Data-level sampling for dealing with imbalanced datasets: better protection against membership inference attacks

Karla F. C. da Silva[1] **, Antônio de Abreu Batista-Jr**[1] **, Jesús P. Mena-Chalco**[2] **, Luciano R. Coutinho**[1]

[1] Programa de Pós-Graduação em Ciência da Computação
Universidade Federal do Maranhão (UFMA) – São Luís – MA – Brasil

karlafelicia.ti@gmail.com, luciano.rc@ufma.br, antonio.batista@ufma.br

[2]Programa de Pós-Graduação em Ciência da Computação
Universidade Federal do ABC (UFABC) – Santo André – SP – Brasil.

jmenac@gmail.com

Abstract. *The use of machine learning models trained on imbalanced datasets with sensitive information has raised privacy concerns. One significant threat is the Membership Inference Attack (MIA), which tries to figure out if a particular data point was included in the training set. This paper investigates whether algorithm-level cost-sensitive learning poses a greater privacy leakage risk than data-level sampling. We conducted experiments using two datasets: the UCI Adult dataset, which focuses on predicting income, and the APS dataset, which focuses on predicting scientific productivity. Our results indicate that models trained with cost-sensitive learning are more vulnerable to MIAs. This supports the hypothesis that correcting for imbalances at the algorithm level can reveal more private information.*

1. Introduction

The increasing integration of machine learning systems into critical domains has triggered an important debate about their fairness [Aragão et al. 2024]. However, the discussion on trustworthy artificial intelligence (AI) [Kaur et al. 2022] — which encompasses AI systems that are developed, deployed and used in an ethical, robust and socially beneficial way and have attributes such as fairness, transparency, security, privacy and accountability — would be incomplete without addressing the privacy dimension.

This is because machine learning models naturally internalize patterns from the training data and can unintentionally remember sensitive information. Membership Inference Attacks (MIAs) [Niu et al. 2024], for example, demonstrate the ability to infer data points from the original datasets used to train the model. This leak is not just a technical issue, but also has profound implications for fairness. The disclosure of sensitive attributes — such as health, demographic or financial information — can enable targeted discrimination, undermining efforts to build equitable systems.

To illustrate this: In a recidivism prediction application, demographic groups (e.g., black and white defendants) should be treated equally, i.e., receive similar prediction

accuracy. At the same time, participation in the training data implies that the person has already committed an offense, which is very sensitive and should be kept confidential. Disclosing information about participation in training data can enable targeted discrimination. One consequence of this is the increase in discrimination based on certain personal characteristics, known as protected characteristics. Therefore, both social justice [Pinheiro et al. 2023] and privacy [Nguyen et al. 2024] protection are necessary for the ethical use of machine learning, and understanding the interplay between these two aspects is crucial.

In this context, a well-known challenge arises with recidivism datasets, where an unbalanced distribution — often containing significantly more examples of black recidivists than white recidivists, reflecting social and systemic inequalities — impairs the performance of standard learning algorithms that expect a balanced class distribution [Japkowicz and Stephen 2002]. This imbalance directly leads to unfair performance of the classifier, as its accuracy is demonstrably lower for the class with fewer examples, disproportionately affecting the already vulnerable groups. Fortunately, several approaches have been developed to mitigate this class imbalance problem and its consequences on the fairness of the model: at the data level [Johnson and Khoshgoftaar 2019], by adding or removing random instances from the training dataset to restore balance; at the algorithm level [Thai-Nghe et al. 2010], by internally modifying the classifier by manipulating its objective function to increase the importance of the minority class; and by using hybrid methods that combine both algorithm-level and data-level approaches.

In this paper, we investigate which of the previously discussed approaches to reducing the effects of class imbalance leads to models that are more susceptible to membership inference attacks (MIAs). Our hypothesis is that models trained using a cost-sensitive learning technique are more susceptible to these attacks than models that use data sampling, thereby increasing the probability of MIA success.

To demonstrate this, we attack neural network models using information about the model loss function and the model error distribution. We compare the attack success of the traditional supervised learning approach using the data undersampling approach with the alternative scenario where there are four times more examples for one class than the other, and this imbalance is corrected at the algorithm level. Our results point to the confirmation of our hypothesis.

This paper argues that while strategies to address class imbalances are often motivated by the pursuit of fairness (and aim to eliminate these performance gaps), they should be carefully scrutinized from a privacy perspective in cases where the dataset contains sensitive personal information. Paradoxically, these strategies may increase susceptibility to MIAs and consequently jeopardize fairness through the disclosure of sensitive data.

2. Theorical background

This section presents the main concepts for understanding the proposed study, including the fundamentals of supervised learning and the main strategies to address class imbalance in machine learning tasks.

2.1. Supervised learning

We use $z = (x, y)$ to denote an element or data sample, where x is the feature vector and y is the class or label. Let $\mathcal{Z}$ the element space, i.e. $z \in \mathcal{Z}$. We use $S \in \mathcal{Z}^n$ to denote a training set that contains n elements $z \in \mathcal{Z}$. Let $\mathcal{D}$ be a probability distribution over $\mathcal{Z}$; $z \sim \mathcal{D}$ means that z is a random sample of $\mathcal{D}$, and $S \sim \mathcal{D}^n$ means that S consists of n independent samples of $\mathcal{D}$. With $\mathcal{D}$ we denote a probability distribution over $\mathcal{Z}^n$, i.e. a joint distribution for all samples in a data set; $S \sim \mathcal{D}$ means that the data set S is a sample of $\mathcal{D}$.

A training algorithm $\mathcal{A}$ is a (possibly randomized) function that takes a training set $S \in \mathcal{Z}^n$ and outputs a trained model $a \in \mathcal{A}$, where $\mathcal{A}$ is the space of trained models. We use $a = \mathcal{A}(S)$ to denote that a is the model that results from applying the training algorithm $\mathcal{A}$ to the training set S. The aim of the trained model a is to solve a classification task, i.e. to assign a label y to a feature vector x.

2.2. Strategies for addressing class imbalance

To deal with imbalanced datasets, researchers have so far focused on the data level and the classifier level. At the data level, the common task is to change the class distribution. At the classifier level, many techniques have been introduced, such as internal manipulation of classifiers.

- **Modifying class distribution**: Random Oversampling (ROS) is a non-heuristic method used to equalize the class distribution by randomly duplicating the minority class examples, while Random Undersampling (RUS) randomly eliminates the majority class examples.
- **Cost-Sensitive Learning**: This approach assumes that the misclassification costs (false-negative and false-positive costs) are not equal. The cost of misclassifying a non-terrorist as a terrorist is much lower than the cost of misclassifying an actual terrorist carrying a bomb on a flight.
  This study focuses on binary classification problems; we denote the positive class (+ or +1) as the minority and the negative class (- or -1) as the majority. Let $C(i, j)$ be the cost of predicting an example that belongs to class i although it actually belongs to class j; the cost matrix is shown in Table 1.

Table 1. Cost matrix for binary classification.

	Actual Class	
	-1	+1
Predicted Class -1	$C(-1,-1)$	$C(-1,+1)$
Predicted Class +1	$C(+1,-1)$	$C(+1,+1)$

The purpose of cost-sensitive learning is to build a model with minimum misclassification costs (total cost):

$$\text{TotalCost} = C(-,+) \times FN + C(+,-) \times FP$$

where FN and FP are the number of false negative and false positive examples respectively.

3. Related Work

Data protection has become a central issue in machine learning [Mariscal et al. 2024]. The reason for this is that data protection and transparency are two important pillars for reliable machine learning. In this context, Membership Inference Attacks (MIA) pose a serious problem.

[Shokri et al. 2017] have shown that models can be vulnerable to MIA. Popular attacks are the loss threshold and the likelihood ratio attack (LRT). The loss threshold detects whether a data point was included in the training set by comparing the model's error rate to a threshold value, which requires access to labels and model details [Sablayrolles et al. 2019]. LRT, on the other hand, uses shadow models to compare the confidence levels of data that are in or out of the training set and calculates a likelihood ratio to predict membership without requiring direct access to the model [Carlini et al. 2022].

[Chang and Shokri 2021] investigated the relationship between algorithmic fairness and privacy in trustworthy machine learning [Ali et al. 2023]. Based on the systematic methodology developed by [Agarwal et al. 2018], the authors apply this approach to achieve fairness in a binary classification setting. They show that imposing fairness on models can significantly increase the risk of privacy disclosure, especially for unprivileged subgroups. Furthermore, the work shows that the greater the bias in the training data, the higher the cost of preserving privacy for these groups.

In addition, other studies have also examined various facets of data privacy risks in machine learning. For example, [Shokri et al. 2021] investigated the privacy leakage risks of explainable AI (XAI) methods, employing a threshold-based attack that infers data membership based on the model's confidence or its explanatory output. Similarly, [Pawelczyk et al. 2023] developed a distance-based attack, where a data point is considered a member of the training set if its loss falls below a determined threshold.

Building on the theme of fairness in machine learning, other researchers have also conducted similar analyzes. For example, [Sena and Machado 2024] conducted a study using the UCI Adult Dataset that specifically looked at biases related to sensitive features such as ethnicity and gender. Their approach was to reduce the dimensionality of the dataset to mitigate these biases. They evaluated the performance and fairness of three common machine learning models— - logistic regression, random forest and gradient boosting — both when including and excluding sensitive attributes. The main results show that while the performance metrics remained stable, the fairness metrics provided important insights, highlighting the need to consider fairness alongside performance in real-world applications of machine learning.

Increasing concerns about data privacy and security have led to a growing interest in Machine Unlearning (MU) [Shaik et al. 2025]. Since ML models naturally internalize training data patterns and can inadvertently remember sensitive information, MU is emerging as an important and increasingly promising area of research. It aims not only to remove or modify predictions, but also to directly mitigate the risk of data leakage, a major concern when integrating ML systems in critical domains.

In this paper, we quantitatively investigate how machine learning models reveal information about the individual datasets on which they were trained. We focus on ma-

chine learning approaches for dealing with data imbalances — a neglected area.

4. Methodological Procedures

We compare the success of an inference attack on an unbalanced dataset. We want to know which of the two techniques for dealing with imbalance the attack performs better. The standard machine learning approach, where we first apply data-level techniques to balance the dataset, or the alternative, where this problem is tackled at the algorithm level by adjusting the loss function.

Our steps for the data sampling approach are as follows. We randomly select samples from the majority class in the same number as from the minority class, so that the minority and majority classes contain (approximately) the same number of samples. Following this procedure, we apply the standard supervised learning approach. In the cost-sensitive learning approach, we adjust the losses to achieve fairness by weighting the samples from the minority class more heavily, thus modifying the objective function of the standard supervised learning approach.

4.1. Membership Experiment

We use membership experiments to evaluate MIA. A membership experiment is a theoretical game between a data owner and an opponent in which the opponent tries to guess the membership of a data sample. The game determines how the data owner generates the training data and what information is available to the attacker.

We have used a variation of [Yeom et al. 2018]'s procedure. In our method, this experiment does not receive the training set S as input, but the test set D. The data owner trains a with S and then selects a bit $b \in \{0, 1\}$ uniformly at random. This bit determines whether the opponent receives an element (z randomly selected from S) or a non-element (a new sample $z \sim \mathcal{D}$). The opponent receives the element z and the trained model a, the training algorithm $\mathcal{A}$ and the test dataset $\mathcal{D}$. The attacker uses this information to carry out an attack $\text{Att}(z, a, \mathcal{A}, \mathcal{D})$, which outputs a bit indicating the predicted membership of z. The attacker is successful if the attack correctly infers the membership status of z.

4.2. Membership inference attack (Att)

The attack exploits the different behavior of a classifier on data points it has seen during training compared to those it has not seen. The attack works as follows:

First, for a given input z, the attacker observes the classifier's prediction and determines whether it resulted in a **misclassification (error)** or a **correct classification (no error)**. Let E_1 denote the event of a misclassification and E_0 denote the event of a correct classification for input z.

The attacker's objective is to infer whether this input z belongs to the model's original **training dataset** (event B_{train}) or the **test dataset** (event B_{test}). This inference is performed by computing the posterior probability of z's origin given the observed classification outcome, using Bayes' Theorem.

For instance, to calculate the probability that z is from the training set given a misclassification (E_1), the attacker computes:

$$\Pr(B_{\text{train}} \mid E_1) = \frac{\Pr(E_1 \mid B_{\text{train}}) \cdot \Pr(B_{\text{train}})}{\Pr(E_1)}$$

where:

- $\Pr(E_1 \mid B_{\text{train}})$ is the conditional probability of a misclassification if the input z is from the training set. This represents the **classifier's error rate on its training data**.
- $\Pr(B_{\text{train}})$ is the prior probability that a given input z originates from the training set. This is typically calculated as the **proportion of training examples relative to all examples** (i.e., training + test examples).
- $\Pr(E_1)$ is the overall probability of a misclassification for any given input z, regardless of its origin. This corresponds to the **classifier's total error rate across the entire dataset** (training and test examples combined).

The probabilities $\Pr(B_{\text{train}} \mid E_0)$, $\Pr(B_{\text{test}} \mid E_1)$, and $\Pr(B_{\text{test}} \mid E_0)$ are calculated analogously, using the respective error/accuracy rates and prior probabilities. For example, $\Pr(E_0 \mid B_{\text{train}})$ would be the classifier's accuracy on the training data.

Attack Decision Phase: Once these posterior probabilities are computed, the attacker infers the membership of input z.

- If a misclassification (E_1) is observed for input z, the attacker assigns z to the training set (i.e., infers B_{train}) if $\Pr(B_{\text{train}} \mid E_1)$ is above a certain threshold or significantly higher than $\Pr(B_{\text{test}} \mid E_1)$.
- Similarly, if a correct classification (E_0) is observed, the attacker infers z belongs to the training set if $\Pr(B_{\text{train}} \mid E_0)$ is high, or assigns z to the test set otherwise. The attacker's output is this inferred binary membership.

4.3. Metrics for assessing Membership inference attacks

The membership advantage in experiment 4.1 is defined as:

$$\text{Advantage} = \Pr(\text{Att} = 0 \mid b = 0) - \Pr(\text{Att} = 0 \mid b = 1).$$

In this equation:

- The first term, $\Pr(\text{Att} = 0 \mid b = 0)$, represents the True Positive Rate (TPR).
- The second term, $\Pr(\text{Att} = 0 \mid b = 1)$, represents the False Positive rate (FPR).

We assume that $b = 0$ indicates a positive example, i.e. the attack described in section 4.2 correctly predicts that the network has been trained on the specific input.

The membership advantage, is defined as follows $Adv = TPR - FPR$. This metric is 0 if the attacker randomly determines z's membership, and it is 1 if the attacker always guesses z's membership correctly.

4.4. Datasets

We conducted our experiments using two publicly available datasets, each representing a distinct application scenario and set of features. A brief description of both datasets is provided below:

- **UCI Adult dataset** (Census Income) [1]
  This data set comprises 48,842 data records with 14 attributes such as age, gender, education, marital status, occupation, working hours and native country. The (binary) classification task is to predict whether a person earns more than 50,000 dollars per year based on the census attributes. We use this dataset to train the target models.
- **APS dataset** [2] We used the dataset managed by the APS as the data source for our analyzes and used the author names from the article metadata to identify the authors. We defined the academic age of authors based on the years between their first and last publication. We selected authors who started publishing after 1985 and had an academic age of at least 8 years. To calculate the value of the predictors, we used their publications up to three years before their peak age and their remaining publications as future publications. We used all citations for a given article within the dataset and therefore did not define a time window for the accumulation of citations.
  The (binary) classification task consists of predicting whether a scientist will publish more than 3 publications, each with the same number of citations, in the years following the prediction year. We use the following bibliometric indices of the scientist as predictor features: H, the H-index of the scientist u from the specific year y of career to the year of prediction; co, the number of different co-authors of the scientist u from the year y of career to the year of prediction; c, the total number of citations of the scientist u from the year y of career to the year of prediction; I_{iu}, the total number of articles by scientist u, each with i citations, from year y of career to year of prediction; and v, the total number of publications in which scientist u published from year y of career to year of prediction.

To accurately describe the imbalance within our experimental design, we provide the exact sampling distributions for the UCI Adult and APS datasets used in our study.

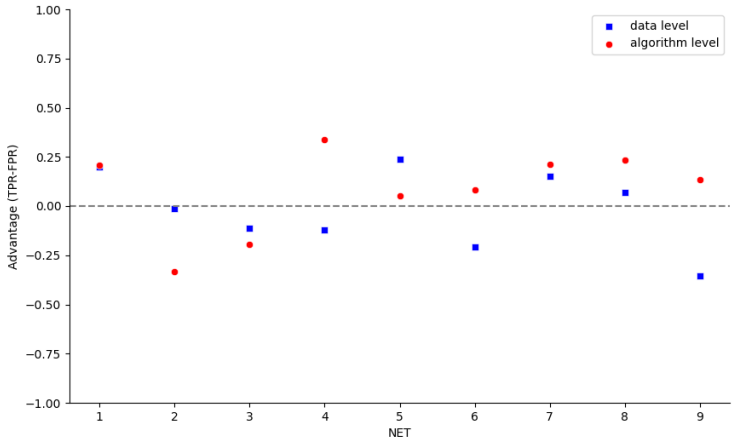
- **UCI Adult Dataset:** This dataset comprises a total of 47,621 instances. The majority class consists of 39,780 samples, while the minority class has 7,841 samples, resulting in an approximate imbalance ratio of 1:5.07.
- **APS Dataset:** The APS dataset used in our experiments contains 28,481 instances. Specifically, the majority class includes 23,168 samples, and the minority class is represented by 5,313 samples. This results in an imbalance ratio of approximately 1:4.36, indicating a significant disparity between the classes.

5. Results and discussion

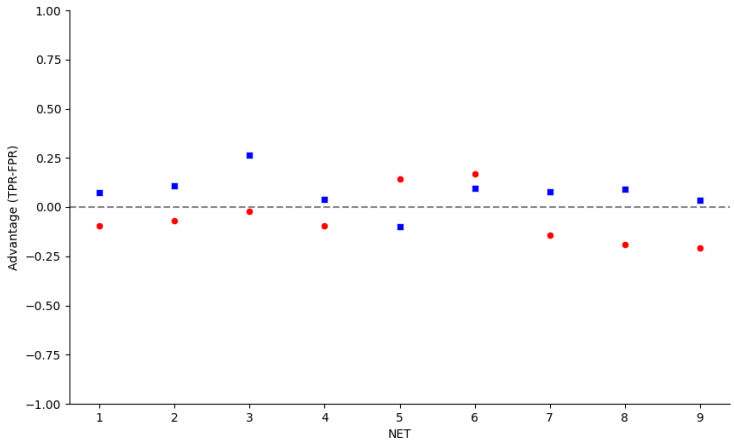
To test the hypothesis that the algorithm-level cost-sensitive learning approach reveals more than the data sampling approach and therefore provides a higher value for the membership advantage metric, we conducted a controlled experiment. We compared the performance of Attack 1 in a scenario with a traditional supervised learning approach, where the dataset is first balanced at the data level (the data sampling approach), with the performance observed with a supervised learning approach that addresses this problem at the algorithm level and acts directly on the loss function (the algorithm-level cost-sensitive

[1] https://archive.ics.uci.edu/dataset/2/adult
[2] https://journals.aps.org/datasets

learning approach). Figure 1 shows the performance of the NETs(Table 2) attack on the adult dataset and Figure 2 shows the performance of the NETs (Table 3) attack on the APS dataset. We use loss BCEWithLogitsLoss$(weight)$ with $weight = 4.5$ for the cost-sensitive approach and $weight = 1$ for the alternative approach. Adult NET input size 100 and output 1 and APS NET input size 13 and output 1.



(a) adult dataset

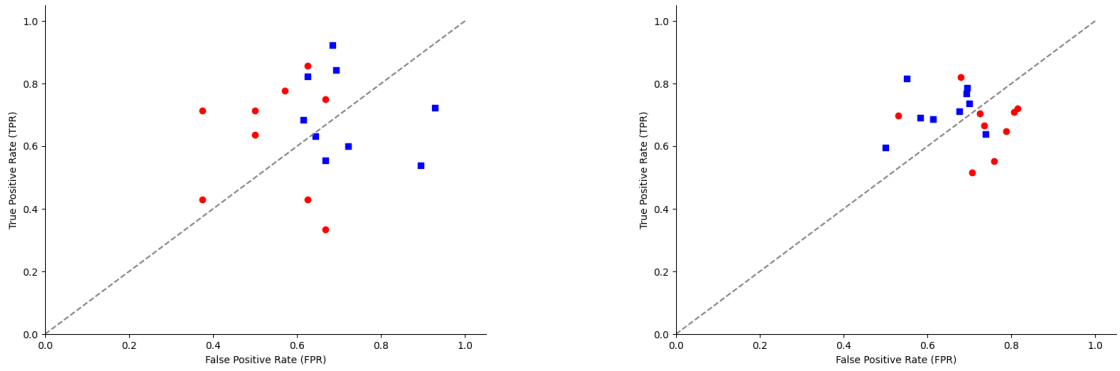


(b) APS dataset

Figure 1. Evaluation of MIA performance in different networks, depending on the technique used for the advantage metric.

Our results reveal that the algorithm-level technique leads to a higher disclosure of private data than the data-level approach. As shown in Figure 1, which depicts the Advantage (Adv) metric of the attack, and in Figure 2, which shows the True Positive Rate (TPR) versus the False Positive Rate (FPR), the performance of the attack is consistently superior to the algorithm-level technique.

In both figures, the diagonal line (gray dashed line) represents the performance of a random attack. In Figure 1 (advantage), values above this line indicate that the attacker



(a) adult dataset (b) APS dataset

Figure 2. Evaluation of Membership Inference Attack (MIA) performance (TPR vs. FPR) in different networks, grouped by technique.

gained some information about whether an input was part of the training data. In Figure 2 (FPR vs. TPR), points that are above this line and tend to be positioned in the upper left corner (indicating fewer false positives at a given TPR) also indicate a more effective attack.

It is worth noting that in the algorithm-level technique, there are more networks (NETs) whose results are above the diagonal line in both graphs and tend towards the upper left corner in the TPR vs. FPR graph in particular. This clearly shows that this approach favors private data disclosure and outperforms both a random attack and the alternative approach, regardless of which metric is used to visualize MIA performance.

Our results reveal that the Membership Inference Attack (MIA) against networks trained with the algorithm-level technique achieves higher average Adv values, up to **0.3**, than the alternative approach (data-level technique). As defined, a Adv value of 0 means that the attacker randomly determines whether an input is part of the training dataset, while a value of 1 represents perfect inference. The observed performance of 0.3, which is well above chance, clearly shows that the attacker has gained a non-trivial advantage in determining whether a data point belongs to the training set.

This result clearly confirms the existence of information leakage of the models with respect to their training data. Although the attacker did not achieve perfect inference, a Adv of 0.3 is already a worrying indicator of privacy vulnerability. It suggests that the models reveal recognizable patterns about whether an input is part of the training data, which could be exploited by an attacker. Even such partial disclosure poses a significant privacy risk as it allows for better-than-random identification of individuals whose sensitive data may have been used in training, opening the door to discrimination or other privacy violations.

It is important to note that the attack on the APS dataset performs poorly and shows less variation between techniques. However, there were cases where the attack on some networks trained with the algorithm-level approach revealed more private data than the data-level approach.

Our experiments confirm previous findings [Le et al. 2022, Shokri et al. 2017, Pawelczyk et al. 2023] that it is possible to gain some insight into whether a network has been trained on a particular input. However, we have shown that this becomes clearer when the algorithm-level cost-sensitive learning approach is used to deal with an imbalanced dataset.

On the other hand, caution is needed, as only feedforward neural network models were tested on two datasets. A better attack could better separate the performance of each technique. It is also important to experiment with other methods for dealing with imbalanced data. We have provided evidence that points to confirmation of our hypothesis. However, further testing is needed.

Table 2. Configuration of the networks trained on the UCI adult dataset.

NET	Hidden 1	Hidden 2
1	50	100
2	50	50
3	100	75
4	200	50
5	150	50
6	100	150
7	200	100
8	100	100
9	200	200

Table 3. Configuration of the networks trained on the APS dataset.

NET	Hidden 1	Hidden 2
1	25	25
2	50	25
3	100	25
4	25	50
5	50	50
6	100	50
7	25	100
8	50	100
9	100	100

6. Conclusions and Future Work

This study explored how different strategies for handling class imbalance affect the risk of private data exposure through membership inference attacks (MIAs). We compared data-level sampling with algorithm-level cost-sensitive learning in neural networks trained on two datasets. Our findings suggest that the cost-sensitive approach is more vulnerable to privacy leakage, particularly in the UCI Adult dataset, where attack success was significantly higher.

These results highlight a potential trade-off between fairness and privacy: while algorithm-level techniques may improve performance on underrepresented classes, they

can also increase the model's susceptibility to privacy attacks. This observation raises important considerations for the design of machine learning systems in sensitive applications.

Future work should evaluate whether these results generalize to other model architectures and datasets, and investigate defenses that jointly address class imbalance and privacy risks. It is also worth exploring whether certain classes or demographic groups are disproportionately affected by MIAs, an aspect that remains largely unexamined in the literature.

7. Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) (Finance Code 001). We also thank the Fundação de Amparo à Pesquisa do Estado do Maranhão (FAPEMA) (processes nº APP-09405/22 and BM-02169/23).

References

[Agarwal et al. 2018] Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR.

[Ali et al. 2023] Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., and Herrera, F. (2023). Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99:101805.

[Aragão et al. 2024] Aragão, L. R., Silva, M., and Machado, J. (2024). The inefficiency of achieving fairness with protected attribute suppression. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 813–819, Porto Alegre, RS, Brasil. SBC.

[Carlini et al. 2022] Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramèr, F. (2022). Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE.

[Chang and Shokri 2021] Chang, H. and Shokri, R. (2021). On the Privacy Risks of Algorithmic Fairness . In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 292–303, Los Alamitos, CA, USA. IEEE Computer Society.

[Japkowicz and Stephen 2002] Japkowicz, N. and Stephen, S. (2002). The class imbalance problem: A systematic study. *Intell. Data Anal.*, 6(5):429–449.

[Johnson and Khoshgoftaar 2019] Johnson, J. M. and Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):27.

[Kaur et al. 2022] Kaur, D., Uslu, S., Rittichier, K. J., and Durresi, A. (2022). Trustworthy artificial intelligence: A review. *ACM Comput. Surv.*, 55(2).

[Le et al. 2022] Le, T.-T.-H., Liu, Z., Li, R., Miao, D., Ren, L., and Zhao, Y. (2022). Membership inference defense in distributed federated learning based on gradient differential privacy and trust domain division mechanisms. *Security and Communication Networks*, 2022:1615476.

[Mariscal et al. 2024] Mariscal, C., Galante, R., and Cordeiro, W. (2024). Predicting the next transaction on anonymized payment datasets with deep learning models. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 639–651, Porto Alegre, RS, Brasil. SBC.

[Nguyen et al. 2024] Nguyen, T. T., Huynh, T. T., Ren, Z., Nguyen, T. T., Nguyen, P. L., Yin, H., and Nguyen, Q. V. H. (2024). Privacy-preserving explainable ai: a survey. *Science China Information Sciences*, 68(1):111101.

[Niu et al. 2024] Niu, J., Liu, P., Zhu, X., Shen, K., Wang, Y., Chi, H., Shen, Y., Jiang, X., Ma, J., and Zhang, Y. (2024). A survey on membership inference attacks and defenses in machine learning. *Journal of Information and Intelligence*, 2(5):404–454.

[Pawelczyk et al. 2023] Pawelczyk, M., Lakkaraju, H., and Neel, S. (2023). On the privacy risks of algorithmic recourse. In Ruiz, F., Dy, J., and van de Meent, J.-W., editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 9680–9696. PMLR.

[Pinheiro et al. 2023] Pinheiro, M. V., Silva, M. M., and Machado, J. (2023). Strategies selection for a fair classification in logistic regression: A comparative analysis. In *Anais Estendidos do XXXVIII Simpósio Brasileiro de Bancos de Dados*, pages 15–21, Porto Alegre, RS, Brasil. SBC.

[Sablayrolles et al. 2019] Sablayrolles, A., Douze, M., Schmid, C., Ollivier, Y., and Jegou, H. (2019). White-box vs black-box: Bayes optimal strategies for membership inference. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5558–5567. PMLR.

[Sena and Machado 2024] Sena, L. and Machado, J. (2024). Evaluation of fairness in machine learning models using the uci adult dataset. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 743–749, Porto Alegre, RS, Brasil. SBC.

[Shaik et al. 2025] Shaik, T., Tao, X., Xie, H., Li, L., Zhu, X., and Li, Q. (2025). Exploring the landscape of machine unlearning: A comprehensive survey and taxonomy. *IEEE Transactions on Neural Networks and Learning Systems*, 36(7):11676–11696.

[Shokri et al. 2021] Shokri, R., Strobel, M., and Zick, Y. (2021). On the privacy risks of model explanations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 231–241, New York, NY, USA. Association for Computing Machinery.

[Shokri et al. 2017] Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18.

[Thai-Nghe et al. 2010] Thai-Nghe, N., Gantner, Z., and Schmidt-Thieme, L. (2010). Cost-sensitive learning methods for imbalanced data. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

[Yeom et al. 2018] Yeom, S., Giacomelli, I., Fredrikson, M., and Jha, S. (2018). Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting . In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282, Los Alamitos, CA, USA. IEEE Computer Society.