MCTI Project Final Phase Report

Yu Leng, University Of Guelph
Student ID: 0939686

ABSTRACT

Privacy and data security have emerged as critical challenges in large language models (LLMs), particularly in the face of membership inference attacks (MIAs). In these attacks, adversaries attempt to determine whether a specific data sample was part of a model's training set, potentially exposing sensitive information. While techniques such as Differential Privacy (DP) and its implementation via DP-SGD have been proposed to counter MIAs, these methods often rely on noise injection, which can degrade model performance. This trade-off between privacy and utility remains a major concern.

RelaxLoss has been proposed as a mitigation strategy to reduce the vulnerability to membership inference attacks (MIA) while maintaining competitive predictive performance. This project systematically assesses the efficacy of RelaxLoss across a broad spectrum of datasets and application domains. Beyond conventional benchmark datasets such as CIFAR-10, Purchase, and Texas, our evaluation encompasses the MIMIC-IV clinical dataset, as well as structurally heterogeneous real-world datasets including Covertype, Adult (binary classification), and STL-10 (image classification). These datasets present distinct challenges due to characteristics such as high dimensionality, feature sparsity, and severe class imbalance, all of which complicate model generalization. Although not all datasets entail high privacy risk, their diverse feature structures offer critical perspectives on the architectural robustness and adaptability of RelaxLoss. For STL-10, we additionally implement a custom black-box attack pipeline to assess the defense efficacy of RelaxLoss in a vision-based context. Collectively, these experiments provide a more holistic evaluation of RelaxLoss's generalization performance and resilience in privacy-relevant and structurally complex scenarios.

Dataset Access Statement: This study was conducted using the MIMIC-IV dataset, accessed with credentialed approval through PhysioNet. The following files were used for data extraction and analysis: `labevents.csv.gz` (lab test results), `prescriptions.csv.gz` (medication prescriptions), `chartevents.csv.gz` (vital signs and observations), `icustays.csv.gz` (ICU admissions), `inputevents.csv.gz`, `outputevents.csv.gz` (fluid balance), and `procedureevents.csv.gz` (clinical procedures).

1 INTRODUCTION

In general, privacy and data security concerns in large language models can often be traced back to optimization strategies employed during the early stages of model training. To improve the model's accuracy and precision, a variety of loss minimization techniques are applied from the outset. Depending on the model type, such as linear regression, binary classification, or multi-class classification, several widely used loss functions are adopted, including Mean Squared Error (MSE), Mean Absolute Error (MAE), Cross-Entropy, Hinge Loss, and Huber Loss. While these loss functions differ in their mathematical formulations, they all serve the common goal of improving predictive performance by reducing error rates.

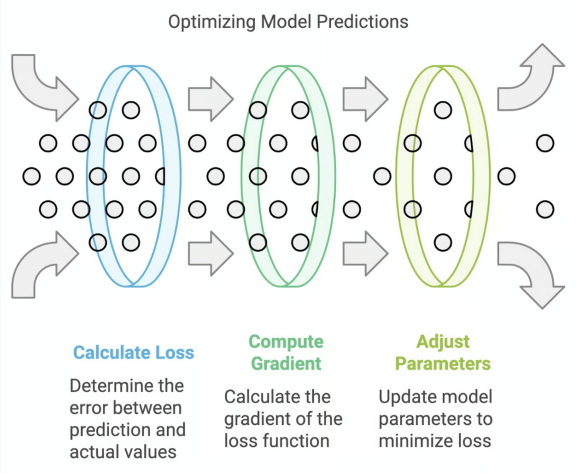

Figure 1: Optimizing Model Predictions [2]

However, such aggressive optimization can unintentionally introduce privacy vulnerabilities. As the model becomes increasingly confident and tightly fit to the training data, the distinction between training (member) and non-training (non-member) samples becomes more pronounced. This heightened separability can be exploited by Membership Inference Attacks (MIAs), where adversaries attempt to infer whether a specific sample was part of the training dataset. In essence, the very mechanisms that boost model performance may also amplify privacy risks, highlighting a fundamental tension between utility and security in model design.

To obscure the characteristics of training samples, mainstream methods typically introduce noise into the model or data to blur the boundary between member and non-member instances, which has been considered highly effective. However, this approach raises concerns about the model's utility.

This project investigates how recent defense mechanisms can address this privacy-utility trade-off, with a particular focus on RelaxLoss[5], a promising strategy that modifies model loss behavior rather than injecting noise. We aim to evaluate its effectiveness under both controlled (CIFAR-10) and real-world (MIMIC-III) settings. A detailed review of related works, including Differential Privacy (DP) and RelaxLoss, is presented in the next section. [2][9] [14] [4]

To complement our evaluation on standard datasets like CIFAR-10 and to address the structural limitations encountered with MIMIC-IV, we additionally include the Covertype dataset in our study. Although it does not raise explicit privacy concerns, Covertype offers structurally diverse input features—combining continuous variables with sparse categorical encodings—that serve as a practical testbed for assessing the generalization of RelaxLoss beyond typical benchmarks.

2 RELATED WORKS

2.1 Membership Inference Attacks and Defense Categories

Membership Inference Attacks (MIAs) have become a central threat in the field of privacy-preserving machine learning. In a typical MIA, the attacker attempts to determine whether a specific data sample was included in the training set of a target model, exploiting the model's tendency to memorize or overfit certain data points. Based on the amount of information accessible to the attacker, MIAs can be broadly categorized into several types, including white-box attacks, black-box attacks, label-only attacks, and those based on logit outputs or member-specific loss and gradient distributions. Each category differs in its assumptions regarding the attacker's access to internal model information such as gradients, confidence scores, or output probabilities.

To counter these threats, a range of defense mechanisms have been proposed. Among the most well-established is Differential Privacy (DP) and its practical implementation via DP-SGD[1]. Differential Privacy (DP) and its practical variant, DP-SGD, are designed to mitigate privacy risks particularly those arising from model overfitting by adding mathematically grounded Laplace or Gaussian noise to model outputs or gradients. DP-SGD additionally applies per-sample gradient clipping and introduces a configurable privacy budget to bound the total privacy loss, thereby reducing the likelihood that individual training samples can be inferred from the model's behaviour.

Laplace noise is generally more suitable for scenarios involving a small number of queries or single-release mechanisms, while Gaussian noise is better suited for complex models and repeated computations. As the number of operations increases, Gaussian noise tends to provide more stable and effective privacy guarantees compared to Laplace noise. [12] [3] [7]

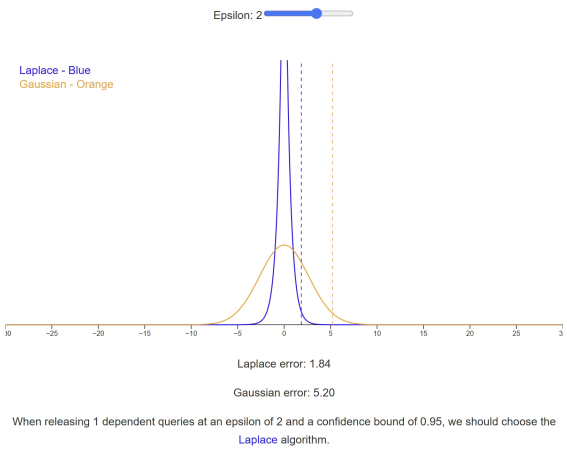

Figure 2: Comparison between Laplace and Gaussian mechanisms under differential privacy. At $\epsilon = 2$, the Laplace mechanism (blue) yields lower error (1.84) compared to the Gaussian mechanism (orange, error = 5.20), making it the preferred choice for single-query releases with high confidence.[12]

In theory, even the most carefully designed noise injection mechanisms grounded in solid mathematics inevitably impact model utility. This trade-off becomes especially problematic in high-stakes applications where predictive accuracy is critical for instance, in medical diagnosis or financial risk assessment. In such contexts, the privacy-utility dilemma becomes an ethical "trolley problem": Should we prioritize saving a life, or protecting the individual's data privacy?

2.2 The Principle of RelaxLoss

To address the trade-off between privacy and utility in membership inference defense, RelaxLoss [5] was proposed by Chen, Yu, and Fritz as a utility-preserving alternative that avoids explicit noise injection. Rather than obfuscating gradients or outputs with randomness, RelaxLoss reduces the distinguishability between member and non-member samples by dynamically altering the training dynamics and output confidence structure. RelaxLoss consists of two key mechanisms:

Relaxed Loss Target. Instead of minimizing the training loss to zero, RelaxLoss introduces a configurable loss threshold α and applies gradient ascent when the batch loss $\mathcal{L}_{\text{avg}}$ falls below this threshold. This approach increases the variance of the training loss distribution, thereby weakening the overfitting signal typically exploited by MIAs:

$$\nabla \mathcal{L}_{\text{Relax}} = \begin{cases} \nabla \mathcal{L} & \text{if } \mathcal{L}_{\text{avg}} > \alpha \\ -\nabla \mathcal{L} & \text{if } \mathcal{L}_{\text{avg}} \leq \alpha \end{cases}$$

Posterior Flattening. To preserve classification utility while suppressing prediction overconfidence, RelaxLoss retains the confidence score for the ground-truth class and uniformly redistributes the remaining probability mass across all other classes:

$$p_i' = \begin{cases} p_y & \text{if } i = y \\ \frac{1-p_y}{C-1} & \text{otherwise} \end{cases}$$

This label transformation reduces sharp prediction patterns that MIAs can exploit while maintaining the accuracy of the ground-truth prediction.

Together, these two mechanisms blur the statistical indicators (e.g., loss gap, prediction confidence) commonly targeted by MIA algorithms, achieving effective privacy protection without significantly harming model performance.

2.3 Limitations in Applying RelaxLoss to Clinical Data

Although MIMIC-IV represents a highly privacy-sensitive and clinically relevant dataset, its application posed several challenges for our project. The dataset's structural heterogeneity and severe class imbalance hindered straightforward adaptation of RelaxLoss, which was originally developed for image classification using MLPs. As a result, detailed preprocessing and feature engineering efforts were required, which we describe in Section 4. Despite these efforts, the combination of high dimensionality and sparse feature coverage ultimately prevented us from completing RelaxLoss experiments on MIMIC-IV within the project timeline.

2.4 Use the Covertype Dataset

Compared to Purchase-100 and Texas-100—both consisting of dense, normalized numerical features, the Covertype dataset presents a structurally distinct challenge. It contains a mix of continuous features and high-dimensional one-hot encoded categorical variables, such as soil type and wilderness area, many of which are extremely sparse and highly imbalanced. For instance, some one-hot categories appear in only a few hundred records out of 50,000+, resulting in extreme feature sparsity and irregular class distributions.

This complexity has practical implications for privacy-preserving learning. RelaxLoss is designed for MLP classifiers with cross-entropy loss and depends on mechanisms such as posterior flattening and relaxed loss targeting. These mechanisms are sensitive to the distribution of confidence scores and batch-level loss variance, both of which may behave differently in structurally heterogeneous datasets like Covertype. Although the dataset itself does not raise privacy concerns, it offers a compelling proxy for evaluating the structural robustness and generalization potential of RelaxLoss beyond vision tasks or homogeneous tabular domains.

3 PROBLEM DEFINITION

RelaxLoss offers a theoretically grounded and empirically effective strategy for defending against membership inference attacks. However, its real-world applicability remains uncertain due to the structural complexity of practical datasets. In domains such as healthcare, datasets like MIMIC-IV often involve high-dimensional, sparse, and imbalanced features, as well as multi-modal inputs and heavy preprocessing requirements. These characteristics differ significantly from the balanced and homogeneous benchmarks (e.g., CIFAR-10, Purchase-100) on which RelaxLoss was originally evaluated.

Although the mechanism of RelaxLoss—modulating training dynamics and reducing overconfident predictions—has demonstrated promising results, its implementation is currently limited to neural network architectures trained with cross-entropy loss. This constrains its compatibility with clinical models such as LightGBM or transformer-based encoders, which are commonly preferred for structured medical data. As such, evaluating the effectiveness and limitations of RelaxLoss under these structural and operational constraints remains a critical and largely unexplored research question.

Our initial goal was to evaluate RelaxLoss on the MIMIC-IV clinical dataset, a large-scale, privacy-sensitive collection of ICU patient records. However, due to the structural complexity of the data—including sparse and imbalanced one-hot features, multiple modalities, and high preprocessing demands—we encountered substantial difficulty in building a usable feature matrix suitable for MLP-based models. Furthermore, RelaxLoss is currently only implemented for neural network architectures with cross-entropy loss, limiting its compatibility with other common clinical models such as LightGBM or transformer-based encoders.

To address this gap from a different angle, we initially shifted our focus to the Covertype dataset. Although not privacy-sensitive, Covertype presents structurally complex input distributions, including sparse categorical encodings and significant class imbalance. This made it a suitable candidate for assessing whether RelaxLoss maintains its privacy-preserving properties and classification utility under structurally challenging conditions.

To further evaluate the generalization of RelaxLoss beyond standard image or example benchmarks, we extended our experiments to include the UCI Adult dataset (a tabular binary classification task) and STL-10 (a multi-class image dataset). These datasets differ in modality, label distribution, and input dimensionality. For STL-10 in particular, we developed a custom black-box attack framework to enable detailed membership inference evaluation. Collectively, these experiments provide a more comprehensive view of RelaxLoss performance across structurally diverse and practically relevant scenarios.

4 METHODOLOGY

4.1 Baseline Reproduction on CIFAR-10

To enable reproduction experiments, we first restored and modified portions of the original RelaxLoss source code to address compatibility and runtime issues in our current environment. During setup, we encountered several dependency-related challenges, including deprecated NumPy syntax (e.g., `np.isfinite`) and tight coupling to specific PyTorch versions. These modifications were made with caution to preserve the original training dynamics and avoid introducing unintended side effects.

We successfully reproduced the original RelaxLoss experiments on the CIFAR-10 dataset using the official codebase [1]. A ResNet model was trained as the target model, and multiple membership inference attacks, including confidence-based, entropy-based, and gradient-based methods—were implemented through a shadow model framework. The training hyperparameters and attack settings were kept as close as possible to the original paper. Evaluation metrics included classification accuracy, attack success rate, and AUC, serving as baselines for our extended experiments.

4.2 Feature Engineering for MIMIC-IV

To prepare the MIMIC-IV dataset for modeling, we developed a modular feature extraction pipeline covering key clinical sources, including laboratory results (`labevents`), chart events, input/output records, ICU stays, and procedure codes. Each source was processed using custom Python scripts to generate structured tabular features.

For numeric time-series data, we grouped values by hospital admission ID (`hadm_id`) and computed statistical summaries (mean, std, min, max, count). Categorical variables like care units and procedures were one-hot encoded. Features with less than 5% coverage were dropped to control dimensionality. Missing values were imputed with column medians, and binary missing indicators were appended. The resulting tables (e.g., `lab_features.csv`,

`input_features.csv`) were joined by `hadm_id` to form a unified patient-feature matrix.

Binary mortality labels were created from `admissions.csv.gz` based on the presence of a `deathtime` entry. This entire pipeline was built to enable multi-class classification tasks such as ICD or procedure prediction using MLP-based models under RelaxLoss. Despite these efforts, extreme feature sparsity and class imbalance

[1] https://github.com/DingfanChen/RelaxLoss

limited downstream model performance. We describe these challenges in the next Section.

4.3 Challenges in Applying RelaxLoss to MIMIC-IV

Despite preliminary feature engineering, subsequent limitations emerged due to the constrained adaptability of the RelaxLoss framework when applied to MIMIC-based clinical data. These limitations can be classified into two domains: data structure constraints and model architecture incompatibility.

First, the extracted feature matrix exhibited pronounced sparsity and distributional skewness. A large proportion of categorical variables, particularly one-hot encoded indicators corresponding to specific laboratory tests and clinical procedures, were extremely infrequent, occurring in fewer than 0.5 percent of patient records. The overall feature distribution was heavily imbalanced, with a small subset of high-frequency variables exerting a disproportionate influence on the input space. Such sparsity and redundancy substantially reduced the capacity of the multilayer perceptron to learn stable and generalizable representations, especially under the batch-dependent optimization dynamics inherent to RelaxLoss.

Second, the targeted multi-class classification tasks, such as ICD code prediction and procedural code prediction, were affected by extreme label imbalance. Certain labels contained in excess of ten thousand instances, whereas others contained fewer than fifty, resulting in unstable convergence and degraded training performance.

4.4 Preprocessing Covertype for Structural Evaluation

To test the structural generalization of RelaxLoss in a non-clinical but complex tabular setting, we selected the UCI Covertype dataset, which contains 54 features describing forest cover types. These include 10 continuous terrain-related variables and 44 sparse one-hot encoded categorical variables, such as soil type and wilderness area. Although the dataset is not privacy-sensitive, its structural heterogeneity, including high-dimensional categorical inputs and severe class imbalance, makes it a suitable proxy for evaluating RelaxLoss in difficult non-image domains.

For preprocessing, we retained all original features and performed z-score normalization on the continuous variables. One-hot encoded features were left unchanged to preserve their inherent sparsity. No feature selection or dimensionality reduction was applied in order to fully expose the model to structural irregularities. The target label, representing forest cover type, is a multiclass variable with seven categories, some of which are underrepresented. We split the data into 80% training and 20% testing subsets, ensuring stratification to preserve class proportions. To avoid data leakage, all normalization parameters were computed on the training set only and then applied to the test set.

The resulting dataset preserves both sparse and dense characteristics and was used directly as input to a multilayer perceptron (MLP) with RelaxLoss. This allowed us to assess the robustness of the loss function when applied to high-dimensional, imbalanced, and partially sparse data structures.

4.5 Model Design for Covertype

To evaluate the performance of RelaxLoss on the structurally complex Covertype dataset, we implemented a custom multilayer perceptron (MLP) classifier tailored to the dataset's mixed input types. The architecture, shown in Figure 3, consists of five fully connected layers with decreasing dimensions: 1024, 512, 256, 128, and 64 units, respectively. Each layer is followed by `LayerNorm` and `ReLU` activation. The final hidden representation is passed to a classification head composed of a dropout layer (rate = 0.1) and a linear output layer predicting 7 forest cover types.

We selected this architecture to balance model capacity and regularization, given the dataset's mixture of continuous and sparse categorical features. Compared to smaller MLPs used in prior RelaxLoss experiments on Purchase or Texas, this deeper structure helps the model accommodate more complex feature interactions. The base loss function is cross-entropy, which is modified by RelaxLoss during training to incorporate posterior flattening and relaxed loss thresholding (α = 2.0).

```python
class CovertypeClassifier(nn.Module):
    def __init__(self, input_size=54, num_classes=7, droprate=0.1):
        super(CovertypeClassifier, self).__init__()

        self.features = nn.Sequential(
            nn.Linear(input_size, 1024),
            nn.LayerNorm(1024),
            nn.ReLU(),

            nn.Linear(1024, 512),
            nn.LayerNorm(512),
            nn.ReLU(),

            nn.Linear(512, 256),
            nn.LayerNorm(256),
            nn.ReLU(),

            nn.Linear(256, 128),
            nn.LayerNorm(128),
            nn.ReLU(),

            nn.Linear(128, 64),
            nn.LayerNorm(64),
            nn.ReLU(),
        )

        self.classifier = nn.Sequential(
            nn.Dropout(droprate),
            nn.Linear(64, num_classes)
        )

    def forward(self, x):
        x = self.features(x)
        x = self.classifier(x)
        return x
```

Figure 3: Architecture of the CovertypeClassifier used with RelaxLoss.

4.6 RelaxLoss Modification for Adult Dataset

The original RelaxLoss implementation was primarily designed for multi-class classification using cross-entropy loss. To enable evaluation on the binary-class UCI Adult dataset, we extended the RelaxLoss codebase to support binary classification with BCE loss. These modifications were implemented in both the training logic and dataset pipeline.

At the training level, we introduced a new flag `-binary` in the argument parser, allowing users to toggle between binary and multi-class settings. When this flag is enabled, the loss function switches from `CrossEntropyLoss` to `BCEWithLogitsLoss`, and the final activation changes from softmax to sigmoid. Additionally, we modified the loss redistribution behavior: the posterior flattening mechanism

was adapted for binary logits by scaling output confidence scores via sigmoid and clamping them with a configurable upper threshold α.

For the dataset, we implemented a new function `prepare_adult()` in `dataset.py`, which loads the preprocessed Adult dataset from a local NPZ file. This function partitions the data into member and non-member samples using a ground-truth `members` indicator. We split the member data into 25% target train, 50% shadow train, and the remaining 25% for pseudoattack training. Non-member samples were similarly divided into shadow and target test sets. All subsets were converted into PyTorch `TensorDataset` format for model training and evaluation.

These code-level adjustments (not present in the original RelaxLoss codebase) enable us to evaluate defence mechanisms in a binary classification setting using real-world tabular data. Our modified implementation maintains compatibility with the entire RelaxLoss training framework while extending its applicability to a wider range of data types.

4.7 STL Baseline & RelaxLoss Black-box Test Design

To evaluate RelaxLoss under the different image classification tasks, we conducted experiments on the STL-10 dataset using a custom black-box MIA testing pipeline. Both the baseline and RelaxLoss models adopted the same CNN architecture (SmallCNN), and the dataset was split evenly: 2500 samples for the target model and 2500 for the shadow model, with each further divided into training and testing sets (1250 each).

In the baseline setup, models were trained using standard cross-entropy loss. For the RelaxLoss version, we implemented a dynamic training strategy based on alternating epochs: even-numbered epochs enforced a target loss level α (e.g., 2.0), while odd-numbered epochs applied posterior flattening to reduce overconfidence, with confidence scores clipped by a threshold (e.g., 0.6).

After training, we extracted sample-level features such as softmax probabilities, loss, confidence, entropy, and correctness from both models. A black-box attack model was then trained using shadow model outputs to predict membership on the target model. This allowed us to compare the privacy leakage between baseline and RelaxLoss settings under the same attack conditions.

5 EXPERIMENTAL RESULTS

5.1 Initial Reproduction

In order to make the subsequent experiments successful, we conducted a reproduction experiment. We have successfully cloned and built the official RelaxLoss repository from GitHub, and are currently analyzing its structure to adapt it to our own models and datasets. As an initial test, we selected the CIFAR-10 dataset and trained a baseline ResNet-20 model to ensure that the training process works as expected. For faster iteration and efficiency, we reduced the number of training epochs from 300 to 100 (`num_epochs = 100`). The figure below shows the early-stage training output and directory structure of the RelaxLoss implementation. (see figure-1) We then conducted a series of membership inference attacks (MIA) to evaluate the privacy leakage of the RelaxLoss-protected model. The attacks were implemented using the shadow model methodology to closely simulate the behaviour of the target model. Both black-box and white-box attacks were performed. See Table 1 and Table 2.

Table 1: Black-box MIA Results on RelaxLoss Model (CIFAR-10)

Attack Method	Accuracy	AUC	AP
Correctness	0.535	–	–
Confidence (pre-threshold)	0.501	–	–
Entropy (pre-threshold)	0.500	–	–
Modified Entropy (pre)	0.501	–	–
Loss (pre-threshold)	0.501	–	–
Confidence	–	0.548	0.533
Entropy	–	0.513	0.514
Modified Entropy	–	0.549	0.533
Loss	–	0.548	0.533

Table 2: White-box MIA Results via Gradient-Based Methods

Gradient Feature	Accuracy	AUC	AP
grad_wrt_x_l1	0.503	0.548	0.535
grad_wrt_x_l2	0.503	0.548	0.534
grad_wrt_w_l1	0.504	0.550	0.536
grad_wrt_w_l2	0.503	0.551	0.536

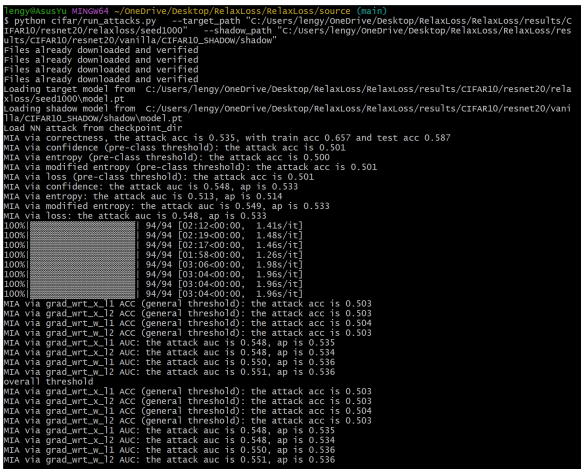

Figure 4: Training logs and MIA output of RelaxLoss experiment.

To better understand the training dynamics and model sensitivity, we report four gradient-based metrics. `grad_wrt_x_l1` and `grad_wrt_x_l2` denote the L1 and L2 norms of gradients with respect to the input, which reflect how sensitive the model predictions are to small perturbations in the input data. Similarly,

`grad_wrt_w_l1` and `grad_wrt_w_l2` represent the L1 and L2 norms of gradients with respect to the model weights, indicating the scale of parameter updates during training. Higher values generally imply greater sensitivity or more aggressive learning behaviour.[11][13]

In our current experimental setup, the model was trained for only 100 epochs, yet it already achieved a validation accuracy of 73.33%. Moreover, the accuracy was still steadily increasing by the end of training, indicating that the model was in a healthy learning phase and had not yet converged. This suggests that with extended training (e.g., up to 300 epochs as in the original paper), the model is expected to reach accuracy levels comparable to or even exceeding those reported in previous studies. Based on our current experiments, all membership inference attacks maintained low performance, close to random guessing (around 0.50), which demonstrates the strong effectiveness of the RelaxLoss defence.

The loss curve and accuracy curve (see Figure 13 and Figure 14) indicate stable training behavior. Additionally, the histogram (see Figure 6) shows that the predicted score distributions of training and validation data largely overlap, implying weak distinguishability and thus effective privacy protection.

0.100000	1.012601	1.160148	71.083333	62.200000	97.558333	95.800000
0.100000	1.022437	1.374716	71.858333	55.933333	96.966667	93.941667
0.100000	1.011752	1.184665	72.191667	60.341667	97.250000	95.083333
0.100000	1.002509	1.196778	72.983333	63.208333	97.541667	96.308333
0.100000	1.000559	1.145067	71.708333	61.458333	97.191667	95.741667
0.100000	1.016116	1.107431	72.416667	65.975000	97.491667	96.375000
0.100000	1.016657	1.230712	68.500000	57.375000	96.950000	94.841667
0.100000	1.013540	1.133220	72.500000	65.208333	97.291667	96.108333
0.100000	1.013687	1.324167	71.650000	56.358333	97.241667	92.858333
0.100000	1.017927	1.248617	72.941667	61.625000	97.308333	94.708333
0.100000	1.018883	1.186241	70.366667	61.008333	97.275000	93.816667
0.100000	0.999517	1.253021	73.333333	58.683333	97.408333	95.700000

Ln 101, Col 10 | 7,896 characters | 100% | Windows (C

Figure 5: Model training RelaxLoss epochs:100

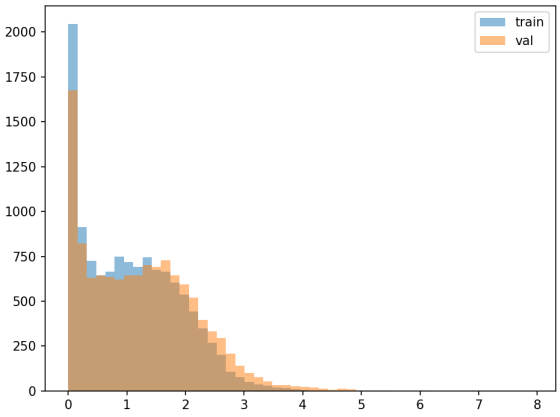

Figure 6: Loss Distribution Visualization

In addition to CIFAR-10, we also reproduced RelaxLoss experiments on the Purchase-100 and Texas-100 datasets. The results closely matched those reported in the original paper, confirming the correctness of our implementation.

5.2 RelaxLoss on MIMIC-IV (MLP Model)

Figure 8 shows the architecture of the neural network used in our MIMIC-IV experiments. After feature engineering, the final input matrix consisted of 8,390 features per hospital admission, derived from a combination of laboratory measurements, chart events, medication records, procedures, fluid balances, and demographic information. Each feature represents a statistical summary (e.g., mean, count) or one-hot encoded clinical code, resulting in a highly sparse and high-dimensional design matrix.

The target label was constructed as a multi-class classification problem over the top-100 most frequent ICD-9 diagnosis codes, using data from the MIMIC-IV clinical database (v2.2) accessed via PhysioNet. This makes the task a 100-class structured classification problem on tabular health records. We implemented a compact feedforward neural network with two hidden layers (256 and 128 units), dropout regularization, and ReLU activation, as shown in the figure. This model, referred to as `SparseMIMICClassifier`, served as the base model for applying RelaxLoss under realistic clinical data conditions.

```
class SparseMIMICClassifier(nn.Module):
    def __init__(self, input_size=8390, num_classes=100, droprate=0.5):
        super(SparseMIMICClassifier, self).__init__()

        self.net = nn.Sequential(
            nn.Linear(input_size, 256),
            nn.ReLU(),
            nn.Dropout(droprate),

            nn.Linear(256, 128),
            nn.ReLU(),
            nn.Dropout(droprate),

            nn.Linear(128, num_classes)
        )

    def forward(self, x):
        return self.net(x)
```

Figure 7: Architecture of the SparseMIMICClassifier used for MIMIC-IV Top-100 ICD classification.

Despite these design choices, the model failed to achieve satisfactory performance, reaching only 20% test accuracy after 120 training epochs (see Figure 8). Throughout training, we observed high variance in both training loss and validation performance, further indicating poor convergence. Since similar models perform well in similar tasks, it indirectly proves that we still have huge room for improvement in data processing, especially feature engineering and targeted tasks.

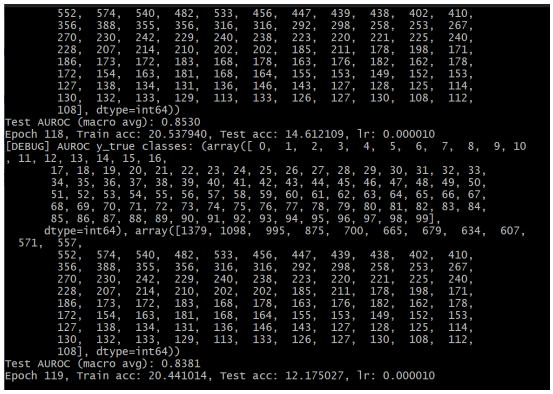

Figure 8: Architecture of the SparseMIMICClassifier used for MIMIC-IV Top-100 ICD classification.

5.3 Related Research on MIMIC

Since the original RelaxLoss implementation primarily employs an MLP architecture to evaluate a wide range of benchmark datasets, our experiments also largely followed a similar design. However, this choice may introduce data compatibility issues when applied to datasets with substantially different structures and feature distributions, such as the MIMIC series. Multiple studies from the U.S. National Library of Medicine have demonstrated that adopting model architectures better suited for structured clinical data can significantly improve performance. For example, in the study on ICU patient readmission risk prediction, the LightGBM model achieved an AUC of 0.736 on the validation set, outperforming various baseline methods [10]. In another study on high-risk chest pain patient prediction based on the MIMIC-IV dataset, researchers compared multiple models, including Logistic Regression, SVM, LightGBM, and XGBoost. The results showed that LightGBM achieved the best performance on the independent test set, with an AUC of approximately 0.89 (95% CI 0.878–0.902), an accuracy of about 95%, and a macro-average F1-score of approximately 0.94, significantly outperforming Logistic Regression, linear SVM, and MLP [6]. In the ICD diagnostic code prediction task using the MIMIC-III dataset, Logistic Regression was employed as the baseline model to predict the 50 most frequent ICD-9 diagnosis codes. Even when compared to a deep learning CNN model, Logistic Regression demonstrated strong competitiveness [8].

Although our study did not conduct a more in-depth evaluation of these alternative models due to time constraints, these findings provide an important insight: selecting model architectures that are more compatible with the structural characteristics of specific medical datasets may yield a better balance between privacy preservation and predictive performance.

5.4 RelaxLoss on Covertype

To evaluate RelaxLoss under structurally complex tabular conditions, we applied it to the UCI Covertype dataset using the same five-layer MLP architecture used in baseline experiments. The task was to classify forest cover types into seven categories, based on 54 structured features, including both continuous variables and sparse one-hot encoded categorical attributes.

The vanilla model achieved strong performance, reaching a test accuracy of 88.89% after 120 epochs (see Figure 15 and Figure 19). The training and validation curves converged smoothly, and no major overfitting was observed.

When RelaxLoss was enabled ($\alpha = 2.0$, posterior flattening active), the same model structure exhibited a clear decline in test performance. The final accuracy dropped to approximately 78.85%, representing a 10% absolute loss in predictive utility (see Figure 16 and Figure 17). Moreover, the training dynamics became less stable, with higher validation loss and slower convergence. These patterns were consistent across multiple seeds and configurations.

While RelaxLoss reduced the effectiveness of black-box membership inference attacks (e.g., lowering AUC from 0.72 to 0.60), the defence was not as strong as reported in the original paper. The sharp trade-off between privacy and utility on Covertype suggests that RelaxLoss may be sensitive to structural irregularities such as feature sparsity, class imbalance, and heterogeneous input types.

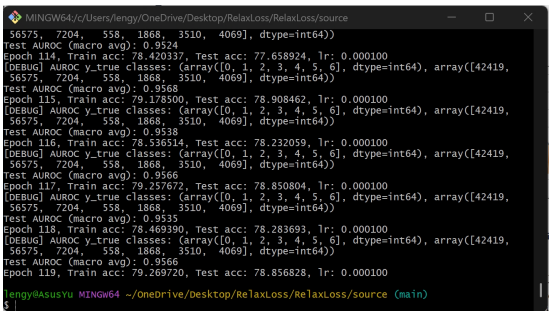

Figure 9: Covertype relaxloss training screenshot

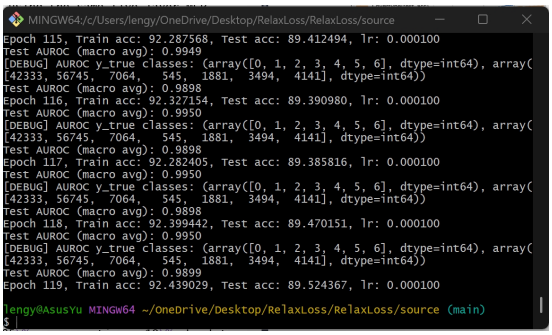

Figure 10: Covertype vanilla training screenshot

5.5 Membership Inference Attack Evaluation

We performed a series of membership inference attacks on both the vanilla and RelaxLoss-protected Covertype models, using white-box (gradient-based), black-box (confidence/loss-based), and neural network shadow attacks. As shown in Table 3, neither configuration exhibited high attack success: most AUC values remained between 0.50 and 0.52, indicating near-random guessing.

Nonetheless, we observed a consistent and measurable gap between the two setups. The vanilla model yielded slightly elevated

AUC values around 0.520 across all attack types, while RelaxLoss effectively suppressed these signals to values near half. Although this difference may seem numerically small, it is statistically consistent across attack modalities, and indicative of a meaningful reduction in member/non-member distinguishability.

This subtle shift reflects RelaxLoss's intended function to reduce overconfident and overfit model behaviour, even in tasks where privacy leakage is inherently low. The result also implies that shadow modelling did not succeed in capturing exploitable membership signals in this dataset, likely due to a combination of strong model regularization, tabular sparsity, and balanced label distribution.

Table 3: Membership Inference Attack AUC on Covertype

Attack Method	Vanilla AUC	RelaxLoss AUC
Gradient ∇_x (L1)	0.519	0.502
Gradient ∇_x (L2)	0.519	0.502
Gradient ∇_w (L1)	0.520	0.502
Gradient ∇_w (L2)	0.520	0.502
Neural Net (logits)	0.510	0.502
Confidence (black-box)	0.520	0.505
Entropy	0.511	0.501
Modified Entropy	0.520	0.505
Loss	0.520	0.505

Overall, the limited effectiveness of membership inference attacks, particularly on the vanilla model, suggests that Covertype is inherently resistant to privacy leakage under current conditions. Since the training process, data pipeline, and shadow model setup closely follow standard procedures, we attribute this weakness primarily to the structural properties of the dataset and the complexity of the model architecture. Unlike the dense, user-specific signals in datasets like Purchase or Texas, Covertype contains geospatial and ecological features that are less prone to memorization. Moreover, our MLP is deeper and more regularized than those used in prior MIA work, which may further reduce the emergence of member-specific patterns. A more detailed discussion of these limitations and their implications will be provided in Section 6.

5.6 Adults Experiment

We applied RelaxLoss to the UCI Adult dataset using a 4-layer MLP trained with binary cross-entropy loss. The dataset was split into 25% target train, 25% target test, and 50% for shadow modelling.

Black-box membership inference attacks based on loss, entropy, and confidence were conducted. Under baseline training, AUC scores reached around 0.74, while under RelaxLoss, they dropped close to 0.50, showing a decrease of approximately 0.30 across all metrics. This confirms that RelaxLoss effectively reduced membership leakage on structured binary data.

5.7 STL Black-Box MIA Results

We further evaluated RelaxLoss on the STL-10 image classification dataset using a custom black-box MIA pipeline. Under baseline training, all single-feature attacks achieved high AUCs above 0.95, indicating severe membership leakage.

Table 4: Black-box MIA AUC scores on Adult dataset

Feature	Baseline AUC	RelaxLoss AUC	AUC Drop
Loss	0.740	0.507	-0.233
Confidence	0.712	0.502	-0.210
Entropy	0.681	0.500	-0.181

After applying RelaxLoss, the AUC scores of all features dropped significantly. For example, the AUC of the loss-based attack dropped from 0.9563 to 0.6678, and the AUC of the confidence-based attack dropped from 0.9535 to 0.5627, with the average AUC decreasing by about 0.38.

Table 5: STL-10 MIA AUC scores (black-box)

Feature	Baseline	RelaxLoss	Drop
Loss (neg)	0.9563	0.6678	-0.2885
Confidence	0.9535	0.5627	-0.3908
Softmax Max Conf	0.9535	0.5627	-0.3908

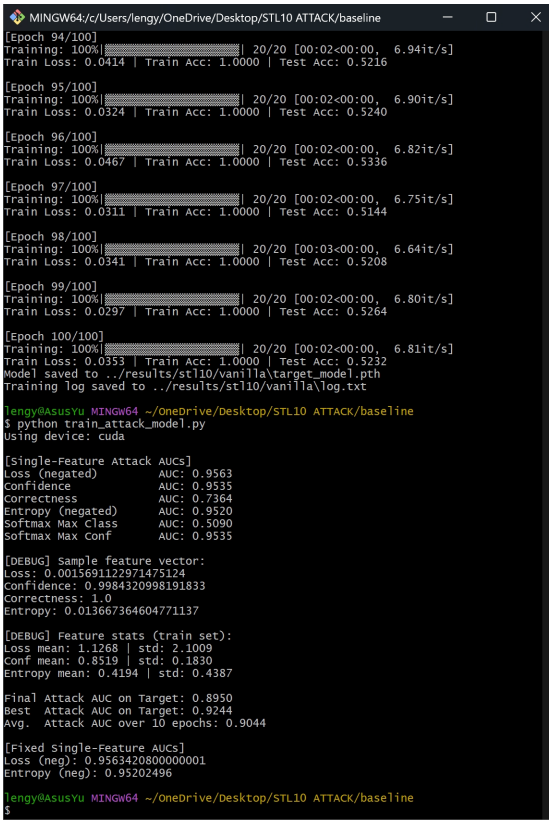

Figure 11: Covertype relaxloss training screenshot

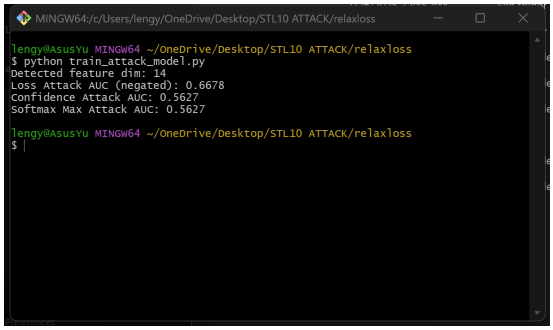

Figure 12: Covertype relaxloss training screenshot

These results confirm that RelaxLoss effectively suppresses membership leakage in high-dimensional image classification settings under black-box threats.

6 CONCLUSION AND FUTURE WORKS

In this study, we conducted a comprehensive evaluation of the RelaxLoss defence mechanism across a wide range of datasets, tasks, and attack settings. Starting from a faithful reproduction of the official implementation on CIFAR-10, Texas-100, and Purchase-100, we gradually extended our experiments to include structurally diverse datasets such as MIMIC-IV (clinical), Covertype (geospatial), Adult (tabular), and STL-10 (image).

Our experiments confirm that RelaxLoss consistently reduces the effectiveness of membership inference attacks, particularly in settings with overconfident or highly separable models. In the CIFAR-10 and STL-10 image classification tasks, black-box attack AUCs dropped by 0.30 to 0.40, indicating substantial mitigation of privacy leakage. Similarly, for tabular binary classification tasks like Adult, the attack success rate was suppressed to near-random levels.

However, our results also reveal important limitations. On high-dimensional, sparse datasets such as MIMIC-IV, RelaxLoss alone was insufficient to ensure both privacy and utility. The MLP-based classifier struggled to learn effectively, reaching only 20% accuracy on a 100-class ICD prediction task. This suggests that applying RelaxLoss to clinical data may require more advanced modelling strategies, such as tree-based models, hybrid embeddings, or improved feature engineering pipelines.

Overall, while RelaxLoss remains a promising defence mechanism, its effectiveness is highly dependent on the data structure, task complexity, and model design. Future work should explore its integration with domain-specific architectures, investigate its compatibility with other regularization techniques, and evaluate its performance under more advanced white-box and adaptive attacks. Expanding its applicability to imbalanced, noisy, or privacy-critical datasets remains an open and important direction.

The experiment can be found from the Git Hub repositories: https://github.com/ColdenLeng/MCTI-project-phase-report

REFERENCES

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 308–318. https://doi.org/10.1145/2976749.2978318

[2] Richmond Alake. 2024. Loss Functions in Machine Learning: An Introduction. https://www.datacamp.com/tutorial/loss-function-in-machine-learning. (2024). DataCamp Tutorial, last updated December 2024. Accessed: 2025-07-03.

[3] Joel Daniel Andersson, Rasmus Pagh, and Sahel Torkamani. 2024. Count on Your Elders: Laplace vs Gaussian Noise. *arXiv preprint arXiv:2408.07021* (2024). https://arxiv.org/abs/2408.07021 Accessed: 2025-07-03.

[4] Eric Aubinais, Elisabeth Gassiat, and Pablo Piantanida. 2023. Fundamental Limits of Membership Inference Attacks on Machine Learning Models. *arXiv preprint arXiv:2310.13786* (2023). https://arxiv.org/abs/2310.13786

[5] Dingfan Chen, Ning Yu, and Mario Fritz. 2022. RelaxLoss: Defending Membership Inference Attacks Without Losing Utility. In *Proceedings of the Tenth International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=FEDfGWVZYIn

[6] Hongyi Chen, Haiyang Song, Hongyu Huang, Xiaojun Fang, Huang Chen, Qingqing Yang, Junyu Zhang, Wenjun Ding, Zheng Gong, and Jun Ke. 2025. Machine learning prediction and interpretability analysis of high-risk chest pain: a study from the MIMIC-IV database. (2025). https://doi.org/10.3389/fphys.2025.1594277 PMID: 40661663, PMCID: PMC12256431.

[7] Amalie Dyda, Michael Purcell, Stephanie Curtis, Emma Field, Priyanka Pillai, Kieran Ricardo, Haotian Weng, Jessica C. Moore, Michael Hewett, Graham Williams, and Colleen L. Lau. 2021. Differential privacy for public health data: An innovative tool to optimize information sharing while protecting data confidentiality. *Patterns* 2, 12 (2021), 100366. https://doi.org/10.1016/j.patter.2021.100366

[8] Shuyuan Hu, Fei Teng, Lufei Huang, Jun Yan, and Haibo Zhang. 2021. An explainable CNN approach for medical codes prediction from clinical text. (2021). https://pmc.ncbi.nlm.nih.gov/articles/PMC8596896/ PMID: 34789241, PMCID: PMC8596896.

[9] Scikit learn Developers. 2024. sklearn.metrics.log_loss — Scikit-learn Documentation (Version 1.3.2). https://scikit-learn.org/stable/modules/generated/sklearn.metrics.log_loss.html. (2024). Accessed: 2025-07-03.

[10] Jinfeng Miao, Chengchao Zuo, Huan Cao, Zhongya Gu, Yaqi Huang, Yu Song, and Furong Wang. 2024. Predicting ICU readmission risks in intracerebral hemorrhage patients: Insights from machine learning models using MIMIC databases. (2024). Available at: https://pubmed.ncbi.nlm.nih.gov/38147802/.

[11] Milad Nasr, Reza Shokri, and Aria Houmansadr. 2019. Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning. In *IEEE Symposium on Security and Privacy (SP)*.

[12] Lukáš Panavás. 2021. Mechanism Comparison: Laplace vs Gaussian (Interactive Tool). (2021). https://lpanavas.github.io/mechanism-comparison/ Accessed: 2025-07-03.

[13] Shahbaz Rezaei and Xin Liu. 2021. On the Difficulty of Membership Inference Attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR46437.2021.00750

[14] PyTorch Team. 2024. torch.nn.BCELoss — PyTorch Documentation (Version 2.3). https://pytorch.org/docs/stable/generated/torch.nn.BCELoss.html. (2024). Accessed: 2025-07-03.

7 APPENDIX

7.1 Experiment Result Loss Visualization CIFAR

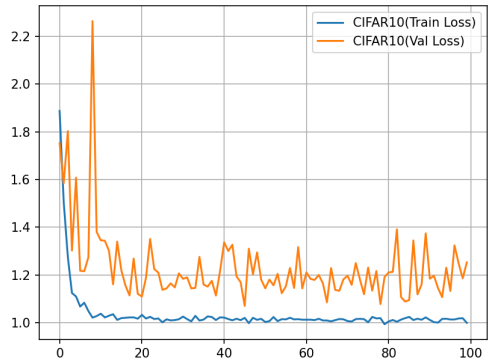

Figure 13: Model training loss comparison

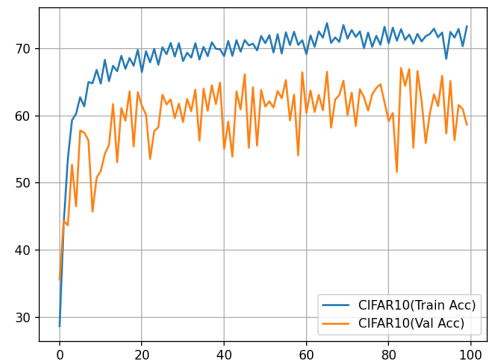

Figure 14: Model training ACC comparison

7.2 Experiment Result Loss Visualization COVERTYPE

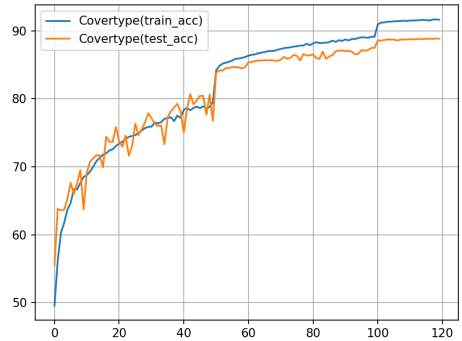

Figure 15: Training and validation accuracy for the vanilla MLP model on Covertype.

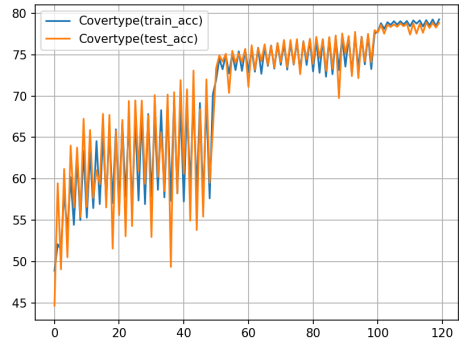

Figure 16: Training and validation accuracy with RelaxLoss on Covertype.

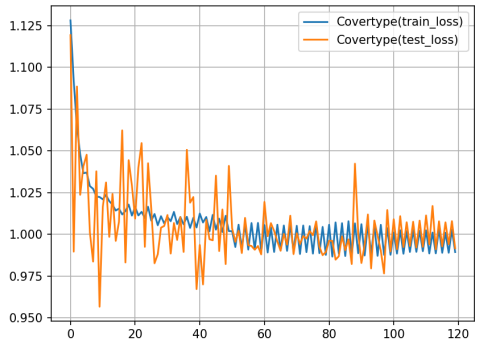

Figure 17: Training and validation loss with RelaxLoss on Covertype.

7.3 Experiment Result Loss Visualization Adult

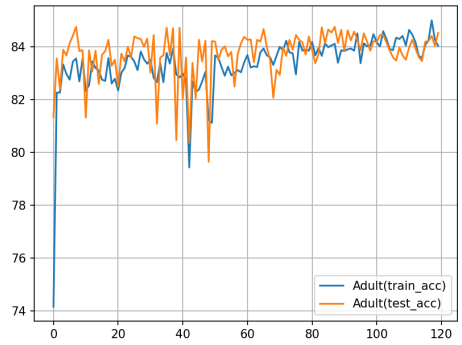

Figure 18: Training and validation acc for the Relaxloss MLP model on Adult.

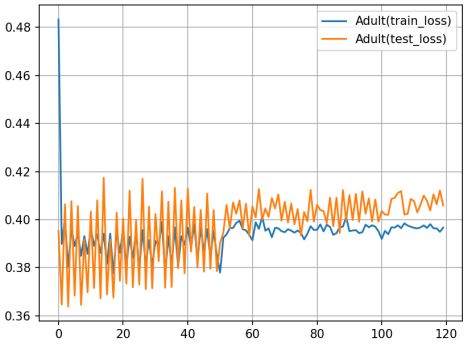

Figure 19: Training and validation loss for the Relaxloss MLP model on Adult.