Can we estimate privacy vulnerability of individual records? Towards Mitigating Attribute Inference Attacks on ML Models

Ehsanul Kabir
Pennsylvania State University

Najrin Sultana
Pennsylvania State University

Ninghui Li
Purdue University

Shagufta Mehnaz
Pennsylvania State University

Abstract

Machine learning (ML) has brought transformative applications across various sectors, including sensitive fields like healthcare, finance, and customer analytics. However, ML models are susceptible to privacy leaks, especially through attribute inference and model inversion attacks, raising concerns for data confidentiality in privacy-critical domains. Existing defenses pursue much broader objectives than specifically preventing privacy leakage from attribute inference attacks, and as a result often fail to provide fine-grained, vulnerability-aware protection without significant utility costs. Motivated by this need, we first investigate record-level vulnerability estimation through `NeighVE`, an adversary-side tool designed to identify which individual records are more exposed to inference. Insights from `NeighVE` reveal that the record-level risk of privacy leakage is largely agnostic to model architectures and attack strategies and is instead governed by dataset-level characteristics, particularly the distribution of sensitive attributes in the local neighborhood of each record. Building on this insight, we propose `VESL`, a subspace-learning–inspired defense that mitigates attribute-inference leakage while keeping utility loss to a bare minimum. As a byproduct of its balancing mechanism, `VESL` also improves fairness across sensitive attributes and prevents `NeighVE` from reliably identifying vulnerable records. As a supporting contribution, we introduce `AttriVET`, an estimator that predicts which individual records are vulnerable with over 90% accuracy across diverse scenarios, enabling risk-aware defense design and auditing.

1 Introduction

Machine learning (ML) has unlocked innovative data applications across a range of domains, including sensitive fields like personalized healthcare [2, 31], finance [16, 19], and customer analytics [5, 15]. Although powerful, recent findings reveal that ML models are prone to privacy leaks, with the potential for inadvertent exposure of their training data. For instance, adversaries may determine if specific records were part of training [12, 34, 41, 42] or, through model inversion, extract sensitive training data from the model [13, 21, 36, 50]. Such vulnerabilities hinder the adoption of ML in privacy-critical sectors, where data confidentiality is non-negotiable.

Model inversion attacks can generally be divided into two key types: class representative reconstruction [49, 53] and attribute inference [21, 36]. The first type, class representative reconstruction, involves adversaries attempting to reconstruct typical or "average" examples of specific classes within the training data, thereby generating a form of "class template." For example, if an adversary has query access to a face recognition model, they may attempt to reconstruct an image similar to a target individual's actual face. Attribute inference, however, is more intrusive for models that use tabular data, where attackers can target specific attributes within the data to recover sensitive details. In this type of attack, the adversary's objective is to infer sensitive attributes using model predictions and publicly accessible data of individuals included in the training. For example, in the healthcare domain, an adversary might query using known information about an individual and uncover sensitive attribute(s) related to their medical history. This presents a serious privacy concern for tabular data. Alarmingly, despite tabular datasets' widespread use in many applications, privacy vulnerabilities in tabular data are still under-examined [18, 21, 28, 30, 36] compared to other data types like images [3, 4, 6, 17, 34, 37, 51, 53].

Recent work [30] shows that by focusing on vulnerable subsets of data instead of treating all records equally, targeted attribute inference attacks achieve far greater success than untargeted ones. These vulnerable subsets are identified through a group-level vulnerability ranking method based on angular difference. This finding prompts a critical research question– *Can attribute inference vulnerability be predicted at the granularity of individual records rather than coarse groups?* Such predictive capability would allow adversaries to pinpoint the records for which inference is likely to succeed, thereby enabling stronger targeted attacks and, more importantly, generating insight into the causes of record-level vulnerability

that can guide the design of proactive, vulnerability-aware defenses. However, moving from group-level to record-level vulnerability prediction poses a key challenge: group-based approaches assume access to multiple records from the same subgroup—an assumption irrelevant for an adversary targeting a *single* individual's record. To overcome this limitation, we develop `NeighVE` (**Neigh**borhood-based Attribute Inference **V**ulnerability **E**stimator), which constructs synthetic neighbors of the target record and computes angular difference within this constructed neighborhood to predict the target record's vulnerability. Our results show that `NeighVE` provides a practical means of identifying records more susceptible to attribute inference.

Importantly, our use of synthetic neighborhoods in `NeighVE` points to a deeper insight: the local neighborhood of a record may be an underlying contributing factor to its vulnerability to attribute inference. Our intuition is that the distribution of sensitive attributes within this neighborhood closely influences its vulnerability. We formalize this idea by introducing the metric of *neighborhood similarity* and demonstrate that vulnerable and non-vulnerable records exhibit distinct distributions of neighborhood similarity, which persist across datasets. Neighborhood similarity is defined solely in terms of data and does not refer to models or attacks, yet it proves effective in vulnerability estimation across diverse models and attack strategies. This consistency suggests that record-level vulnerability is driven primarily by dataset-level characteristics. Consequently, neighborhood similarity serves as a valuable indicator of attribute inference vulnerability, providing a foundation for understanding, auditing, and defending against privacy leakage.

Building on this key finding, we ask: *Can vulnerability estimation be leveraged to mitigate privacy leakage from attribute inference attacks without incurring prohibitive utility loss?* Pursuing this objective, however, presents unique challenges. Existing defenses, including differential privacy [20] and mutual information regularization [47], apply coarse privacy mechanisms uniformly across all records. While they reduce attack performance, they do so only at the cost of substantial model utility [18, 22]. Furthermore, it is essential to ensure that these defenses do not disproportionately impact the performance of the ML model on the vulnerable subset of records compared to the non-vulnerable subset, which could lead to fairness concerns. To address these challenges, we propose `VESL` (**V**ulnerability **E**stimation-driven **S**ubspace **L**earning), a novel technique inspired by subspace learning [27, 48]. `VESL` constructs multiple balanced subsets of the original training data, maintaining an optimal proportion between vulnerable and non-vulnerable records. Our findings show that `VESL` reduces attribute inference attack performance to that of imputation, an attack that adversaries can mount without access to the trained model. In other words, models trained with `VESL` do not leak more information than what is already available through simple data statistics. In addition, `VESL` renders `NeighVE` ineffective in identifying vulnerable records. Furthermore, `VESL` maintains utility while also improving fairness, a byproduct of its dataset balancing mechanism that distributes vulnerable and non-vulnerable records more evenly across sensitive attribute values.

Finally, we develop `AttriVET` (**Attri**bute **I**nference **V**ulnerability **E**stimation **T**ool), a defender-side tool for estimating record-level vulnerability to attribute inference attacks. `AttriVET` provides a systematic way to audit privacy risk by directly estimating which records are most susceptible, offering a practical means to quantify potential leakage before deploying models in sensitive domains. We empirically show that `AttriVET` is capable of accurately predicting a record's vulnerability to various attribute inference attacks across a broad range of scenarios.

Summary of contributions. Our work makes the following contributions:

❒ We identify neighborhood similarity as a key dataset-level factor underlying record-level vulnerability. We show that vulnerable and non-vulnerable records exhibit distinct neighborhood similarity distributions, and that this metric remains effective across datasets, models, and attack strategies, establishing dataset characteristics as the primary driver of attribute inference risk.

❒ We develop `VESL`, a preemptive defense inspired by subspace learning. `VESL` reduces attribute inference risk while preserving model utility, and as a byproduct of its dataset balancing mechanism, also improves fairness across sensitive attributes. Models trained with `VESL` are significantly more resilient: attacks perform no better than imputation, and the precision of `NeighVE` in identifying vulnerable records drops sharply.

❒ We introduce `AttriVET`, a defender-side tool for estimating record-level vulnerability to attribute inference attacks. `AttriVET` provides a systematic way to audit privacy risk in datasets and models by identifying which records are most susceptible, and we empirically show that it accurately estimates vulnerability across diverse attack strategies, models, and datasets.

2 Preliminaries

Notations

Table 1 lists the notations used throughout the paper for convenience.

Attribute Inference Attack

Let $\mathcal{M}$ represent an ML model trained on the dataset $\mathbb{D}$. The objective of an attribute inference attack is to predict $s(\mathbf{x})$, the sensitive attribute value of $\mathbf{x}$, a record/instance in $\mathbb{D}$. The adversary has query access to $\mathcal{M}$ and knowledge of the non-sensitive portion, $\mathcal{N}(\mathbf{x})$, of the record $\mathbf{x}$. Certain attack vari-

Symbol	Description
$\mathbb{D}$	The original training dataset
$\mathcal{S}$	Set of all possible sensitive attribute values
$\mathcal{M}$	Target model trained on $\mathbb{D}$
$Pr(\mathcal{M}(.))$	Confidence score output by $\mathcal{M}$
$s(.)$	Sensitive attribute value of a record
$\mathcal{N}(.)$	Non-sensitive attribute values of a record
$s_1, \ldots, s_l$	All possible sensitive attribute values
$\mathcal{V}_k(.)$	k-bounded Neighborhood of a record
$\mathcal{E}(.)$	Neighborhood similarity of a record
$\Gamma[.]$	Marginal distribution of sensitive attribute values
$\mathbb{D}_V$	Set of vulnerable records
$\mathbb{D}_N$	Set of non-vulnerable records
c	Correlation between sensitive attribute and output

Table 1: Notations used and their descriptions

ations require the adversary to have additional information, such as –

Auxiliary Data. D_{aux} is a set of records sampled from the same or a different distribution as $\mathbb{D}$. The adversary knows both $\mathcal{N}(\mathbf{x})$ and $s(\mathbf{x})$ for each record $\mathbf{x} \in D_{aux}$.

Marginal Distribution. $\Gamma[\cdot]$ denotes the fraction of records in $\mathbb{D}$ with each sensitive value, i.e., $\Gamma[s_i] = \frac{|\{\mathbf{x} \in \mathbb{D} \mid s(\mathbf{x}) = s_i\}|}{|\mathbb{D}|}$ for a sensitive attribute value s_i.

Confidence Score-based Model Inversion Attack (CSMIA)

In this attribute inference technique introduced in [36], an adversary aims to predict the sensitive value of record $\mathbf{x}$ with class label y by querying the model multiple times with $\mathbf{x}_i$ where $\mathcal{N}(\mathbf{x}) = \mathcal{N}(\mathbf{x}_i)$ and $s(\mathbf{x}_i) = s_i$, with s_i representing the i-th sensitive value. The model returns predictions $\hat{y}_i$ and confidence scores $conf_i = Pr[\mathcal{M}(\mathbf{x}_i) = \hat{y}_i]$ for $i \in [1, l]$. Based on these confidence scores and predictions, $\mathbf{x}$ is assigned to one of the three following cases –

Case-1 If only one $\hat{y}_i \in \{\hat{y}_1, \ldots, \hat{y}_l\}$ matches y, $\mathbf{x}$ is a case-1 instance. In this case, CSMIA predicts the corresponding s_i as the sensitive attribute.

Case-2 If multiple $\hat{y}_i \in \{\hat{y}_1, \ldots, \hat{y}_l\}$ match y, $\mathbf{x}$ is a case-2 instance. In this case, the one with the highest $conf_i$ among them is chosen and CSMIA predicts the corresponding s_i as the sensitive attribute.

Case-3 If none of the predictions $\hat{y}_i$ matches y, $\mathbf{x}$ is a case-3 instance. In this case, CSMIA outputs s_j as prediction where $j = \underset{j \in [1,l]}{argmin}\,(conf_j)$. The rationale is that the actual sensitive attribute value is likely to have the least confidence on the incorrect prediction.

Label Only Model Inversion Attack (LOMIA)

CSMIA requires the ability to query the target model and obtains the predicted confidence scores for all labels. In [36], LOMIA is introduced for the case when one gets only a predicted label from querying the target model. In LOMIA, one creates an attack dataset $\mathbb{D}_{\mathcal{A}}$ from all $\mathbf{x}$ from the training data that returned a true prediction for only one $\mathbf{x}_i$ (case-1 instances), adding $\mathbf{x}$ along with label y as input and s_i as output. Formally, $\mathbb{D}_{\mathcal{A}} = \{(\mathbf{x} \circ y, s_i) \mid \mathbf{x} \text{ is a case-1 instance}\}$ where $\circ$ denotes the concatenation operation. Subsequently, an attack model is trained on $\mathbb{D}_{\mathcal{A}}$ and used to infer the sensitive attribute value on the remaining records.

Imputation Attack

The adversary constructs an attack dataset similar to that used in LOMIA, but based on auxiliary set D_{aux}, and then trains an attack model to infer the sensitive attribute values of records. Depending on the distributional relationship between D_{aux} and the training set, we consider two variants of the imputation attack [30]: *ImpI*, the ideal imputation attack, assumes that D_{aux} follows the same distribution as the training set, whereas *ImpP*, the practical imputation attack, assumes a distributional shift between the two. Empirically, ImpP performs significantly worse than both ImpI and existing attribute inference attacks [30].

Angular Difference

To quantify group-level differences in model behavior that stem from correlations between sensitive attributes and outputs, the metric known as *angular difference* has been introduced in prior work [30]. Given a dataset $\mathbb{D}$ with n records and a trained model $\mathcal{M}$, a *confidence matrix* $\mathcal{C} \in \mathbb{R}^{n \times |\mathcal{S}|}$ is first constructed, where $\mathcal{S}$ denotes the set of all possible sensitive attribute values. For each record $x \in \mathbb{D}$, the model is queried with synthetic variants $\mathcal{T}(x)$ obtained by varying the sensitive attribute while keeping the non-sensitive attributes fixed. The resulting confidence matrix is defined as:

$$\mathcal{C} = \Big[\big[\Pr(\mathcal{M}(x')) : x' \in \mathcal{T}(x)\big]^T \quad \forall x \in \mathbb{D}\Big]^T$$

where $Pr(\mathcal{M}(x'))$ denotes the probability for the predicted class. Let $\mathbb{Y}$ be the set of output labels, and for each $y \in \mathbb{Y}$, let $\mathcal{C}_y$ denote the submatrix of $\mathcal{C}$ corresponding to records with predicted label y. A regression line L_y is then fitted through the $|\mathcal{S}|$-dimensional points in $\mathcal{C}_y$. The *angular difference* is defined as the average angle between all pairs of such regression lines:

$$\Delta = \frac{1}{\binom{|\mathbb{Y}|}{2}} \sum_{\substack{y_1, y_2 \in \mathbb{Y} \\ y_1 \neq y_2}} \angle(L_{y_1}, L_{y_2})$$

This metric captures the divergence in confidence score distributions across output classes when the sensitive attribute is perturbed. Larger angular differences indicate greater disparity in model behavior across groups and have been shown to correlate strongly with increased vulnerability to attribute

inference attacks. Importantly, angular difference can be computed without access to ground truth sensitive attributes, making it especially useful in black-box settings where such information is unavailable.

Mutual Information Regularization

Given a target model $\mathcal{M}$ with input distribution X and predicted output $\hat{Y} = \mathcal{M}(X)$, mutual information between X and $\hat{Y}$ is defined as $I(X;\hat{Y}) = H(\hat{Y}) - H(\hat{Y}|X)$, where $H(\cdot)$ denotes the entropy function. The goal of the mutual information regularization defense [47] is to minimize this mutual information and limit the adversary's ability to infer private information about X from $\hat{Y}$. During neural network training, the loss function $\mathcal{L}$ is modified to $\mathcal{L} = \mathcal{L}_{classification} + \beta \times I(X;\hat{Y})$ to incorporate this information and loss minimization is performed through gradient descent over all training records. The parameter β acts as a hyperparameter that balances both minimization objectives, achieving an optimal privacy-utility tradeoff. In neural networks, $I(X;\hat{Y})$ is approximated as $-I(Z;Y) + \lambda I(Z;X)$, where Z is the latent representation of X in the network. For attribute inference attacks, $I(Z;X)$ is replaced by $I(Z;s(X))$ to limit information leakage from the sensitive data portion $s(X)$.

3 Threat Model

We assume the adversary has the following capabilities:

❏ Access to the black-box target model through an ML-as-a-service (MLaaS) system, i.e., the adversary can query the model with $\mathbf{x}$ and obtain the output label y and the corresponding confidence scores.

❏ Full knowledge of the non-sensitive attributes of the target records.

❏ Knowledge of the possible values of the sensitive attribute and non-sensitive attributes.

The capabilities described above are standard assumptions in attribute inference attacks [21, 36, 50]. The adversary's goal is to predict the sensitive attribute for a set of target records. Our objective is to design a vulnerability estimation method that can predict whether a record in the training set will be correctly inferred by this adversary and then to create a vulnerability-informed defense to mitigate privacy leakage from such attacks. We also assume that the adversary cannot poison the training data, a requirement in some attribute inference attacks [46]. Although data poisoning is feasible in crowdsourced scenarios, training data is generally not crowdsourced in privacy-sensitive domains like healthcare and finance, making our assumption realistic. We further limit the adversary from having white-box access, as the MLaaS provider has no incentive to publicize model parameters.

4 Estimating Record-Level Vulnerability

4.1 Vulnerability Estimation from Adversary's Perspective

We argue that if a record is vulnerable to attribute inference, then the records in its immediate neighborhood in the training data are also likely to be vulnerable. We define a neighborhood as the set of records whose non-sensitive attributes differ from the target within a bounded distance in the attribute space. Prior work shows that vulnerability is not uniform but instead varies across groups [30,36]. Because neighbors share group membership across multiple attributes, they are subject to the same group-level risk; thus, when a record is vulnerable, its neighborhood is also more likely to contain other vulnerable records. Since angular difference has been effective for ranking vulnerable groups [30], we hypothesize that the angular difference of the neighborhood of a record similarly signals its vulnerability: higher n.a.d. indicates a higher probability of correct inference. To analyze this hypothesis rigorously, we first formalize the notions of neighborhood and neighborhood angular difference.

Definition 4.1 (k-bounded Neighborhood)**.** *Given a training dataset $\mathbb{D}$, the k-bounded Neighborhood for each record $x \in \mathbb{D}$ is defined as-*

$$\mathcal{V}_k(\mathbf{x}) = \{\mathbf{x}' \in \mathbb{D} \mid (\mathbf{d}(\mathcal{N}(\mathbf{x}), \mathcal{N}(\mathbf{x}')) < k) \ \& \ \mathbf{x} \textit{ and } \mathbf{x}' \textit{ share the same output label } \}$$

where d measures the distance between two records.

Definition 4.2 (Neighborhood Angular Difference (n.a.d.))**.** *Let $\mathcal{V}_k(\mathbf{x})$ denote the k-bounded neighborhood of the target record $\mathbf{x}$, and let $\mathcal{C}$ be the confidence matrix generated by querying the target model $\mathcal{M}$ with $\mathcal{V}_k(\mathbf{x})$. Let $\mathbb{Y}$ represent the set of output labels corresponding to records in $\mathcal{V}_k(\mathbf{x})$. Then, the Neighborhood Angular Difference (n.a.d.) is defined as the average angle between all pairs of lines L_{y_1} and L_{y_2}, where $y_1, y_2 \in \mathbb{Y}$. Here, L_y denotes the regression line fitted through the $|\mathcal{S}|$-dimensional points from $\mathcal{C}_y$, the submatrix of $\mathcal{C}$ containing rows corresponding to records with label y.*

Synthetic Neighborhood Generation. The definitions above formalize the concept of n.a.d., but applying this measure in practice requires determining the neighborhood of the target record. Since an adversary does not have access to the true neighborhood in the training dataset, we employ a synthetic neighborhood generation technique detailed in Algorithm 1 that randomly perturbs the target record to create its neighborhood. Because these neighbors are synthetically generated, they do not have ground-truth labels and must be assigned labels for n.a.d. computation. For each synthetic neighbor $\mathbf{x}'$, if any of its sensitive-attribute variants is predicted as the target record's label, then $\mathbf{x}'$ is assigned that label; otherwise, it is assigned the label predicted for the variant least likely

Algorithm 1 Synthetic Neighborhood Generation

Input: $\mathcal{M}$ (trained model), $\mathbf{x}$ (target record), k (neighborhood bound), m (number of candidates)
Output: $\mathcal{V}'_k(\mathbf{x})$ (labeled synthetic neighborhood of $\mathbf{x}$)

$\mathbb{P} \leftarrow \{\mathbf{x}' \mid \mathbf{x}'$ is generated by perturbing $\mathbf{x}$ $s.t.$ $d(\mathcal{N}(\mathbf{x}), \mathcal{N}(\mathbf{x}')) < k\}$
$\mathcal{V}'_k(\mathbf{x}) \leftarrow \varnothing$
for each $\mathbf{x}'$ in $\mathbb{P}$ **do**
  $\mathcal{X} \leftarrow \{\mathbf{x}'' \mid \mathcal{N}(x') = \mathcal{N}(x'') \wedge s(x') \in \mathcal{S}\}$
  if $\exists\, \mathbf{x}'' \in \mathcal{X}$ $s.t.$ $\mathcal{M}(x'') = y$ **then**
    $y' \leftarrow y$
  else
    $\mathbf{x}'' \leftarrow \min_{\mathbf{x}''} [Pr(\mathcal{M}(\mathbf{x}'')) = y]$
    $y' \leftarrow \mathcal{M}(\mathbf{x}'')$
  end if
  $\mathcal{V}'_k(\mathbf{x}) \leftarrow \mathcal{V}'_k(\mathbf{x}) \cup \{(\mathbf{x}', y')\}$
end for
return $\mathcal{V}'_k(\mathbf{x})$

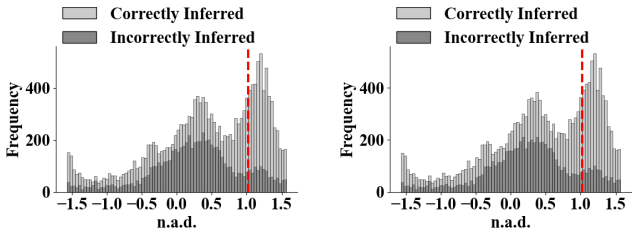


Figure 1: Distributions of n.a.d. values for correctly versus incorrectly inferred records when launching CSMIA (left) and LOMIA (right) on the Adult dataset.

to be classified as the target record's label. Finally, the n.a.d. of the target record is computed on this labeled synthetic neighborhood.

To examine the effectiveness of n.a.d. in identifying vulnerable records, we compute n.a.d. values for all the records from the Adult dataset using the synthetic neighborhood generation technique above and launch CSMIA and LOMIA to plot the n.a.d. distributions of correctly and incorrectly inferred records, as shown in Figure 1. These distributions differ noticeably, with the separation most apparent in the high-n.a.d. region. Specifically, above 1.02 (the 75th percentile of n.a.d. values), 83.45% of records under CSMIA and 83.54% under LOMIA are correctly inferred. This distinction demonstrates that n.a.d. values are effective in separating vulnerable records (i.e., those correctly inferred) from non-vulnerable records. Importantly, even though the neighborhoods used for this computation are synthetically generated, the results indicate that local neighborhood characteristics play a central role in attribute inference vulnerability.

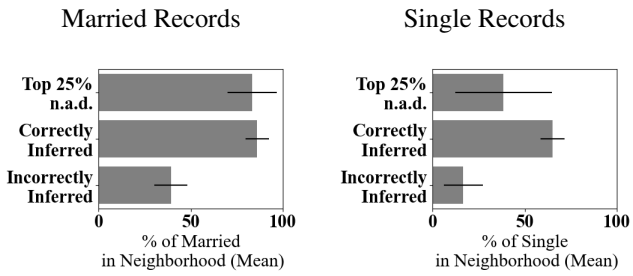


Figure 2: Mean proportion of married and single records in the k-bounded neighborhoods (k=5) of records grouped by top 25% n.a.d., correctly inferred, and incorrectly inferred by CSMIA in the Adult dataset.

4.2 Neighborhood and Attribute Inference Risk

Our analysis with synthetic neighborhoods revealed a deeper insight: a record's vulnerability to attribute inference is strongly influenced by the characteristics of its neighborhood, particularly the distribution of sensitive attributes. To examine this effect, we partition married and single records in the Adult dataset into three groups—top 25% by n.a.d., correctly inferred, and incorrectly inferred—and for each group compute the mean proportion of married and single records in their k-bounded neighborhoods ($k = 5$), with the results shown in Figure 2. The results demonstrate that vulnerable records—those in the top 25% n.a.d. group or correctly inferred—are surrounded by neighborhoods containing higher proportions of the same sensitive attribute, whereas incorrectly inferred records exhibit markedly weaker concentrations. Hence, the fraction of neighbors sharing the same sensitive attribute value, which we define as neighborhood similarity, serves as an indicator of a record's vulnerability level. Formally, neighborhood similarity is defined as:

Definition 4.3 (Neighborhood Similarity)**.** *Let $\mathbb{D}$ be a training dataset, $\mathbf{x} \in \mathbb{D}$ a record with sensitive attribute value $s(\mathbf{x})$, and $\mathcal{V}_k(\mathbf{x})$ the k-bounded neighborhood of $\mathbf{x}$. The neighborhood similarity of $\mathbf{x}$ is defined as*

$$\mathcal{E}[\mathbf{x}] = \frac{|\{\mathbf{x}' \in \mathcal{V}_k(\mathbf{x}) \mid s(\mathbf{x}') = s(\mathbf{x})\}|}{|\mathcal{V}_k(\mathbf{x})|}.$$

The finding that vulnerable records are surrounded by neighborhoods with disproportionately high concentrations of the same sensitive attribute is consistent with long-standing principles in machine learning. Regularization techniques such as L2 regularization [45], dropout [44], early stopping [40], and manifold regularization [10] are designed to smooth model behavior in the vicinity of input points, encouraging neighboring records to incur similar loss values. These principles suggest that the elevated vulnerability of records

Algorithm 2 VESL

Input: $\mathbb{D}$ (Training Dataset), I (Inference strategy)
Output: $\mathcal{M}$ (Trained Model)

1: $\mathbb{V}, \mathbb{N} \leftarrow$ AttriVET($\mathbb{D}$) ▷ $\mathbb{V}$ is the set of vulnerable records, $\mathbb{N}$ is the set of non-vulnerable records
2: **for** each t in $[1, T]$ **do**
3: **for** each i in $[1, n]$ **do**
4: $\mathbb{D}_i \leftarrow \varnothing$
5: **end for**
6: **for** each i in $[1, l]$ **do**
7: $\mathbb{V}^i \leftarrow \{x \in \mathbb{V} \mid s(x) = s_i\}$
8: $\mathbb{N}^i \leftarrow \{x \in \mathbb{N} \mid s(x) = s_i\}$
9: **if** $|\mathbb{V}^i| > |\mathbb{N}^i|$ **then**
10: Split $\mathbb{V}^i$ into n equal size non-overlapping subsets $\mathbb{D}^i_1, \ldots, \mathbb{D}^i_k$
11: $\mathbb{D}' \leftarrow \mathbb{N}^i$
12: **else**
13: Split $\mathbb{N}^i$ into n equal size non-overlapping subsets $\mathbb{D}^i_1, \ldots, \mathbb{D}^i_k$
14: $\mathbb{D}' \leftarrow \mathbb{V}^i$
15: **end if**
16: **for** each j in $[1, n]$ **do**
17: Generate the subset $\mathbb{D}'_j$ by sampling $|\mathbb{D}^i_j|$ elements from $\mathbb{D}'$
18: $\mathbb{D}_j \leftarrow \mathbb{D}_j \cup \mathbb{D}^i_j \cup \mathbb{D}'_j$
19: **end for**
20: **end for**
21: Initialize θ_0
22: **for** each i in $[1, n]$ **do**
23: $\theta_i \leftarrow Train(\theta_0, \mathbb{D}_i)$
24: **end for**
25: $\varepsilon_t \leftarrow \frac{1}{n} \sum_{i=1}^{n} \theta_i$
26: **end for**
27: **if** $I =$ majority-voting **then**
28: **Define** $\mathcal{M}$ so that for any input $\mathbf{x}$:
29: $\hat{y} \leftarrow \text{mode}\left(\{\varepsilon_t(\mathbf{x})\}_{t=1}^{T}\right)$ ▷ majority vote over ensembles predicted labels $\varepsilon_t(\mathbf{x})$
30: $V(\mathbf{x}) \leftarrow \{t \mid \varepsilon_t(\mathbf{x}) = \hat{y}\}$ ▷ indices of majority-voting ensembles
31: $Pr[\mathcal{M}(\mathbf{x}) = \hat{y}] \leftarrow \frac{1}{|V(\mathbf{x})|} \sum_{t \in V(\mathbf{x})} Pr[\varepsilon_t(\mathbf{x}) = \hat{y}]$ ▷ confidence = average of majority voters
32: **else if** $I =$ random-selection **then**
33: **Define** $\mathcal{M}$ *s.t.* $\mathcal{M}(\mathbf{x}) = \varepsilon_t(\mathbf{x})\ \forall \mathbf{x}$ where $t \sim Uniform([1, T])$
34: **end if**
35: **return** $\mathcal{M}$

with homogeneous neighborhoods is not incidental, but rather reflects the very mechanisms through which ML models are trained to generalize [25].

5 Defense Methodology

5.1 Vulnerability Estimation-driven Subspace Learning (VESL)

Existing defenses against attribute inference attacks, such as differential privacy [20] and mutual information regularization [47], apply coarse privacy mechanisms uniformly across all records. While effective at reducing attack performance, they often do so only at the cost of substantial model utility [18,22]. Because these defenses apply privacy mechanisms uniformly across all records without distinguishing between vulnerable and non-vulnerable ones, they impose privacy protection at a coarse granularity that results in greater utility loss. This gap motivates the need for defenses that explicitly take record-level vulnerability into account.

VESL addresses this challenge through two key insights. First, it is unrealistic to expect that all records can be made to exhibit the same neighborhood similarity, since neighborhoods overlap and adjusting one record's neighbors inevitably affects others. Instead, VESL applies a coarser balancing strategy: for each sensitive attribute value, we construct training subsets that contain equal proportions of vulnerable and non-vulnerable records. This balancing reduces the overall skew that drives attribute inference risk while avoiding large utility losses. Vulnerable records become less concentrated in extremely skewed neighborhoods, while non-vulnerable records are only minimally affected, leading to a more uniform distribution of vulnerability across the dataset.

Second, rather than training a single model on the entire dataset, VESL splits the training data into multiple subsets and trains multiple submodels on different balanced subsets. In subspace learning, each record is placed in only one subset, while its neighbors may be distributed across others. As a result, the neighborhood of a record seems more balanced, in the sense that across subspaces it reflects the global marginal distribution of the sensitive attribute rather than a consistent local skew.

Together, these two mechanisms—dataset balancing and subspace learning—systematically reduce attribute inference risk without the utility losses characteristic of existing defenses.

Design

Algorithm 2 describes the VESL defense. Line 1 performs the AttriVET algorithm (described in section 5.2) and finds the vulnerable and non-vulnerable subset from the training set. Lines 3-5 initialize the split subsets used for subset learning as empty with n denoting the number of split subsets. The loop from line 6 to line 20 loops over all the sensitive attribute values and populates the split subsets such that there is an equal number of samples from both the vulnerable subset and non-vulnerable subset with the specific sensitive attribute value achieving the desired balancing from our first key insight. Note that when balancing two unequal-sized sets, there are two options: either reduce the size of the larger set or increase the size of the smaller set. The first approach would result in a loss of training records potentially impacting model utility, so we opt for the latter by oversampling from the smaller set. Lines 21 to 23 train a submodel on each of the split subsets and aggregate them to produce an aggregate model corresponding to that particular data split. A total of T such aggregate models are trained using different random splits of the same training dataset. An ensemble model is then constructed from these aggregate models. VESL supports two inference strategies. In the random selection strategy, a query is answered by choosing one aggregate model uniformly at random. This

randomness provides additional privacy because a record may appear vulnerable in some aggregates but non-vulnerable in others, forcing the adversary into inconsistent outcomes. However, relying on a single aggregate can skew predictions and reduce utility. To mitigate this, the majority voting strategy determines predictions by majority vote across all aggregates and assigns confidence as the average of the majority voters, providing more stable utility. We refer to these two variants as `VESL-RS` (random selection) and `VESL-MV` (majority voting).

5.2 `AttriVET` Design

Building on our formulation of neighborhood similarity, we now present `AttriVET`, whose objective is to identify whether a record is vulnerable to attribute inference. Let Γ denote the marginal distribution of the sensitive attribute values in the training dataset, and let $\mathcal{E}[\mathbf{x}]$ denote the neighborhood similarity of $\mathbf{x}$. `AttriVET` classifies $\mathbf{x}$ as vulnerable or non-vulnerable according to:

$$\Delta(\mathbf{x}) = \begin{cases} \text{Vulnerable}, & \text{if } \mathcal{E}[\mathbf{x}] > \Gamma[s(\mathbf{x})], \\ \text{Non-vulnerable}, & \text{otherwise.} \end{cases}$$

Intuitively, a record is marked vulnerable if the proportion of neighbors sharing its sensitive attribute value exceeds its marginal probability in the dataset, indicating an elevated local concentration of that attribute.

While the described approach applies to categorical sensitive attributes, we follow Mehnaz et al. [36] for non-discrete attributes by binning value ranges and treating the set of bin means as categorical substitutes.

6 Experiments

In this section, we describe our experimental setup, datasets, machine learning models, and performance metrics. We then evaluate the performance of our proposed vulnerability estimation tools and defense.

6.1 Experimental Setup

Datasets. We use the following three datasets in our experiments: Census19 [11], Texas-100X [38], and Adult [9].

(1) Census-19. Derived from the 2019 US Census Bureau Database [11], the Census-19 dataset includes over 1.6 million records and 12 variables, capturing a wide range of personal and demographic details of US residents, such as age, gender, race, marital status, education level, occupation, working hours, country of origin, and various disability statuses. The dataset's objective is to classify individuals by annual income, with a threshold set at over $90,000. This threshold is adjusted from the Adult dataset's $50,000 benchmark to reflect inflation. In this dataset, marital status is chosen as the sensitive attribute. While it can take multiple values, we convert it to binary by labeling all non-married statuses as single. This grouping of marital status into married and single is done as part of initial preprocessing.

(2) Texas-100X. The Texas-100X dataset builds on the Texas-100 hospital dataset [38], originally introduced by Shokri et al. [42], and contains 925,128 records from 441 hospitals. This expanded dataset includes detailed demographic and medical information for each patient, such as age, gender, race, ethnicity, length of hospital stay, admission type and source, diagnosis at admission, discharge condition, total medical costs, and main surgical procedure. We use the PRINC_SURG_PROC_CODE column as the output attribute; however, it is a categorical column with 100 possible values. Since the other columns lack enough information to train a model effectively for 100-class classification, we reduce the 100 values into two categories: the top-10 most common surgeries and all other procedures, making it a binary classification problem.

(3) Adult. This dataset [9] is used to predict whether an individual earns over 50,000 a year. The dataset contains 48,842 instances and has 14 attributes. Following the preprocessing technique in [36] we merge the marital status attribute into two distinct clusters: Married, which includes 'Married-civ-spouse,' 'Married-spouse-absent,' and 'Married-AF-spouse'; and Single, which includes 'Divorced,' 'Never-married,' 'Separated,' and 'Widowed.' After removing records with missing values, the final dataset consists of 45,222 records. We split the dataset and use 35,222 records to train the target models, and the remaining 10,000 records to evaluate attacks on data from the same distribution but not in the training set.

Sampling Technique. The Census-19 and Texas-100X datasets used in our analysis each contain approximately a million records and we follow the same sampling technique as used in [30] to get 50,000 records from each. The details of this sampling technique is provided in Appendix section A. We set the correlation to -0.4 for Census-19 and 0.4 for Texas-100X, as these values enable both CSMIA and LOMIA to achieve strong attack accuracy [30] (around 70%). For the majority of experiments, we sample an equal number of records for each sensitive attribute value. In addition, we experiment with the Adult dataset without using controlled sampling to create an even more realistic setting.

Sensitive Attribute Selection. We select the `MAR` column and `marital` column, denoting marital status, as the sensitive attribute for the Census-19 and Adult dataset, respectively. For the Texas-100X dataset, we pick `SEXCODE` as the sensitive attribute to be inferred.

Model Training. In our experiments with the Texas-100X and Census-19 datasets, 50,000 records are used to create the training set, and another 50,000 records are randomly sampled as the test set, ensuring mutual exclusivity. For the majority of experiments, we use Scikit-learn's [39] Multilayer Perceptron (MLP) implementation as the default model.

The neural network has three hidden layers with 32, 16, and 8 neurons, respectively, and ReLU activation. A 4-layer MLP adds a 64-neuron layer near the input, while a 2-layer MLP removes the 32-neuron layer. The output layer is a softmax with one neuron per output class. The Adam optimizer is used with a 0.001 learning rate, and training runs for 500 iterations.

Hyperparameters. AttriVET requires a distance threshold, k, to define a record's neighborhood, which we set at 5. In section 16, we perform an ablation study to assess the impact of varying k on AttriVET's performance. For distance calculation, we use the following method: individual attribute distances are summed, with categorical attribute mismatches contributing a distance of 1. For numerical attributes, the attribute's standard deviation is calculated, and the numerical difference between two records is normalized by dividing by this standard deviation; the resulting value represents the distance for that attribute. For NeighVE, we generate synthetic neighbors by randomly selecting up to k non-sensitive categorical features and replacing each with a value sampled uniformly from that feature's categories. In our experiments, we set the number of subspaces in VESL to 5, corresponding to the number of subsets constructed during each split. We use 5 random splits of the training dataset, resulting in an ensemble of 5 aggregate models. In our evaluation of NeighVE, we use 2000 neighbors and a distance threshold of 4.

Metrics. We evaluate our approach along three dimensions: attack performance, model utility, and fairness. Attack performance is measured using accuracy, precision, recall, and F1-score, with precision, recall, and F1 averaged across all sensitive attribute values. Model utility is assessed by test accuracy on the target task. Fairness is evaluated using two standard measures: equalized odds difference (EOD) [14] and demographic parity difference (DPD) [14].

6.2 Performance of AttriVET

Because VESL depends on AttriVET to partition records into vulnerable and non-vulnerable sets, we begin by validating AttriVET before turning to the evaluation of VESL. We evaluate the effectiveness of AttriVET by comparing it against the correctness of predictions made by the attacks which we refer to as *prediction validity*. For a target record $\mathbf{x}$, the prediction validity under attack $\mathcal{A}$ is the boolean value of $s(\mathbf{x}) = s_{\mathcal{A}}$, where $s_{\mathcal{A}}$ is the sensitive attribute value predicted by the adversary. We then use accuracy, precision, recall, and F1-score to measure how closely the vulnerability predicted by AttriVET matches prediction validity. Since AttriVET is the first vulnerability predictor specifically addressing attribute inference attacks, we have no baseline for direct comparison.

The performance of AttriVET in predicting vulnerability against CSMIA and LOMIA is shown in Table 2. The vulnerable status determined by AttriVET aligns well with prediction validity for the sensitive attribute in all datasets. Under the LOMIA attack, performance is essentially near-perfect, with almost no misclassified records. Although CSMIA produces slightly lower scores than LOMIA, AttriVET continues to perform at a level that is both robust and practically meaningful for detecting record-level vulnerability. Importantly, AttriVET's high performance is consistent across Census19, Adult, and Texas100, demonstrating that the tool generalizes well to datasets with different characteristics. The confusion matrices further reinforce this observation: the number of false negatives and false positives remains extremely small compared to the large volume of correct predictions. This shows that AttriVET does not only excel at detecting vulnerable records, but is equally effective at recognizing non-vulnerable ones, making it a reliable mechanism for comprehensive record-level vulnerability auditing.

The effectiveness of AttriVET is rooted in its reliance on neighborhood similarity. By directly measuring how closely a record's neighborhood reflects its own sensitive attribute, AttriVET captures the very local concentration effects that make some records more exposed to attribute inference. The near-perfect accuracy under LOMIA and consistently strong results under CSMIA suggest that neighborhood similarity is not just a correlating factor but a primary contributor to record-level vulnerability. In other words, records that AttriVET identifies as vulnerable are precisely those where neighborhood composition amplifies inference risk, underscoring neighborhood similarity as the fundamental driver behind attribute inference vulnerability.

6.3 Effectiveness of VESL

We launched CSMIA and LOMIA on models trained by both variants of VESL to evaluate the effectiveness of our proposed defense. Because VESL-RS selects an ensemble model at random for each inference and may exhibit performance variation, we report results averaged over five runs. In addition to the no-defense baseline (i.e., models trained without any privacy-preserving techniques), we implement a modified version of mutual information regularization [47], called Selected Mutual Information Regularization (SMIR), to serve as a comparative baseline. SMIR calculates information loss exclusively on vulnerable records rather than across all records. Further details on this baseline are provided in Appendix section B.

An effective defense should meet the objectives of mitigating attack performance, maintaining model utility, and preserving fairness, all of which VESL achieves, as demonstrated in the following paragraphs.

Mitigating Attack Performance

Tables 3 and 4 report the accuracy, precision, recall and F1 scores of CSMIA and LOMIA attacks on models trained with VESL-MV and VESL-RS, as well as the no-defense and SMIR baselines. An effective defense should reduce the per-

Dataset	CSMIA								LOMIA							
	Accuracy	Precision	Recall	F1-score	TP	TN	FP	FN	Accuracy	Precision	Recall	F1-score	TP	TN	FP	FN
Census-19	87.93	88.80	93.64	91.16	31101	12865	3922	2112	98.54	98.80	99.11	98.96	34603	14668	420	309
Adult	97.43	97.76	98.58	98.17	24296	10020	557	349	97.85	98.51	98.45	98.48	24483	9983	370	386
Texas-100X	94.28	94.93	96.83	95.87	33226	13912	1774	1088	99.30	99.82	99.19	99.50	34936	14716	64	284

Table 2: Accuracy, Precision, Recall, F1-score, and confusion matrix of AttriVET in identifying vulnerable records when the adversary launches CSMIA and LOMIA on models trained with Census-19, Texas-100X, and Adult datasets.

Defense	Census-19				Adult				Texas-100X			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
No Defense	66.42	66.58	66.42	66.34	69.96	75.64	68.95	67.48	68.63	68.68	68.63	68.61
Random	50.06	50.06	50.06	50.06	49.96	49.96	49.96	49.94	49.73	49.73	49.73	49.73
Imputation	65.84	65.85	65.84	65.84	66.28	74.91	65.02	61.99	69.03	69.03	69.03	69.03
SMIR	50.09	50.09	50.09	50.09	49.86	49.86	49.86	49.83	49.98	49.98	49.98	49.98
VESL-MV	37.03	36.99	37.03	36.98	62.76	65.16	61.82	60.15	55.58	55.67	55.58	55.41
VESL-RS	46.12	42.61	46.12	38.87	61.19	61.64	61.47	61.11	53.12	58.01	53.12	44.67

Table 3: Performance of CSMIA on the Census-19, Texas-100X, and Adult datasets under both variants of VESL, a baseline model with no defense, and the SMIR defense. Imputation and random attack performance is also added for comparison.

formance of attribute inference attacks, ideally bringing it down to or below the performance of baseline attacks that do not rely on access to the target model—such as the practical imputation attack (imputation) [28,30], which exploits statistical correlations between non-sensitive and sensitive attributes, or the random attack, which predicts sensitive attributes uniformly at random. The performance of these baseline attacks reflects the level of inference possible without access to the model. Thus, if a defense reduces the attack performance to that of imputation or the random attack, it implies that the release of the model does not contribute to additional privacy leakage. Accordingly, we also report the performance of imputation and random attacks for comparison.

The results show that both variants of VESL consistently reduce the performance of attribute inference attacks below that of the imputation baseline across all evaluation settings. In particular, VESL-RS drives attack performance close to, or even below, that of a random attack in most cases, underscoring its effectiveness in mitigating privacy leakage. While the reduction achieved by VESL-MV is less pronounced than that of VESL-RS, it still consistently lowers attack performance beneath the imputation baseline—an attack strategy available to an adversary without model access. This demonstrates that models trained with VESL-MV do not leak more information than what an adversary could already obtain through imputation. Importantly, defenders can adopt VESL-RS when near-zero privacy leakage is required.

Although the SMIR defense can reduce attack performance to nearly that of a random attack, this privacy gain comes at the expense of substantial utility loss, as shown in the next section, limiting its practicality in real-world applications. This indicates that, although both SMIR and VESL leverage AttriVET to identify vulnerable records, VESL is more effective at utilizing the insights from vulnerability estimation to defend against inference attacks.

Can the adversary benefit by flipping predictions when accuracy falls below random? In some cases, most notably Census-19, the attack performance on models trained with VESL falls below that of a random attack. Given the binary sensitive attribute used in this evaluation, one might expect the adversary to improve performance by simply flipping predictions, thereby exceeding random attack performance. However, even under this adjustment, attack performance remains below the imputation baseline. In the worst case, observed for VESL-MV on Census-19 under LOMIA, the accuracy, precision, recall, and F1-score are 64.89, 64.92, 64.89, and 64.97, respectively—all lower than the corresponding imputation attack. Thus, the adversary gains no advantage beyond what could already be achieved without model access. Therefore, our proposed defense, VESL, effectively mitigates the risk of privacy leakage resulting from model release.

Maintaining Model Utility

Table 5 reports the test accuracy of models trained using both variants of our proposed defense VESL, the SMIR baseline, and the no-defense baseline on the Census-19, Adult, and Texas-100X datasets. The results indicate that both variants of VESL incurs only a slight reduction in test accuracy compared to the no-defense baseline in all scenarios. This minor utility loss is a reasonable tradeoff for the privacy protection against attribute inference attacks provided by VESL. While VESL-RS sacrifices more utility, VESL-MV offers a stronger balance, enabling defenders to preserve high task accuracy while reducing privacy leakage. In contrast, SMIR experiences substantial utility degradation, with test accuracy dropping to around 50% in all cases except the Adult dataset—effectively reducing the model to random guessing. Other existing defenses, such as Differential Privacy [1], similarly suffer from

Defense	Census-19				Adult				Texas-100X			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
No Defense	69.82	69.83	69.82	69.82	70.61	75.42	69.67	68.49	70.44	70.44	70.44	70.44
SMIR	49.59	49.58	49.59	49.51	49.84	49.84	49.84	49.82	50.14	50.14	50.14	50.14
Random	49.93	49.93	49.93	49.93	50.10	50.09	50.09	50.07	49.77	49.77	49.77	49.77
Imputation	65.84	65.85	65.84	65.84	66.28	74.91	65.02	61.99	69.03	69.03	69.03	69.03
VESL-MV	35.11	35.08	35.11	35.08	65.76	66.28	65.27	65.01	55.33	55.35	55.33	55.28
VESL-RS	36.34	36.30	36.34	36.29	62.89	63.69	62.24	61.59	52.42	52.42	52.42	52.41

Table 4: Performance of LOMIA on the Census-19, Texas-100X, and Adult datasets under both variants of VESL, a baseline model with no defense, and the SMIR defense. Imputation and random attack performance is also added for comparison.

Defense	Census-19	Adult	Texas-100X
No Defense	79.37	84.48	80.48
SMIR	50.29	75.65	50.30
VESL-MV	76.35	79.13	76.08
VESL-RS	75.60	77.59	75.08

Table 5: Test accuracy of both versions of our proposed defense (VESL), SMIR, and the no defense baseline on the Census-19, Texas-100X, and Adult datasets.

Defense	Census-19		Adult		Texas-100X	
	EOD	DPD	EOD	DPD	EOD	DPD
No Defense	0.310	0.484	0.307	0.338	0.379	0.562
VESL-MV	0.034	0.206	0.068	0.193	0.038	0.234
VESL-RS	0.029	0.208	0.092	0.190	0.029	0.218

Table 6: Fairness impact of both variants of VESL on the Census-19, Texas-100X, and Adult datasets. Results are shown in terms of EOD and DPD.

significant performance loss when applied in this context [47]. Overall, VESL offers the most favorable utility-privacy trade-off among all evaluated defenses against attribute inference attacks.

Preserving Fairness

Because VESL involves strategic sampling during preprocessing, it is important to assess whether this process introduces any fairness concerns. Table 6 reports the EOD and DPD for the proposed defense on the Census-19 and Texas-100X datasets. In all cases, both versions of VESL reduce EOD and DPD, indicating improved fairness after applying the defense. This improvement appears to result from the dataset balancing mechanism, suggesting that fairness arises as a byproduct of the privacy protection measure.

6.4 Evaluation of NeighVE

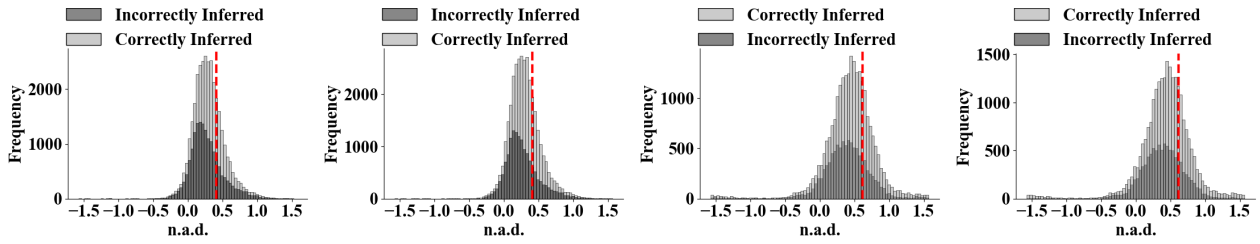
In Section 4.1, we observed clear separation in the n.a.d. distributions of correctly and incorrectly inferred records when launching CSMIA and LOMIA on Adult. To assess whether NeighVE generalizes, we repeated the experiment on Census-19 and plotted the corresponding n.a.d. distributions in Figures 3(a–b). In this case, the differences between the distributions are much smaller for both attacks. Whereas roughly 84% of records in the top 25% of n.a.d. values were correctly inferred in Adult, the proportion drops to 69.95% under CSMIA and 73.98% under LOMIA in Census-19. The overall attack accuracies in Census-19 are 66.42% (CSMIA) and 69.82% (LOMIA). Thus, NeighVE fails to identify subsets with attack accuracy substantially exceeding the global average, indicating that its effectiveness is not consistent across datasets. We attribute this limitation to the use of synthetically generated neighbors, which may be of lower quality in certain datasets. In future work, we plan to explore improved techniques for record-level vulnerability estimation from the adversary's perspective to better capture such risks.

VESL can prevent NeighVE from identifying correctly inferred records. While VESL lowers overall attack accuracy, it remains important to ask whether an adversary using NeighVE might still gain an advantage by focusing on records with high n.a.d. values. To investigate this, we plot n.a.d. histograms for correctly and incorrectly inferred records on Adult models trained with VESL-MV, shown in Figures 3(c–d). The resulting distributions are nearly indistinguishable, with no region where correctly inferred records dominate. In the top 25% of records ranked by n.a.d., only 74.34% and 74.90% of records are correctly inferred for CSMIA and LOMIA respectively, compared to roughly 84% without defense. This indistinguishability is also observed for VESL-RS, indicating that both variants are robust against targeted attacks. These findings demonstrate that NeighVE no longer isolates subsets with elevated attack performance under VESL, confirming the defense's ability to mitigate targeted inference.

6.5 Ablation Study

Effect of Model Architecture on AttriVET

Although AttriVET is fundamentally dataset-driven, an important question is whether its effectiveness depends on the model architecture, since adversarial success rates can vary across model types. To examine this, we evaluate AttriVET on Census-19 models trained with three architectures: MLP, XGBoost, and Random Forest, with results shown in Table 7. Overall, AttriVET achieves strong performance across all



(a) Attack = CSMIA (b) Attack = LOMIA (c) Attack = CSMIA (d) Attack = LOMIA

Figure 3: Histogram of n.a.d. on model trained on Census-19 dataset with no defense (a-b) and on Adult dataset with VESL-MV (c-d). The distribution for correctly and incorrectly inferred records by CSMIA (a,c) and LOMIA (b,d) are plotted separately. The red dotted lines indicate the 75th percentile threshold for respective scenarios.

	CSMIA				LOMIA			
Model	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
MLP	87.93	88.80	93.65	91.16	98.54	98.80	99.11	98.96
XGB	97.63	98.35	98.27	98.31	99.43	99.55	99.63	99.59
RF	77.60	95.73	77.55	85.69	95.74	97.84	96.15	96.99

Table 7: Accuracy, Precision, Recall, F1-score, and confusion matrix of AttriVET in identifying vulnerable records when the adversary launches CSMIA and LOMIA on different architecture models trained on Census-19.

Dataset →	Census-19			Texas-100X		
Scenario →	S-1	S-2	S-3	S-1	S-2	S-3
CSMIA	98.91	87.93	89.94	99.27	94.28	93.45
LOMIA	99.86	98.54	98.71	99.98	99.30	99.87

Table 8: Accuracy of AttriVET across models trained on Census-19 and Texas-100X subsets with varying distribution of sensitive attribute values. S-1, S-2, and S-3 contains 25%, 50%, and 75% single records for Census-19 and male records for Texas-100X respectively.

settings, with accuracy, precision, and F1-scores consistently high. XGBoost yields the strongest results, with near-perfect performance under both CSMIA and LOMIA. MLP also performs very well, particularly under LOMIA, though recall dips slightly under CSMIA (93.65%). Random Forest exhibits lower overall accuracy (77.60% under CSMIA) and reduced recall (77.55%), however, even in this weaker case, AttriVET maintains strong F1-scores of 85.69 and 96.99 under CSMIA and LOMIA, respectively, underscoring its robustness. Taken together, these results suggest that model architecture has only a limited effect on record-level vulnerability estimation. Instead, dataset-specific factors, particularly neighborhood similarity, play a much larger role in determining vulnerability.

Effect of Distributional Skew on AttriVET

Thus far, we have evaluated AttriVET under balanced scenarios where the training datasets contain equal numbers of records for each sensitive attribute value. A natural question is whether skewed distributions affect its ability to estimate vulnerability. To study this, we vary the distribution of sensitive attributes in Census-19 and Texas-100X, constructing three subsets for each with 25%, 50%, and 75% of records having marital status *single* for Census-19 and gender *male* for Texas-100X (S-1, S-2, S-3 in Table 8). Across all scenarios, AttriVET maintains consistently high accuracy under both CSMIA and LOMIA, with results nearly identical to those obtained in the balanced case. Even under severe skew, accuracies remain above 87.9% for CSMIA and above 98.7% for LOMIA in Census-19, and near-perfect across all scenarios in Texas-100X. These results demonstrate that distributional skew has little to no effect on AttriVET's capability, reinforcing the conclusion that record-level vulnerability estimation is governed primarily by neighborhood similarity rather than distribution of sensitive attribute.

Causality Between Neighborhood Similarity and Inference Success

To test whether neighborhood similarity causally influences prediction validity, we performed a counterfactual intervention. We selected 100 random records from the Census-19 dataset. Further, we altered the local neighborhood (within a radius of 5) for each to contain an equal mix of Married and Single records. Each modified dataset was used to train a new model, resulting in 100 models. Using CSMIA and LOMIA to infer the sensitive attribute for each target record, we observed success in 51 and 49 cases, respectively—consistent with random guessing. By reducing neighborhood similarity to match the relative frequency, we removed the attacker's advantage. These findings support the claim that neighborhood similarity has a causal effect on a record's vulnerability.

Effectiveness of VESL on Non-Binary Sensitive Attributes

To evaluate the performance of VESL on non-binary sensitive attributes, we considered race (4 classes: White, Black, Asian, Others) as the sensitive attribute on Census-19. We

Defense	CSMIA		LOMIA	
	Accuracy	F1-score	Accuracy	F1-score
No Defense	51.25	35.94	71.61	40.85
VESL-MV	46.04	33.89	40.81	23.54
VESL-RS	41.66	32.57	40.39	22.05
Imputation	67.61	46.96	67.61	46.96

Table 9: Accuracy and F1-score of VESL in identifying vulnerable records under CSMIA and LOMIA on Census-19 when race (4 classes) is used as a non-binary sensitive attribute, comparing no defense, VESL variants. Imputation attack performance is also added for comparison.

trained models with no defense and both variants of VESL, and launched both CSMIA and LOMIA, as reported in Table 9. Both versions of VESL reduce LOMIA's accuracy from 71.61% to about 40% on this 4-class race attribute. In this setting, AttriVET identifies vulnerable records with 84.41% accuracy. Even before applying VESL, CSMIA underperforms compared to imputation on race; applying VESL further diminishes its effectiveness. Overall, the results demonstrate that VESL remains effective in mitigating attribute inference attacks on non-binary sensitive attributes.

Effectiveness of VESL on Other Existing Attribute Inference Attacks

A defense should not be tuned only to a few specific types of attribute inference attack. Our goal in designing VESL is to mitigate the underlying leakage, rather than CSMIA's/LOMIA's specific heuristic. We demonstrate that vulnerability is driven by neighborhood similarity—a dataset-level property that is independent of a particular attack—and that AttriVET, which utilizes only this quantity, accurately predicts vulnerable records across models. VESL reshapes neighborhood of records and reduces sensitivity of output to the sensitive attribute. We chose CSMIA/LOMIA [36] in our evaluation because prior work reports that they outperform earlier methods: attacks that rely on auxiliary datasets [28] often achieve accuracy close to or below imputation when the auxiliary and target distributions differ [30] and other methods based on simpler heuristics [21,50] are subpar sometime underperforming than random baseline [36]. Nevertheless, we evaluate the WCAI attack [50] on Census-19. Under no defense, WCAI achieves 64.35% accuracy and 63.50% F-1 score, while imputation yields comparable performance with 64.05% accuracy and 63.07% F-1 score. This marginal improvement over imputation illustrates WCAI's limited inference capability in this setting. Applying VESL-MV and VESL-RS yields similar WCAI performance, with accuracies of 64.04% and 64.00% respectively. Despite not using auxiliary data, AttriVET predicts vulnerable records with 72.92% accuracy.

Evidence that VESL Learns from the Sensitive Attribute

We trained models on all three datasets without the sensitive attribute and observed similar performance to models trained with VESL variants (see Appendix C.3). This raises the question of whether VESL reduces leakage by suppressing use of the sensitive attribute altogether. To validate this, we computed SHAP-based [35] global feature-importance rankings for all three scenarios. In Census-19 containing 13 attributes, MAR remains the fourth most important both with and without defense. In Adult (10 features), marital shifts from first to third but remains highly influential. In Texas-100X (10 features), SEXCODE moves from second to roughly seventh (VESL-RS) or eighth (VESL-MV), yet continues playing a meaningful predictive role. This confirms that models still learn from the sensitive attribute, sometimes to a reduced extent that ensures downstream leakage is low. Since removing sensitive attributes is not always feasible in practice, VESL allows models to retain this information while reducing privacy leakage. The detailed results from SHAP-based analysis is also presented in Appendix C.3.

Fairness Analysis on Vulnerable Records

We further evaluate whether VESL preserves utility on vulnerable records. Across Census-19, Adult, and Texas-100X, accuracy on vulnerable records closely tracks overall accuracy: on Census-19 it is 76.12% versus 76.78% overall, on Adult 79.31% versus 79.79%, and on Texas-100X 77.69% versus 76.87%. This indicates that applying VESL does not degrade predictive performance on vulnerable records.

7 Related Works

Notable Works in Attribute Inference

In [22], Fredrikson et al. introduced model inversion attacks for linear regression models, later expanding these attacks to non-linear ML models in [21]. The latter study defined two categories of model inversion attacks: attribute inference, in which the adversary tries to uncover sensitive attributes in the dataset, and class representative reconstruction, aimed at recreating instances similar to those in the training data. Gong et al. and Jia et al. [23, 24, 29] applied attribute inference attacks in social media contexts, where adversaries infer private attributes from public information. These attacks rely on users publicly disclosing private attributes, making them applicable only when private-public attribute pairs can be collected to match the target distribution. Mehnaz et al. [36] introduce CSMIA and LOMIA, with LOMIA the first to show that attribute inference attacks can be launched without requiring confidence scores, demonstrating the limitations of score-masking defenses. Jayaraman et al. [28] propose a white-box attribute inference attack that identifies the hidden neurons most correlated with the sensitive attribute, using their output

for predictions. Kabir et al. [30] demonstrate that targeted attribute inference attacks can achieve significantly higher accuracy by leveraging angular difference, a group-level metric that serves as an effective indicator of collective vulnerability. Motivated by this, we incorporate a modified version of angular difference in our proposed tool NeighVE, enabling the adversary to identify subsets with a high concentration of vulnerable records.

Vulnerability Estimation Works in Inference Attacks

Existing research highlights various techniques for assessing privacy risks, focusing predominantly on membership inference attacks , which attempt to identify whether specific data samples were part of a model's training dataset. Kumar and Shokri [33] provides a tool for measuring privacy risk by simulating membership inference attacks, which helps organizations evaluate their data's vulnerability. However, this approach can be misleading, as it may not lead to effective defenses in all attack scenarios due to the model owner's lack of knowledge about the exact attack strategy and hyperparameters. Bai et al. [7] offers a method based on clustering to identify which data points are at higher risk, enabling proactive privacy assessments before data is shared. However, this method applies only to class representative reconstruction attacks and cannot be extended to attribute inference attacks on individual records. Song and Mittal [43] introduces a "privacy risk score" to capture individual sample vulnerabilities against membership inference attacks, offering a detailed approach to risk assessment by capturing sample-specific vulnerabilities. Yeom et al. [50] investigate how the influence, defined as the extent to which changes in a sensitive attribute affect predictions, impacts vulnerability to attribute inference attacks. Their analysis reveals that while attacker advantage grows with initial increases in influence, it decreases as the influence becomes more significant. This analysis, however, is applicable at the dataset level rather than the individual record level.

Defenses against Attribute Inference Attacks

Wang et al. [47] propose mutual information regularization as a defense against model inversion attacks, including attribute inference attacks. However, this approach results in a substantial decrease in model utility, rendering it an unacceptable tradeoff. Dibbo et al. [18] investigate fairness constraints as a defense against attribute inference attacks, but find them to be ineffective.

8 Limitation and Future Work

Our proposed defense, VESL, provides the strongest privacy protection against attribute inference attacks while incurring the least utility loss among existing defenses. However, in many ML applications, even minimal utility loss is undesirable. As future work, we plan to explore advanced techniques such as fine-tuning [52] and knowledge diffusion [26] to eliminate utility degradation entirely. A key limitation of the adversary-side vulnerable records identification tool NeighVE is its inconsistent effectiveness across datasets, as the quality of synthetically generated neighbors can vary; future work will explore more advanced techniques for constructing neighborhoods to achieve stable performance. Our proposed defender-side vulnerability estimation tool, AttriVET, is designed for centralized machine learning settings, where a single party has access to the entire dataset. In decentralized frameworks such as federated learning [32], each client only has access to its local data, making it infeasible to compute record-level vulnerability using AttriVET. This limitation also poses challenges for applying VESL in federated settings. As part of future work, we aim to extend the core ideas behind AttriVET and VESL to develop effective vulnerability estimation and privacy defense mechanisms tailored for federated learning environments.

9 Conclusion

We present a novel framework to address the growing privacy risks associated with attribute inference attacks in machine learning models. Motivated by the need for vulnerability-aware defenses, we first developed NeighVE, an adversary-side estimator that uses synthetic neighborhoods and angular difference to estimate record-level vulnerability without requiring sensitive attribute labels. Although its performance is not consistent across all datasets, NeighVE revealed the central role of neighborhood similarity as a key dataset-level factor underlying record-level vulnerability. Building on this insight, we proposed VESL, a subspace-learning–inspired defense that balances vulnerable and non-vulnerable records across sensitive attributes. VESL reduces attribute inference risk to the level of imputation attacks—ensuring no more leakage than an adversary could achieve without model access—while preserving model utility and improving fairness as a byproduct of its balancing mechanism. Moreover, VESL nullifies NeighVE, preventing adversaries from isolating vulnerable records for targeted attacks. Finally, we introduced AttriVET, a defender-side tool that predicts record-level vulnerability with over 90% accuracy across diverse scenarios, enabling systematic auditing and risk-aware defense design. Together, these contributions establish neighborhood similarity as the primary driver of record-level vulnerability and provide both auditing and defense mechanisms that advance the state of privacy-preserving machine learning.

Acknowledgments

This work was supported in part by NSF grant 2442825. We thank the anonymous reviewers for their feedback and suggestions.

Ethical Considerations

We consider the ethical considerations of this work through a stakeholder-based analysis, explicitly considering the Menlo Report principles [8] in the context of both the research process and publication.

Stakeholders. The primary stakeholders impacted by this work include individuals whose records appear in the public datasets used for evaluation, machine learning practitioners and organizations that train and deploy models on tabular data, and the broader research community studying privacy risks and defenses. The datasets used in this work are publicly available and de-identified; as a result, no individuals are directly affected by the research procedures, and no personally identifiable information is accessed or processed. Practitioners and deploying organizations can use the proposed methods to assess and reduce record-level privacy leakage prior to model deployment. While the paper also introduces an adversary-side vulnerability estimation tool to characterize attack feasibility, its potential misuse is mitigated by the accompanying defense, which is shown to nullify the effectiveness of such attacks. The research community is impacted through the introduction of methods for estimating record-level vulnerability from attribute inference attacks and reducing such risk, which may shape future works in record-level vulnerability estimation and vulnerability-guided defense design.

Impact. Before proceeding with this research, we considered the potential benefits and harms of both the research process and publication across all identified stakeholders, following the ethical principles articulated in the Menlo Report [8]—Beneficence, Respect for Persons, Justice, and Respect for Law and Public Interest. For individuals represented in the datasets, the principle of Respect for Persons is satisfied by the exclusive use of publicly available, de-identified data, ensuring that no privacy expectations, consent requirements, or individual rights are implicated. For practitioners and deploying organizations, Beneficence applies through the provision of methods that enable assessment and reduction of record-level privacy leakage prior to deployment. For the research community and society at large, Beneficence is reflected in improving understanding of attribute inference risks and effective defenses, supporting safer use of machine learning in sensitive domains. Consideration of the principles of Justice and Respect for Law and Public Interest did not reveal additional concerns, as the work does not differentially impact any group and does not involve unlawful data access or violations of terms of service. We identify no tangible harms arising from the research or its publication, as the work does not involve human subjects, private data, or exposure to harmful content. Potential indirect harms related to adversarial misuse are limited and addressed through the defensive mechanisms evaluated in this work.

Mitigations. The potential harms relevant to this work primarily concern privacy leakage through attribute inference attacks. These risks are directly addressed by the core contributions of the paper, which are explicitly designed to identify, quantify, and mitigate such leakage. The research process itself introduces no additional harms, as all experiments are conducted on publicly available, de-identified datasets and do not involve human subjects, private data, or interaction with individuals. While the paper includes an adversary-side vulnerability estimation tool for analysis purposes, any potential misuse is mitigated by the accompanying defense, which is shown to substantially reduce or nullify the effectiveness of such adversarial capabilities. As a result, the work does not introduce unmitigated harms beyond those already present in existing machine learning deployments and instead reduces the overall risk of privacy leakage.

Decision. The decision to proceed with this research and to publish its results was reached by considering both potential harms and benefits, as well as the avoidance of any violation of individual rights. From a Beneficence perspective, the work provides concrete methods for identifying and mitigating attribute inference risks, thereby reducing privacy leakage in machine learning systems and supporting safer deployment in sensitive domains. From a Respect for Persons perspective, the research does not implicate individual rights, as it relies exclusively on publicly available, de-identified datasets and does not involve human subjects, consent, or expectations of privacy. These analyses lead to the same conclusion: conducting and publishing the research introduces minimal risk while offering clear privacy-preserving benefits. Given this, we determined that proceeding with and disseminating the work is ethically appropriate.

Open Science

We fully support the principles of open science as outlined in the USENIX Security 2025 Open Science Policy. To promote the principles of reproducibility and transparency, we provide the full codebase including all scripts, binaries and datasets required to replicate our experiments at https://zenodo.org/records/17905133. Because the Texas-100X dataset is subject to licensing restrictions, the attached artifact does not contain the original dataset but detailed instructions for obtaining and using it.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang.

Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[2] Farhad Ahamed and Farnaz Farid. Applying internet of things and machine-learning for personalized healthcare: Issues and challenges. In *2018 International Conference on Machine Learning and Data Engineering (iCMLDE)*, pages 19–21. IEEE, 2018.

[3] Tiago A. O. Alves, Felipe M. G. França, and Sandip Kundu. Mlprivacyguard: Defeating confidence information based model inversion attacks on machine learning systems. In *Proceedings of the 2019 on Great Lakes Symposium on VLSI*, GLSVLSI '19, page 411–415, New York, NY, USA, 2019. Association for Computing Machinery.

[4] Shengwei An, Guanhong Tao, Qiuling Xu, Yingqi Liu, Guangyu Shen, Yuan Yao, Jingwei Xu, and Xiangyu Zhang. Mirror: Model inversion for deep learning network with high fidelity. In *Proceedings of the 29th Network and Distributed System Security Symposium*, 2022.

[5] Omer Artun and Dominique Levin. *Predictive marketing: Easy ways every marketer can use customer analytics and big data*. John Wiley & Sons, 2015.

[6] Ulrich Aïvodji, Sébastien Gambs, and Timon Ther. Gamin: An adversarial approach to black-box model inversion. *arXiv*, 2019.

[7] Yang Bai, Mingyu Fan, Yu Li, and Chuangmin Xie. Privacy risk assessment of training data in machine learning. In *ICC 2022 - IEEE International Conference on Communications*, pages 1015–1015, 2022.

[8] Michael Bailey, David Dittrich, Erin Kenneally, and Doug Maughan. The menlo report. *IEEE Security & Privacy*, 10(2):71–75, 2012.

[9] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

[10] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(11), 2006.

[11] US Census Bureau. Pums data— census.gov. https://www.census.gov/programs-surveys/acs/microdata/access.2019.html#list-tab-735824205, 2019. [Accessed 29-05-2024].

[12] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914. IEEE, 2022.

[13] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.

[14] Jaewoong Cho, Gyeongjo Hwang, and Changho Suh. A fair classifier using kernel density estimation. *Advances in neural information processing systems*, 33:15088–15099, 2020.

[15] Ping Chou, Howard Hao-Chun Chuang, Yen-Chun Chou, and Ting-Peng Liang. Predictive analytics for customer repurchase: Interdisciplinary integration of buy till you die modeling and machine learning. *European Journal of Operational Research*, 296(2):635–651, 2022.

[16] Robert Culkin and Sanjiv R Das. Machine learning in finance: the case of deep learning for option pricing. *Journal of Investment Management*, 15(4):92–100, 2017.

[17] Sayanton V Dibbo, Adam Breuer, Juston Moore, and Michael Teti. Improving robustness to model inversion attacks via sparse coding architectures. In *European Conference on Computer Vision*, pages 117–136. Springer, 2025.

[18] Sayanton V Dibbo, Dae Lim Chung, and Shagufta Mehnaz. Model inversion attack with least information and an in-depth analysis of its disparate vulnerability. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 119–135. IEEE, 2023.

[19] Matthew F Dixon, Igor Halperin, and Paul Bilokon. *Machine learning in finance*, volume 1170. Springer, 2020.

[20] Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006.

[21] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015.

[22] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *23rd USENIX Security*

Symposium (USENIX Security 14), pages 17–32, San Diego, CA, August 2014. USENIX Association.

[23] Neil Zhenqiang Gong and Bin Liu. You are who you know and how you behave: Attribute inference attacks via users' social friends and behaviors. In *25th USENIX Security Symposium (USENIX Security 16)*, pages 979–995, Austin, TX, August 2016. USENIX Association.

[24] Neil Zhenqiang Gong and Bin Liu. Attribute inference attacks in online social networks. *ACM Trans. Priv. Secur.*, 21(1), January 2018.

[25] Ian Goodfellow. Deep learning, 2016.

[26] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[27] Samyak Jain, Sravanti Addepalli, Pawan Kumar Sahu, Priyam Dey, and R Venkatesh Babu. Dart: Diversify-aggregate-repeat training improves generalization of neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16048–16059, 2023.

[28] Bargav Jayaraman and David Evans. Are attribute inference attacks just imputation? *arXiv preprint arXiv:2209.01292*, 2022.

[29] Jinyuan Jia, Binghui Wang, Le Zhang, and Neil Zhenqiang Gong. Attriinfer: Inferring user attributes in online social networks using markov random fields. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1561–1569, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee.

[30] Ehsanul Kabir, Lucas Craig, and Shagufta Mehnaz. Disparate privacy vulnerability: Targeted attribute inference attacks and defenses. In *34th USENIX Security Symposium (USENIX Security 25)*, 2025.

[31] Balaram Yadav Kasula. Harnessing machine learning for personalized patient care. *Transactions on Latest Trends in Artificial Intelligence*, 4(4), 2023.

[32] Jakub Konečnỳ. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[33] Sasi Kumar and Reza Shokri. Ml privacy meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning. In *Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs)*, 2020.

[34] Jiacheng Li, Ninghui Li, and Bruno Ribeiro. Membership inference attacks and defenses in classification models. In *Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy*, CODASPY '21, page 5–16, New York, NY, USA, 2021. Association for Computing Machinery.

[35] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[36] Shagufta Mehnaz, Sayanton V Dibbo, Roberta De Viti, Ehsanul Kabir, Björn B Brandenburg, Stefan Mangard, Ninghui Li, Elisa Bertino, Michael Backes, Emiliano De Cristofaro, et al. Are your sensitive attributes private? novel model inversion attribute inference attacks on classification models. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 4579–4596, 2022.

[37] Ngoc-Bao Nguyen, Keshigeyan Chandrasegaran, Milad Abdollahzadeh, and Ngai-Man Cheung. Re-thinking model inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16384–16393, 2023.

[38] Texas Department of State Health Services. Texas Inpatient Public Use Data File (PUDF) | Texas DSHS — dshs.texas.gov. https://www.dshs.texas.gov/texas-health-care-information-collection/health-data-researcher-information/texas-inpatient-public-use, 2006. [Accessed 29-05-2024].

[39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[40] Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 2002.

[41] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*. The Internet Society, 2019.

[42] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

[43] Liwei Song and Prateek Mittal. Systematic evaluation of privacy risks of machine learning models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2615–2632, 2021.

[44] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[45] Andrey Nikolayevich Tikhonov. Solutions of ill-posed problems. *VH Winston and Sons*, 1977.

[46] Florian Tramèr, Reza Shokri, Ayrton San Joaquin, Hoang Le, Matthew Jagielski, Sanghyun Hong, and Nicholas Carlini. Truth serum: Poisoning machine learning models to reveal their secrets. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 2779–2792, 2022.

[47] Tianhao Wang, Yuheng Zhang, and Ruoxi Jia. Improving robustness to model inversion attacks via mutual information regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11666–11673, 2021.

[48] Mitchell Wortsman, Maxwell C Horton, Carlos Guestrin, Ali Farhadi, and Mohammad Rastegari. Learning neural network subspaces. In *International Conference on Machine Learning*, pages 11217–11227. PMLR, 2021.

[49] Ziqi Yang, Jiyi Zhang, Ee-Chien Chang, and Zhenkai Liang. Neural network inversion in adversarial setting via background knowledge alignment. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 225–240, 2019.

[50] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE, 2018.

[51] Hongxu Yin, Arun Mallya, Arash Vahdat, Jose M Alvarez, Jan Kautz, and Pavlo Molchanov. See through gradients: Image batch recovery via gradinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16337–16346, 2021.

[52] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

[53] Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 253–261, 2020.

A Sampling Technique

1. Let n represent the desired number of samples for the training data. Additionally, let m denote the ratio of samples with negative sensitive values to samples with positive sensitive values, making it the inverse of the marginal prior. Let c represent the desired correlation between the sensitive attribute and the output.

2. Let n_+^+ represent the number of samples with positive sensitive values and positive output values. Let n_+^- represent the number of samples with positive sensitive values and negative output values. Similarly, let n_-^+ and n_-^- represent the number of samples with negative sensitive values and positive and negative output values, respectively. Here, the subscript represents the sensitive value and the superscript represents the output value. $n_-^+ = \lfloor \frac{\sqrt{m}\times(\sqrt{m}-c)\times n}{2\times(m+1)} \rfloor$, $n_-^- = \lfloor \frac{\sqrt{m}\times(\sqrt{m}+c)\times n}{2\times(m+1)} \rfloor$ $n_+^- = \lceil \frac{n}{2} - n_-^- \rceil$, $n_+^+ = \lceil \frac{n}{2} - n_-^+ \rceil$

A.1 Correctness Proof of Sampling Technique

Let, s and y denote sensitive attribute and output respectively and n denote total number of records. According to definition,

$$c = \frac{n\sum sy - (\sum s)(\sum y)}{\sqrt{(n\sum s^2 - n(\sum s)^2)\,(n\sum y^2 - n(\sum y)^2)}} \tag{1}$$

Now, $\sum y$ essentially means the sum of output values of n records. Since, the output can either take 0 or 1 and takes 1 a total of $n_+^+ + n_-^+$ times, $\sum y = n_+^+ + n_-^+$ and $\sum y^2 = n_+^+ \times 1^2 + n_-^+ \times 1^2 = n_+^+ + n_-^+$. Similarly, $\sum s^2 = \sum s = n_+^+ + n_+^-$. Substituting these values in equation 1 and performing a few algebraic operations we get,

$$c = \frac{n_+^+ \times n_-^- - n_+^- \times n_-^+}{\sqrt{(n_+^+ + n_+^-)(n_+^+ + n_-^+)(n_-^+ + n_-^-)(n_+^- + n_-^-)}} \tag{2}$$

According to our requirement, the number of positive samples $(n_+^+ + n_-^+)$ is equal to the number of negative samples $(n_+^- + n_-^-)$ and the ratio between positive and negative samples $(\frac{n_+^+ + n_+^-}{n_-^+ + n_-^-})$ is m. If we combine $n_+^+ + n_+^- + n_-^+ + n_-^- = n$ with the above we get,

$$n_+^+ + n_-^+ = n_+^- + n_-^- = \frac{n}{2}$$
$$n_+^+ + n_+^- = \frac{n}{m+1}, \quad n_-^+ + n_-^- = \frac{mn}{m+1}$$

Dataset →	Census-19			Texas-100X		
MLP Depth →	4	3	2	4	3	2
CSMIA	82.27	93.04	92.60	77.74	93.89	99.31
LOMIA	96.44	98.68	99.76	97.12	99.61	99.96

Table 10: Vulnerability prediction accuracy of AttriVET on the Census-19 dataset for 4-layer, 3-layer, and 2-layer MLP architectures with CSMIA and LOMIA.

Dataset →	Census-19			Texas-100X		
MLP Depth →	4	3	2	4	3	2
CSMIA	63.99	67.10	63.43	61.36	66.05	68.71
LOMIA	69.46	69.00	68.82	69.68	68.99	68.85

Table 11: Attack accuracy of CSMIA and LOMIA on the Census-19 dataset for 4-layer, 3-layer, and 2-layer MLP architectures.

Substituting all these in equation 2 we get,

$$n_+^+ \times n_-^- - n_+^- \times n_-^+ = c\sqrt{m} \times \frac{n}{m+1} \times \frac{n}{2} \tag{3}$$

Substituting n_+^+ with $\frac{n}{2} - n_-^+$ and n_+^- with $\frac{n}{2} - n_-^-$ we get,

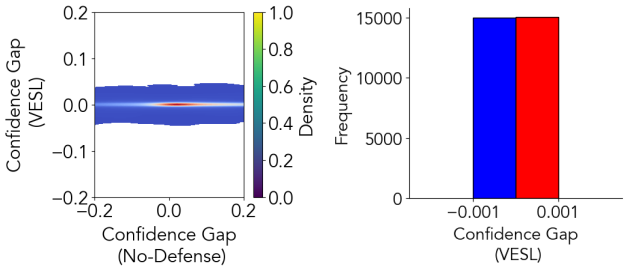
$$n_-^- - n_-^+ = c\sqrt{m} \times \frac{n}{m+1} \tag{4}$$

Now we can solve to get the values of n_-^- and n_-^+ first, and using those to get the values of n_+^+ and n_+^- as shown in section 6.1. Note that, these values are integers which is why the fractions are rounded up using floor and ceiling. Therefore, the sampled records may not have the exact correlation as c but the difference would be negligible for our purpose.

This sampling approach ensures that the desired correlation and marginal prior are achieved, while maintaining an equal number of positive and negative output samples.

B Selective Mutual Information Regularization Details

With the added capability of computing vulnerability level of individual records, we introduce a key modification to the mutual information regularization defense [47]: mutual information loss is calculated only for records within the vulnerable subset, minimizing mutual information between the sensitive attribute and output for these records alone. We refer to this modified defense approach as selective mutual information regularization (SMIR). In our experiments, we set the loss balancing hyperparameter β to 0.5, as lower values render SMIR ineffective in reducing attack performance across dataset scenarios.



(a) (b)

Figure 4: (a) Contour plots of confidence gaps from models trained with no defense and with VESL-RS on the Census-19 dataset (marginal prior 0.5). (b) Histogram showing the distribution of records with confidence gaps in the range $[-0.001, 0.001]$ for the VESL-RS model. Red (blue) bars correspond to records correctly (incorrectly) inferred by CSMIA.

C Ablation Study of AttriVET

C.1 Impact of Model Architecture on AttriVET

Since AttriVET is computed on the training dataset without model training, it is important to investigate whether model architecture influences AttriVET's vulnerability estimation. To examine this, we use the Census-19 dataset and train 4-layer, 3-layer, and 2-layer MLPs (detailed architecture in section 6.1). This setup represents a scenario with semi-skewed sensitive attribute values. After training, we launch both CSMIA and LOMIA on each model. AttriVET then computes the vulnerability of each record, and we evaluate its alignment with prediction validity from each attack model pair. Table 10 presents AttriVET's vulnerability prediction accuracy for these scenarios. The results indicate that AttriVET's vulnerability prediction accuracy is high in all cases except with the 4-layer model under CSMIA, where there is a slight decrease. Table 11 presents the attack performance of CSMIA and LOMIA across all models, showing that CSMIA's accuracy is also low in the 4-layer MLP case. This observation reinforces our earlier point: when attack performance is poor, vulnerability prediction accuracy also drops slightly but still high enough to not cause any concern. Notably, LOMIA's performance in the 4-layer MLP case remains stable. In this scenario, the added depth causing overfitting and reduced smoothness in a record's vicinal area, reducing the correlation between confidence gap and neighborhood similarity.

C.2 How Does VESL Mitigate Attribute Inference Attacks?

Let, the *confidence gap* of a record be the difference between the model's confidence in predicting its true label and the

Defense	Census-19	Adult	Texas-100X
No Defense	79.37	84.48	80.48
WOUT-SENS	78.47	81.97	76.56
VESL-MV	76.35	79.13	76.08
VESL-RS	75.35	78.06	75.02

Table 12: Test accuracy of both variants of VESL compared to the no-defense baseline and a model trained without the sensitive attribute (WOUT-SENS).

highest confidence it assigns to any version of that record with a different sensitive attribute value. CSMIA typically infers records correctly when the confidence gap is positive, incorrectly when it is negative, and randomly when it is zero. To counteract this, the dataset rebalancing and model aggregation steps in VESL are designed to drive the confidence gap toward zero across the dataset. During training, each data split assigns each record to only one subset. The confidence gap for a given record under any aggregate model from a split is, on average, lower in magnitude than under the no-defense baseline. However, this reduction alone is insufficient to ensure that the performance of attribute inference attacks approaches that of a random guess. In the final stage of VESL, an ensemble model is constructed that either randomly selects one of the aggregate models during inference or performs majority voting among the aggregate models. Because a record's neighbors differ across splits, the confidence gap values it receives from each aggregate model vary—some being positive, others negative. This symmetry results in an equal probability of a record having a positive or negative confidence gap in the ensemble. Consequently, the expected number of correctly inferred records by CSMIA matches that of incorrectly inferred ones, and similarly for LOMIA, effectively neutralizing the attack advantage.

To illustrate this, we compare confidence gap distributions from models trained with and without VESL in Figure 4(a). The contour plot shows a high density of points along the x-axis, indicating that many records trained under VESL have confidence gaps near zero. Indeed, approximately 60% of the training set have a confidence gap magnitude below 0.001. Figure 4(b) plots the histogram of these records' confidence gaps, revealing nearly equal numbers of positive and negative values. This near-symmetry implies that VESL makes it equally likely for records to be correctly or incorrectly inferred. Since LOMIA relies on case-1 instances from CSMIA to construct its attack dataset, the diminished quality of CSMIA predictions under VESL leads to weaker LOMIA performance as well.

C.3 Evidence that VESL Learns from the Sensitive Attribute

Here, we present detailed results from evaluating models trained without the sensitive attribute, as well as SHAP-based feature-importance analyses with and without defense. Table 12 reports test accuracy when the sensitive attribute is removed from training, compared against the no-defense baseline and both variants of VESL. Table 13 reports SHAP-based global feature-importance rankings for Census-19, Table 14 reports the corresponding results for Adult, and Table 15 reports results for Texas-100X under the target model and VESL-MV.

	Target Model		VESL-MV	
Feature	Mean	Std	Mean	Std
SCHL	0.095813	0.000256	0.121295	0.000272
AGEP	0.056643	0.001207	0.044278	0.001181
WKHP	0.043789	0.001162	0.046429	0.001463
ST	0.022865	0.000186	0.026032	0.000182
SEX	0.019333	0.000697	0.038655	0.000667
COW	0.016425	0.000347	0.010927	0.000301
MAR	0.016144	0.000757	0.029084	0.000679
RAC1P	0.013247	0.000574	0.022751	0.000647
WAOB	0.005496	0.000197	0.002948	0.000200
DREM	0.002399	0.000248	0.005531	0.000258
DEAR	0.000775	0.000225	0.000869	0.000201
DPHY	0.000613	0.000183	0.002368	0.000225
DEYE	0.000471	0.000120	0.001053	0.000154

Table 13: SHAP feature importance for Census-19 under the target model without defense and VESL-MV.

	Target Model		VESL-MV	
Feature	Mean	Std	Mean	Std
marital	0.043387	0.002063	0.043387	0.002063
capitalgain	0.037287	0.001735	0.037287	0.001735
sex	0.027870	0.001638	0.027870	0.001638
education	0.021647	0.001474	0.021647	0.001474
occupation	0.017387	0.000881	0.017387	0.000881
work	0.014383	0.000828	0.014383	0.000828
capitalloss	0.009247	0.001243	0.009247	0.001243
hoursperweek	0.008097	0.001892	0.008097	0.001892
race	0.002253	0.000604	0.002253	0.000604
fnlwgt	0.001633	0.001634	0.001633	0.001634

Table 14: SHAP feature importance for Adult under the target model without defense and VESL-MV.

C.4 Effect of Distance Threshold on Vulnerability Prediction

In this experiment, we vary the hyperparameter k, which represents the neighborhood radius in AttriVET, from 2 to 5 and measure its effect on AttriVET's vulnerability prediction accuracy to determine the ideal value of k. Table 16 shows the vulnerability prediction accuracy for these scenarios when CSMIA and LOMIA are launched on a model trained on Census-19 dataset. The results indicate that higher k values yield higher vulnerability prediction accuracy, though with a diminishing rate of increase. For instance, the accuracy increase from $k = 2$ to $k = 3$ is more pronounced than from

Feature	Target Model		VESL-MV	
	Mean	Std	Mean	Std
ADMITTING_DIAGNOSIS	0.221641	0.000264	0.276771	0.000260
SEX_CODE	0.044995	0.001057	0.005612	0.000613
PAT_STATUS	0.014371	0.000151	0.012123	0.000158
SOURCE_OF_ADMISSION	0.009835	0.000264	0.013557	0.000312
TYPE_OF_ADMISSION	0.008265	0.000310	0.012828	0.000367
ETHNICITY	0.007750	0.000501	0.006581	0.000592
RACE	0.004299	0.000512	0.000653	0.000567

Table 15: SHAP feature importance for Texas-100X (0.5) under the target model without defense and VESL-MV

Distance threshold, k	2	3	4	5
CSMIA	86.97	89.23	89.86	89.94
LOMIA	93.47	97.28	98.44	98.71

Table 16: Vulnerability prediction accuracy of AttriVET with varying distance threshold on the Census-19 dataset with CSMIA and LOMIA.

$k = 4$ to $k = 5$. However, increased prediction accuracy is beneficial in constructing a more robust defense, as it allows the defender to identify the set of vulnerable records more accurately. Therefore, we select $k = 5$ as the standard distance threshold across the other experiments in the literature.