Balancing Explainability Strength and Privacy in Counterfactual Explanations for Insider Threat Detection

Naila Azam
Centre for Sustainable Cyber Security
University of Greenwich
London, UK
n.naila@greenwich.ac.uk

George Loukas
Centre for Sustainable Cyber Security
University of Greenwich
London, UK
g.loukas@greenwich.ac.uk

Manos Panaousis
Centre for Sustainable Cyber Security
University of Greenwich
London, UK
e.panaousis@greenwich.ac.uk

***Abstract*—Explainable AI (XAI) methods, such as counterfactual explanations, help make model decisions more transparent and actionable. However, they can unintentionally reveal sensitive information, which is a critical concern in security settings such as insider threat detection, where both transparency and privacy are vital, particularly when explanations are shared with external analysts or decision-makers. We propose a privacy-aware counterfactual framework that integrates Differential Privacy (DP) to add only the minimum adjusted noise required to satisfy user-defined privacy caps, reducing information leakage while preserving explanation fidelity. Two task-specific metrics are introduced: XAIStrength, which measures how faithfully a counterfactual follows the model's decision boundary, and XAILeakage, which quantifies privacy risk based on movement into sparse feature regions. Experiments on the CERT insider threat dataset show that initially generated counterfactuals were highly faithful for most users but frequently violated privacy requirements. After optimisation, leakage was reduced to meet strict, moderate, and lenient privacy caps ($\tau = 0.05/0.10/0.20$) while maintaining practical fidelity. More than 60% of users retained $S_{\text{opt}} \geq 0.70$ under strict privacy constraints, while over 60% maintained $S_{\text{opt}} \geq 0.85$ under moderate privacy settings. Comparative evaluation against standard and privacy-aware baselines shows that the proposed framework achieves a stronger balance between explanation utility and disclosure risk. To the best of our knowledge, this is the first study to formally quantify and optimise the XAIStrength–XAILeakage trade-off in counterfactual explanations using statistically grounded risk measures and DP-calibrated optimisation.**

***Index Terms*—Explainable AI (XAI), Counterfactual, XAIStrength, XAILeakage, Optimisation, Differential Privacy**

I. Introduction

Explainable AI (XAI) techniques, particularly counterfactual explanations, have gained significant attention because they describe the smallest changes to an input that would alter a machine learning model's decision [1]. Unlike many other explanation methods, counterfactuals provide explanations in a form that is quite intuitive for human users. Rather than explaining the internal structure of a model, they demonstrate how a prediction can be altered by small changes to the input features. They are therefore especially valuable in areas where analysts and decision makers need explanations that are understandable and actionable.

However, the transparency provided by counterfactual explanations may also introduce privacy risks. Counterfactuals often reveal how predictions change with respect to sensitive attributes such as behavioural patterns, psychometric traits, or system-usage profiles [2]. In insider threat detection, these attributes are directly related to user behaviour and organisational activity. Consequently, even small perturbations can move explanations into statistically sparse regions of the data space, increasing the risk of disclosing information about individuals or the underlying cohort. This concern becomes more important in operational cybersecurity settings where explanations may be shared with investigators, auditors, or external analysts. Even if explicit identifiers are removed, unusual combinations of behavioural features may still reveal sensitive patterns that enable linkage or attribute inference attacks [3]. Therefore, explanations that are very informative from an interpretability point of view may also increase the disclosure risk.

To date, most XAI research has focused on improving transparency and explanation quality, while comparatively little attention has been given to the privacy implications of the explanations themselves. Differential Privacy (DP) has been widely studied for protecting sensitive information during model training [4], but its integration into post-hoc explanation techniques such as counterfactual generation remains limited. Existing counterfactual methods mainly optimise for validity, sparsity, and proximity to the original instance, without explicitly considering whether the generated explanations leak sensitive statistical information about the underlying dataset. Another limitation in the current literature is the absence of a clear mechanism to evaluate the trade-off between explanation transparency and privacy preservation. In practice, highly faithful explanations may also carry greater disclosure risk, especially when explanations are externally shared or repeatedly queried. Therefore, explanations should remain useful and interpretable while limiting the possibility of exposing sensitive information.

In this paper, we address this problem through a privacy-aware counterfactual explanation framework based on dif-

ferential privacy, evaluated in the context of insider threat detection. The framework introduces a controlled optimisation process that applies only the minimum level of DP noise required to satisfy a predefined privacy constraint. This allows the framework to reduce privacy leakage while preserving explanation utility. The main contributions of this paper are as follows:

- An optimisation method that determines the minimal DP noise required to satisfy a user-defined privacy cap, protecting sensitive information without heavily degrading the quality of explanation.
- The introduction of *XAIStrength* to measure how closely counterfactuals follow the model's decision boundary and *XAILeakage* to quantify privacy risk based on movement into sparse regions of the data space.
- A comprehensive evaluation on the CERT insider threat dataset, demonstrating the ability of the proposed framework to enforce strict, moderate, and lenient privacy caps while preserving high explanation fidelity, together with a comparative analysis against both standard and privacy-aware counterfactual baselines.

The proposed framework is intended for practical settings where both interpretability and privacy preservation are important requirements. By explicitly incorporating privacy constraints into the counterfactual generation process, the framework provides a balanced approach for releasing explanations in sensitive cybersecurity applications. To the best of our knowledge, this is the first work to formally define, measure, and optimise the *XAIStrength–XAILeakage* trade-off, demonstrating that privacy-compliant explanations can be achieved with limited loss in interpretability while accounting for both statistical plausibility and disclosure risk.

II. Background and Related Work

A. Counterfactual Explanations in Explainable AI

Counterfactual explanations are widely used post-hoc methods that explain model predictions by identifying the minimal changes required in an input to achieve a different outcome [1]. Instead of describing the internal workings of a model, they provide intuitive "what-if" insights, such as what needs to change for a prediction to be reversed. This makes them particularly useful in high-stakes domains, including finance, healthcare, and decision support systems, where interpretability and user trust are essential [5].

Early work by Wachter et al. [1] formulated counterfactual generation as an optimisation problem, introducing two key criteria: *proximity*, which measures how close a counterfactual is to the original instance, and *sparsity*, which captures how few features are modified. While these criteria ensure minimality, they do not guarantee realism. As a result, later work introduced *plausibility* and *manifold constraints* to ensure that counterfactuals remain consistent with the data distribution [6]. Another important development is the notion of *diversity*, which recognises that multiple valid counterfactuals may exist for a single instance. Methods such as DiCE generate diverse and feasible alternatives, allowing users to explore different decision pathways [7]. DiCE is particularly useful in practice due to its model-agnostic design and support for mixed data types.

More recent research has highlighted additional considerations, including *actionability*, where suggested changes must be feasible in real-world settings, and *causal consistency*, which ensures that relationships between features are respected [8]. These aspects are critical when deploying counterfactual explanations in realistic environments. Despite these advances, important challenges remain. Many methods assume full access to feature information and underlying data distributions, which may not hold in privacy-sensitive contexts. Moreover, counterfactual explanations can inadvertently reveal information about the training data or population structure, raising concerns when explanations are shared across organisations. In summary, proximity, sparsity, plausibility, and diversity remain the core evaluation criteria for counterfactual explanations. However, recent work increasingly emphasises actionability, causality, and privacy, reflecting the need for explanations that are not only interpretable but also realistic and secure.

B. Differential Privacy and Noise Calibration for Privacy-Preserving AI

Differential Privacy (DP) has become one of the most widely accepted frameworks for protecting individual-level information in data analysis and machine learning. At its core, DP ensures that the inclusion or removal of any single individual's data does not significantly affect the outcome of a computation [9]. This property provides a strong and quantifiable privacy guarantee, making it particularly suitable for sensitive domains such as healthcare, finance, and government data systems.

The key idea behind DP is to introduce carefully calibrated randomness into computations. Instead of releasing exact results, a mechanism adds noise in such a way that the output remains statistically similar regardless of whether a particular individual's data is present. The strength of this guarantee is controlled by the *privacy budget* ϵ. Smaller values of ϵ provide stronger privacy by making outputs more indistinguishable, but at the cost of reduced accuracy. Conversely, larger ϵ values retain higher utility while weakening privacy protection [10], [11]. This trade-off between privacy and utility is central to the practical deployment of DP systems. In practice, the amount of noise added depends on the sensitivity of the function being computed. Two commonly used mechanisms are the *Laplace mechanism*, which is suitable for functions with bounded ℓ_1 sensitivity, and the *Gaussian mechanism*, which is often used when ℓ_2 sensitivity is considered or when approximate DP guarantees are acceptable [4], [9]. The calibration of noise to global sensitivity ensures that the privacy guarantees are mathematically rigorous while remaining adaptable to different application settings.

An important advantage of DP is its robustness under composition and post-processing. Multiple DP mechanisms can be applied sequentially while maintaining an overall

privacy bound, and any further processing of DP outputs does not weaken the original guarantee [10]. This makes DP particularly appealing for explainable AI and model auditing scenarios, where outputs may be reused, shared, or combined across different stages of analysis. DP has also moved beyond theory into large-scale real-world deployment. It underpins systems such as the U.S. Census Bureau's 2020 Disclosure Avoidance System and has been adopted by major technology companies for privacy-preserving analytics and machine learning [11]. More recent work has further explored how explanation methods themselves, including model explanations and interpretability tools, can introduce privacy risks, highlighting the need to integrate DP into explainable AI pipelines [2].

Overall, DP provides a flexible and principled approach to balancing data utility with strong privacy guarantees. Its ability to offer tunable protection, combined with solid theoretical foundations and practical applicability, makes it a cornerstone of modern privacy-preserving AI systems.

C. Related Work

Recent research has shown that explainable AI (XAI) methods, particularly counterfactual explanations, can unintentionally expose sensitive information about individuals or the underlying training data. Goethals et al. [12] demonstrated that counterfactual explanations may contain quasi-identifiers that can be linked with external information to re-identify individuals, introducing what they describe as *explanation linkage attacks*. Similarly, Shokri et al. [13] showed that model explanations can leak sensitive attributes and even reveal membership information, highlighting that interpretability methods themselves may create additional privacy risks.

Privacy concerns related to machine learning explanations are also connected to broader attacks on model outputs. Fredrikson et al. [14] demonstrated that confidence scores and prediction outputs can enable model inversion attacks capable of reconstructing sensitive information about individuals. Likewise, Hayes et al. [15] showed that generative models are vulnerable to membership inference attacks, reinforcing concerns that information exposed through model behaviour or explanations may reveal details about the training data distribution. Several studies have explored ways to measure disclosure risks in explanations. Traditional privacy measures such as k-anonymity, pureness [12], and proximity-based distances such as L_1 norms [2] are often used as indicators of potential leakage. However, these metrics are usually considered independently and do not fully capture the relationship between privacy protection and explanation quality. Tools such as ML Privacy Meter [16] can estimate inference risks in machine learning models, but they are not specifically designed for counterfactual explanations and therefore cannot directly evaluate the privacy–utility trade-off in XAI settings.

To reduce disclosure risks, recent work has started incorporating DP into explanation generation. Patel et al. [17] investigated differentially private local explanations and analysed how different privacy budgets (ϵ) influence explanation fidelity. Yang et al. [18] proposed Differentially Private Counterfactuals (DPC), where calibrated noise is added to preserve privacy while maintaining explanation usefulness. More recently, Pentyala et al. [19] proposed privacy-preserving algorithmic recourse methods that use differentially private clustering to generate realistic and actionable counterfactual paths while reducing disclosure risks.

Although these studies improve privacy protection, several limitations remain. Most existing approaches apply privacy mechanisms in a general-purpose manner and do not explicitly optimise the balance between privacy preservation and explanation quality under user-defined constraints. In addition, current methods rarely provide dedicated metrics that jointly evaluate explanation strength and privacy leakage. This issue becomes particularly important in insider threat detection, where counterfactual explanations may involve highly sensitive behavioural, psychometric, or demographic information. Existing insider threat detection studies either rely on black-box predictions or apply explainability methods without systematically examining the associated privacy risks [20].

Our work addresses this gap by introducing a privacy-aware counterfactual framework that dynamically calibrates DP noise according to user-defined privacy thresholds while preserving explanatory quality as much as possible. In addition, we introduce two task-specific metrics, namely *XAIStrength* and *XAILeakage*, to jointly quantify explanation utility and disclosure risk. To the best of our knowledge, this is the first empirical study that systematically evaluates privacy-optimised counterfactual explanations in the context of insider threat detection.

III. Privacy-Aware Counterfactual Optimisation

A. Proposed Metrics

We introduce two novel metrics, XAIStrength and XAILeakage, to jointly assess counterfactual explanations by capturing both fidelity to the model's decision boundary and their statistical privacy risk. Unlike prior work that focuses mainly on proximity, sparsity, or heuristic plausibility measures [5], our approach provides a unified, measurable way to evaluate transparency while accounting for privacy.

B. XAIStrength

XAIStrength quantifies how faithfully a generated counterfactual captures the *minimal feature perturbation* required to alter the model's prediction. Following the fidelity-error view of White and d'Avila Garcez [21], we define the deviation between the estimated perturbation $\hat{b}$ and the actual decision-boundary shift b as error $\mathrm{E} = |\hat{b} - b|$. In practice, the true boundary shift b is approximated using local linearisation of the classifier around the factual instance, while $\hat{b}$ is computed from the displacement induced by the candidate counterfactual. This formulation allows the metric to evaluate whether the generated counterfactual remains consistent with the local behaviour of the underlying model rather than only measuring surface-level feature similarity.

To make this error comparable across instances, we normalise it by the magnitude of the true perturbation and invert it so that higher values indicate stronger explanations:

$$\mathrm{XAIStrength} = 1 - \frac{|\hat{b} - b|}{|b| + \epsilon}, \qquad 0 \leq \mathrm{XAIStrength} \leq 1, \tag{1}$$

where ϵ is a small constant to avoid division by zero.

Here, the numerator represents the fidelity error, the denominator scales it relative to the true boundary shift, and subtracting from one ensures that smaller errors yield higher strength. Thus, $\mathrm{XAIStrength} = 1$ denotes perfect alignment with the model's true decision boundary, while values near 0 correspond to counterfactuals that deviate substantially. In practice, higher XAIStrength values indicate that the generated explanation captures the key feature changes required for the prediction flip without introducing unnecessary distortion. This is particularly important in security applications where analysts rely on counterfactual explanations to understand the behavioural factors contributing to a model decision.

Unlike sparsity or proximity-based scores, Eq. 1 directly measures whether the counterfactual reflects the causal boundary changes needed for class flip [22]. As a result, the metric provides a more direct assessment of explanatory fidelity by evaluating how closely the generated counterfactual follows the classifier's local decision structure.

C. XAILeakage

Beyond fidelity, counterfactual explanations must remain plausible and avoid exposing statistically implausible or sensitive feature combinations. We therefore introduce *XAILeakage*, a metric that quantifies the disclosure risk of a counterfactual relative to the empirical cohort distribution. The metric is designed to evaluate whether a generated counterfactual remains consistent with the statistical characteristics of the population or moves toward atypical regions that may increase privacy risk. In cybersecurity applications such as insider threat detection, counterfactuals located far from the normal behavioural distribution may unintentionally expose unusual user activity patterns or rare feature combinations.

Let $\Delta = x' - x$ denote the perturbation vector and let Σ be the covariance matrix of the cohort. The squared Mahalanobis distance $D_M^2 = \Delta^\top \Sigma^{-1} \Delta$ [23] measures how far x' lies from the cohort mean in variance-weighted units. Unlike Euclidean distance, the Mahalanobis formulation accounts for both feature scale and correlations between features, making it more appropriate for structured behavioural data where dependencies between variables are common. Under the assumption of multivariate normality, D_M^2 follows a χ^2 distribution with p degrees of freedom. We therefore map this distance to a cumulative probability $P_{MD} = F_{\chi_p^2}(D_M^2)$, which represents the probability mass enclosed by the ellipsoid defined by D_M^2. We use the cumulative probability rather than the tail probability to enforce strict plausibility constraints, ensuring that counterfactuals remain within the highest-density regions of the empirical distribution.

We define

$$\mathrm{XAILeakage} = P_{MD}, \tag{2}$$

such that larger values of *XAILeakage* correspond to counterfactuals in low-density, statistically atypical regions of the cohort distribution. As the Mahalanobis distance D_M^2 increases, the cumulative probability P_{MD} approaches 1, indicating reduced plausibility and increased disclosure risk. Counterfactuals with lower leakage values therefore remain closer to the dominant data distribution and are less likely to expose statistically distinctive behavioural characteristics.

By accounting for feature scale and correlation and mapping distance to a bounded probability in $[0, 1]$, *XAILeakage* provides an interpretable and monotonic measure of privacy risk, where higher values reflect greater deviation from the data distribution and higher potential for privacy leakage [24]. The metric also enables direct integration of privacy constraints into the optimisation process, allowing counterfactual explanations to be evaluated not only by their fidelity but also by their statistical plausibility within the cohort.

D. Optimising the XAIStrength–XAILeakage Trade-off

XAIStrength (Eq. 1) captures how closely a counterfactual matches the minimal change required to cross the model's decision boundary, while XAILeakage (Eq. 2) reflects how atypical that counterfactual is with respect to the underlying data distribution. These two objectives are often at odds. Counterfactuals that closely follow the true decision boundary tend to move into low-density regions of the feature space, which increases privacy risk. Conversely, enforcing stronger privacy constraints introduces additional randomisation, which can reduce fidelity to the decision boundary while limiting the disclosure risk associated with statistically atypical counterfactuals. This trade-off becomes particularly important in sensitive applications such as insider threat detection, where explanations must remain informative for analysts while avoiding exposure of unusual behavioural patterns or rare cohort characteristics. Under the cumulative probability formulation of XAILeakage, enforcing $P_{MD} \leq \tau$ ensures that selected counterfactuals remain within the most probable regions of the cohort distribution.

We therefore formulate counterfactual generation as a constrained optimisation problem that seeks to maximise fidelity while enforcing an explicit bound on privacy leakage:

$$\max_{x'}; f(x') = 1 - \frac{|\hat{b}(x') - b|}{|b| + \epsilon}, \tag{3}$$

$$\text{s.t. } g(x') = P_{MD}(x, x') - \tau \leq 0, \quad x' \in \mathcal{X}. \tag{4}$$

Here, x denotes the original instance and x' a candidate counterfactual. The value b represents the true minimal shift needed to cross the model's decision boundary, while $\hat{b}(x')$ denotes the shift induced by x'. The objective function rewards counterfactuals whose induced shift closely matches b, thereby promoting high-fidelity explanations. As a result, the optimisation favours explanations that remain aligned with the

classifier's local decision structure while avoiding unnecessary feature perturbations..

The constraint $g(x') \leq 0$ enforces a privacy cap by requiring the cumulative probability P_{MD}, derived from the Mahalanobis distance between x and x', to remain below a user-defined threshold τ. This limits counterfactuals to statistically plausible, high-density regions of the data distribution and reduces the risk of privacy leakage. The feasible set $\mathcal{X}$ encodes domain-specific validity constraints, such as allowable feature ranges or semantic restrictions. These constraints ensure that generated counterfactuals remain realistic and consistent with the underlying application domain. Overall, th optimisation framework provides a controlled mechanism for balancing explanation fidelity and privacy preservation. By jointly considering XAIStrength and XAILeakage during generation, the framework produces counterfactual explanations that remain interpretable while satisfying explicit privacy requirements.

E. Differential Privacy as a Configurable PET

To control disclosure risk, we integrate a DP–based privacy-enhancing mechanism into the counterfactual generation process. Privacy protection is achieved through additive Gaussian noise, rather than deterministic scaling, providing probabilistic guarantees under a formal (ε, δ) privacy budget. The noise scale is calibrated using the standard Gaussian mechanism based on the ℓ_2-sensitivity of the boundary-aligned perturbation, ensuring formal differential privacy guarantees.

Counterfactuels are generated by perturbing the boundary-aligned shift with noise calibrated to the sensitivity of the perturbation and adjusted according to the statistical atypicality of the instance, such that higher-risk cases receive stronger randomisation. This noise impacts both evaluation metrics: increasing the noise level typically reduces *XAIStrength* by distorting alignment with the decision boundary, while altering *XAILeakage* through changes in statistical deviation. By solving the privacy-constrained optimisation problem across different privacy budgets, the framework traces a privacy–utility frontier and selects counterfactuals that balance explanation fidelity with acceptable disclosure risk.

IV. Optimisation of XAIStrength–XAILeakage Trade-offs in Insider Threat Detection

This section describes the implementation of *XAIStrength* and *XAILeakage* and shows how they are integrated into an optimisation framework to balance these metrics in counterfactual explanations.

A. Implementation of XAIStrength and XAILeakage

Algorithm 1 (lines#1–5) computes *XAIStrength* by comparing the estimated perturbation $\hat{b}$ with the true minimal shift b. Line#1 calculates the absolute error e, and line#2 normalises it by the true boundary magnitude $|b| + \epsilon$ to form a fidelity score S_{XAI}. Line#3 checks whether the score becomes negative and clips it to zero if needed (line#4). The final value returned in line#5 is bounded in $[0, 1]$: higher scores mean the counterfactual closely follows the true decision boundary; lower scores show greater deviation. This formulation allows the framework to directly evaluate how accurately a generated counterfactual reflects the underlying model behaviour. Counterfactuals with higher *XAIStrength* therefore provide explanations that remain closer to the minimum perturbation required to alter the prediction, making them more reliable for interpretation and analysis.

Algorithm 2 (lines 1–5) computes the privacy leakage associated with a counterfactual explanation. Line 1 constructs the perturbation vector $\Delta = x' - x$, while line 2 computes its squared Mahalanobis distance D_M^2, which accounts for both feature variance and correlations within the cohort. In line 3, this distance is mapped to the cumulative distribution function of the χ_p^2 distribution, yielding the probability P_{MD}. Line 4 assigns the leakage score as $\mathcal{L}_{\text{XAI}} = P_{MD}$, and line 5 returns the final value. Larger leakage scores indicate counterfactuals that lie further in low-density, statistically atypical regions of the cohort distribution and therefore correspond to higher disclosure risk. By incorporating the covariance structure of the dataset, the metric captures not only the magnitude of the perturbation but also whether the resulting counterfactual deviates from common behavioural patterns observed within the population. This makes the leakage estimation more suitable for sensitive cybersecurity applications where unusual feature combinations may increase privacy risk.

Together, the two algorithms are complementary: *XAIStrength* measures fidelity to the model's decision boundary, while *XAILeakage* captures statistical plausibility and privacy risk, enabling an explicit balance between interpretability and privacy leakage. The two metrics therefore provide a practical mechanism and for evaluating whether a counterfactual explanation remains sufficiently informative while satisfying predefined privacy constraints.

Algorithm 1 XAIStrength Computation

Input: True perturbation b, estimated perturbation $\hat{b}$, stability constant ϵ
Output: S_{XAI} (XAIStrength score)
1: $e \leftarrow |\hat{b} - b|$
2: $S_{\text{XAI}} \leftarrow 1 - \dfrac{e}{|b| + \epsilon}$
3: **if** $S_{\text{XAI}} < 0$ **then**
4: $\quad S_{\text{XAI}} \leftarrow 0$
5: **return** S_{XAI}

Algorithm 2 XAILeakage Computation

Input: Factual instance x, counterfactual x', cohort covariance Σ, degrees of freedom p
Output: $\mathcal{L}_{\text{XAI}}$ (leakage score)
1: $\Delta \leftarrow x' - x$
2: $D_M^2 \leftarrow \Delta^\top \Sigma^{-1} \Delta$
3: $P_{MD} \leftarrow F_{\chi_p^2}(D_M^2)$
4: $\mathcal{L}_{\text{XAI}} \leftarrow P_{MD}$
5: **return** $\mathcal{L}_{\text{XAI}}$

B. Implementation of DP-Augmented Optimisation

Algorithm 3 implements a DP-augmented mechanism for balancing explanation fidelity and privacy risk through additive random noise. The optimisation is performed independently for each factual instance, providing instance-level differential privacy and reducing the possibility of cross-user information leakage. This is particularly important in insider threat detection, where behavioural characteristics may differ substantially across users and therefore require different levels of privacy protection.

Line 1 computes the squared Mahalanobis distance D_0 of the boundary-aligned shift b, capturing how statistically atypical the original counterfactual direction is with respect to the cohort distribution. Line 2 derives the chi-square cutoff c associated with the predefined leakage cap τ using the cumulative distribution formulation of XAILeakage. Unlike Euclidean distance, the Mahalanobis formulation incorporates both feature variance and correlations, making the leakage estimation more representative of the underlying behavioural distribution. Lines 3–5 calculate the baseline Gaussian noise scale under the standard (ε, δ) differential privacy mechanism and adjust it using a risk-adaptive scaling factor derived from D_0. Consequently, counterfactuals associated with higher statistical risk receive stronger perturbation, while lower-risk instances remain closer to the original decision boundary. This avoids introducing unnecessary distortion into explanations that already satisfy the privacy constraint.

Gaussian noise is then added to the boundary-aligned shift to generate the perturbed counterfactual. Line 9 evaluates XAILeakage by mapping the Mahalanobis distance of the perturbed instance to the χ^2_p cumulative distribution, while line 10 computes XAIStrength based on the deviation from the true minimal boundary shift. The algorithm finally returns the perturbed counterfactual together with its fidelity and leakage scores. Overall, the optimisation process provides a controlled way to balance interpretability and privacy preservation. By jointly considering both XAIStrength and XAILeakage, the framework can generate explanations that remain practically informative while limiting exposure to statistically sensitive behavioural patterns.

Algorithm 3 DP-Augmented Optimisation of XAIStrength-XAILeakage

Input: $x \in \mathbb{R}^p$ (factual), $b \in \mathbb{R}^p$ (boundary shift), $\Sigma \in \mathbb{R}^{p\times p}$ (cohort covariance), cap $\tau \in (0,1)$, (ε, δ), $\epsilon > 0$
Output: $\sigma^\star$, x', S_{XAI}, L_{XAI}

1: $D_0 \leftarrow b^\top \Sigma^{-1} b$
2: $c \leftarrow F^{-1}_{\chi^2_p}(\tau)$
3: $\Delta_2 \leftarrow \|b\|_2$
4: $\sigma_{\text{DP}} \leftarrow \dfrac{\Delta_2 \sqrt{2\ln(1.25/\delta)}}{\varepsilon}$
5: $\alpha \leftarrow \max\left(1, \sqrt{\dfrac{D_0}{c}}\right)$
6: $\sigma^\star \leftarrow \alpha \cdot \sigma_{\text{DP}}$
7: $\xi \sim \mathcal{N}(0, (\sigma^\star)^2 I)$
8: $x' \leftarrow x + b + \xi$
9: $L_{\text{XAI}} \leftarrow F_{\chi^2_p}\Big((x'-x)^\top \Sigma^{-1}(x'-x)\Big)$
10: $S_{\text{XAI}} \leftarrow 1 - \dfrac{\left|\hat{b}(x') - b\right|}{\|b\|_2 + \epsilon}$
11: **return** $\sigma^\star$, x', S_{XAI}, L_{XAI}

C. Optimisation Framework for Insider Threat Detection

To demonstrate the implementation of our optimisation framework, we evaluated it on the CERT insider threat dataset[1] (R6.2), which simulates enterprise activity such as web access, email, file operations, device use, and psychometric traits.

From the raw logs, we engineered features across three dimensions: **(i)** psychometric traits (Openness (0), Conscientiousness (C), Extraversion (E), Agreeableness (A), Neuroticism (N)); **(ii)** file-transfer behaviour (`FCwh`, `FCowh`, `FCwke`); **(iii)** domain access patterns (suspicious: `SDwh`, `SDowh`, `SDwke`; cloud: `CDwh`, `CDowh`, `CDwke`; job-related: `JDwh`, `JDowh`, `JDwke`). Features were normalised, aggregated over time, and used to train a neural network and an ensemble classifier for insider threat detection.

1) Counterfactual Explanation Generation: Post-hoc counterfactuals were generated with the DiCE framework using L_1-regularisation and domain-aware constraints. To ensure consistency with the proposed optimisation framework, DiCE-generated counterfactuals were used as initial solutions before applying privacy-aware refinement. Each counterfactual proposes the smallest plausible feature changes needed to flip a Model Prediction (MP) from *Malicious* to *Benign* (Table I). For example, for user **DMD1197** (initially *Malicious*), the counterfactual suggested reducing weekend file copies (`FCwke` : $0.0 \rightarrow 1.0$) and slightly lowering cloud-domain access outside work (`CDowh` : $1.0 \rightarrow 0.9$) while leaving other features unchanged.

2) Evaluating XAIStrength and XAILeakage: We computed the baseline squared Mahalanobis distance $D_0 = 5.064$ for the true perturbation $b = [1.0, 0.1]^\top$ under cohort covariance Σ. For a two-feature space ($p = 2$) and strict privacy cap $\tau = 0.05$, the chi-square cutoff was $c = 5.991$. The initial counterfactual yielded a leakage score $P_{MD} = 0.0438 < \tau$ and fidelity $\text{XAIStrength} = 0.8889$, indicating strong alignment with the decision boundary but near the privacy limit.

3) DP-Augmented Optimisation: We then applied our DP-augmented optimisation (Algorithm 3), solved using MATLAB's `fmincon`. The optimisation determined the minimal Gaussian noise scale $\sigma^\star$ required to ensure that the privacy constraint $P_{MD}(x, x') \leq \tau$ was satisfied after noise injection. Because the baseline Mahalanobis distance D_0 was already close to the chi-square cutoff, only a small noise scale $\sigma^\star \approx 0.04$ was required. As a result, $\text{XAIStrength} = 0.8889$ remained effectively unchanged while the leakage score stayed safely within the specified privacy cape.

V. System's Performance and Results Analysis

This section evaluates the DP-augmented optimisation framework by analysing its effect on the XAIStrength–XAILeakage trade-off and comparing metrics before and after optimisation.

[1] https://github.com/mel0778/dsa4263

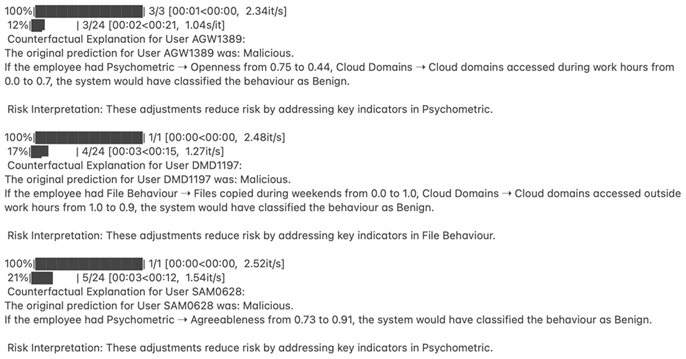

Fig. 1: CounterfactualExamples

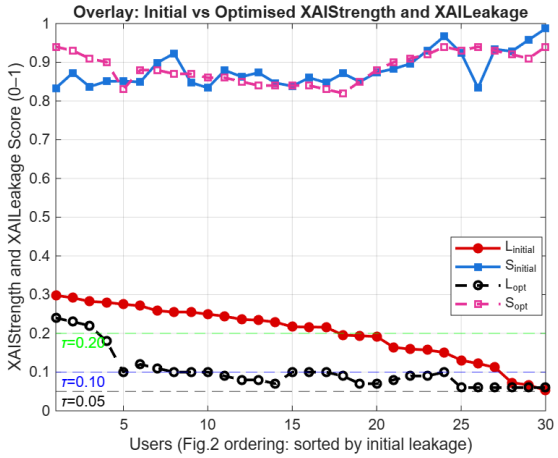

Fig. 2: Initial and Optimised XAIStrength & XAILeakage

A. Experimental Evaluation

We evaluated the impact of our DP-augmented optimisation on *XAIStrength* and *XAILeakage* for insider threat detection. Experiments were run on a 64-bit Windows 11 machine with an Intel(R) Core(TM) i9-10900KF CPU @ 3.7 GHz and 32 GB RAM. Using the DiCE framework on the CERT r6.2 dataset, we generated minimally perturbed counterfactuals to flip model predictions from *Malicious* to *Benign* (Fig. 1). XAIStrength and XAILeakage were first measured for 30 users; the optimisation then added DP noise η to find the minimal scale $\eta^\star$ satisfying our proposed three privacy caps: strict ($\tau = 0.05$), moderate ($\tau = 0.10$), and lenient ($\tau = 0.20$).

The optimisation used MATLAB's `fmincon` to maximise fidelity $f(x')$ (by minimising its negative) while enforcing the privacy constraint $g(x') \leq 0$, ensuring the Mahalanobis tail probability stayed within the specified threshold τ. Results show that the framework consistently reduces privacy leakage while preserving high explanation fidelity across all privacy levels. Implementation and configuration templates are available on GitHub.[2]

B. Comparative Analysis

Figure 2 compares the proposed metrics before and after applying the DP-augmented optimisation. Initially, most counterfactuals exhibited high fidelity, with the majority of users achieving $S_{\text{init}} \geq 0.85$. However, privacy leakage was substantial, as most users exceeded $L_{\text{init}} > 0.10$, and many violated the moderate and strict privacy caps. This indicates that highly faithful explanations may still expose cohort-specific information. The results show that counterfactuals located close to the model's decision boundary can still correspond to statistically uncommon regions of the feature space, increasing the possibility of disclosure risk in sensitive cybersecurity settings.

After optimisation, leakage is substantially reduced while fidelity remains high. Most optimised counterfactuals lie below the moderate leakage threshold, and only a small fraction violate the strict cap, while the majority retain $S_{\text{opt}} > 0.80$ and many remain above 0.90. The introduction of controlled DP perturbation reduces privacy risk while preserving the practical usefulness of the generated explanations. As a result, the framework shifts counterfactuals from highly faithful but privacy-violating to privacy-compliant with minimal loss of explanation. Compared with standard counterfactual approaches, the proposed framework achieves a more balanced trade-off between interpretability and privacy preservation by explicitly incorporating leakage constraints into the optimisation process.

C. Analysis and Discussion

The results reveal a clear trade-off between *XAIStrength* and *XAILeakage*. As shown in Fig. 2, many users initially achieved high fidelity ($S_{\text{init}} \geq 0.85$), but their perturbations lay in low-density regions, resulting in high leakage. This observation indicates that explanations closely aligned with the model's decision boundary may still correspond to statistically uncommon feature combinations, increasing the risk of exposing sensitive behavioural information. In the context of insider threat detection, such counterfactuals may unintentionally reveal atypical user activity patterns or rare behavioural characteristics associated with specific individuals or cohorts.

The proposed optimisation addresses this by injecting only the minimal amount of differential privacy noise required to satisfy the leakage cap τ. If $L_{\text{init}} \leq \tau$, fidelity is preserved; if the cap is exceeded, noise is added adaptively to reduce leakage while minimising strength degradation. Figure 2 illustrates this behaviour, showing that high-risk counterfactuals receive stronger perturbations, whereas low-risk cases remain largely unchanged. This adaptive mechanism avoids unnecessary modification of explanations that already satisfy the

[2]https://github.com/HKn-007/Insider-Threat-Detection-System

TABLE I: Factual and Counterfactual Instances: Example from Insider Threat Detection

Instances	(O)	(FCwke)	(CDowh)	(SDowh)	MP
Factual	0.72	0.0	1.0	0.6	Malicious
Counterfactual	0.72	1.0	0.9	0.6	Benign

privacy requirement and focuses perturbation on statistically higher-risk instances.

As a result, the framework enforces privacy constraints while maintaining high explanatory fidelity. The findings demonstrate that privacy-aware counterfactual explanations can remain practically informative while reducing the likelihood of disclosure risk. Overall, the results support the effectiveness of combining differential privacy with risk-aware optimisation for generating explanations in sensitive cybersecurity applications.

D. Experimental Comparison with Baseline Counterfactual Methods

To benchmark the proposed privacy-aware optimisation, we compared our approach with two representative counterfactual generation methods: the gradient-based method of Wachter *et al.* [1] and the privacy-aware learning-based approach of Vo *et al.* (L2C) [25]. Wachter's method searches for the smallest perturbation required to flip the model's prediction, focusing primarily on proximity and fidelity, while L2C incorporates representation learning to promote diversity and reduce privacy risks. Unlike standard counterfactual methods, L2C explicitly considers privacy preservation during generation. This selection enables comparison with both a widely adopted non-privacy baseline and a contemporary privacy-aware counterfactual method, addressing concerns regarding baseline fairness. The inclusion of both methods also allows assessment of whether existing privacy-aware approaches are sufficient under stricter privacy constraints in sensitive cybersecurity environments.

All methods were applied to the same neural network and ensemble classifiers trained on the CERT r6.2 insider-threat dataset, generating one counterfactual per instance for 30 representative users. Counterfactual quality was evaluated using *Proximity* $\left(\frac{1}{d}\sum_{j=1}^{d}\frac{|x'_j - x_j|}{\sigma_j}\right)$, *Sparsity* $\left(\sum_{j=1}^{d}\mathbf{1}[x'_j \neq x_j]\right)$, and the proposed *XAIStrength* and *XAILeakage*. Lower proximity, lower sparsity, and lower leakage, together with higher XAIStrength, indicate more desirable explanations. All methods were configured using recommended hyperparameters and evaluated under identical data splits to ensure a fair and controlled comparison. This consistent evaluation setup ensures that the observed differences are associated with the counterfactual generation strategies rather than variations in model training or data partitioning.

Table II reports the mean $\pm$ standard deviation across the 30 users under a strict privacy cap ($\tau = 0.05$). Wachter's method produced highly faithful counterfactuals with minimal edits but exhibited substantial privacy leakage, violating the strict privacy constraint for most users. The results indicate that optimising primarily for fidelity and proximity may unintentionally generate explanations located in statistically sparse regions of the feature space. The L2C method reduced leakage compared to Wachter's baseline, demonstrating improved privacy awareness, but still exceeded the strict privacy cap in many cases and incurred higher proximity costs.

In contrast, the proposed DP-optimised framework consistently satisfied the privacy constraint by injecting only the minimal calibrated noise $\eta^\star$ required to meet the leakage bound. This resulted in a modest increase in proximity and sparsity while preserving high XAIStrength. Although slight reductions in fidelity were observed compared with the baseline methods, the explanations remained highly interpretable and achieved substantially lower leakage scores. Overall, the comparison shows that our approach achieves a more favourable balance between explanation fidelity and privacy protection than both standard and existing privacy-aware baselines. These findings suggest that existing privacy-aware approaches may reduce leakage to some extent but do not explicitly enforce strict privacy guarantees during optimisation.

TABLE II: Performance of baseline counterfactual methods vs. the proposed DP-optimised.

Method	Proximity	Sparsity	XAIStrength	XAILeakage
Wachter [1]	0.16 ± 0.05	1.7 ± 0.6	$\mathbf{0.94 \pm 0.03}$	0.25 ± 0.08
Vo *et al.* [25]	0.7 ± 0.07	1.8 ± 0.4	$\mathbf{0.93 \pm 0.04}$	0.24 ± 0.06
DP-Optimised	0.21 ± 0.05	2.2 ± 0.8	0.90 ± 0.05	$\mathbf{0.05 \pm 0.03}$

VI. Conclusion

This work addressed the trade-off between transparency and privacy in counterfactual explanations with a particular case study in insider threat detection. We proposed a privacy-aware framework that integrates DP to add only the minimal noise needed to satisfy a user-defined privacy cap while preserving explanation strength. The framework was designed to support scenarios where explanations must remain informative for analysts and investigators while limiting the disclosure of sensitive behavioural information. Two metrics were introduced: *XAIStrength* for explanatory accuracy and *XAILeakage* for privacy risk, and the framework was evaluated on the CERT insider threat dataset. Results show that although most initial counterfactuals were highly faithful, over half exceeded moderate privacy limits ($L_{\text{init}} > 0.10$) and nearly half surpassed 0.20. These findings highlight that high-quality explanations do not necessarily guarantee privacy preservation and may still reveal statistically sensitive information about users or behavioural patterns. Our optimised differential privacy approach reduced leakage to strict, moderate, and lenient caps ($\tau = 0.05/0.10/0.20$) while maintaining strength of explainability with more than 60% of users retaining $S_{\text{opt}} \geq 0.70$ under strict privacy and over 60% keeping $S_{\text{opt}} \geq 0.85$ under moderate privacy. Comparative evaluation against both standard and privacy-aware counterfactual baselines further demonstrated that the proposed framework achieves a more consistent balance between fidelity and privacy than existing methods. Overall, the results indicate that privacy-aware counterfactual explanations can be generated without substantially reducing their practical interpretability or decision-support value in cybersecurity applications. In future work, we will assess the generalisability of our approach with different PETs and different applications and methods of XAI. We also plan

to investigate the applicability of the proposed framework to other high-risk domains where explanations involve sensitive behavioural or personal data, including healthcare, finance, and social media analysis.

ACKNOWLEDGMENT

This work has received funding from the European Union's Horizon Europe Research and Innovation Programme through the GANNDALF project under Grant Agreement No. 101167951.

REFERENCES

[1] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[2] R. Shokri, M. Strobel, and Y. Zick, "On the privacy risks of model explanations," 2021. [Online]. Available: https://arxiv.org/abs/1907.00164

[3] B. Z. H. Zhao, A. Agrawal, C. Coburn, H. J. Asghar, R. Bhaskar, M. A. Kaafar, D. Webb, and P. Dickinson, "On the (in) feasibility of attribute inference attacks on machine learning models," *arXiv preprint arXiv:2103.07101*, 2021.

[4] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, Oct. 2016. [Online]. Available: http://dx.doi.org/10.1145/2976749.2978318

[5] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, Aug. 2018. [Online]. Available: https://doi.org/10.1145/3236009

[6] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das, "Explanations based on the missing: Towards contrastive explanations with pertinent negatives," *Advances in neural information processing systems*, vol. 31, 2018.

[7] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," ser. FAT* '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 607–617. [Online]. Available: https://doi.org/10.1145/3351095.3372850

[8] A.-H. Karimi, B. Schölkopf, and I. Valera, "Algorithmic recourse: from counterfactual explanations to interventions," in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 353–362.

[9] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*. Springer, 2006, pp. 265–284.

[10] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[11] J. M. Abowd, "The u.s. census bureau adopts differential privacy," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 2867. [Online]. Available: https://doi.org/10.1145/3219819.3226070

[12] S. Goethals, K. Sörensen, and D. Martens, "The privacy issue of counterfactual explanations: explanation linkage attacks," *ACM Transactions on Intelligent Systems and Technology*, vol. 14, no. 5, pp. 1–24, 2023.

[13] R. Shokri, M. Strobel, and Y. Zick, "On the privacy risks of model explanations," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 231–241. [Online]. Available: https://doi.org/10.1145/3461702.3462533

[14] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," ser. CCS '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 1322–1333. [Online]. Available: https://doi.org/10.1145/2810103.2813677

[15] B. Hitaj, G. Ateniese, and F. Perez-Cruz, "Deep models under the gan: Information leakage from collaborative deep learning," in *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, 2017, pp. 603–618.

[16] S. K. Murakonda and R. Shokri, "ML privacy meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning," *CoRR*, vol. abs/2007.09339, 2020. [Online]. Available: https://arxiv.org/abs/2007.09339

[17] N. Patel, R. Shokri, and Y. Zick, "Model explanations with differential privacy," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1895–1904.

[18] F. Yang, Q. Feng, K. Zhou, J. Chen, and X. Hu, "Differentially private counterfactuals via functional mechanism," *arXiv preprint arXiv:2208.02878*, 2022.

[19] S. Pentyala, S. Sharma, S. Kariyappa, F. Lecue, and D. Magazzeni, "Privacy-preserving algorithmic recourse," *arXiv preprint arXiv:2311.14137*, 2023.

[20] S. Yuan and X. Wu, "Deep learning for insider threat detection: Review, challenges and opportunities," *Computers & Security*, vol. 104, p. 102221, 2021.

[21] A. White and A. S. d'Avila Garcez, "Measurable counterfactual local explanations for any classifier," *CoRR*, vol. abs/1908.03020, 2019. [Online]. Available: http://arxiv.org/abs/1908.03020

[22] S. Verma, V. Boonsanong, M. Hoang, K. Hines, J. Dickerson, and C. Shah, "Counterfactual explanations and algorithmic recourses for machine learning: A review," *ACM Comput. Surv.*, vol. 56, no. 12, Oct. 2024. [Online]. Available: https://doi.org/10.1145/3677119

[23] S. Prykhodko, N. Prykhodko, L. Makarova, and A. Pukhalevych, "Application of the squared mahalanobis distance for detecting outliers in multivariate non-gaussian data," in *2018 14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET)*, 2018, pp. 962–965.

[24] E. Cabana, R. E. Lillo, and H. Laniado, "Multivariate outlier detection based on a robust mahalanobis distance with shrinkage estimators," *Statistical papers*, vol. 62, no. 4, pp. 1583–1609, 2021.

[25] V. Vo, T. Le, V. Nguyen, H. Zhao, E. V. Bonilla, G. Haffari, and D. Phung, "Feature-based learning for diverse and privacy-preserving counterfactual explanations," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 2211–2222. [Online]. Available: https://doi.org/10.1145/3580305.3599343