

# A Hierarchical Approach for Assessing the Vulnerability of Tree-Based Classification Models to Membership Inference Attack

Richard J. Preen, Jim Smith

Department of Computer Science and Creative Technologies, University of the West of England, Bristol, UK.

E-mail: {richard2.preen, james.smith}@uwe.ac.uk

Received 13 February 2025; received in revised form 12 June 2025; accepted 28 December 2025

**Abstract.** Machine learning models can inadvertently expose confidential properties of their training data, making them vulnerable to membership inference attacks (MIA). While numerous evaluation methods exist, many require computationally expensive processes, such as training multiple shadow models. This article presents two new complementary approaches for efficiently identifying vulnerable tree-based models: an *ante-hoc* analysis of hyperparameter choices and a *post-hoc* examination of trained model structure. While these new methods cannot certify whether a model is safe from MIA, they provide practitioners with a means to significantly reduce the number of models that need to undergo expensive MIA assessment through a hierarchical filtering approach.

More specifically, it is shown that the rank order of disclosure risk for different hyperparameter combinations remains consistent across a range of datasets, enabling the development of simple, human-interpretable rules for identifying relatively high-risk models before training. While this *ante-hoc* analysis cannot determine absolute safety since this also depends on the specific dataset, it allows the elimination of unnecessarily risky configurations during hyperparameter tuning. Additionally, computationally inexpensive structural metrics serve as indicators of MIA vulnerability, providing a second filtering stage to identify risky models after training but before conducting expensive attacks. Empirical results show that hyperparameter-based risk prediction rules can achieve high accuracy in predicting the most at risk combinations of hyperparameters across different tree-based model types, while requiring no model training. Moreover, target model accuracy is not seen to correlate with privacy risk, suggesting opportunities to optimise model configurations for both performance and privacy.

**Keywords.** Privacy, Machine Learning, Membership Inference Attack, Tree Models

## 1 Introduction

An increasing body of work has shown that machine learning (ML) models may expose confidential properties of the data on which they are trained [41]. This has resulted in a wide range of proposed attack methods with varying assumptions that exploit the model structure and/or behaviour to infer sensitive information. For example, to perform data reconstruction, model extraction, attribute inference, and membership inference [14, 20].

The General Data Protection Regulation (GDPR) (and other laws, according to the data and jurisdiction) mean that before an ML model trained on confidential data can be released into the public domain, its vulnerability to attack should be assessed [51]. One of the most fundamental techniques to assess ML privacy disclosure is the membership inference attack (MIA) [45], which attempts to identify whether an individual record was part of the original training data. Since many of the other attacks have been shown to be more difficult to perform [57], vulnerability to MIA can be seen as a strong indicator of potential privacy risks, and its presence suggests that further, more sophisticated attacks might also be feasible, even if more challenging to execute. Therefore, a model’s robustness against MIA is a crucial first step in demonstrating its privacy-preserving capabilities before public deployment.

However, some of the currently most effective MIAs require considerable computational effort. For example, running a Likelihood Ratio Attack (LiRA) [5] typically involves sampling (and hence training) 50–100 shadow models. Therefore, from the perspective of people training models, and those tasked with that risk assessment, there is significant value in being able to rapidly identify vulnerable models or reduce the number that require computationally expensive evaluation.

In this paper we investigate two approaches to this, which also cast light on the closely related issue of what are the underlying causes of model vulnerability, and what can be measured as proxies for those. First we explore the relationship between model hyperparameters and risk. That is, we use an empirical approach to learn simple human-interpretable rules that identify whether a model is *relatively* high risk based on the hyperparameters independent of the training dataset and before any model fitting. These rules may provide additional insight into the causes of disclosure risk and enable researchers to avoid training unnecessarily high risk models, e.g., such combinations could be removed from a typical hyperparameter search. Furthermore, they may provide additional information to support making an assessment of whether a model should be publicly released.

However, since the risk of information disclosure depends on characteristics of both model and dataset, ascertaining the *absolute* risk posed by a model must be performed after training. Therefore in the second strand of work we leverage decades of statistical disclosure control (SDC) practice for risk-assessing tables and regression models to design a number of computationally cheap structural measures based on the trained model and how it partitions the training data.

The contributions of this paper are:

1. In terms of *ante-hoc* analysis of hyperparameter choices, we show that:
  - (a) The rank order of risk for different hyperparameter combinations is preserved across different datasets, i.e., is dataset independent.
  - (b) Risk is not correlated with target accuracy.

Therefore we can identify a region of unnecessary risk—which can be characterised by simple human-readable models, and used to help researchers avoid creating risky models in advance.

2. In terms of *post-hoc* analysis of trained model structure and how it partitions the decision space we show that:
  - (a) We can design useful cheap-to-compute structural metrics based on properties of target models and their behaviour with training sets.

- (b) Empirical results suggest a combination of these structural metrics constitutes a sufficient but not necessary indicator of whether a target model is vulnerable to MIA attacks. In other words, a classifier using these structural metrics to predict MIA output has high precision but lower recall—although the latter improves if we take into account the statistical significance of MIA observations.

This suggests that the structural metrics are proxies for some as yet unquantified latent property of the models, similar to the conditioned loss metric at the heart of LiRA and other state-of-the-art attacks.

3. We make clear the fundamental difference between being able to rule out release of a model (as being disclosive) and being able to say that it is safe.

These contributions allow us to avoid computationally costly MIA attacks and provide actionable insights into how to reduce disclosure risk. The remainder of this paper is organised as follows. We begin in Section 2 by discussing MIAs, traditional SDC within TREs, and the implications for ML privacy. Section 3 then details the methodology used to investigate the privacy risks of hyperparameters and structural metrics in tree-based ML models. We present our empirical results in Section 4, and finally, Section 5 summarises and discusses our findings.

## 2 Background

### 2.1 Membership Inference Attacks

Recognising different sources of uncertainty in the model training process, the MIA can be framed as the question: *what are the probabilities that a model  $M$  comes from the population  $\mathcal{M}_{in}$  (or  $\mathcal{M}_{out}$ ) of models trained with (or without) a particular record  $X$ ?* For this to take a computational form requires a mapping from some set of metrics describing the structure and/or behaviour of models, and records, onto a probability distribution. Given the current state of knowledge, a key issue is how to identify one or more suitable metrics as proxies for underlying risk causes.

A common early approach to MIA exploits the difference in confidence scores between predictions for samples that were in the training set and those that were not [45]. However, this approach fails to take into account that (hopefully) models will correctly identify regions where high probability records have the same label—whether they are from training set or not. The conceptual and methodological shortfalls of this approach have been well documented [39].

More recent attacks implicitly focus on decision boundaries by considering the record-wise difficulty of making correct predictions for the original task. For example, LiRA uses as proxy the prediction error (loss) for a record  $X$  from the original target model  $M$ , comparing it to the distributions of loss values from sampled subsets from  $\mathcal{M}_{in}$  and  $\mathcal{M}_{out}$  [5]. Similarly, the quantile regression [3] and robust MIA (RMIA) [56] condition the confidence on that for other records (and models). Relative measures of MIA vulnerability have also been suggested, computing risk based on the difference in prediction loss for member samples relative to a predefined reference model [2]. Recent work has also sought to improve performance with target model knowledge distillation [11].

MIAs typically assess whether a given data point exactly matches a training record, which potentially neglects similar or overlapping memorised data. For example, an image from a

different angle may still be recognisable and evoke an exploitable behaviour from the model even though it would be incorrectly assessed as a false positive with a binary exact matching metric. Consequently, Tao and Shokri [50] propose a broader assessment framework that evaluates whether the model was trained on any data within a specified range using domain-specific distance measures.

Typically, MIAs assume that the model's confidence scores (or loss values) are exposed. However, even when this information is hidden or unavailable, label-only based attacks can still identify whether a record was in the original training data. This is achieved by perturbing records to assess model robustness with the assumption that records exhibiting high robustness belong to the training dataset [7]. Moreover, in some cases training membership can be accurately inferred with just a single query [32] by exploiting the fact that member samples exhibit more resistance to adversarial perturbations than non-member samples.

One of the most significant drawbacks for assessing ML privacy with MIAs is their computational cost, particularly when using shadow models, which can require training many models with the same architecture and algorithm as the original target model. This can mean that they may not be feasible for large models or datasets. Consequently, a growing body of work is exploring ways to reduce the computational burden, for example maintaining MIA accuracy with only a few shadow models [56]. Interestingly, a recent theoretical analysis of likelihood ratio-based MIAs on stochastic gradient descent [23] has shown how various hyperparameters influence vulnerability, without needing to train any shadow models.

A number of risk factors and proxy measures have been established, with overfitting (quantified by generalisation error) recognised as a significant cause of information leakage [54, 55]. Increasing model complexity often entails a trade-off between privacy and utility, as observed in neural networks where more parameters can enhance performance but compromise privacy [49]. While explainable ML models are often proposed to boost transparency and trust, recent studies caution that they may inadvertently increase vulnerability by revealing feature importance information, which can be exploited [24, 27]; perturbing these features typically results in a bigger loss in confidence for member samples.

Implementing robust ML practices to prevent overfitting and ensure generalisation is therefore an important first step in mitigating MIA vulnerabilities [9]. Additionally, a number of defences have been proposed [15, 18]. The most common of these is to incorporate information perturbation within the learning algorithm, such as through differential privacy (DP) [10]. However, the use of DP and knowledge distillation techniques may only mitigate some inference attacks [25]. Indeed it has been suggested that model distillation may not provide any effective privacy defence [16], and the level of DP necessary to protect a model from MIA may render its utility unusable [37].

Moreover, it has recently been highlighted that in practice DP implementations suffer from a number of problems and standard anti-overfitting techniques can often achieve better results [4]. For example, standard regularisation techniques have been shown to help mitigate the risk posed by an increasing number of model parameters [48].

As the use of ML continues to proliferate, it is becoming increasingly common to release multiple updated versions of a model. However, doing so may further expose confidential properties of the data, strengthening reconstruction attacks [44]. This has also been shown to be true for improving MIA success [17], which may be a particular risk for secure data facilities such as TREs that may egress many models trained on the same data, and is a similar concern to differencing attacks in traditional output SDC.

## 2.2 Trusted Research Environments

Trusted research environments (TREs)<sup>1</sup> offer a means for researchers to analyse confidential datasets and publish their findings. The globally recognised Five Safes<sup>2</sup> approach to privacy preservation and regulatory compliance [13] requires TRE staff to independently check the disclosure risk posed by any proposed output before release is approved.

This process of SDC is well understood for outputs of traditional statistical analyses. However, while the use of ML methods has rapidly expanded, a recent study of TREs by Kavianpour et al. [22] in the UK revealed a shortfall of SDC understanding and tools to support the use of ML. In particular, while researchers can already use ML methods within TREs, the inability to export those trained models precludes reproducibility of results, a critical tenet of scientific practice [1].

Regulatory authorities such as the UK Information Commissioner's Office recognise the risks of ML models, but offer general advice rather than practical guidance and tools. Jefferson et al. [19] have recently presented a detailed discussion of the implications of ML model release and proposed a set of guidelines for TREs. A recent report by DARE UK [8] highlighted the need to *"Where possible, automate the review of outputs to support and focus the use of skilled personnel on the areas of most significant risk"*.

## 2.3 Relationship to Statistical Disclosure Control

We briefly introduce the notions of SDC and TREs, before turning our attention to the implications of this body of work for the risk assessment of ML models in practice.

### 2.3.1 TREs and SDC

For all the apparent novelty of risk assessing ML models, it is important to place these within a wider context of assessing the risk of privacy leakage from any outputs created from confidential data. From that perspective, risks such as attribute inference date back to at least the instructions for the 1850 US census:

*"to consider the facts intrusted [sic] to them as... not to be used in any way to the gratification of curiosity, the exposure of any man's business or pursuits..."* [43]

Nowadays, national statistics institutes (and other holders of confidential information such as health authorities) routinely conduct SDC of all outputs they produce. They also offer access to researchers via TREs. TREs almost universally employ the Five Safes framework to provide multiple layers of assurance to data owners, within which the final stage of checking for 'safe outputs' is a vital part.

Notably, most TREs have long employed principles-based SDC [42] with human output checkers involved to facilitate nuance and assess the 'discreditability of an attack' in the interpretation of risk analysis. Similar ideas have recently been discussed for ML models [40].

### 2.3.2 Implications for ML Privacy Research

Thus for traditional forms of analyses, there is decades of well understood theory and practice surrounding the SDC of outputs such as (multi-dimensional) tables, regression models, survival statistics and a range of other outputs, which have recently been consolidated into a

<sup>1</sup>Also known as safe havens, secure data environments, and secure research services.

<sup>2</sup>Safe people, safe projects, safe settings, safe data, safe outputs.

coherent ‘statbarns’ framework [12]. From these we can identify three key concepts relevant to ML models:

- Residual degrees of freedom risk (DoF): for regression models such as linear, logistic or probit, TREs typically require the number of records used to create the model exceeds the number of trainable parameters by at least 10.

The equivalent of this can be directly calculated for ML models; for example, the number of values involved in specifying branches in (forests of) trees, or the number of trainable weights in neural networks.

- $k$ -anonymity: queries should not report on groups of fewer than  $k$  people. In practice, for tables this means that each cell must reach a minimum threshold count.

Since tree-based models (and by extension forests) create a partition of the input space in exactly the same way as tables do, this can be measured directly for a model if we have the training data and the model. For more complex cases we can derive a proxy for this for each training record: the number of other training records for which the model gives the same output (to some precision).

- Class disclosure risk. It should not be possible to make inference about all members of a subgroup via the publication of query results containing zeros, e.g., “*in our fictitious survey no ML researchers were aware of the Secure Data Access Professionals manual*”<sup>3</sup>.

For ML models we can interpret class disclosure risks as occurring when the model’s predicted probability for *any* class falls below a threshold ( $k$ /size of training set) for *any* record. In other words, a model that is completely confident in its predictions presents a class disclosure risk. While this might be seen excessively restrictive, the threshold will be small in practice, and we note that this is somewhat similar to the effects produced by the defence strategies proposed by MemGuard [21].

We note in passing that protection against class disclosure is at best poorly catered for by DP training methods since the guarantees ( $\epsilon$ ) multiply with the size of the subgroup, which may not be known in advance.

### 3 Methodology

This section presents the methodology used to investigate MIA risks in tree-based classification models. We first state our hypotheses for disclosure risk based on model hyperparameters in Section 3.1 and structural model properties in Section 3.2. Section 3.3 provides further details of the statistical framework used for determining the significance of observed MIA success. We then describe our experimental setup in Section 3.4. This includes details of the specific MIA and metrics explored, including a novel metric  $FDIF_z$ , as well as target models and datasets used for evaluation.

#### 3.1 Disclosure Risk and Model Hyperparameters

It is here hypothesised that while absolute disclosure risk is dataset-dependent, the *relative* risk ranking of different model hyperparameter combinations may generalise across datasets.

<sup>3</sup>In some cases it may be argued that zeros may be structural (people born without certain organs are unlikely to develop cancers of those organs) or may be ‘evidential’ and scientifically important: “*no people with gene XYZ had disease A*”. Hence the practice of applying principles-based SDC rather than strictly applying rules.

However, since many different risk metrics have been proposed, it is first necessary to (i) select a single representative measure to analyse the relative risk; and (ii) establish that this metric is likely to generalise across different datasets. Subsequently, we test whether a simple, interpretable model can predict if a hyperparameter combination leads to relatively high or low risk as defined by this chosen metric.

That is, we hypothesise:

- H1.1: Since MIA metrics measure different aspects of the same phenomenon (risk), hyperparameter rank orderings for one metric will be predictive of rank orderings for another. To test this hypothesis, for each target model algorithm, pair-wise rank correlation statistics are computed to compare the relative risk **between metrics** across a range of hyperparameter combinations for each dataset.
- H1.2: This rank ordering will generalise **across datasets**. To test this hypothesis, we select a single representative MIA metric and compute Kendall  $\tau$  rank correlation coefficients of hyperparameter risk to perform pair-wise comparisons for each dataset.
- H1.3: We can exploit H1.1 and H1.2 to produce a **risk prediction model** for ‘unnecessary vulnerability’ that can be applied to different datasets without the need for sampling MIA scores on these new datasets. To test this hypothesis, we assess the performance of models trained to predict whether a given hyperparameter combination is relatively high or low risk as a binary classification task. These models are trained on a single dataset (mimic2-iaccd) that has been sampled for MIA risk, and the rules are then extracted and evaluated on the remaining (unseen) datasets to test generalisation performance.

That is, for each target model type, a Decision Tree classifier is used to learn a simple human-interpretable ruleset that predicts whether a hyperparameter combination is relatively high or low risk. Risk is here defined as high if the previously chosen metric is in the top 20% of all hyperparameter combinations. While this is a somewhat arbitrary threshold, the aim is to prove that a simple set of rules can be identified that predict whether a hyperparameter combination poses unnecessarily high risk.

In order to extract rules that are easy to interpret, the Decision Trees are restricted to a maximum depth of 5; using a simple grid search for the other hyperparameters. Stratified 10-fold cross-validation was applied to the single training dataset (mimic2-iaccd) and the best performance was found in each case to be a minimum number of samples per split of 2, a minimum number of samples per leaf of 2, the entropy criterion, and a balanced class weight.

### 3.2 Disclosure Risk and Structural Model Properties

To investigate the relationship of disclosure risk and the structural model properties described in Section 2.3.2, we take the following approach. For a range of algorithm-dataset-hyperparameter combinations sampled, we analyse the model produced and its predictions for the training set, and use these to capture the structural metrics.

Note that different to the *ante-hoc* ‘unnecessary risk’ metric explored above, this is a *post-hoc* analysis, since the minima (e.g., samples per leaf) and maxima (e.g., depth of a tree) defined via an algorithm’s hyperparameters will be lower/upper bounds that may not be reached if the data does not merit it.

We then join these with the various metrics captured from the MIAs and use this combined data to examine two complementary hypotheses:

- H2.1: Structural risks are **necessary** conditions for model vulnerability. In other words, if a given structural metric is 0 then there will be no MIA risk. In the following analysis we insist on 100% support for hypothesis 2.1 because this is a statement about **safety**.
- H2.2: The presence of structural risks are **sufficient** conditions for model vulnerability. In other words, if (the product of a combination of) structural metric(s) is 1 there will be non-zero MIA risk. We insist on high, but not 100% support for hypothesis 2.2 because this is a statement about **risk**.

### 3.3 Statistical Significance of MIA Vulnerabilities

#### 3.3.1 Preliminaries

Assume a researcher has access to a dataset  $\mathcal{D}$  of size  $N = |\mathcal{D}|$  which they split into disjoint training and test sets  $\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}$  of sizes  $N_{\text{train}}$  and  $N_{\text{test}}$  respectively. They use these to train and test a *target* model  $M$  for a classification problem with  $L$  classes, so  $M : \text{Domain}(\mathcal{D}) \rightarrow [0, 1]^L$ . For simplicity we assume the outputs of  $M$  have been normalised (e.g., via a softmax function) so that they represent the predicted class probabilities for each record.

MIAs may be created using synthetic datasets which approximate the underlying distribution from which  $\mathcal{D}$  is drawn. It is well-established that the success of MIAs often increases as uncertainties in approximating the data distribution are reduced [5, 45, 53].

Thus, in recognition of the pace at which new MIAs are being invented, in our scenario, we remove those uncertainties to estimate an upper bound on the risk. Hence we explicitly use the sets of outputs (either directly as probabilities or converted to losses) as the basis for conducting different MIAs.

For the purpose of evaluating the statistical significance of attack vulnerabilities, we consider a generic MIA regressor  $A$  derived from the target model  $M$ 's outputs. That is:

- Without loss of generality, we assume that an attack creates an MIA regression model  $A : M(\mathcal{D}) \rightarrow \mathcal{R}$ . Given a target model output  $x' = M(x), x \in \mathcal{D}$ , we condition the outputs so that  $A(x')$ : represents the estimated probability that  $x \in \mathcal{D}_{\text{train}}$ .
- These  $N$  membership estimates can then be sorted into decreasing order and used to produce a ROC curve, which via interpolation can be probed to produce estimates of the true positive rate (TPR) for different thresholds of false positive rate (FPR).
- A specific MIA classifier is created by applying a threshold to the MIA regression model's predictions, above which the label '*member*' is output. We note that the likely accuracy of different MIA classifiers will vary since the difficulty of inferring membership will vary between records, as reported elsewhere [5, 56]. Therefore the estimation procedure described below will need to be repeated for different thresholds.

#### 3.3.2 From True Positive Rates to Attack Vulnerability

Following Carlini et al. [5] it has become standard practice to compare different MIAs based on values for the TPR at different low FPR regimes. We will denote these  $\text{TPR}@z$ , where typically the FPR  $z \in \{0, 0.001, 0.01\}$ .

However, from the perspective of an SDC output checker tasked with making a decision on whether to approve the release of the trained model  $M$ , what would be more helpful is to quantify the likelihood of observing different numerical MIA classifier results (TPR@ $z$ ) by chance.

We assess this via a null hypothesis:

- $H_0(z)$ : The classifier produced by choosing a threshold  $t_z$  such that FPR =  $z$  is making uniform random selections (without replacement) from  $\mathcal{D}$  when constructing  $\mathcal{X}$  the set of records it predicts to be members:  $\mathcal{X} \subset \mathcal{D} \mid \forall x \in \mathcal{X}, A(x) > t_z$ .

To assess the likelihood of this null hypothesis  $H_0(z)$  we proceed by first deriving an expression for the number of records  $n = |\mathcal{X}|$  the MIA classifier will predict as ‘members’ based on the threshold  $t_z$  and on the number of true positives it predicts, TPR@ $z$ .

Let  $TP_z, FP_z$  be the number of true/false positive records in  $\mathcal{X}$ , and assume that when the (floating point) true/false positive rates are interpolated from empirical data we take the floor of those values to map onto specific records included in  $\mathcal{X}$ . By definition:

$$\begin{aligned} n &= TP_z + FP_z \\ &= \lfloor (TPR@z \cdot N_{\text{train}}) \rfloor + \lfloor (z \cdot N_{\text{test}}) \rfloor \end{aligned} \quad (1)$$

We then apply a one-tail test for  $H_0(z)$ , assessing the over-representation of true positives in  $\mathcal{X}$ , using a hypergeometric distribution to calculate the probability of observing  $TP_z$  or more true positives when using a process that makes selections without replacement from  $\mathcal{D}$  and stops as soon as  $FP_z$  incorrect ‘member’ predictions have been made. The null hypothesis  $H_0(z)$  is rejected at confidence level  $\alpha$  if  $\alpha > P(TP_z^+ | H_0, z)$ , where

$$P(TP_z^+ | H_0, z) = 1 - \sum_{i=0}^{i < TP_z} P(i | \mathcal{D}, n) \quad (2)$$

$$= 1 - \sum_{i=0}^{i < TP_z} \frac{C(N_{\text{train}}, i) C(N_{\text{test}}, FP_z)}{C(N, (i + FP_z))} \quad (3)$$

$$= 1 - \sum_{i=0}^{i < TP_z} \frac{\frac{N_{\text{train}}!}{i!(N_{\text{train}}-i)!} \frac{N_{\text{test}}!}{FP_z!(N_{\text{test}}-FP_z)!}}{\frac{N!}{(i+FP_z)!(N-i-FP_z)!}} \quad (4)$$

where Equation 3 recognises that it is  $FP_z$  that is fixed rather than  $n$  when applying the hypergeometric distribution, and  $C(a, b) = \frac{a!}{b!(a-b)!}$  is the number of ways of selecting  $b$  items from a total of  $a$  without replacement.

Thus after some rearrangement we assert that with  $(1 - \alpha) \times 100\%$  confidence, a model  $M$  is MIA vulnerable at a FPR of  $z$  if:

$$1 - \alpha < \frac{N_{\text{train}}! N_{\text{test}}!}{N! FP_z! (N_{\text{test}} - FP_z)!} \sum_{i=0}^{i < TP_z} \frac{(i + FP_z)!(N - i - FP_z)!}{i!(N_{\text{train}} - i)!} \quad (5)$$

## 3.4 Experimental Setup

### 3.4.1 Membership Inference Attack

Here we use the LiRA [5] implementation from the open source SACRO-ML framework [47] to perform membership inference and assess the disclosure risk presented by a range of fitted tree-based classification models. LiRA is a fine-grained black-box evaluation metric-based MIA [30] and (especially when using sufficient shadow models) is representative of the current state-of-the-art. Furthermore, it places an emphasis on an attacker scenario that corresponds to TRE concerns: *“If a membership inference attack can reliably violate the privacy of even just a few users in a sensitive dataset, it has succeeded”*. For each target model hyperparameter combination, 10 random training and test data splits are performed to fit 10 different models, and a LiRA attack is executed for each. The attack uses 100 shadow models to estimate the loss distributions.

### 3.4.2 Disclosure Metrics

Since the ML privacy community has yet to reach a consensus on a single disclosure risk metric, a range of standard MIA success metrics are computed from the LiRA outputs; for example, TPR at low FPR, and area under the curve (AUC). These metrics are averaged over the 10 runs to produce a set of risk metrics for a given target model hyperparameter combination on a dataset.

Moreover, we note that attackers will almost always have access to some records they know are non-members of the training set—either because they have artificially created them, or perhaps because they may be their own records. We therefore posit that armed with that information, and aware (from the literature) that in most cases the MIA cannot make a confident prediction, attackers will prefer a model that: (i) does not make high confidence member predictions for records they know are non-members; (ii) is correct when it predicts non-member with high confidence.

From these premises we derive a novel frequency difference metric  $FDIF_z$ , calculated as follows:

1. For each training record we note: the probability (softmax value) of the MIA’s prediction for the class member; and the true (binary) value.
2. This list is sorted by decreasing confidence of the member prediction.
3. The difference between the proportion of true positives in the extreme subsets; that is, the top and bottom  $z \times N$  records respectively is calculated. The more positive this  $FDIF_z$  value, the better the MIA is at discriminating when it is confident. Here we report scores for  $z \in \{0.02, 0.01, 0.001\}$ .
4. Finally we calculate  $P(FDIF_z)$ : the probability of observing this (or greater) value as an effect of stochastic sampling if the MIA was making purely random guesses at the extrema.

In addition to these MIA-specific metrics, several target model metrics (e.g., generalisation error) are also examined for correlations. All metrics considered here are described in Table 1.

**Table 1:** MIA and target model metrics.

Metric	Description
MIA AUC	Area under the receiver operating characteristic curve (AUC) of MIA success.
MIA $P(\text{AUC})$	Probability of observing MIA AUC with a random attack.
MIA $\text{FDIF}_z$	Difference between MIA TPR in the top and bottom $z$ records.
MIA $P(\text{FDIF}_z)$	Probability of observing $\text{FDIF}_z$ with a random attack.
MIA $\text{TPR}@z$	MIA TPR at a FPR of $z : z \in \{0.1, 0.001\}\%$ .
MIA Advantage	Maximum FPR minus TPR [54].
Target Train ACC	Target model training accuracy.
Target Test ACC	Target model test accuracy.
Target AUC	Target model test AUC.
Target GE	Target model generalisation error (GE). While overfitting has been shown to be a sufficient condition for MIA vulnerability in popular ML algorithms, it is not a necessary condition [54]. GE can therefore be considered as a lower bound for disclosure risk.

### 3.4.3 Target Models

The target models assessed here are Random Forest and Decision Tree classifiers from the widely used open source Python scikit-learn [31] package, and the XGBoost classifier [6]. The hyperparameter space is sampled with a range of values for the most commonly used hyperparameters; all others are set to their implementation defaults.

The sets of values used to define all the possible combinations of hyperparameter settings for the Decision Tree classifier, Random forest classifier and XGBoost are given in Table 2.

Tree-based models have been shown to provide state-of-the-art performance for tabular data classification tasks [46, 6] and therefore a selection of tabular datasets has been chosen here to provide a range of different landscape complexities for target model fitting. Binary classification problems are used here to aid comparison; increasing class number is known to increase the vulnerability to MIA [45]. Each dataset is split into three partitions, stratified by target label. The first subset is used for target model fitting, and the second is used for assessing target model accuracy. The LiRA attack then uses a combination of both the known training set and the third (unseen) subset to evaluate the MIA success.

### 3.4.4 Datasets

The mimic2-iaccd dataset was created for the purpose of investigating the effectiveness of indwelling arterial catheters in hemodynamically stable patients with respiratory failure for mortality outcomes. The dataset is derived from MIMIC-II [36], the publicly accessible critical care database. It contains summary clinical data and outcomes for 1776 patients. The following repetitive and uninformative features were removed: 'service unit', 'day icu intime', 'hosp exp flg', 'icu exp flg', 'day 28 flg' and 'sepsis flg'. The target label selected was 'sensor flg'.

The in-hospital-mortality dataset was created to characterise the predictors of in-hospital mortality for intensive care unit admitted patients. Data were extracted from the MIMIC-III [58] database from 1177 heart failure patients. Features 'ID' and 'group' were removed and the target label selected was 'outcome'.

The Indian liver dataset [38], contains 416 liver patient records and 167 non liver patient records. The dataset was collected from north east of Andhra Pradesh, India. This dataset

**Table 2:** Target models and hyperparameters sampled.

Model	Name	Values
Decision Tree	criterion	gini, entropy
	max_depth	1, 2, 5, 10, 20
	min_samples_leaf	1, 2, 5, 10, 20
	min_samples_split	2, 5, 10, 20
	max_features	None, sqrt, log2
	class_weight	None, balanced
	splitter	best, random
Random Forest	criterion	gini, entropy
	max_depth	1, 2, 5, 10, 20
	n_estimators	2, 5, 10, 20, 50, 100
	min_samples_split	2, 5, 10, 20
	class_weight	None, balanced
	min_samples_leaf	1, 2, 5, 10, 20
	bootstrap	True, False
XGBoost	n_estimators	1, 2, 5, 10, 15, 25, 50, 75, 90, 100
	max_depth	1, 2, 5, 10, 20
	reg_alpha	0, 0.001, 0.005, 0.01, 0.05
	min_child_weight	1, 2, 4, 8, 12

contains 441 male patient records and 142 female patient records. The feature ‘gender’ was transformed to numerical values and the target label selected was ‘Selector’, which identifies whether patients have liver disease or not.

The synth-ae dataset [28] was created by a Bayesian statistical model of de-identified patient information collected in the English National Health Service from 2014 to 2018. The following features were removed: ‘AE Arrive Date’, ‘AE Arrive HourOfDay’, ‘Admission Method’, ‘ICD10 Chapter Code’, ‘Treatment Function Code’, ‘Length Of Stay Days’, ‘Provid’. Any patients with missing data were also removed. Non-numerical features were transformed to numerical. Only the first 5000 rows were kept to generate the models. The target label selected was ‘Admitted flag’.

The sick and mammography datasets are open datasets available from OpenML<sup>4</sup> with ID 41946 and 310, respectively.

## 4 Results

### 4.1 Results for Model Hyperparameter Risks

#### 4.1.1 Risk Correlation Across Metrics

Table 3 shows the Kendall  $\tau$  rank correlation statistics for each metric with pair-wise comparisons for the XGBoost classifier averaged over all datasets. Strong correlations between all of the MIA metrics can be seen. Target model generalisation error also positively correlates with risk, however target model test accuracy does not, indicating that it is strongly dataset dependant. These correlations are also seen on the other datasets with both Random

<sup>4</sup><https://openml.org>

**Table 3:** Kendall  $\tau$  rank correlation statistics of different MIA and target model metrics for XGBoost target models. Mean over all datasets. Darker shades of grey indicate stronger correlation.

	MIA							Target Accuracy			
	AUC	$P(\text{AUC})$	$P(\text{FDIF}_{0.1})$	$\text{FDIF}_{0.1}$	TPR @0.1	TPR @0.001	Adv.	Train	Test	AUC	GE
MIA AUC		-0.90	0.86	0.87	0.86	0.72	0.90	0.83	0.27	0.30	0.80
MIA $P(\text{AUC})$	-0.90		-0.81	-0.82	-0.82	-0.67	-0.85	-0.79	-0.28	-0.31	-0.76
MIA $P(\text{FDIF}_{0.1})$	0.86	-0.81		0.96	0.86	0.74	0.82	0.83	0.28	0.26	0.80
MIA $\text{FDIF}_{0.1}$	0.87	-0.82	0.96		0.88	0.75	0.84	0.83	0.28	0.26	0.80
MIA TPR@0.1	0.86	-0.82	0.86	0.88		0.73	0.83	0.81	0.27	0.26	0.78
MIA TPR@0.001	0.72	-0.67	0.74	0.75	0.73		0.71	0.73	0.28	0.20	0.70
MIA Advantage	0.90	-0.85	0.82	0.84	0.83	0.71		0.82	0.28	0.30	0.78
Target Train	0.83	-0.79	0.83	0.83	0.81	0.73	0.82		0.32	0.27	0.90
Target Test	0.27	-0.28	0.28	0.28	0.27	0.28	0.28	0.32		0.61	0.22
Target AUC	0.30	-0.31	0.26	0.26	0.26	0.20	0.30	0.27	0.61		0.22
Target GE	0.80	-0.76	0.80	0.80	0.78	0.70	0.78	0.90	0.22	0.22	

Forest and Decision Tree classifiers (not shown). Since these metrics are highly correlated, this provides strong evidence in support of hypothesis 1.1 and the MIA AUC is used for simplicity in the remainder of the analysis. Regardless of the chosen metric, the following results should hold.

#### 4.1.2 Risk Correlation Across Datasets

Table 4 shows the correlation of risk across different datasets to test hypothesis 1.2. That is, the Kendall  $\tau$  rank correlation statistics for dataset pair-wise comparisons of the MIA AUC for each target model hyperparameter combination. As can be seen, there is a strong correlation of risk with respect to hyperparameter combination across the different datasets for each of the target model types. For Decision Tree models the mean is 0.64 (min=0.45, max=0.81); for Random Forest models the mean is 0.72 (min=0.51, max=0.88); and for XGBoost models the mean is 0.67 (min=0.48, max=0.84). This suggests that it may be possible to train a classification model that identifies the relative risk of a hyperparameter combination for most datasets.

#### 4.1.3 Risk Prediction Across Datasets

Table 5 shows the precision, recall, and accuracy of the simple rulesets at predicting whether a given target model hyperparameter combination is in the top 20% highest MIA AUC (high risk) or in the bottom 80% (low risk). Note that the risk prediction rules for each target model were derived by training a Decision Tree classifier on the results from the mimic2-iaccd dataset then distilling the tree into a set of rules. The mean accuracy on the unseen datasets is 91.4% (min=89.1%, max=93.9%) for Decision Tree target models; 95.2% (min=92.9%, max=96.2%) for Random Forest; and 94.1% (min=90.8%, max=97.2%) for XGBoost target models, showing that the very simple ruleset is able to accurately differentiate the hyperparameter combinations with the highest relative risk on each dataset, supporting hypothesis 1.3.

Scatter plots showing the relationship between target model accuracy (test set AUC) and privacy risk (MIA AUC) for different hyperparameter combinations of the Decision Tree classifier can be seen in Figure 1. Similar plots are shown for the Random Forest classifier target model in Figure 2 and for the XGBoost classifier target model in Figure 3. The high accuracy of the risk rulesets is illustrated by colouring the points red where the rules predict

**Table 4:** Kendall  $\tau$  rank correlation statistics for MIA AUC across different datasets for each target model type. Darker shades of grey indicate stronger correlation.

Model	Dataset	Mimic	Liver	Hospital	Synth	Sick	Mammo.
Decision Tree	Mimic		0.81	0.76	0.79	0.56	0.62
	Liver	0.81		0.73	0.73	0.53	0.60
	Hospital	0.76	0.73		0.68	0.51	0.55
	Synth	0.79	0.73	0.68		0.56	0.68
	Sick	0.56	0.53	0.51	0.56		0.45
	Mammography	0.62	0.60	0.55	0.68	0.45	
Random Forest	Mimic		0.88	0.71	0.73	0.80	0.76
	Liver	0.88		0.71	0.74	0.78	0.76
	Hospital	0.71	0.71		0.51	0.67	0.61
	Synth	0.73	0.74	0.51		0.68	0.73
	Sick	0.80	0.78	0.67	0.68		0.71
	Mammography	0.76	0.76	0.61	0.73	0.71	
XGBoost	Mimic		0.84	0.61	0.79	0.75	0.76
	Liver	0.84		0.74	0.68	0.72	0.69
	Hospital	0.61	0.74		0.48	0.53	0.52
	Synth	0.79	0.68	0.48		0.61	0.65
	Sick	0.75	0.72	0.53	0.61		0.73
	Mammography	0.76	0.69	0.52	0.65	0.73	

**Table 5:** Accuracy of MIA (high/low) risk prediction rules extracted from the mimic dataset and applied to other datasets for each target model type.

Model	Dataset	Low Risk		High Risk		Accuracy
		Precision	Recall	Precision	Recall	
Decision Tree	Mimic (train)	1.00	0.93	0.79	1.00	0.95
	Liver	0.99	0.92	0.75	0.95	0.93
	Hospital	0.96	0.90	0.68	0.86	0.89
	Synth	0.99	0.93	0.77	0.98	0.94
	Sick	0.97	0.91	0.71	0.90	0.91
	Mammography	0.97	0.91	0.71	0.90	0.91
Random Forest	Mimic (train)	1.00	0.98	0.91	0.99	0.98
	Liver	0.99	0.96	0.87	0.94	0.96
	Hospital	0.98	0.96	0.86	0.93	0.96
	Synth	0.97	0.94	0.80	0.87	0.93
	Sick	0.99	0.97	0.87	0.95	0.96
	Mammography	0.98	0.96	0.86	0.94	0.96
XGBoost	Mimic (train)	1.00	0.99	0.97	0.99	0.99
	Liver	0.98	0.98	0.92	0.94	0.97
	Hospital	0.96	0.96	0.82	0.84	0.93
	Synth	0.97	0.97	0.86	0.88	0.95
	Sick	0.94	0.94	0.76	0.78	0.91
	Mammography	0.97	0.96	0.85	0.87	0.94

a hyperparameter combination to have an MIA AUC in the top 20% highest risk. Overall, it can clearly be seen that removing the high risk combinations would eliminate a large

portion of significantly vulnerable models.

Interestingly, the plots show that removing the hyperparameter combinations predicted as high risk results in little to no loss of maximum achievable target AUC for most datasets. In the few cases where there is a hyperparameter combination classified as high risk that results in a higher target AUC, the difference is small, yet the reduction in MIA AUC is much larger. For example, the Decision Tree target model hyperparameter combinations with the highest accuracy are categorised as low risk for every dataset except the hospital dataset. Moreover, on the hospital dataset, the loss of accuracy (AUC) that would be experienced by removing the high risk combinations is only 0.0011. As another example, removing the high risk hyperparameters for Random Forest target models on the sick dataset, would result in a reduction in the maximum achievable accuracy (AUC) of only 0.005, but a reduction in MIA AUC of 0.13.

Beyond overfitting, MIA vulnerability is significantly influenced by model type and structure [45]. This is confirmed here with the rules learned from the mimic dataset for identifying high risk hyperparameter combinations in tree-based models. As might be expected, these rules consistently associate 'high risk' MIA vulnerability with model configurations that promote higher complexity. For example, across all model types (Decision Tree, Random Forest, and XGBoost), a greater `max_depth` is a recurring condition for high risk. Similarly, a larger number of `n_estimators` in ensemble methods (Random Forest and XGBoost) also points to increased vulnerability.

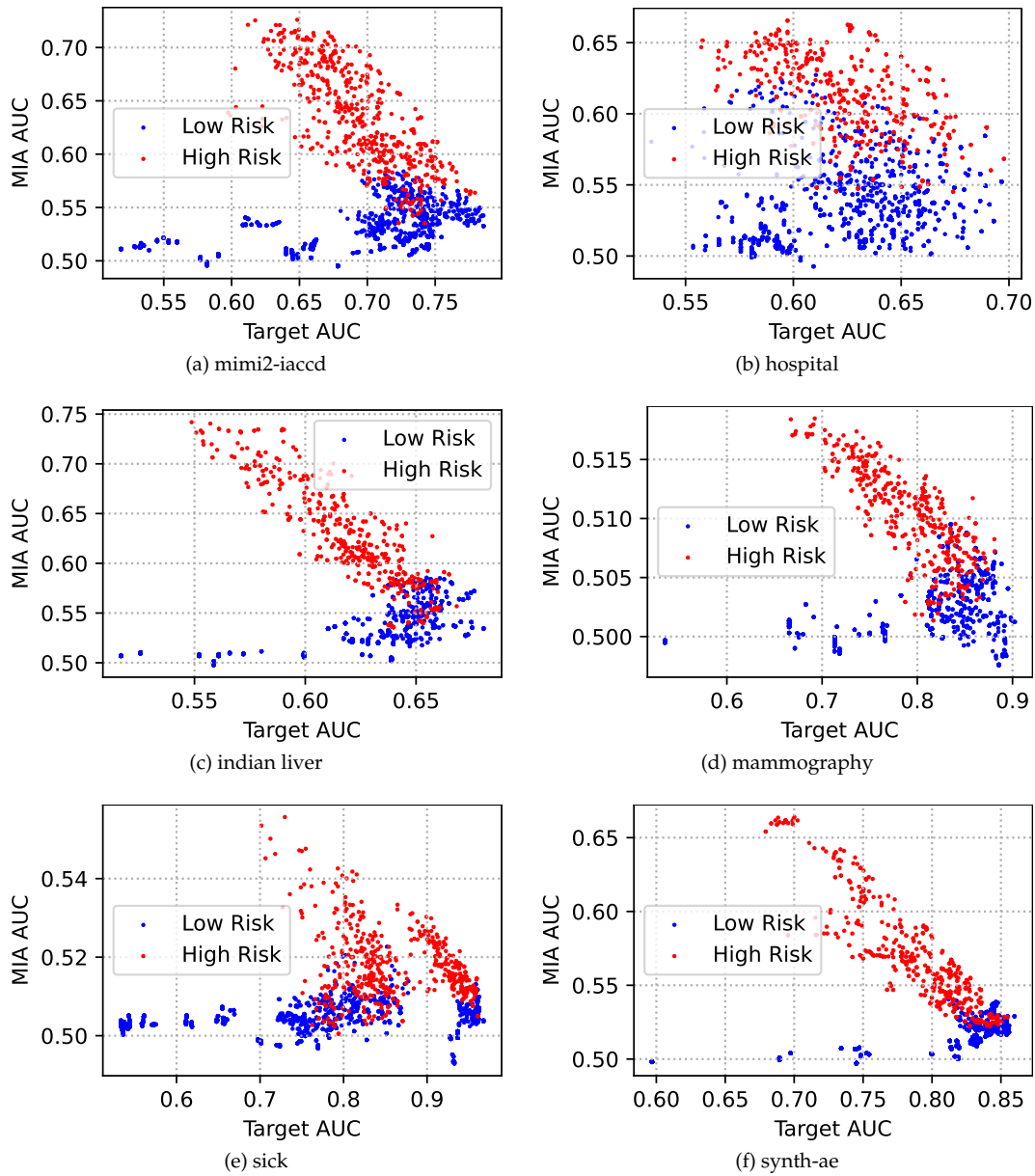
However, the rules reveal nuances beyond sheer size. Specific hyperparameter choices that dictate the model's learning capacity and flexibility are critical:

- For Decision Trees, risk factors include: utilising `splitter=best` rather than `random`; allowing small values for `min_samples_leaf` or `min_samples_split`; or considering all features for splits (`max_features=None`).
- In Random Forests, conditions such as `bootstrap=False` (where each tree is trained on the entire dataset, potentially reducing diversity) or specific `min_samples_split` thresholds when combined with other hyperparameters contribute to high risk.
- For XGBoost, a low `min_child_weight`, which permits the creation of tree nodes with fewer samples, is a key indicator of increased vulnerability.

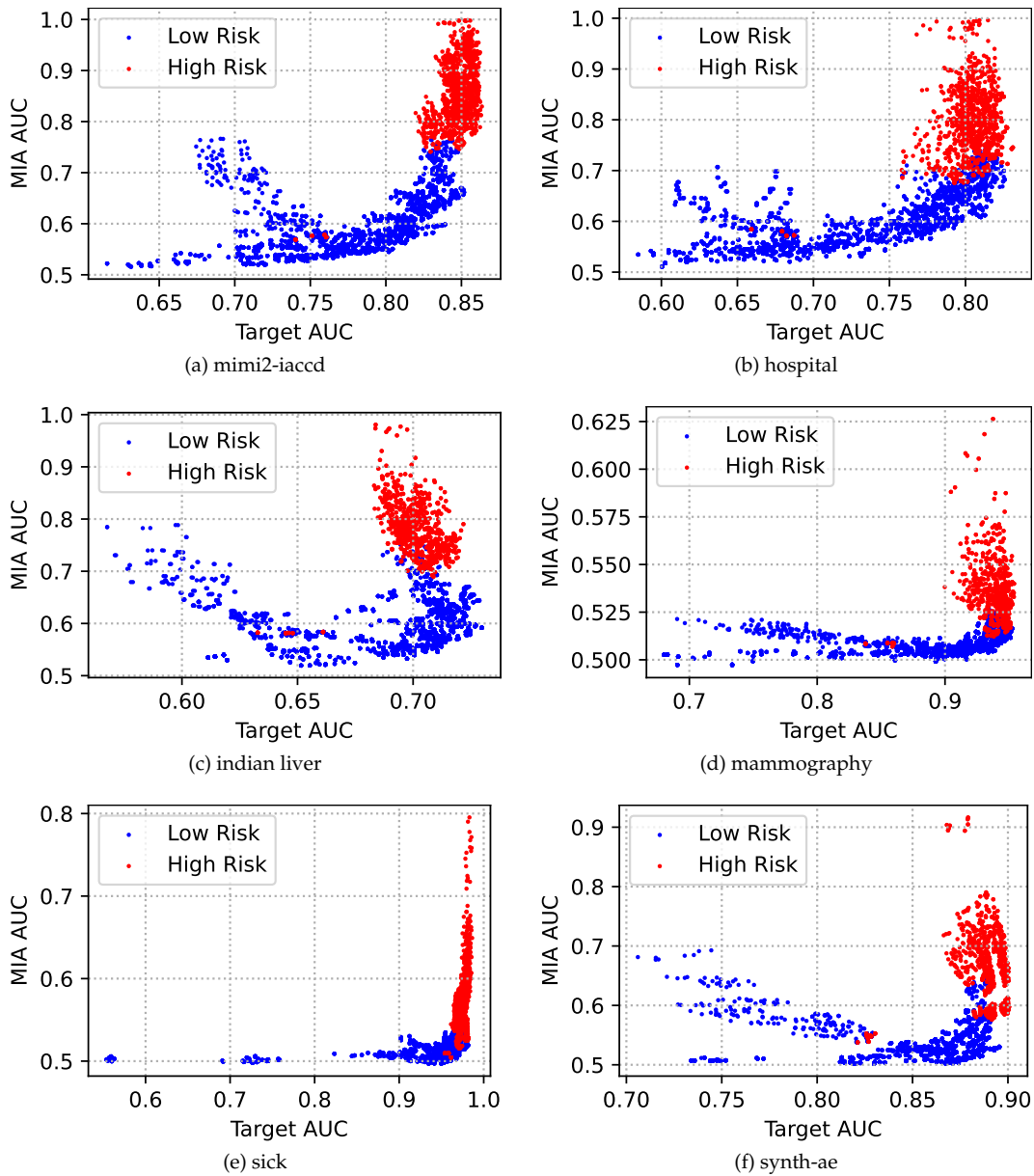
These specific hyperparameter settings effectively enable the models to create more intricate and numerous decision boundaries. This increased capacity to fit the training data closely not only reflects a higher Vapnik-Chervonenkis (VC) dimension (hence the link to generalisation error) but also, as the rules suggest, heightens the model's susceptibility to MIA.

The greediness of the search algorithm also significantly contributes to risk with `splitter` and `bootstrap` controlling how the trees are built. The `splitter` parameter determines the strategy used to choose the best split point for each node in a tree and a greedy approach evaluates all possible splits and chooses the one that maximises the chosen criterion. This can lead to trees that are highly sensitive to the specific training data and may not generalise well. Similarly, `bootstrap` controls whether the individual trees are trained on random subsets of the training data, which can introduce diversity into the ensemble, reducing overfitting. If bootstrapping is used in a way that reduces diversity, the trees may become too similar and prone to the same overfitting tendencies, again increasing risk.

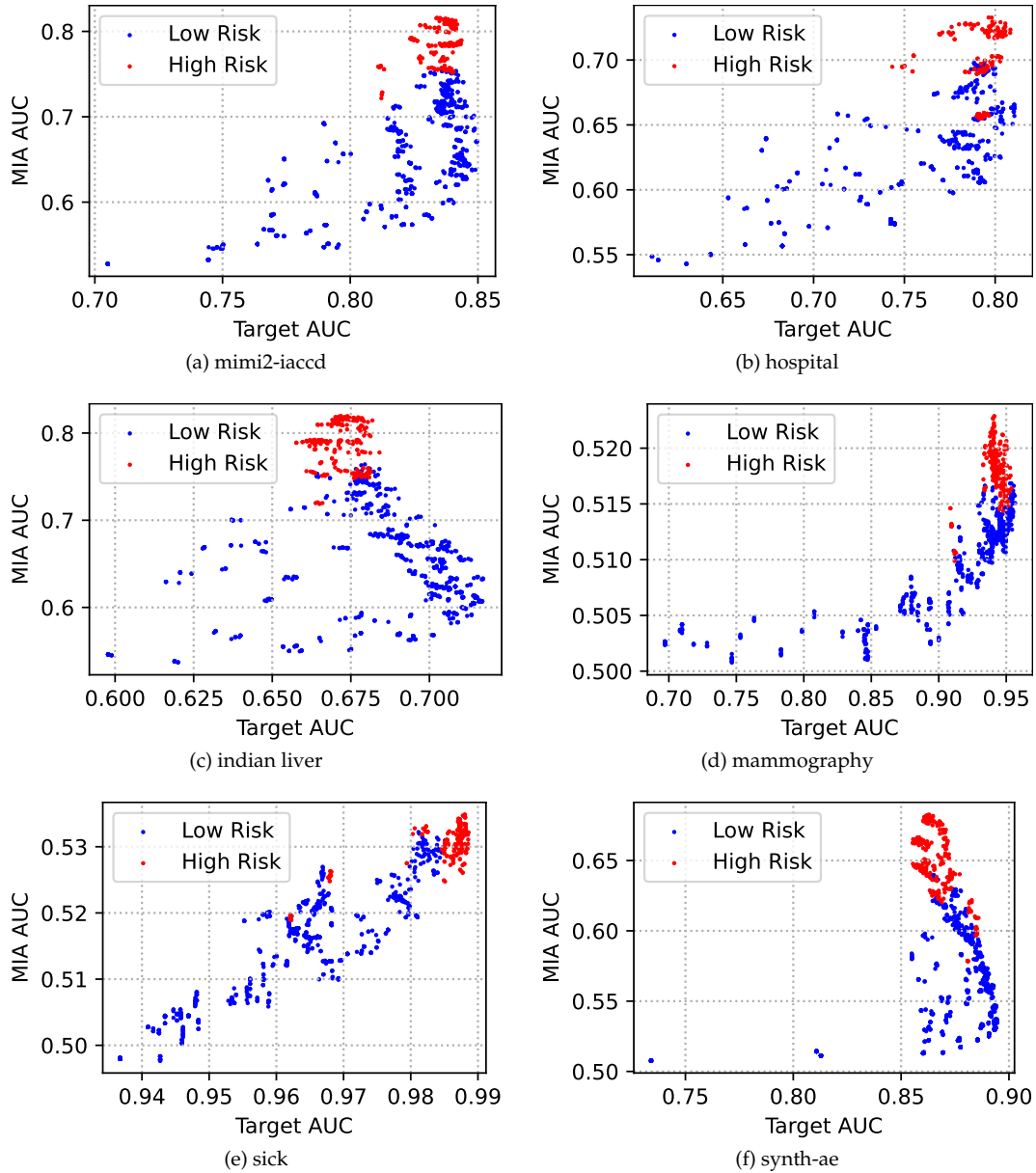
A combination of higher *potential* model complexity and greedy search algorithms are therefore much more likely to lead to models that are highly susceptible to overfitting and thus pose a higher disclosure risk in practice. The risk prediction rules may be found in Appendix A.1.



**Figure 1:** Privacy risk (MIA AUC) vs. accuracy (target AUC) for Decision Tree target models. Points in red denote hyperparameter combinations predicted to be in the top 20% (high risk) most vulnerable to MIA.



**Figure 2:** Privacy risk (MIA AUC) vs. accuracy (target AUC) for Random Forest target models. Points in red denote hyperparameter combinations predicted to be in the top 20% (high risk) most vulnerable to MIA.



**Figure 3:** Privacy risk (MIA AUC) vs. accuracy (target AUC) for XGBoost target models. Points in red denote hyperparameter combinations predicted to be in the top 20% (high risk) most vulnerable to MIA.

## 4.2 Results for Structural Risks

To examine support for the two different hypotheses stated in Section 3.2 we aggregated the results from LiRA and structural attacks on all (73610) combinations of dataset-algorithm-hyperparameters. We also created a `Combined_AND` structural risk as the logical conjunction (AND) of the degrees of freedom (`DoF`),  $k$ -anonymity (`K_anon`) and class disclosure (`lowvals_cd`) structural measures.

Figure 4 shows the histograms of the observed values of different MIA metrics (columns) for different datasets (rows) separated into models with combined structural risk 0 (low risk; green bars) and 1 (high risk; red bars).

As can be seen:

- The risk metrics have different distributions; higher where there is an identified combined structural risk (red).
- Except for the `TPR@0.001` metric and the two biggest datasets, there are always cases with no combined structural risk but non-zero MIA risk (green bars).
- In contrast, only in a tiny number of cases is there evidence of models having 0 value for MIA risk metrics but 1 for the combined structural risk metric (red bars).

Hypothesis 2.1 proposed that when there was no structural risk, the LiRA risk metrics would also be near zero. The second observation above is illustrated in Figure 5 (left). This shows conclusive evidence that this hypothesis can be rejected. In other words, **structural risk is not necessary** for vulnerability to MIA, at least in the form in which structural risk is currently quantified.

Hypothesis 2.2 took the alternative position that the presence of structural risk was predictive of a model being vulnerable to LiRA. As Figure 5 (right) shows there are only a handful of such cases that do not also display MIA vulnerability.

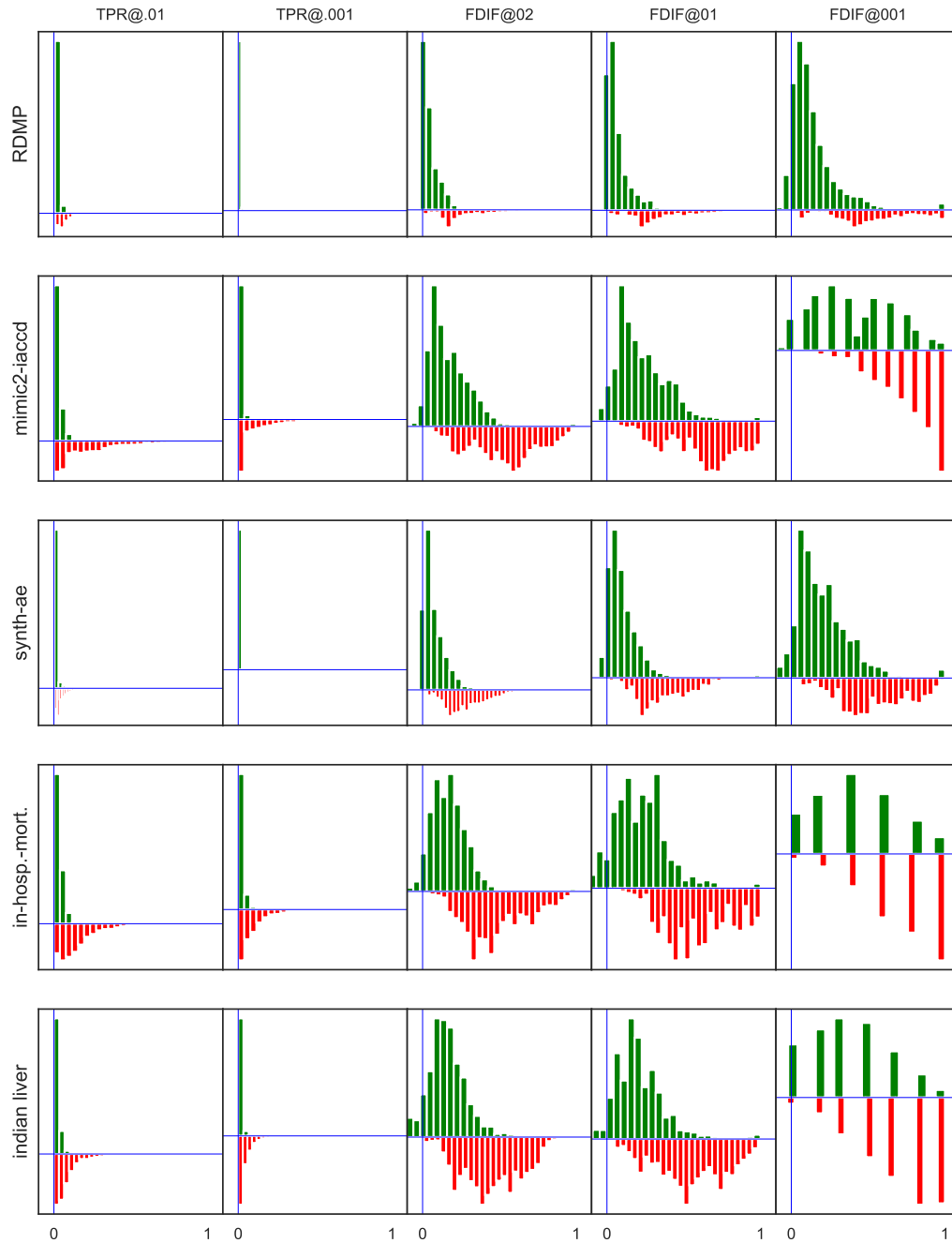
While these counterexamples cannot be ignored; from the perspective of an analyst making pragmatic decisions about whether a trained model poses a privacy risk, there is a clear message. The combined structural measure, `Combined_AND` presents **sufficient but not necessary** evidence of MIA vulnerability, without the need to run computationally expensive attacks. A risk prediction classifier based on this simple rule has:

- high precision (0.927–0.993, depending on the specific MIA risk metric being predicted)
- but relatively low recall (0.461–0.514, depending on the attack metric).

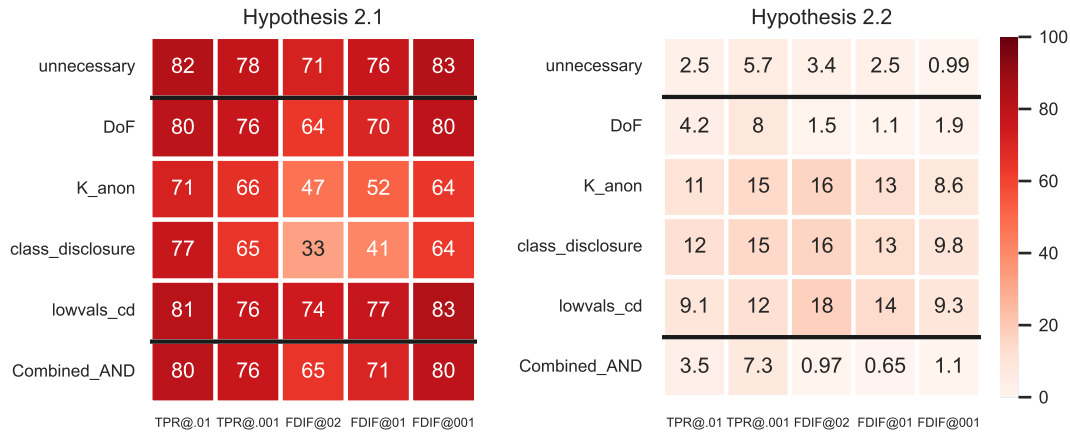
It is worth reiterating that this is only a filtering approach. A prediction of ‘not unsafe’ should not be interpreted as meaning the target is safe, rather that further investigation is needed (because the AND rule has low recall). This is somewhat akin to the distinction between ‘innocent’, ‘guilty’, and ‘not proven’ in some legal jurisdictions.

### 4.2.1 Analysis of Where MIA Metrics are Statistically Significant

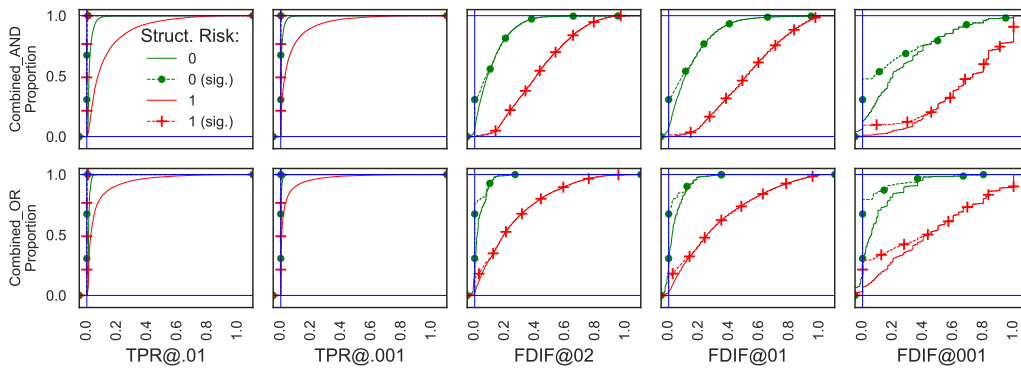
Section 3.3 describes a procedure for calculating whether the TPR at a given FPR is in fact statistically significant. For the `FDIF` metric we also have the equivalent calculation of probability that the result is observed by chance. Figure 6 shows the results (95% confidence level) of using this analysis to mask the results (i.e., metric values are set to 0 if the null hypothesis cannot be rejected) and then plotting the cumulative distribution of observed metric values. From these plots we can summarise that:



**Figure 4:** Frequencies of observed risk metrics for different datasets when combined structural risk is 0 (above, green) or 1 (below, red). Note each subplot has different  $y$  scales. Blue lines indicate frequency or metric value of 0. Note that while TPR metrics are always  $\geq 0$ , a poor membership attack model may have  $FDIF < 0$ .



**Figure 5:** Heatmaps showing percentages of cases providing evidence against the two hypotheses, cross-tabulated by structural (rows) and LiRA (columns) vulnerability metrics.



**Figure 6:** Empirical cumulative distribution functions for proportion of cases ( $y$ -axis) vs. value of different metrics ( $x$ -axis). Top row shows vulnerability to *all* structural metrics, where red lines correspond to hypothesis 2.2. Bottom row shows vulnerability to *any* of the structural metrics, where green lines relate to hypothesis 2.1. Green lines are when the structural risk metric is 0 (low risk) and red lines when risk is 1 (high risk) with solid lines for all cases, and dashed lines with markers showing only those cases where the reported MIA metric is statistically significant with  $>95\%$  confidence. In an ideal case, the green lines would rise immediately to 1 at the origin; that is, all models with no structural risk would have zero MIA risk. By contrast, ideally the red lines (structural risk is present) would begin low then rise only if significant MIA metrics were observed. If all observed results were statistically significant then the paired lines (i.e., with/without cross markers) would be identical.

- In the overwhelming majority of cases, the numbers of true positive cases identified by LiRA for a fixed (low) proportion of false positives are not high enough to reject the null hypothesis that the attack is just making random guesses. Of the 73600 different

algorithm-dataset-hyperparameter combinations tested, only 451 were statistically significant for  $\text{TPR}@0.01$  and 97 for  $\text{TPR}@0.001$ .

In Figure 6 this is evidenced in the two left hand plots. The distributions of observed  $\text{TPR}@z$  metrics for raw data (ignoring statistical significance) show a clear difference between whether structural risk is present (plain red line) or not (plain green). However that difference disappears for the significant results (lines with markers).

- By way of contrast, the FDIF metrics are almost always significant, i.e., the pairs of plain and cross-marked lines are close. The main difference is for  $\text{FDIF}_{0.001}$  when only the top and bottom 0.1% of records (ranked by attack confidence) are considered. Even in this case the difference between the extrema is significant in more than 55% of cases where the structural metric is 0 and 90% of cases where the structural risk metric is 1.
- For all cases, the cumulative frequency of cases where the structural metric is 1 (red line) grows more slowly than when it is 0. In other words, LiRA risk metrics are higher when the structural risk is present.

## 5 Conclusion

This article has presented the first empirical analysis of tree-based classification model hyperparameters in terms of their susceptibility to MIA. Surprisingly, regardless of MIA success metric, the rank order correlation of disclosure risk is seen to be preserved across datasets. It has been shown that simple human-interpretable rulesets learned on one dataset can predict with approximately 90–95% accuracy whether a given hyperparameter combination is one of the 20% most vulnerable on other unseen datasets. Significantly, it has been shown that for most datasets, high target model accuracy is still obtainable with hyperparameter combinations that are not in this highest risk category.

The use of more complex rules/models such as deep neural networks would likely increase the accuracy of risk prediction, however the use of a simple Decision Tree to extract the ruleset enables easy interpretation by researchers so that they can more easily adjust their preferred hyperparameters into a safer region that maintains sufficient classification accuracy. Future work may explore the use of rule-based learners such as the XCSF learning classifier system [34] to extract interpretable rules with improved accuracy.

While these simple rules offer valuable insights, it is important to acknowledge that they cannot definitively determine ‘safe’ hyperparameter combinations, as absolute risk scores are highly dependent on dataset characteristics such as distribution, imbalances, outliers, and size. Nevertheless, our results suggest that the *relative* risk of hyperparameters can remain surprisingly consistent across diverse datasets. This consistency implies that hyperparameter values associated with the highest risk can reasonably be excluded from the tuning process—and consequently, before running expensive MIAs—without significant loss of performance. Although further dataset testing with the rules extracted here may reveal limitations in generalisation, the existence of dataset classes where rules learned from one dataset successfully generalise to others presents an exciting opportunity to exploit similarities within secure data facilities such as TREs, which often house large numbers of related datasets.

Some datasets/applications may still require some privacy trade-off for sufficient utility, but it seems reasonable that as the first stage of risk assessment it should be incumbent on the researchers wishing to release a model to first show that it is necessary to use these highest risk hyperparameters. In the event that the rules do not appear to be generalising to

a particular dataset after sampling some hyperparameter values, then the use of surrogate modelling to efficiently identify the best privacy-utility trade-off could be used (i.e., generalising within the dataset rather than across datasets), however this may not be possible for very large models and datasets.

Once the ML model has been trained, the use of computationally cheap structural metrics as shown here can serve as sufficient, although not necessary, indicators of MIA vulnerability. Again, this approach for identifying vulnerable models cannot certify a model as safe, but potentially enables a large number of high risk models to be filtered and therefore a significant reduction in the number needing to undergo expensive MIAs such as LiRA.

Current work is integrating the learned rulesets and structural attacks within the open source SACRO Python toolkit [47, 33] to provide a comprehensive hierarchical framework to perform these risk assessments, in addition to running current state-of-the-art MIAs. Future work may explore additional target models, hyperparameters, and datasets. Including, for example, the consideration of non-tree models such as neural networks. Potentially, a similar approach could be undertaken by using feature encoding schemes similar to those used for the hyperparameters when surrogate modelling neural architectures, e.g., hand-crafted features such as the number of parameters, depth, width, number of convolutions, or sequence based representations such as tokenising the architecture [52, 29]. The structural metrics presented here also have similarities with zero-cost proxies for estimating the accuracy of deep neural networks [26] and issues such as overparameterising and lack of regularisation hyperparameters may be indicative of MIA risk.

The data produced from experimentation and analysis code used to produce the figures and statistics are available in [35].

## Acknowledgements

This work was funded by UK Research and Innovation (Grant Numbers MC\_PC\_23006, MC\_PC\_21033, and MC\_PC\_24038) as part of the Data and Analytics Research Environments UK (DARE UK) programme, delivered in partnership with Health Data Research UK (HDR UK) and Administrative Data Research UK (ADR UK).

## References

- [1] B. Arnold et al. The Turing way: A handbook for reproducible data science, Mar. 2019.
- [2] L. Bai et al. RMR: A relative membership risk measure for machine learning models. *IEEE Transactions on Dependable and Secure Computing*, 2025. In press.
- [3] M. Bertran et al. Scalable membership inference attacks via quantile regression. In A. Oh et al., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 314–330, Red Hook, NY, USA, 2023. Curran Associates Inc.
- [4] A. Blanco-Justicia, D. Sánchez, J. Domingo-Ferrer, and K. Muralidhar. A critical review on the use (and misuse) of differential privacy in machine learning. *ACM Computing Surveys*, 55(8):1–16, Dec. 2022. Article number 160.
- [5] N. Carlini et al. Membership inference attacks from first principles. In R. Bobba, editor, *IEEE Symposium on Security and Privacy*, pages 1897–1914, Piscataway, NJ, USA, 2022. IEEE Press.
- [6] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, New York, NY, USA, 2016. ACM.

- [7] C. A. Choquette-Choo, F. Tramer, N. Carlini, and N. Papernot. Label-only membership inference attacks. In M. Meila and T. Zhang, editors, *International Conference on Machine Learning*, volume 139, pages 1964–1974. PMLR, July 2021.
- [8] DARE UK. Paving the way for a coordinated national infrastructure for sensitive data research. <https://doi.org/10.5281/zenodo.7022440>, Aug. 2022.
- [9] A. Dionysiou and E. Athanasopoulos. SoK: Membership inference is harder than previously thought. *Proceedings on Privacy Enhancing Technologies*, 2023(3):286–306, 2023.
- [10] C. Dwork. Differential privacy: A survey of results. In M. Agrawal, D. Du, Z. Duan, and A. Li, editors, *International Conference on Theory and Applications of Models of Computation*, volume 4978 of LNCS, pages 1–9, Berlin, Germany, 2008. Springer-Verlag.
- [11] A. V. Galichin, M. Pautov, A. Zhavoronkin, O. Y. Rogov, and I. Oseledets. GLiRA: Closed-box membership inference attack via knowledge distillation. *IEEE Transactions on Information Forensics and Security*, 20:3893–3906, Mar. 2025.
- [12] E. Green, F. Ritche, and P. White. The statbarn: A new model for output statistical disclosure control. In J. Domingo-Ferrer and M. Önen, editors, *Privacy in Statistical Databases*, pages 284–293, Cham, Switzerland, 2024. Springer Nature.
- [13] E. Green and F. Ritchie. The present and future of the Five Safes framework. *Journal of Privacy and Confidentiality*, 13(2):1–20, Nov. 2023.
- [14] H. Hu et al. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys*, 54(11s):1–37, Sept. 2022. Article number 235.
- [15] L. Hu et al. Defenses to membership inference attacks: A survey. *ACM Computing Surveys*, 56(4):1–34, Nov. 2023. Article number 92.
- [16] M. Jagielski et al. Students parrot their teachers: Membership inference on model distillation. In A. Oh et al., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 44382–44397, Red Hook, NY, USA, 2023. Curran Associates, Inc.
- [17] M. Jagielski, S. Wu, A. Oprea, J. Ullman, and R. Geambasu. How to combine membership-inference attacks on multiple updated machine learning models. *Proceedings on Privacy Enhancing Technologies*, 2023(3):211–232, 2023.
- [18] I. Jarin and B. Eshete. MIAShield: Defending membership inference attacks via preemptive exclusion of members. *Proceedings on Privacy Enhancing Technologies*, 2023(1):400–416, 2023.
- [19] E. Jefferson et al. GRAIMATTER Green Paper: Recommendations for disclosure control of trained Machine Learning (ML) models from Trusted Research Environments (TREs). *arXiv*, 2211.01656, Nov. 2022.
- [20] M. Jegorova et al. Survey: Leakage and privacy at inference time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):9090–9108, July 2023.
- [21] J. Jia, A. Salem, M. Backes, Y. Zhang, and N. Z. Gong. MemGuard: Defending against black-box membership inference attacks via adversarial examples. In *ACM SIGSAC Conference on Computer and Communications Security*, pages 259–274, New York, NY, USA, 2019. ACM Press.
- [22] S. Kavianpour, J. Sutherland, E. Mansouri-Benssassi, N. Coull, and E. Jefferson. Next-generation capabilities in trusted research environments: Interview study. *Journal of Medical Internet Research*, 24(9):e33720, Sept. 2022.
- [23] T. Leemann, M. Pawelczyk, and G. Kasneci. Gaussian membership inference privacy. In A. Oh et al., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 73866–73878, Red Hook, NY, USA, 2023. Curran Associates, Inc.
- [24] H. Liu, Y. Wu, Z. Yu, and N. Zhang. Please tell me more: Privacy impact of explainability through the lens of membership inference attack. In *IEEE Symposium on Security and Privacy*, pages 4791–4809, 2024.
- [25] Y. Liu et al. ML-Doctor: Holistic risk assessment of inference attacks against machine learning

- models. In *USENIX Security Symposium*, pages 4525–4542. USENIX, 2022.
- [26] J. Mellor, J. Turner, A. Storkey, and E. J. Crowley. Neural architecture search without training. In *Proceedings of the International Conference on Machine Learning*. JMLR.org, 2021.
- [27] F. Naretto, A. Monreale, and F. Giannotti. Evaluating the privacy exposure of interpretable global and local explainers. *Transactions on Data Privacy*, 18(2):67–93, May 2025.
- [28] NHS England. A&E Synthetic data. <https://data.england.nhs.uk/dataset/a-e-synthetic-data>, 2022.
- [29] X. Ning et al. Evaluating efficient performance estimators of neural architectures. In M. Ranzato et al., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12265–12277, Red Hook, NY, USA, 2021. Curran Associates, Inc.
- [30] J. Niu et al. Comparing different membership inference attacks with a comprehensive benchmark. *IEEE Transactions on Information Forensics and Security*, 2025. In press.
- [31] F. Pedregosa et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, Oct. 2011.
- [32] Y. Peng, J. Roh, S. Maji, and A. Houmansadr. OSLO: One-shot label-only membership inference attacks. In A. Globerson et al., editors, *Advances in Neural Information Processing Systems*, volume 37, pages 62310–62333, Red Hook, NY, USA, 2024. Curran Associates, Inc.
- [33] R. J. Preen, M. Albashir, S. Davy, and J. Smith. A multi-language toolkit for the semi-automated checking of research outputs. *IEEE Transactions on Privacy*, 2:55–66, May 2025.
- [34] R. J. Preen and D. Pätzel. XCSF. <https://doi.org/10.5281/zenodo.4277257>, Aug. 2024.
- [35] R. J. Preen and J. Smith. Data for: A hierarchical approach for assessing the vulnerability of tree-based classification models to membership inference attack. <https://doi.org/10.5281/zenodo.15620339>, June 2025.
- [36] J. Raffa. Clinical data from the MIMIC-II database for a case study on indwelling arterial catheters (version 1.0). PhysioNet, Oct. 2016. <https://doi.org/10.13026/C2NC7F>.
- [37] M. A. Rahman, T. Rahman, R. Laganière, and N. Mohammed. Membership inference attack against differentially private deep learning model. *Transactions on Data Privacy*, 11(1):61–79, Apr. 2018.
- [38] B. Ramana and N. Venkateswarlu. ILPD (Indian Liver Patient Dataset). UCI Machine Learning Repository, 2022. <https://doi.org/10.24432/C5D02C>.
- [39] S. Rezaei and X. Liu. On the difficulty of membership inference attacks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7888–7896, Piscataway, NJ, USA, 2021. IEEE Press.
- [40] S. Rezaei and X. Liu. On the discredibility of membership inference attacks. *arXiv*, 2212.02701, 2023.
- [41] M. Rigaki and S. Garcia. A survey of privacy attacks in machine learning. *ACM Computing Surveys*, 56(4):1–34, Nov. 2023. Article number 101.
- [42] F. Ritchie and M. Elliot. Principles- versus rules-based output statistical disclosure control in remote access environments. *IASSIST Quarterly*, 39(2):iq778, Dec. 2015.
- [43] S. Ruggles and D. L. Magnuson. “It’s none of their damn business”: Privacy and disclosure control in the U.S. census, 1790–2020. *Population and Development Review*, 49(3):651–679, 2023.
- [44] A. Salem, A. Bhattacharya, M. Backes, M. Fritz, and Y. Zhang. Updates-Leak: Data set inference and reconstruction attacks in online learning. In *USENIX Security Symposium*, pages 1291–1308. USENIX Association, 2020.
- [45] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In K. R. B. Butler, editor, *IEEE Symposium on Security and Privacy*, pages 3–18, Piscataway, NJ, USA, 2017. IEEE Press.

- [46] R. Shwartz-Ziv and A. Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, May 2022.
- [47] J. Smith et al. Safe machine learning model release from trusted research environments: The SACRO-ML package. *arXiv*, 2212.01233, Oct. 2024.
- [48] J. Tan, D. LeJeune, B. Mason, H. Javadi, and R. G. Baraniuk. Blessing of dimensionality in membership inference through regularization. In F. Ruiz, editor, *International Conference on Artificial Intelligence and Statistics*, volume 206, pages 10968–10993. PMLR, 2023.
- [49] J. Tan, B. Mason, H. Javadi, and R. Baraniuk. Parameters or privacy: A provable tradeoff between overparameterization and membership inference. In S. Koyejo et al., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 17488–17500, Red Hook, NY, USA, 2022. Curran Associates, Inc.
- [50] J. Tao and R. Shokri. Range membership inference attacks. In K. Rieck and S. Jha, editors, *IEEE Conference on Secure and Trustworthy Machine Learning*, pages 346–361, Los Alamitos, CA, USA, 2025. IEEE Computer Society.
- [51] M. Veale, R. Binns, and L. Edwards. Algorithms that remember: Model inversion attacks and data protection law. *Philosophical Transactions of the Royal Society A*, 376(2133):1–15, Nov. 2018.
- [52] C. White, A. Zela, B. Ru, Y. Liu, and F. Hutter. How powerful are performance predictors in neural architecture search? In M. Ranzato et al., editors, *International Conference on Neural Information Processing Systems*, volume 34, pages 28454–28469, Red Hook, NY, USA, 2021. Curran Associates Inc. Article number 2180.
- [53] J. Ye, A. Maddi, S. K. Murakonda, V. Bindschaedler, and R. Shokri. Enhanced membership inference attacks against machine learning models. In H. Yin and A. Stavrou, editors, *ACM SIGSAC Conference on Computer and Communications Security*, pages 3093–3106, New York, NY, USA, 2022. ACM.
- [54] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In C. Cremers, editor, *IEEE Computer Security Foundations Symposium*, pages 268–282, Piscataway, NJ, USA, 2018. IEEE Press.
- [55] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha. Overfitting, robustness, and malicious algorithms: A study of potential causes of privacy risk in machine learning. *Journal of Computer Security*, 28(1):35–70, 2020.
- [56] S. Zarifzadeh, P. Liu, and R. Shokri. Low-cost high-power membership inference attacks. *arXiv*, 2312.03262, 2024.
- [57] B. Z. H. Zhao et al. On the (in)feasibility of attribute inference attacks on machine learning models. In L. Bauer and C. Troncoso, editors, *IEEE European Symposium on Security and Privacy*, pages 232–251, Los Alamitos, CA, USA, 2021. IEEE Computer Society.
- [58] J. Zhou et al. Prediction model of in-hospital mortality in intensive care unit patients with heart failure: Machine learning-based, retrospective analysis of the MIMIC-III database. <https://doi.org/10.5061/dryad.0p2ngf1zd>, July 2021.

## A Appendix

### A.1 Rulesets for Predicting Unnecessary Risk

Figure 7 shows the rulesets distilled from Decision Tree classifiers trained to recognise whether a set of hyperparameters will produce a model ranked in the top 20% (high risk) for MIA risk on the mimic dataset.

---

**Decision Tree:**

- Rule 1: **if** ( $\text{max\_depth} > 7.5$ ) and ( $\text{min\_samples\_leaf} \leq 7.5$ ) and ( $\text{min\_samples\_split} \leq 15$ ) **then** high risk
- Rule 2: **if** ( $\text{splitter} = \text{best}$ ) and ( $\text{max\_depth} > 7.5$ ) and ( $\text{min\_samples\_leaf} \leq 7.5$ ) and ( $\text{min\_samples\_split} > 15$ ) **then** high risk
- Rule 3: **if** ( $\text{splitter} = \text{best}$ ) and ( $\text{max\_depth} > 7.5$ ) and ( $7.5 < \text{min\_samples\_leaf} \leq 15$ ) and ( $\text{max\_features} = \text{None}$ ) **then** high risk
- Rule 4: **if** ( $\text{splitter} = \text{best}$ ) and ( $3.5 < \text{max\_depth} \leq 7.5$ ) and ( $\text{max\_features} = \text{None}$ ) and ( $\text{min\_samples\_leaf} \leq 7.5$ ) **then** high risk
- Rule 5: **if** ( $\text{splitter} = \text{random}$ ) and ( $\text{max\_depth} > 7.5$ ) and ( $\text{min\_samples\_leaf} \leq 7.5$ ) and ( $\text{max\_features} = \text{None}$ ) **then** high risk
- 

**Random Forest:**

- Rule 1: **if** ( $\text{max\_depth} > 3.5$ ) and ( $\text{n\_estimators} > 35$ ) and ( $\text{max\_features} \neq \text{None}$ ) **then** high risk
- Rule 2: **if** ( $\text{max\_depth} > 3.5$ ) and ( $\text{n\_estimators} > 35$ ) and ( $\text{min\_samples\_split} \leq 15$ ) and ( $\text{max\_features} = \text{None}$ ) and ( $\text{bootstrap} = \text{True}$ ) **then** high risk
- Rule 3: **if** ( $\text{max\_depth} > 7.5$ ) and ( $15 < \text{n\_estimators} \leq 35$ ) and ( $\text{min\_samples\_leaf} \leq 15$ ) and ( $\text{bootstrap} = \text{False}$ ) **then** high risk
- 

**XGBoost:**

- Rule 1: **if** ( $\text{max\_depth} > 3.5$ ) and ( $3.5 < \text{n\_estimators} \leq 12.5$ ) and ( $\text{min\_child\_weight} \leq 1.5$ ) **then** high risk
- Rule 2: **if** ( $\text{max\_depth} > 3.5$ ) and ( $\text{n\_estimators} > 12.5$ ) and ( $\text{min\_child\_weight} \leq 3$ ) **then** high risk
- Rule 3: **if** ( $\text{max\_depth} > 3.5$ ) and ( $\text{n\_estimators} > 62.5$ ) and ( $3 < \text{min\_child\_weight} \leq 6$ ) **then** high risk
- 

Figure 7: MIA (high/low) risk rules for each target model; extracted from the mimic dataset.