USENIX Security '26 Artifact Appendix: 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

A Artifact Appendix

A.1 Abstract

Attribute inference attacks pose a serious privacy risk to machine learning (ML) models by enabling adversaries to infer sensitive attributes of individual records from model behavior. This paper investigates whether privacy vulnerability from attribute inference can be estimated at the level of individual records and used to better understand and mitigate attribute inference attacks. From an adversary's perspective, we introduce NeighVE, a vulnerability estimation method that quantifies record-level susceptibility by computing neighborhood angular difference over synthetically generated neighbors. From the defender's perspective, we propose AttriVET, a vulnerability estimation technique that leverages neighborhood similarity to identify records that are vulnerable to attribute inference. Building on these insights, we introduce VESL, a defense mechanism that uses vulnerability-aware subspace learning to mitigate attribute inference attacks while preserving model utility. For artifact evaluation, our goals are to ensure reproducibility of the experiments, validate the implementation of the proposed vulnerability estimation methods and defenses, and enable further exploration of attribute inference risks in ML systems. The provided artifacts include datasets, a modular codebase, and detailed Jupyter notebooks that reproduce the paper's key findings on the neighborhood-driven nature of record-level privacy vulnerability, the use of vulnerability estimation from adversarial and defensive perspectives, and the effectiveness of VESL in mitigating attribute inference attacks. These artifacts aim to support open science practices, ensuring accessibility, transparency, and reusability for the research community.

A.2 Description & Requirements

This section outlines the experimental setup, including security considerations, access instructions, software dependencies, hardware requirements, and benchmark datasets for reproducibility.

A.2.1 Security, privacy, and ethical concerns

The datasets used in the attacks are anonymized and publicly available; therefore, there is no direct security or privacy risk arising from the release or execution of this artifact. As discussed in the paper's ethical considerations, the techniques implemented in this artifact—particularly vulnerability estimation and attribute inference attacks—could be misused if applied irresponsibly. However, the artifact is released strictly for research, auditing, and defensive purposes, and does not introduce new attack capabilities beyond those already established in prior work. The artifact adheres to the FAIR principles: it is findable via a public repository, accessible without restrictive licensing, interoperable through standard data formats and widely used ML libraries, and reusable through comprehensive documentation and clear modular design. We encourage evaluators and future users to follow responsible research practices and applicable ethical guidelines when using this artifact in privacy-sensitive contexts.

A.2.2 How to access

The artifact is available at `https://doi.org/10.5281/zenodo.17905132`. Please refer to 'Version v2' of the artifact for the most up-to-date version.

A.2.3 Hardware dependencies

None

A.2.4 Software dependencies

The artifact requires a Linux, macOS, or Windows operating system with Python 3.9.16 or later. A virtual environment can be set up using Conda or Python's `venv`. All necessary dependencies are listed in `requirements.txt` and can be installed using `pip install -r requirements.txt`. Running Jupyter notebooks requires installing `notebook` via `pip`

`install notebook`. No proprietary software is required, and all dependencies are open-source and can be installed using standard package managers.

A.2.5 Benchmarks

The experiments in this artifact rely on the following datasets: `census19.csv`, which contains the Census-19 dataset, and `Adult_35222.csv` and `Adult_10000.csv`, which contain the Adult dataset partitions for training and testing, respectively. Additionally, the Texas-100X dataset is required. To ensure reproducibility, we have attached a preprocessed version of the dataset with the filename `texas_100_cleaned.csv`.

A.3 Set-up

A.3.1 Installation

To install and set up the artifact, first create a virtual environment using Conda or Python's `venv` and activate it. Then, install all required dependencies using `pip install -r requirements.txt`. The specific versions of the listed packages that were used in implementation are included in `requirements_w_version.txt` for reproducibility. To install those exact versions, run: `pip install -r requirements_w_version.txt`.
To run Jupyter notebooks, install `notebook` via `pip install notebook`.

A.3.2 Basic Test

To verify the setup, open and run all cells in the Jupyter notebook `basic_test.ipynb`. This notebook loads and preprocesses the Adult dataset for different experiment scenarios. A successful run will produce the first five rows of the training data.

A.4 Evaluation workflow

This section outlines the steps required to evaluate the artifact, validate its functionality, and reproduce the key results presented in the paper. It consists of two parts: Major Claims, which enumerate the core findings supported by the experiments, and Experiments, which describe the operational steps needed to test these claims. Each experiment is linked to a corresponding claim and includes details on execution, estimated compute time, and expected outcomes.

A.4.1 Major Claims

(C1): AttriVET can accurately identify vulnerable records across all three datasets under CSMIA and LOMIA. This is proven by the experiment (E1) described in section 6.2 of the main paper whose results are presented in Table 2.

(C2): VESL can mitigate attribute inference attacks across all three datasets and reduce it below that of imputation attack performance while incurring a small loss in utility and preserving fairness. This is proven by the experiment (E2) described in section 6.3 of the main paper whose results are presented in Tables 3 to 6. Tables 3 and 4 reports attack performance with defense under CSMIA and LOMIA respectively. Table 5 reports the accuracy of the model on test set and Table 6 reports the fairness metrics Equalized Odds Difference (EOD) and Demographic Parity Difference (DPD) across various senstive attribute values.

(C3): NeighVE can identify subset of training records with much higher attack performance than global average for Adult dataset but not for Census-19 dataset. Furthermore, when NeighVE is applied on a VESL-trained model, it becomes ineffective in identifying a more vulnerable attack subset. This is proven by the experiment (E3) described in section 4.1 and section 6.4 of the paper whose results are presented in Figures 1 to 3.

A.4.2 Experiments

(E1): [1 compute-hour]: This experiment evaluates the performance of the proposed individual record-level vulnerability estimation tool, AttriVET, on the Census19, Adult, and Texas-100X datasets under attribute inference attacks. The results are expected to demonstrate high accuracy, precision, recall, and F1-score in identifying vulnerable records.
Preparation: No additional preparation required.
Execution: Run all cells in the Jupyter notebook `attriVET_evaluation.ipynb` or run the python script from terminal: `python attriVET.py`
Results: The last two cells of the notebook and the python script are expected to generate results similar to Table 2.

(E2): [5 compute-hour]: This experiment evaluates the performance of the proposed defense against attribute inference attacks, VESL, on the Census19, Adult, and Texas-100X datasets. Performance is measured across three categories—attack performance reduction, model accuracy on the test set, and fairness across sensitive attribute values. In terms of attack performance reduction, the results are expected to show that both CSMIA and LOMIA achieve lower performance than the imputation attack in the presence of both VESL variants. The results are also expected to show that VESL-trained models achieve test accuracy comparable to models trained without any defense. Finally, the results are expected to show no degradation in the fairness metrics EOD and DPD.
Preparation: No additional preparation required.
Execution: Run all cells in the Jupyter notebook `vesl_evaluation.ipynb` or run the python script from

terminal: `python vesl.py`

Results: The python script is expected to generate results similar to Tables 3 to 6. The notebook cell 8 is expected to generate attack performance results similar to Tables 3 and 4. Cell 9 is expected to generate model test accuracy results similar to Table 5. Cell 10 is expected to generate fairness evaluation results similar to Table 6.

(E3): [4 compute-hour]: This experiment evaluates the performance of the adversar-side vulnerability estimation tool, NeighVE, on the Census19 and Adult datasets on models trained with no defense and models trained with VESL. The performance is reported as the proportion of correctly inferred records in the subset of records with top 25% n.a.d. values. The results are expected to show a high proportion of correctly inferred records for Adult dataset but not for Census-19 dataset. The results are also expected to show that when n.a.d. values are computed on a VESL-trained model, the proportion of correctly inferred records for Adult dataset is reduced to the global average.

Preparation: No additional preparation required.

Execution: Run all cells in the Jupyter notebook `neighve_evaluation.ipynb` or run the python script from terminal: `python neighve.py`

Results: The python script is supposed to generate plots similar to Figures 1 to 3 and save them to the folder `Figures`. Notebook cells 8 and 9 are expected to generate n.a.d. histogram plots similar to Figure 1. Cell 10 is expected to generate barplots similar to Figure 2. Cells 11 and 12 are expected to generate n.a.d. histogram plots using CSMIA and LOMIA respectively with no defense models on Census-19 dataset similar to Figures 3(a) and 3(b). Cells 13 and 14 are expected to generate n.a.d. histogram plots using CSMIA and LOMIA with VESL-trained models on Adult dataset similar to Figures 3(c) and 3(d).

A.5 Notes on Reusability

In future work, our implemented codebase can serve as a foundation for exploring additional aspects of attribute inference attacks and for developing and evaluating new defense strategies. The artifact README includes a detailed mapping of the codebase, outlining the role and functionality of each file and directory.

A.6 Version

Based on the LaTeX template for Artifact Evaluation V20231005. Submission, reviewing and badging methodology followed for the evaluation of this artifact can be found at https://secartifacts.github.io/usenixsec2026/.