The Explainability-Privacy-Utility Trade-Off for Machine Learning-Based Tabular Data Analysis

Wisam Abbasi[a], Paolo Mori[b] and Andrea Saracino[c]
Istituto di Informatica e Telematica, Consiglio Nazionale delle Ricerche, Pisa, Italy

Keywords: Data Privacy, Data Utility, Explainable AI, Privacy-Preserving Data Analysis, Trustworthy AI.

Abstract: In this paper, we present a novel privacy-preserving data analysis model, based on machine learning, applied to tabular datasets, which defines a general trade-off optimization criterion among the measures of data privacy, model explainability, and data utility, aiming at finding the optimal compromise among them. Our approach regulates the privacy parameter of the privacy-preserving mechanism used for the applied analysis algorithms and explainability techniques. Then, our method explores all possible configurations for the provided privacy parameter and manages to find the optimal configuration with the maximum achievable privacy gain and explainability similarity while minimizing harm to data utility. To validate our methodology, we conducted experiments using multiple classifiers for a binary classification problem on the Adult dataset, a well-known tabular dataset with sensitive attributes. We used (ε, δ)-differential privacy as a privacy mechanism and multiple model explanation methods. The results demonstrate the effectiveness of our approach in selecting an optimal configuration, that achieves the dual objective of safeguarding data privacy and providing model explanations of comparable quality to those generated from real data. Furthermore, the proposed method was able to preserve the quality of analyzed data, leading to accurate predictions.

1 INTRODUCTION

The volume of data generated and collected across various domains and industries has been increasing exponentially (Jaseena et al., 2014), accompanied by an impressive advancement in data analysis methods aimed at providing deep insights and revealing hidden patterns and correlations for better decision-making. However, as data becomes high-dimensional and analysis methods are of increasing complexity, they may expose sensitive data and/or make unfair or wrong decisions (Jakku et al., 2019). Thus requiring privacy-preserving and explainable analysis models.

A privacy-preserving and explainable model enhances the analysis function by protecting sensitive data from being disclosed while providing explanations for the predictions made. Various data *anonymization* methods have been proposed in the literature to ensure privacy and prevent re-identification, like the (ε, δ)-*differential privacy* (Dwork, 2008) mechanism. *Decision explainability* implementation in machine learning-based data analysis models has also become a hot topic (Rasheed et al., 2021), being a key requirement for Artificial Intelligence (AI) systems to be trustworthy from ethical and technical perspectives, as pointed out in the EU proposal for the Artificial Intelligence Act[1] (Budig et al., 2020).

However, adopting any of these concepts may compromise the others. For example, *decision explainability* may cause security breaches like inference and reconstruction attacks (Shokri et al., 2019), and privacy-preserving techniques may compromise explainability (Budig et al., 2020). Therefore, it is crucial to consider all these elements in a comprehensive research approach (Hleg, 2019).

To address this, we propose a machine learning-based data analysis model applied to tabular data, which defines a general *Trade-Off* criterion for *Data Privacy*, *Data Utility*, and *Model Explainability* aimed at finding the optimal compromise among them. In detail, the model explores all possible configurations for the provided privacy parameter values and finds the best configuration with the maximum

[a] https://orcid.org/0000-0002-6901-1838
[b] https://orcid.org/0000-0002-6618-0388
[c] https://orcid.org/0000-0001-8149-9322

[1] Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence: https://bit.ly/3y5wf6e

Abbasi, W., Mori, P. and Saracino, A.
The Explainability-Privacy-Utility Trade-Off for Machine Learning-Based Tabular Data Analysis.
DOI: 10.5220/0012137800003555
In *Proceedings of the 20th International Conference on Security and Cryptography (SECRYPT 2023)*, pages 511-519
ISBN: 978-989-758-666-8; ISSN: 2184-7711

achievable *Privacy Gain* and *Explainability Similarity* causing the least harm to *Data Utility*. Specifically, our approach defines the metrics of *Privacy Gain*, *Explainability Similarity*, and *Utility Loss* with a general optimization *Trade-Off* criterion and *compatibility matrix*.

The proposed methodology is validated through experiments for tabular dataset classification using multiple machine learning models and the (ε, δ)-differential privacy (DP) mechanism. The approach regulates the privacy parameter and measures the obtained *Privacy Gain*, *Utility Loss*, and *Explainability Similarity* to find the best *Trade-Off* score. It can be extended to other data types and analysis methods.

This paper provides the following contributions:

- we propose a novel approach to trade-off *Data Privacy* and *Model Explainability* while maintaining *Data Utility* in ML-based tabular data analysis;
- we define and explain the concepts of *Privacy Gain*, *Explainability Similarity*, and *Utility Loss*, which are extracted from the implementation of (ε, δ)-*differential privacy* , *model performance*, *variable importance*, *partial dependence profiles*, and *accumulated local profiles*;
- we define a *Trade-Off* criterion for privacy, explainability, and *Data Utility* optimization, which combines *Privacy Gain*, *Explainability Similarity*, and *Utility Loss* to find the best *Trade-Off* score;
- we evaluate the proposed approach through experiments on a tabular dataset, where the privacy parameter is regulated, and the effect on the *Trade-Off* score is measured to reach the optimal score.

The paper's structure is as follows: Section 2 provides an overview of privacy-preserving mechanisms and Explainable Artificial Intelligence (XAI). Section 3 presents privacy, explainability, and *Data Utility* measures, Section 4 outlines the proposed methodology. Section 5 reports conducted use cases and experiments, with results discussion. Section 6 compares related work, while Section 7 concludes the paper.

2 BACKGROUND

This section covers background concepts of privacy-preserving mechanisms and explainability methods.

2.1 Data Privacy Preserving Techniques

Data privacy-preserving techniques aim to protect sensitive data while being shared or analyzed. These methods perform anonymization operations on the data to satisfy the privacy requirement, including *Generalization*, *Suppression*, *Anatomization*, *Perturbation*, and *Permutation* operations (Fung et al., 2010). Some of the most well-known anonymization-based techniques that exploit these operations are *k-anonymity* (Samarati, 2001; Sweeney, 2002), *t-closeness* (Li et al., 2007), *l-diversity*, *Distinct l-diversity*, *Entropy l-diversity*, *Recursive* (c,l)-*diversity* (Machanavajjhala et al., 2007), *differential privacy* (Dwork, 2008), and *Generative Adversarial Networks (GANs)*. We will focus on the last two techniques in this work.

(ε, δ)-*differential privacy (DP)* is a privacy-preserving technique that protects individual data from being identified or reconstructed by ensuring that the output of a differential private analysis on two datasets differing by only one record is indistinguishable. Thus, individual records do not contribute to the results in a way that causes the model to remember identifying individual instances, and the original data cannot be reverse-engineered from the analysis results. DP incorporates random *Laplace* or *Gaussian distribution* noise to data. The degree to which these data are indistinguishable depends on the *sensitivity parameter* ℓ_2 sensitivity and the *privacy budget* parameter ε. DP formula is shown in Equation (1) (Dwork, 2008).

$$Pr[R(D_1) \in S] \leq Pr[R(D_2) \in S] \times exp(\varepsilon) + \delta \quad (1)$$

Where Pr is the probability, ε is the privacy budget, δ is the failure probability, R is the randomized function that incorporates (ε, δ)-differential privacy for datasets D_1 and D_2 which differ in at least one record, and $S \subseteq Range(R)$.

For privacy analysis, we use the *moments accountant* privacy budget tracking method (Abadi et al., 2016), which uses a Differential Private Stochastic Gradient Descent (DP-SGD) algorithm with an additive Sampled Gaussian Mechanism (SGM) to add Gaussian noise to randomly sampled elements (Dwork et al., 2006; Raskhodnikova et al., 2008) as defined in Equation (2) for a real-valued function f mapping subsets of D to $\mathbb{R}^d$:

$$R(D) \triangleq f(D) + N(0, S_f \sigma^2) \quad (2)$$

Where D is a dataset from which a subset is randomly sampled with a sampling rate $0 < q \leq 1$ to be used by the algorithm f. $N(0, \sigma^2)$ is the Gaussian distribution of the noise added with a mean equals to 0, and σ is the noise added with $S_f \sigma^2$ standard deviation of the noise bounded to ℓ_2 sensitivity.

The accounting procedure of the moments accountant allows to prove that an algorithm is (ε, δ)-differential private for appropriately selected configurations of the parameters for any $\varepsilon < c_1 q^2 T$ and for

any ℓ_2 sensitivity > 0 if the noise multiplier σ was defined as in Equation (3) proposed in (Abadi et al., 2016):

$$\sigma \geq c_2 \frac{q\sqrt{Tlog(1/\delta)}}{\varepsilon} \quad (3)$$

where c_1 and c_2 are constants so that given the sampling probability $q = L/n$, L is the sampling ratio, n is the size of the dataset, and T is the number of training steps, and $T = \frac{E}{q}$ and E is the number of epochs. The relationship between the noise multiplier and the privacy budget ε is negative, which implies better privacy protection when increasing the value of the noise multiplier.

2.1.1 DP-WGAN

Generative Adversarial Networks (GANs) are used to generate synthetic data with a similar distribution of original data, but with high quality (Goodfellow et al., 2014). *Wasserstein GAN (WGAN)* is a variant of GANs, which was proposed to generate data with better training performance by minimizing the distance between the original data distribution and the synthesized distribution considering using the *Wasserstein-1 distance* concept (Arjovsky et al., 2017). *Differential Privacy* is used to protect the privacy of synthetic data generated using this method, resulting in the *DP-WGAN* variant (Xie et al., 2018).

2.2 Explainable Artificial Intelligence

As AI algorithms get more complex, it becomes challenging for humans to interpret their predictions, leading to a lack of trust in the model's accuracy and transparency. On the one hand, some AI models are interpretable by design (inherently interpretable), meaning that their results can be easily explained due to their simple structure, such as decision trees (Weisberg, 2005). These models are called *glass box models* and *intrinsic models* (Biecek and Burzykowski, 2021). On the other hand, more powerful ML algorithms are less interpretable due to their complexity. To address this, *Explainable AI (XAI)* has emerged, aiming to produce human-level explanations for complex AI models (Rai, 2020). XAI techniques are applied *pre-model*, *in-model*, or *post-model*, and can be *model specific* or *model agnostic*, producing either *local explanations* or *global explanations* (Linardatos et al., 2021).

Our focus is on *model agnostic*, *global*, and *post-model XAI* methods, which examine the model used for the entire dataset. These techniques aim to produce *dataset-level explanations* (Biecek and Burzykowski, 2021) and are applied at different levels such as *Model Performance* exploration techniques using the performance measures of *Recall*, *Precision*, *F1 Score*, *Accuracy*, and *The Area Under the Curve (AUC)* (Biecek and Burzykowski, 2021). Moreover, *Variable Importance* explanations are used to quantify the impact of each variable on the final prediction made by the model (Breiman, 2001). Finally, explain Model prediction dependency on variable changes using *Partial Dependence (PD) Profiles* and *Accumulated Local (AL) dependent Profiles* (Biecek and Burzykowski, 2021).

3 FORMALISM

This work uses DP with WGAN to generate differential private synthetics, as presented in Section 2.1. The *Privacy Gain* refers to the level of data uncertainty introduced by modifications to the real dataset D to produce a sanitized dataset D'. The degree of privacy is controlled by the *sensitivity parameter* ℓ_2 sensitivity$= 1e-5$ and the *Gaussian noise variance multiplier*, which varies between 0 for no privacy and 1 for full privacy (maximum degree of privacy). *Privacy Gain* measures privacy added by the Differential Privacy mechanism. It is quantified based on the *privacy budget* ε obtained from the *privacy parameter* σ. A lower *privacy budget* implies better privacy. Gained privacy $PG(D,\lambda)$ for dataset D and classifier λ is calculated using Equation (4).

$$PG(D,\lambda) = \frac{1}{\varepsilon(D,\lambda)} \quad (4)$$

Following the approach applied in (Jayaraman and Evans, 2019), the loss of *Data Utility* is calculated by comparing the accuracy of the model applied to the original dataset and the accuracy of the model applied to the anonymized dataset. Thus, *Utility Loss* is represented for dataset D and classifier λ in Equation (5). The value of the *Utility Loss* is in the interval $[-1,1]$, where -1 is the minimum *Utility Loss* if the accuracy of the model for the anonymized dataset is 1 and the model accuracy of the original dataset is 0. And 1 is the maximum *Utility Loss* if the model accuracy for the anonymized dataset is 0 and the model accuracy of the original dataset is 1.

$$UL(D,D\prime,\lambda) = Acc(\lambda(D)) - Acc(\lambda(D\prime)) \quad (5)$$

This work aims to provide explanations for predictions made by analysis models applied to datasets, particularly in terms of how and why a certain decision has been made. The goal is to ensure that the model is fair and not making predictions based on discriminatory parameters such as gender or race. Var-

ious XAI methods presented in Section 2.2 are employed to provide these explanations.

Adopting a privacy mechanism may affect the quality of the explanations produced by XAI methods. Thus, we assess the similarity between explanations generated from the analysis model applied to the original dataset and those from the differential private dataset with varying levels of privacy. A higher similarity indicates a better situation, where the explainability degree is less affected by the privacy mechanism. We use the *Pearson correlation coefficient (PCC)* to quantify the similarity between explanations, as shown in Equation (6).

$$Sim_{PCC} = \frac{\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i-\mu_x)^2\sum_{i=1}^{n}(y_i-\mu_y)^2}} \quad (6)$$

Where x_i is the ith value of the variable x, y_i is the ith value of variable y, μ_x the mean of all x variable values in the dataset, and μ_y the mean of all y variable values in the dataset. The similarity assessment is performed per XAI method as summarized below:

1. *Model performance Similarity*: Correlation assessment between model performance parameters with and without privacy constraints
2. *Variable Importance Similarity*: Correlation assessment of variable importance between anonymized and original datasets
3. *Partial dependence Profiles Similarity*: Correlation assessment between PD profiles of anonymized and original datasets
4. *Accumulated Local (AL) Profiles Similarity*: Correlation assessment between AL profiles of anonymized and original datasets

4 PROPOSED METHODOLOGY

This section covers the privacy preservation and explainability techniques and the proposed strategy for their implementation.

4.1 Problem Statement and Architecture

We use a scenario (Figure 1) where a stakeholder collaborates with an aggregation server to process and share datasets from multiple entities. The server is secure but untrusted, so a privacy-preserving mechanism is needed to protect data and enhance trust. Additionally, an explainability mechanism is needed for transparent and trustworthy predictions. The model architecture starts with a privacy mechanism that produces a differential private synthetic dataset. The sanitized dataset is then shared with the server and analyzed using a machine learning classifier. The predicted result is explained using an explainer based on the specified explainability method.

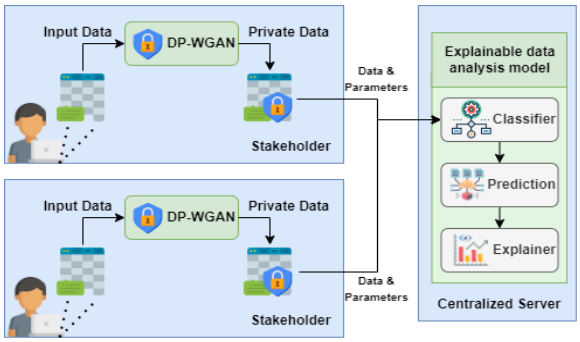


Figure 1: Privacy-preserving Tabular Data Classification Scenario.

To implement privacy and explainability mechanisms in a single solution, we must carefully consider all possible configurations and degrees of privacy to get the maximum possible *Privacy Gain* and *Explainability Similarity* while minimizing the *Data Utility* loss. This is achieved by formulating the problem as linear optimization. However, stakeholders may have specific requirements that do not necessarily result in the best *Trade-Off* score, so they can request a configuration that fits their needs. To achieve this, a *compatibility matrix* is used to find a suitable configuration for all stakeholders. To measure the *Privacy Gain*, *Explainability Similarity*, and *Utility Loss* as discussed in Section 3, we utilize them as below:

Privacy Preserving for Classification. Using *Differential Privacy*, sensitive dataset attributes are protected, with the degree of privacy controlled by the *Gaussian noise variance multiplier*. The *noise multiplier* is an input parameter that ranges between 0 and 1 in increments of 0.1, where 0 represents no privacy and 1 represents maximum privacy. The *Privacy Gain* is calculated as $1/\varepsilon$.

Model Explanation and Similarity for Classification. Our approach uses tabular datasets as input and provides different types of explanations for the predictions made by analysis functions, such as *model performance explanations*, *variable importance explanations*, *PD profiles*, and *AL profiles*. Since *privacy* enforcement affects *explainability*, a similarity assessment is conducted to compare the explanations of the original dataset and the differential private dataset. A higher similarity value indicates a better result, meaning that the explanations for the private data are similar to those of the original data.

Data Utility Loss for Classification. The *Utility Loss* is measured as the model's accuracy difference between the original and sanitized datasets.

4.2 Compatibility Matrix and Trade-Off Score Optimization

Privacy techniques improve *Privacy Gain* in classification models, but they may reduce *Data Utility* and *Explainability Similarity*. To address this, we use a *Trade-Off* formula (Equation (7)) to optimize the *Privacy Gain* and *Explainability Similarity* while minimizing *Utility Loss* and obtain the best *Trade-Off* score $T(D,D',\lambda)$. The *Trade-Off* formula allows us to balance these different objectives by combining them into a single score. The numerator represents the desirable objectives of privacy and explainability, while the denominator represents the undesirable objective of *Utility Loss*. By dividing the desirable objectives by the undesirable objective, we obtain a *Trade-Off* score that reflects the balance between the different objectives. The *Trade-Off* formula and linear optimization offer a systematic and objective method for balancing conflicting objectives like *Utility Loss*, *Privacy Gain*, and *Explainability Similarity* in model development. By combining these objectives into a single score, the equation enables the optimal balance of these objectives by considering a broad range of values for each objective, and linear optimization finds the best *Trade-Off* score that meets all requirements and constraints.

$$T(D,D\prime,\lambda) \quad = \quad \frac{PG(D,\lambda)+ES(D,\lambda)}{2+U(D,D\prime,\lambda))} \quad (7)$$

where $PG(D,\lambda)$ is the *Privacy Gain*, $ES(D,\lambda)$ is the *Explainability Similarity*, $U(D,D\prime,\lambda)$ is the *Utility Loss*, D is the original dataset and $D\prime$ is the sanitized dataset, whilst λ is the analysis model. To determine the optimal values for these parameters, we formulate the following linear optimization problem:

Maximize $T(D,D\prime,\lambda)$ subject to:
$PG(D,\lambda)_{min} \leq PG(D,\lambda) \leq PG(D,\lambda)_{max}$
$ES(D,\lambda)_{min} \leq ES(D,\lambda) \leq ES(D,\lambda)_{max}$
$U(D,D\prime,\lambda))_{min} \leq U(D,D\prime,\lambda)) \leq U(D,D\prime,\lambda))_{max}$

The *Trade-Off* between the privacy mechanism, explainability techniques, and analysis models on a dataset is computed by Equation (7). A tri-dimensional A *compatibility matrix* utilizes the resulting *Trade-Off* scores and aligns them with the stakeholders' requirements to achieve the optimal *Trade-Off* score (Sheikhalishahi et al., 2021), as illustrated in Figure 2. The x-axis of the matrix reflects the privacy mechanism degrees, the y-axis reports the explainability techniques and the z-axis represents the used datasets from different stakeholders with different privacy and explainability requirements. For each classification model used, a *compatibility matrix* is constructed with *Trade-Off* scores for all possible degrees of privacy and explainability mechanisms on all datasets. If the privacy degree of dataset D_i does not meet the requirements set by the owner, the corresponding element on the compatibility matrix is set to 0. This means that it is impossible to compute a *Trade-Off* score for that configuration.

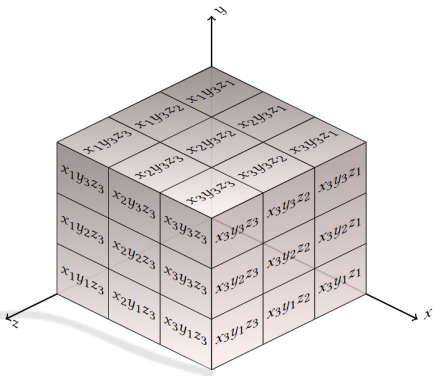


Figure 2: Tri-dimensional compatibility matrix.

For a classification model λ, the optimal *Trade-Off* score is selected based on Equation (8), where for each degree of privacy on the x-axis and the explainability mechanism on the y-axis applied to a specific dataset on the z-axis, a *Trade-Off* score for each configuration is presented in a cell of the *compatibility matrix*. The optimal configuration is the one with the maximum averaged *Trade-Off* score over all datasets.

$$\bar{T}(\bar{D},\bar{D}\prime,\lambda) = \sum_{i=1}^{m} w_i T(D_i,D_i\prime,\lambda) \quad (8)$$

Each *Trade-Off* score in the weighted average calculation has a weight of 1 divided by the count of datasets used, denoted as w_i.

In another scenario, where specific requirements have been defined, the *Trade-Off* score is selected by choosing the defined privacy degree to be equal to or greater than a specific threshold with the preferred explainability mechanism and gets the related *Trade-Off* score based on the *compatibility matrix*. To implement the *Trade-Off* scoring and *compatibility matrix*, we consider a scenario with two stakeholders, where only one satisfies the specified privacy degree and explainability mechanisms. This results in a tri-dimensional matrix with 10 degrees for privacy and 6 options for explainability mechanisms on the x and y axes, respectively, and 1 input dataset. The resulting *compatibility matrix* would be of size $10 \times 6 \times 1$. The

scores can be compared to the *Trade-Off* score when no privacy mechanism is applied.

5 USE CASES AND EXPERIMENTS

This section presents experiments on a tabular dataset using privacy-preserving and explainability techniques to achieve high model accuracy while preserving data privacy and model explainability: *model performance explainability*, *variable importance explainability*, *PD profiles*, and *AL profiles*. *DP-WGAN* is used as the privacy mechanism, and multiple methods are utilized for model explainability. The experiments measure *Privacy Gain*, *Explainability Similarity*, and *Utility Loss*. The *Trade-Off* score is then calculated for all possible settings, and the configuration with the best *Trade-Off* score or the predefined privacy degree is selected.

The UCI Machine Learning Repository's Adult dataset (Blake, 1998) has been used to test our approach as a binary classification problem. It consists of 14 categorical and integer attributes with sensitive social information and $48,842$ instances. The instances have been split into 75% for training and 25% for testing. The dataset aims to predict whether an individual earns more than $50K$ a year.

Experiments involve classification with varying levels of privacy using *DP-WGAN* as a differential private generative model, implementing a *private Wasserstein Generative Adversarial Network (GAN)* with the noisy gradient descent moments accountant. The privacy parameter σ ranges from 0 to 1, where 0 denotes no privacy and 1 is maximum privacy. ℓ_2 sensitivity is set to 10^{-5} for privacy guarantee. Multiple explanation mechanisms were also used.

5.1 Model

The study utilized three classification models: *Logistic Regression (LR)*, *Multilayer Perceptrons (MLP)*, and *Gaussian NB*. The original dataset was transformed with *DP-WGAN* using the Private Data Generation Toolbox[2]. to produce the differential private dataset with varying levels of privacy. Multiple σ values are defined to control privacy, and the model returns the privacy budget ε. The DALEX framework[3] is used for explanatory model analysis, providing methods for global explanations such as *model performance*, *variable importance*, and *variable impact*. The approach computes *Privacy Gain*, four models *explanation assessments*, and *Utility Loss* for the defined privacy. Six *Trade-Off* scores are calculated for each privacy value, and an optimal *Trade-Off* score is selected using a *compatibility matrix* for each explainability method/averaged explainability and each classification model. The *Trade-Off* score for averaged explainability is computed among *Privacy Gain*, *Utility Loss*, and *Explainability Similarity*, which is averaged between *model performance*, *variable importance*, and *PD/AL profiles*.

[2]https://github.com/BorealisAI/private-data-generation

[3]https://github.com/ModelOriented/DALEX

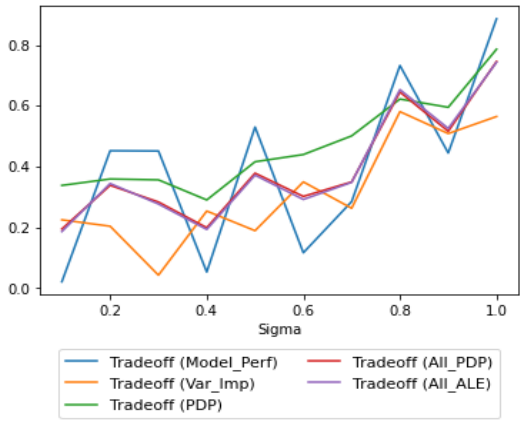
5.2 Differential Privacy with Explainability Results

This sub-section reports the results of experiments using the *DP-WGAN* mechanism and explainability methods presented in Section 2.2. Models results are represented by the *Trade-Off* score, *Privacy Gain*, *Explainability Similarity*, *Utility Loss*, and *Accuracy*.

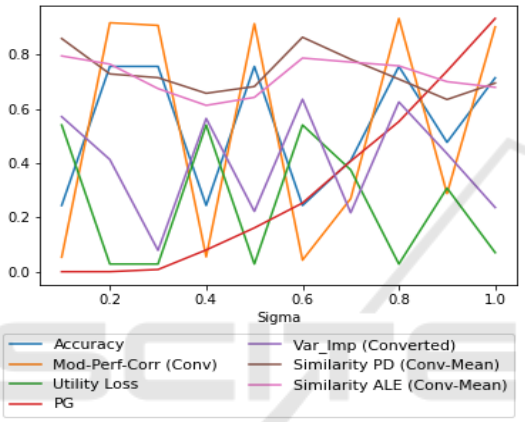
Logistic Regression classifier results are shown in Figure 3. The *Trade-Off* scores in sub-figure 3a with the privacy parameter σ reported on the *x*-axis and the *Trade-Off* scores reported on the *y*-axis for each explainability method and the average *Trade-Off* score for three explainability methods in one *Trade-Off* score each of them represent $1/3$ score weight. The *Trade-Off* scores generally increase with an increase of the privacy degree, and the highest score at maximum $\sigma = 1$, followed by the *Trade-Off* score at $\sigma = 0.8$, except for the *variable importance* explainability method with the highest *Trade-Off* score at $\sigma = 0.8$. All parameters involved in the *Trade-Off* formula are shown in sub-figure 3b on the *y*-axis with the σ on the *x*-axis. *Privacy Gain* increases with no effect on other parameters, while the *Utility Loss* and *model performance correlation* have a negative relationship due to the decrease in model *Accuracy* as σ increases, leading to an increase in *Utility Loss*.

Figure 4 shows the results of the *Gaussian NB* classifier. The *Trade-Off* scores in sub-figure 4a have a positive relationship with the privacy parameter, with the highest score at $\sigma = 1$, except for the *variable importance* and *model performance* methods which fluctuate. Sub-figure 4b represents all parameters of the *Trade-Off* formula, where the *Privacy Gain* increases with the privacy parameter value, and the *Utility Loss* and *model performance correlation* have a negative relationship with the model *Accuracy*.

The *MLP* classifier yields the *Trade-Off* scores presented in Figure 5a Similar to the *Logistic Regression* classifier results, the figure demonstrates a generally positive relationship between the *Trade-Off* scores and the privacy parameter, with the highest

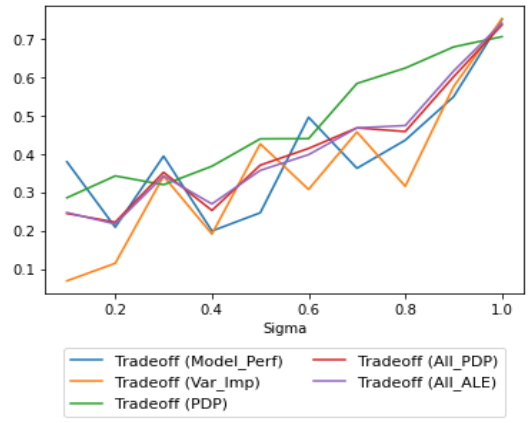


(a) Trade-off Results.

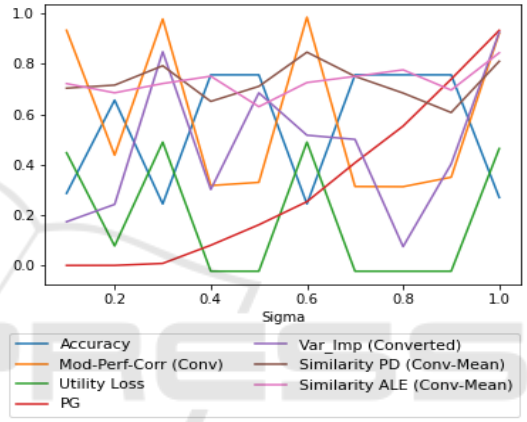


(b) Parameters Correlation values.

Figure 3: Logistic Regression Results.



(a) Trade-off Results.



(b) Parameters Correlation values.

Figure 4: Gaussian NB Results.

scores achieved at $\sigma = 1$. In order to gain further insights into these results, the individual parameters of the *Trade-Off* formula are analyzed in Figure 5b. The *Privacy Gain* is shown to increase proportionally with the privacy parameter. Furthermore, the figure reveals the negative relationship between the *Utility Loss* and the *model performance correlation*, which can be attributed to the inverse relationship between the *Utility Loss* and the *Accuracy* parameters.

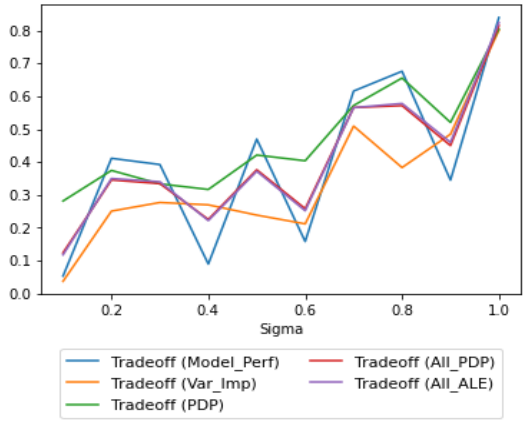
The *compatibility matrix* presents results from applying *DP-WGAN* and explainability mechanisms and is used to select the best configuration with the highest *Trade-Off* score for each classifier or a preferred degree of privacy and explainability mechanism. With 10 privacy degrees, 4 explainability mechanisms, 2 combinations of explainability mechanisms, and one dataset, the *Trade-Off* Equation (8) has a weight of 1, and the matrix dimensions for each classifier are $10 \times 6 \times 1$.

Logistic Regression Classifier: the *compatibility matrix* is represented in Table 1 with the privacy degrees in the rows, explainability mechanisms in the columns, and their *Trade-Off scores (TO)* in the cells. The table displays the *Trade-Off scores (TO)* for 10 privacy degrees, 4 explainability mechanisms, and 2 combinations of explainability mechanisms. The best (maximum) overall *Trade-Off* score is obtained at $\sigma = 1$ with *model performance explanation* of 0.89. Additionally, the best *Trade-Off* score per explainability mechanism is also reported. Specific configurations can be selected, as explained in the problem formulation, such as $\sigma = 0.8$ with all explainability mechanisms, which yields a *Trade-Off* score of 0.65.

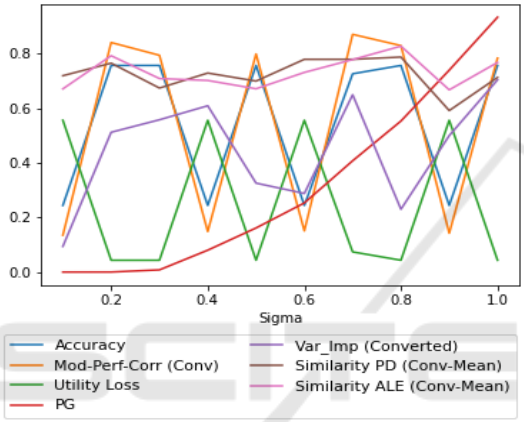
Table 1: Logistic Regression Classifier Results.

σ	TO (Model Perf)	TO (Var Imp)	TO (PD)	TO (AL)	TO (All PD)	TO (All AL)
0.1	0.02	0.22	0.34	0.31	0.19	0.19
0.2	0.45	0.20	0.36	0.38	0.34	0.34
0.3	0.45	0.04	0.36	0.34	0.28	0.28
0.4	0.05	0.25	0.29	0.27	0.20	0.19
0.5	0.53	0.19	0.42	0.40	0.38	0.37
0.6	0.12	0.35	0.44	0.41	0.30	0.29
0.7	0.28	0.26	0.50	0.50	0.35	0.35
0.8	0.73	0.58	0.62	0.65	0.65	0.65
0.9	0.44	0.51	0.59	0.62	0.52	0.53
1	0.89	0.56	0.79	0.78	0.75	0.74

MLP: the *compatibility matrix* is represented in Table 2. The best *Trade-Off* score occurs at $\sigma = 1$



(a) Trade-off Results.



(b) Parameters Correlation values.

Figure 5: MLP Results.

with *model performance explanation* of 0.84. The best *Trade-Off* score per explainability mechanism is shown in the dark blue cells. All the best *Trade-Off* score scores occur at the highest privacy degree.

Gaussian NB: the *compatibility matrix* for the *Gaussian NB* Classifier is in Table 3. The best *Trade-Off* score is at $\sigma = 1$ with *model performance explanation* and *variable importance explainability* method with a score of 0.75. The best *Trade-Off* score per explainability mechanism is in the dark blue cells and all occur at the maximum value of the privacy degree.

The *Logistic Regression* classifier has the best *Trade-Off* score for *model performance explanation* at $\sigma = 1$, while the *MLP* classifier has the best *Trade-Off* scores for *Variable Importance explanation*, *PD Profiles explanation*, *AL Profiles explanations*, *Averaged Trade-off with respect to PD Profiles*, and *Averaged Trade-off with respect to AL profiles* at $\sigma = 1$. These results suggest that *DP* does not significantly impact *data utility*, the *Accuracy* of the model, and the *explainability* of the model for these classifiers.

Table 2: MLP Classifier Results.

σ	TO (Model Perf)	TO (Var Imp)	TO (PD)	TO (AL)	TO (All PD)	TO (All AL)
0.1	0.05	0.04	0.28	0.26	0.12	0.12
0.2	0.41	0.25	0.37	0.39	0.35	0.35
0.3	0.39	0.28	0.33	0.35	0.33	0.34
0.4	0.09	0.27	0.32	0.31	0.23	0.22
0.5	0.47	0.24	0.42	0.41	0.38	0.37
0.6	0.16	0.21	0.40	0.38	0.26	0.25
0.7	0.62	0.51	0.57	0.57	0.57	0.57
0.8	0.68	0.38	0.66	0.68	0.57	0.58
0.9	0.34	0.48	0.52	0.55	0.45	0.46
1	0.84	0.80	0.80	0.83	0.82	0.82

Table 3: Gaussian NB Classifier Results.

σ	TO (Model Perf)	TO (Var Imp)	TO (PD)	TO (AL)	TO (All PD)	TO (All AL)
0.1	0.38	0.07	0.29	0.29	0.25	0.25
0.2	0.21	0.12	0.34	0.33	0.22	0.22
0.3	0.40	0.34	0.32	0.29	0.35	0.34
0.4	0.20	0.19	0.37	0.42	0.25	0.27
0.5	0.25	0.43	0.44	0.40	0.37	0.36
0.6	0.50	0.31	0.44	0.39	0.42	0.40
0.7	0.36	0.46	0.59	0.59	0.47	0.47
0.8	0.44	0.32	0.63	0.67	0.46	0.48
0.9	0.55	0.58	0.68	0.73	0.60	0.62
1	0.75	0.75	0.71	0.72	0.74	0.74

6 RELATED WORK

In (Harder et al., 2020), a method for privacy-preserving data classification using DP and providing model explainability using Locally Linear Maps was proposed. However, adding more noise to preserve privacy can negatively affect prediction accuracy, leading to a trade-off between the two. To address this, the authors used the Johnson-Lindenstrauss transform to decrease the dimensionality of the Locally Linear Maps model. Tuning the number of linear maps allowed for a reasonable trade-off between privacy, accuracy, and explainability on small datasets. Future work could investigate the trade-off on larger datasets and more complex representations.

The privacy-preserving data analysis model proposed by (Patel et al., 2020) lacks global explanations, *Data Utility* measurement, and an optimization criterion to balance privacy, explainability, and accuracy metrics. While various privacy-preserving mechanisms, like federated learning with local explainability, have been proposed in the literature, some of these approaches do not consider *Data Utility* or provide an optimization criterion, and no experiments were conducted. Theoretical methods without optimization techniques have been proposed, such as the method described in (Ramon and Basu, 2020).

7 CONCLUSION

Our work proposes a new method for privacy-preserving data analysis that balances data utility and model explainability with privacy. It also defines a Trade-Off criterion for the use of these measures in an optimal manner. Thus, we provide accurate results without compromising data privacy or model explain-

ability. We validate our approach on the Adult dataset using three classification models, demonstrating the potential for trustworthy data analysis while controlling privacy levels. In future work, we aim to extend our approach to other types of datasets (i.e. non-tabular) and analysis techniques using various privacy preservation techniques.

ACKNOWLEDGEMENTS

This work has been partially funded by the EU-funded project H2020 SIFIS-Home GA ID:952652.

REFERENCES

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR.

Biecek, P. and Burzykowski, T. (2021). *Explanatory model analysis: explore, explain, and examine predictive models*. CRC Press.

Blake, C. (1998). Cj merz uci repository of machine learning databases. *University of California at Irvine*.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Budig, T., Herrmann, S., and Dietz, A. (2020). Trade-offs between privacy-preserving and explainable machine learning in healthcare. In *Seminar Paper, Inst. Appl. Informat. Formal Description Methods (AIFB), KIT Dept. Econom. Manage., Karlsruhe, Germany*.

Dwork, C. (2008). Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. (2006). Our data, ourselves: Privacy via distributed noise generation. In *Annual int. conference on the theory and applications of cryptographic techniques*, pages 486–503. Springer.

Fung, B. C., Wang, K., Chen, R., and Yu, P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys (Csur)*, 42(4):1–53.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Harder, F., Bauer, M., and Park, M. (2020). Interpretable and differentially private predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4083–4090.

Hleg, A. (2019). Ethics guidelines for trustworthy ai. *B-1049 Brussels*.

Jakku, E., Taylor, B., Fleming, A., Mason, C., Fielke, S., Sounness, C., and Thorburn, P. (2019). "if they don't tell us what they do with it, why would we trust them?" trust, transparency and benefit-sharing in smart farming. *NJAS-Wageningen Journal of Life Sciences*, 90:100285.

Jaseena, K., David, J. M., et al. (2014). Issues, challenges, and solutions: big data mining. *CS & IT-CSCP*, 4(13):131–140.

Jayaraman, B. and Evans, D. (2019). Evaluating differentially private machine learning in practice. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pages 1895–1912.

Li, N., Li, T., and Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115. IEEE.

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2021). Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18.

Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3–es.

Patel, N., Shokri, R., and Zick, Y. (2020). Model explanations with differential privacy. *arXiv preprint arXiv:2006.09129*.

Rai, A. (2020). Explainable ai: From black box to glass box. *Journal of the Academy of Marketing Science*, 48(1):137–141.

Ramon, J. and Basu, M. (2020). Interpretable privacy with optimizable utility. In *ECML/PKDD workshop on eXplainable Knowledge Discovery in Data mining*.

Rasheed, K., Qayyum, A., Ghaly, M., Al-Fuqaha, A., Razi, A., and Qadir, J. (2021). Explainable, trustworthy, and ethical machine learning for healthcare: A survey.

Raskhodnikova, S., Smith, A., Lee, H. K., Nissim, K., and Kasiviswanathan, S. P. (2008). What can we learn privately. In *Proceedings of the 54th Annual Symposium on Foundations of Computer Science*, pages 531–540.

Samarati, P. (2001). Protecting respondents identities in microdata release. *IEEE transactions on Knowledge and Data Engineering*, 13(6):1010–1027.

Sheikhalishahi, M., Saracino, A., Martinelli, F., and Marra, A. L. (2021). Privacy preserving data sharing and analysis for edge-based architectures. *International Journal of Information Security*.

Shokri, R., Strobel, M., and Zick, Y. (2019). Privacy risks of explaining machine learning models.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570.

Weisberg, S. (2005). *Applied linear regression*, volume 528. John Wiley & Sons.

Xie, L., Lin, K., Wang, S., Wang, F., and Zhou, J. (2018). Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*.