference on Fairness, Accountability, and Transparency, 2019. 1–9
108 Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, 2015. 1322–1333
109 Yang Z, Zhang J, Chang E-C, et al. Neural network inversion in adversarial setting via background knowledge alignment. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, 2019. 225–240
110 Zhang Y, Jia R, Pei H, et al. The secret revealer: generative model-inversion attacks against deep neural networks. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2020. 253–261
111 Dosovitskiy A, Brox T. Inverting visual representations with convolutional networks. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2016. 4829–4837
112 He Z, Zhang T, Lee R B. Model inversion attacks against collaborative inference. In: Proceedings of the Annual Computer Security Applications Conference, 2019. 148–162
113 Zhao X, Zhang W, Xiao X, et al. Exploiting explanations for model inversion attacks. In: Proceedings of International Conference on Computer Vision, 2021. 682–692
114 Dumoulin V, Visin F. A guide to convolution arithmetic for deep learning. 2016. ArXiv:1603.07285
115 Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of International Conference on Computer Vision, 2017. 618–626
116 Rehman A, Rahim R, Nadeem S, et al. End-to-end trained CNN encoder-decoder networks for image steganography. In: Proceedings of European Conference on Computer Vision Workshops, 2019. 723–729
117 Zhang Y, Tian Y, Kong Y, et al. Residual dense network for image super-resolution. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2018. 2472–2481
118 Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2016. 2921–2929
119 Miller T. Explanation in artificial intelligence: insights from the social sciences. Artif Intelligence, 2019, 267: 1–38
120 Song C, Shmatikov V. Overlearning reveals sensitive attributes. In: Proceedings of International Conference on Learning Representations, 2020
121 Ganju K, Wang Q, Yang W, et al. Property inference attacks on fully connected neural networks using permutation invariant representations. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, 2018. 619–633
122 Melis L, Song C, de Cristofaro E, et al. Exploiting unintended feature leakage in collaborative learning. In: Proceedings of IEEE Symposium on Security and Privacy, 2019. 691–706
123 Zhang W, Tople S, Ohrimenko O. Leakage of dataset properties in multi-party machine learning. In: Proceedings of USENIX, 2021. 2687–2704
124 Duddu V, Boutet A. Inferring sensitive attributes from model explanations. In: Proceedings of ACM International Conference on Information and Knowledge Management, 2022. 416–425
125 Chen J, Song L, Wainwright M, et al. Learning to explain: an information-theoretic perspective on model interpretation. In: Proceedings of International Conference on Machine Learning, 2018. 883–892
126 Salem A, Zhang Y, Humbert M, et al. ML-leaks: model and data independent membership inference attacks and defenses on machine learning models. 2018. ArXiv:1806.01246
127 Tramèr F, Zhang F, Juels A, et al. Stealing machine learning models via prediction APIs. In: Proceedings of USENIX, 2016. 601–618
128 Jagielski M, Carlini N, Berthelot D, et al. High accuracy and high fidelity extraction of neural networks. In: Proceedings of USENIX, 2020. 1345–1362
129 Wang Y, Qian H, Miao C. DualCF: efficient model extraction attack from counterfactual explanations. In: Proceedings of ACM Conference on Fairness, Accountability, and Transparency, 2022. 1318–1329
130 Nguyen D, Bui N, Nguyen V A. Feasible recourse plan via diverse interpolation. In: Proceedings of International Conference on Artificial Intelligence and Statistics, 2023. 4679–4698
131 Artelt A, Hammer B. Convex density constraints for computing plausible counterfactual explanations. In: Proceedings of Artificial Neural Networks and Machine Learning, 2020. 353–365
132 Kumari K, Jadliwala M, Jha S K, et al. Towards a game-theoretic understanding of explanation-based membership inference attacks. 2024. ArXiv:2404.07139
133 Luo X, Wu Y, Xiao X, et al. Feature inference attack on model predictions in vertical federated learning. In: Proceedings of IEEE International Conference on Data Engineering, 2021. 181–192
134 Barocas S, Selbst A D, Raghavan M. The hidden assumptions behind counterfactual explanations and principal reasons. In: Proceedings of ACM Conference on Fairness, Accountability, and Transparency, 2020. 80–89
135 Kasirzadeh A, Smart A. The use and misuse of counterfactuals in ethical machine learning. In: Proceedings of ACM Conference on Fairness, Accountability, and Transparency, 2021. 228–236
136 Hashemi M, Fathi A. PermuteAttack: counterfactual explanation of machine learning credit scorecards. 2020. ArXiv:2008.10138
137 Dwork C, Roth A. The algorithmic foundations of differential privacy. FNT Theor Comput Sci, 2014, 9: 211–407
138 Patel N, Shokri R, Zick Y. Model explanations with differential privacy. In: Proceedings of ACM Conference on Fairness, Accountability, and Transparency, 2022. 1895–1904

139 Yang F, Feng Q, Zhou K, et al. Differentially private counterfactuals via functional mechanism. 2022. ArXiv:2208.02878
140 Hamer J, Valladares J, Viswanathan V, et al. Simple steps to success: axiomatics of distance-based algorithmic recourse. 2023. ArXiv:2306.15557
141 Pentyala S, Sharma S, Kariyappa S, et al. Privacy-preserving algorithmic recourse. 2023. ArXiv:2311.14137
142 Holohan N, Braghin S, Aonghusa P M, et al. Diffprivlib: the IBM differential privacy library. 2019. ArXiv:1907.02444
143 Chaudhuri K, Monteleoni C, Sarwate A D. Differentially private empirical risk minimization. J Mach Learn Res, 2011, 12: 1069–1109
144 Wang D, Ye M, Xu J. Differentially private empirical risk minimization revisited: faster and more general. In: Proceedings of Conference on Neural Information Processing Systems, 2017
145 Joshi D, Thakkar J. k-means subclustering: a differentially private algorithm with improved clustering quality. In: Proceedings of ACM International Conference on Information and Knowledge Management, 2022
146 Lu Z, Shen H. Differentially private k-means clustering with convergence guarantee. IEEE Trans Dependable Secure Comput, 2020, 18: 1541–1552
147 Abadi M, Chu A, Goodfellow I, et al. Deep learning with differential privacy. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, 2016. 308–318
148 Wagner T, Naamad Y, Mishra N. Fast private kernel density estimation via locality sensitive quantization. In: Proceedings of International Conference on Machine Learning, 2023. 339–367
149 Wang G. Interpret federated learning with Shapley values. 2019. ArXiv:1905.04519
150 Watson L, Andreeva R, Yang H-T, et al. Differentially private Shapley values for data evaluation. 2022. ArXiv:2206.00511
151 Naidu R, Priyanshu A, Kumar A, et al. When differential privacy meets interpretability: a case study. 2021. ArXiv:2106.13203
152 Bu Z, Wang Y-X, Zha S, et al. Differentially private optimization on large model at small cost. In: Proceedings of International Conference on Machine Learning, 2023. 3192–3218
153 Hooker S, Erhan D, Kindermans P-J, et al. A benchmark for interpretability methods in deep neural networks. In: Proceedings of Conference on Neural Information Processing Systems, 2019
154 Veugen T, Kamphorst B, Marcus M. Privacy-preserving contrastive explanations with local foil trees. In: Cyber Security, Cryptology, and Machine Learning. Cham: Springer, 2022
155 van der Waa J, Robeer M, van Diggelen J, et al. Contrastive explanations with local foil trees. 2018. ArXiv:1806.07470
156 Lindell Y. Secure multiparty computation. Commun ACM, 2021, 64: 86–96
157 Jia J, Salem A, Backes M, et al. MemGuard: defending against black-box membership inference attacks via adversarial examples. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, 2019. 259–274
158 Olatunji I E, Rathee M, Funke T, et al. Private graph extraction via feature explanations. PoPETs, 2023, 2023: 59–78
159 Montenegro H, Silva W, Cardoso J S. Privacy-preserving generative adversarial network for case-based explainability in medical image analysis. IEEE Access, 2021, 9: 148037–148047
160 Chen J, Konrad J, Ishwar P. VGAN-based image representation learning for privacy-preserving facial expression recognition. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2018. 1570–1579
161 Montavon G, Lapuschkin S, Binder A, et al. Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recogn, 2017, 65: 211–222
162 Gade K, Geyik S C, Kenthapadi K, et al. Explainable AI in industry. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019. 3203–3204
163 Kaur H, Nori H, Jenkins S, et al. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In: Proceedings of CHI Conference on Human Factors in Computing Systems, 2020. 1–14
164 Hu S, Liu X, Zhang Y, et al. Protecting facial privacy: generating adversarial identity masks via style-robust makeup transfer. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2022. 14–23
165 Liu H, Wang Y, Zhang Z, et al. Matrix factorization recommender based on adaptive Gaussian differential privacy for implicit feedback. Inf Process Manage, 2024, 61: 103720
166 Liu Z, Jiang Y, Jiang W, et al. Guaranteeing data privacy in federated unlearning with dynamic user participation. 2024. ArXiv:2406.00966
167 Mi D, Zhang Y, Zhang L Y, et al. Towards model extraction attacks in GAN-based image translation via domain shift mitigation. In: Proceedings of AAAI Conference on Artificial Intelligence, 2024. 902–910
168 Zhang Y, Hu S, Zhang L Y, et al. Why does little robustness help? A further step towards understanding adversarial transferability. In: Proceedings of IEEE Symposium on Security and Privacy, 2024
169 Nguyen T T, Huynh T T, Ren Z, et al. A survey of machine unlearning. 2022. ArXiv:2209.02299
170 Huynh T T, Nguyen T B, Nguyen P L, et al. Fast-FedUL: a training-free federated unlearning with provable skew resilience. In: Proceedings of European Conference on Machine Learning and Knowledge Discovery in Databases, 2024
171 Liu Z, Guo J, Yang W, et al. Dynamic user clustering for efficient and privacy-preserving federated learning. IEEE Trans Dependable Secure Comput, 2024. doi: 10.1109/TDSC.2024.3355458
172 Li Z, Chen H, Ni Z, et al. Balancing privacy protection and interpretability in federated learning. 2023. ArXiv:2302.08044
173 Zhang J, Bareinboim E. Fairness in decision-making—the causal explanation formula. In: Proceedings of AAAI Conference on Artificial Intelligence, 2018
174 Frye C, de Mijolla D, Begley T, et al. Shapley explainability on the data manifold. In: Proceedings of International Conference on Learning Representations, 2021
175 Mittelstadt B, Russell C, Wachter S. Explaining explanations in AI. In: Proceedings of ACM Conference on Fairness, Accountability, and Transparency, 2019. 279–288
176 Gillenwater J, Joseph M, Kulesza A. Differentially private quantiles. In: Proceedings of International Conference on Machine Learning, 2021. 3713–3722
177 Ghosh A, Shanbhag A, Wilson C. FairCanary: rapid continuous explainable fairness. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2022. 307–316
178 Li Z, van Leeuwen M. Explainable contextual anomaly detection using quantile regression forests. Data Min Knowl Disc, 2023, 37: 2517–2563
179 Merz M, Richman R, Tsanakas A, et al. Interpreting deep learning models with marginal attribution by conditioning on quantiles. Data Min Knowl Disc, 2022, 36: 1335–1370
180 Alvarez-Melis D, Jaakkola T. Towards robust interpretability with self-explaining neural networks. In: Proceedings of Conference on Neural Information Processing Systems, 2018

181 Zhang Z, Liu Q, Wang H, et al. ProtGNN: towards self-explaining graph neural networks. In: Proceedings of AAAI Conference on Artificial Intelligence, 2022. 9127–9135
182 Khosla M. Privacy and transparency in graph machine learning: a unified perspective. 2022. ArXiv:2207.10896
183 Tiddi I, Schlobach S. Knowledge graphs as tools for explainable machine learning: a survey. Artif Intell, 2022, 302: 103627
184 Rajabi E, Etminani K. Knowledge-graph-based explainable AI: a systematic review. J Inf Sci, 2024, 50: 1019–1029
185 Qian J, Li X Y, Zhang C, et al. Social network de-anonymization and privacy inference with knowledge graph model. IEEE Trans Dependable Secure Comput, 2019, 16: 679–692
186 Wang Y, Huang L, Yu P S, et al. Membership inference attacks on knowledge graphs. 2021. ArXiv:2104.08273
187 Domingo-Ferrer J, Pérez-Solà C, Blanco-Justicia A. Collaborative explanation of deep models with limited interaction for trade secret and privacy preservation. In: Proceedings of WWW Companion, 2019. 501–507
188 Gaudio A, Smailagic A, Faloutsos C, et al. DeepFixCX: explainable privacy-preserving image compression for medical image analysis. WIREs Data Min Knowl, 2023, 13: e1495
189 Wu Z, Pan S, Chen F, et al. A comprehensive survey on graph neural networks. IEEE Trans Neural Netw Learn Syst, 2021, 32: 4–24
190 Liu Z, Luong N C, Wang W, et al. A survey on blockchain: a game theoretical perspective. IEEE Access, 2019, 7: 47615–47643
191 Yuan H, Yu H, Gui S, et al. Explainability in graph neural networks: a taxonomic survey. IEEE Trans Pattern Anal Mach Intell, 2022, 45: 5782–5799
192 Prado-Romero M A, Prenkaj B, Stilo G, et al. A survey on graph counterfactual explanations: definitions, methods, evaluation, and research challenges. ACM Comput Surv, 2024, 56: 1–37
193 Dai E, Zhao T, Zhu H, et al. A comprehensive survey on trustworthy graph neural networks: privacy, robustness, fairness, and explainability. 2022. ArXiv:2204.08570
194 Ren Z, Qian K, Schultz T, et al. An overview of the ICASSP special session on AI security and privacy in speech and audio processing. In: Proceedings of ACM Multimedia Workshop, 2023
195 Li Z, Shi C, Zhang T, et al. Robust detection of machine-induced audio attacks in intelligent audio systems with microphone array. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, 2021. 1884–1899
196 Carlini N, Wagner D. Audio adversarial examples: targeted attacks on speech-to-text. In: Proceedings of IEEE Security and Privacy Workshops, 2018. 1–7
197 Abdullah H, Warren K, Bindschaedler V, et al. SoK: the faults in our ASRs: an overview of attacks against automatic speech recognition and speaker identification systems. In: Proceedings of IEEE Symposium on Security and Privacy, 2021. 730–747
198 Ren Z, Qian K, Dong F, et al. Deep attention-based neural networks for explainable heart sound classification. Machine Learn Appl, 2022, 9: 100322
199 Ren Z, Baird A, Han J, et al. Generating and protecting against adversarial attacks for deep speech-based emotion recognition models. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2020. 7184–7188
200 Chang Y, Ren Z, Nguyen T T, et al. Example-based explanations with adversarial attacks for respiratory sound analysis. In: Proceedings of Interspeech, 2022. 1–5
201 Liu Z, Guo J, Yang M, et al. Privacy-enhanced knowledge transfer with collaborative split learning over teacher ensembles. In: Proceedings of Secure and Trustworthy Deep Learning Systems Workshop, 2023. 1–13
202 Liu Z, Lin H Y, Liu Y. Long-term privacy-preserving aggregation with user-dynamics for federated learning. IEEE Trans Inform Forensic Secur, 2023, 18: 2398–2412
203 Liu Z, Guo J, Lam K Y, et al. Efficient dropout-resilient aggregation for privacy-preserving machine learning. IEEE Trans Inform Forensic Secur, 2023, 18: 1839–1854
204 Belhadj-Cheikh N, Imine A, Rusinowitch M. FOX: fooling with explanations: privacy protection with adversarial reactions in social media. In: Proceedings of International Conference on Privacy, Security and Trust, 2021. 1–10
205 Jia R, Dao D, Wang B, et al. Towards efficient data valuation based on the Shapley value. In: Proceedings of International Conference on Artificial Intelligence and Statistics, 2019. 1167–1176