Embedding-Space Data Augmentation to Prevent Membership Inference Attacks in Clinical Time Series Forecasting

Marius Fracarolli FRACAROLLI@CL.UNI-HEIDELBERG.DE
Department of Computational Linguistics, Heidelberg University, Germany
Michael Staniek STANIEK@CL.UNI-HEIDELBERG.DE
Department of Computational Linguistics, Heidelberg University, Germany
Stefan Riezler RIEZLER@CL.UNI-HEIDELBERG.DE
Department of Computational Linguistics
& Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University, Germany

Abstract

Balancing strong privacy guarantees with high predictive performance is critical for time series forecasting (TSF) tasks involving Electronic Health Records (EHR). In this study, we explore how data augmentation can mitigate Membership Inference Attacks (MIA) on TSF models. We show that retraining with synthetic data can substantially reduce the effectiveness of loss-based MIAs by reducing the attacker's true-positive to false-positive ratio. The key challenge is generating synthetic samples that closely resemble the original training data to confuse the attacker, while also introducing enough novelty to enhance the model's ability to generalize to unseen data. We examine multiple augmentation strategies — Zeroth-Order Optimization (ZOO), a variant of ZOO constrained by Principal Component Analysis (ZOO-PCA), and MixUp — to strengthen model resilience without sacrificing accuracy. Our experimental results show that ZOO-PCA yields the best reductions in TPR/FPR ratio for MIA attacks without sacrificing performance on test data.

Keywords: Time Series Forecasting, Electronic Health Records, Membership Inference Attack, Privacy, Synthetic Data, Data Augmentation

Data and Code Availability We use public data from physionet.org (MIMIC-III, eICU). Our code is available at GitHub[1].

Institutional Review Board (IRB) IRB approval is not required for public data.

1. https://github.com/MariusFracarolli/ML4H_2025_Data-Augmentation-to-Prevent-MIA-in-TSF/

1. Introduction

Membership Inference Attacks (MIAs) are a "de facto standard" attack scenario on the privacy of machine learning models (Carlini et al., 2022). Such adversarial attempts to determine whether a given sample was part of the training set are a real-world problem in the application of machine learning to sensitive health data such as Electronic Health Records (EHRs). As is known since Dinur and Nissim (2003), removing identifying information such as patients' names from a database is not enough to protect privacy, but instead random perturbations have to be applied to the outputs in order to protect privacy even in the simplest case of "statistical" queries such as averages over databases. If EHR data is used to train machine learning models, e.g., for medical time series forecasting (TSF), the randomization techniques provided in the framework of differential privacy (Dwork, 2006; Dwork and Roth, 2014) allow giving strong guarantees on the information derivable from private training data when querying a machine learning algorithm. However, there is an unavoidable tradeoff between utility and privacy of trained machine learning algorithms (Jayaraman and Evans, 2019; Cai et al., 2021). As shown by Yeom et al. (2018), overfitting of machine learning models is sufficient to allow MIA under the simple assumption that the average training loss of the model, but no information about the training data is leaked.

The goal of this work is to investigate the possibilities of data augmentation techniques to balance privacy protection against MIA and prediction utility for machine learning on EHR data, ideally with improvements on both criteria. As we will show empirically, although the simple loss-based MIA of Yeom

© 2025 M. Fracarolli, M. Staniek & S. Riezler.

et al. (2018) is weaker than attack scenarios based on shadow or reference models (Shokri et al., 2017; Ye et al., 2022; Carlini et al., 2022; Zarifzadeh et al., 2024), it is still effective for complex applications like multivariate TSF on sparsely sampled medical data. Following Carlini et al. (2022), we start from an evaluation of the effectiveness of MIAs by considering the ratio of their true positive rate (TPR) over their false positive rate (FPR). The privacy of a trained machine learning model can be protected by decreasing the TPR/FPR ratio from two sides. First, the FPR rate can be increased by generating synthetic data and re-training the model on the augmented dataset. These data need to be sufficiently different from the training examples in order to allow a decrease in held-out loss of a re-trained model. Second, the TPR rate can be reduced by adding synthetic examples that are similar enough to the original training set such that fewer original training examples are correctly recognized as members by the attacker. The goal is to reduce the TPR/FPR ratio towards 1, where the attack becomes equivalent to random guessing.

The main contribution of our work is the presentation of an algorithm that optimizes a joint objective of privacy protection and utility by zeroth-order optimization (ZOO) in the space of neural embedding matrices of training examples. The goal of this algorithm is to guide data augmentation in the direction of synthesizing examples that lead to decrease of the TPR/FPR ratio. We exemplify our algorithm on the difficult regression problem of 24 hour TSF on multivariate clinical data. Our experimental comparison investigates the power of ZOO versus standard data augmentation algorithms like MixUp (Zhang et al., 2018). Best results are obtained by enhancing ZOO with information about the principal components in embedding space (ZOO-PCA). While MixUp achieves best performance on test data, the best reductions in TPR/FPR ratio for MIA attacks without sacrificing performance on test data are achieved by ZOO-PCA. Since ZOO-PCA directs the augmentation process along the most significant data variations, it also enjoys favorable convergence rates due to effectively performing sparse zeroth-order optimization (Balasubramanian and Ghadimi, 2022).

2. Related Work

Differential privacy has become a de-facto standard for theoretically well-founded privacy preservation in machine learning since it allows giving strong guarantees on the information derivable from private training data when querying a machine learning algorithm (Dwork, 2006, 2008; Dwork and Roth, 2014). A shortcoming of randomization-infusion as done in differential privacy approaches is an unavoidable trade-off between model utility and privacy protection. This privacy-utility gap has been proven to be unavoidable by Cai et al. (2021), and quantified empirically by Abadi et al. (2016). According to Jayaraman and Evans (2019), it leads to either models with limited accuracy loss and meaningless privacy guarantees, or to useless models with strong privacy guarantees.

Data augmentation as alternative method for privacy protection has been discussed quite controversially. Sablayrolles et al. (2019) show that due to a reduction of the gap between training and held-out accuracy of models trained on augmented data, the effectiveness of loss-based attacks can be decreased. An even stronger claim is made by Hintersdorf et al. (2022) who show that using generative adversarial networks to synthesize data, a potentially infinite number of samples will be falsely classified as members of the training set by an attacker. However, Kaya and Dumitras (2021) claim that on complex tasks, data augmentation cannot provide a "free lunch" to defeat membership inference attacks. Besides different experimental settings, a conceptual difference between these approaches is the decision whether synthetic data should count as genuine members of the training set instead of as false positives (Yu et al., 2021). In our case, synthetic data are generated in the latent embedding space and do not even have correlates in the space of input representations. We thus believe synthetic embeddings generated by ZOO or MixUp to be fair and effective distractions of an attacker.

3. Time Series Forecasting (TSF)

TSF with Transformers To convert real-world data – specifically, irregularly sampled medical time series – into a long-term time series forecast (TSF), we use the implementation of Staniek et al. (2024) which is based on a Transformer encoder-decoder architecture (Vaswani et al., 2017). The architecture of this model is illustrated in Figure 1.

In our dataset, each sample $S = \{(t_i, f_i, v_i)\}_{i=1}^{n}$ consists of triplets where:

- $t_i \in \mathbb{R}_{\geq 0}$ is a timestamp,

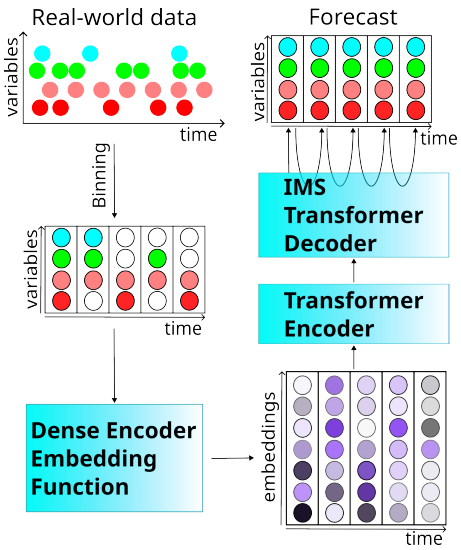

Figure 1: The process of time series forecasting for irregularly sampled medical time series data. Real-world data is binned into hourly buckets. The binned data is then transformed into embeddings using a Dense Encoder Embedding Function, capturing the embeddings relevant for the augmentation process. The embeddings are processed by a Transformer encoder, which learns contextual representations. An iterative multistep forecasting (IMS) decoder with autoregressive properties generates forecasts.

- f_i belongs to a set of clinical variables F,
- $v_i \in \mathbb{R}$ is the corresponding measured value.

Thus, a data point $(t_\alpha, f_\alpha, v_\alpha)$ indicates that variable f_α was measured with value v_α at time t_α. To enhance comparability, all values are standardized on a per-variable basis.

To prepare the data for the transformer, we bin the time series into hourly buckets, retaining only the first observed value per hour and variable. Given the high sparsity of the data, we apply zero-mean imputation and simultaneously generate a masking matrix indicating which values were missing.

Next, we apply an encoder embedding function to transform the samples into embeddings, which are subsequently used for data augmentation because they provide a consistent matrix size without missing or imputed data. This approach exclusively focuses on the right side of Figure 1, where the embeddings of the original data are fixed, as are the weights of the dense encoder embedding function. Generating new data begins at the embedding layer, making this strategy particularly valuable for enhancing model robustness and improving generalization without compromising data integrity.

The embeddings are processed by an encoder, followed by an autoregressive Transformer decoder for forecasting. For the decoder, we employ an Iterative Multi-Step (IMS) approach, where an output vector $\hat{y} \in \mathbb{R}^{|F|}$ is generated using the history of previous predictions.

For long-term TSF, we concatenate the predictions over all time steps $t = 1, \ldots, T$. As demonstrated in Staniek et al. (2024), the autoregressive decoder performs better when applied with student forcing, where previous predictions are treated as ground truth during training.

Performance Evaluation Metrics A data point $x = (e, y, m)$ representing a patient's ICU stay consists of an embedding matrix $e \in \mathbb{R}^{24 \times n}$ representing the first 24 hours, a ground truth time series $y \in \mathbb{R}^{T \times |F|}$ for the subsequent T hours, and a mask $m \in \{0,1\}^{T \times |F|}$ indicating the presence of measurements in y with the count $|m|$. T indicates the number of hours predicted. The dimension n of the embedding is model-dependent.

To evaluate the performance of TSF models, we define a masked mean squared error by

$$\text{mMSE}(x, \theta) := \frac{1}{|m|} \|(f_\theta(e) - y) \odot m\|_2^2 \qquad (1)$$

for a single sample. The mMSE is masked to ensure that only observed values contribute to the loss calculation, effectively ignoring missing data points. It is then normalized by the cardinality of the mask, enhancing the comparability between samples with varying amounts of observed data. For a dataset X, we generalize mMSE to

$$\text{MSE}(X, \theta) := \frac{1}{|X|} \sum_{x \in X} \text{mMSE}(x, \theta). \qquad (2)$$

4. Membership Inference Attack (MIA) and Privacy Loss Evaluation

The loss-based MIA described in Yeom et al. (2018) can be formalized as follows:

Definition 1 (Loss-based MIA) *A loss-based MIA is a privacy attack on a machine learning model f where an adversary $\mathcal{A}$ attributes a positive membership to a datapoint x if the model loss on the*

datapoint $\ell(f(x), y)$ is lower than a threshold τ for a model prediction $f(x)$ and a label y:

$$\mathcal{A}_{loss}(x, y) = \mathbb{I}_{\ell(f(x),y)<\tau} \tag{3}$$

In order to evaluate the attacker's success, we follow Carlini et al. (2022) who replace dataset-wide evaluation metrics for success rates of privacy attacks by an evaluation that measures if a MIA can reliably violate the privacy, even if it affects just a few users. Conversely, an attack that unreliably achieves a high aggregate success rate should not be considered successful. This is achieved by an evaluation of the attacker's True Positive Rate (TPR) at low False Positive Rate (FPR).

TPR/FPR ratio We assume that the attacker has access to both mMSE (Equation 1), for example, via an API that provides access to the model as scoring device, and to MSE (Equation 2), for example, via published results on the model's average MSE on particular datasets. To defend against a loss-based attack with τ set to the average training loss, we aim for a low TPR and a high FPR for an attacker $\mathcal{A}$. TPR and FPR are defined as follows:

$$\mathrm{TPR}(X_{\text{train}}, \tau, \theta) := \frac{1}{|X_{\text{train}}|} \sum_{x \in X_{\text{train}}} \mathbb{I}_{\mathrm{mMSE}(x,\theta)<\tau} \tag{4}$$

$$\mathrm{FPR}(X_{\text{test}}, \tau, \theta) := \frac{1}{|X_{\text{test}}|} \sum_{x \in X_{\text{test}}} \mathbb{I}_{\mathrm{mMSE}(x,\theta)<\tau} \tag{5}$$

In other words, members should ideally not be recognized as members (minimizing TPR), while non-members should be classified as members (maximizing FPR) until the ratio reaches 1 and non-members become indistinguishable from members. This TPR/FPR ratio is advantageous as it remains valid even for unbalanced test sets. Following Yeom et al. (2018), we assume the attacker is aware of the average training loss, which we will use as threshold τ. This yields the following privacy metric:

$$\mathrm{Priv}(X_{\text{train}}, X_{\text{test}}, \tau, \theta) := \frac{\mathrm{TPR}(X_{\text{train}}, \tau, \theta)}{\mathrm{FPR}(X_{\text{test}}, \tau, \theta)} \tag{6}$$

This metric represents the attacker's advantage in correctly identifying members compared to incorrectly identifying non-members. A lower Priv value indicates less advantage for the attacker.

Additionally, we define privacy loss (PL) as the positive membership prediction of a single data point:

$$\mathrm{PL}(x, \tau, \theta) := \mathbb{I}_{\mathrm{mMSE}(x,\theta)<\tau} \tag{7}$$

In our implementation, the threshold τ is set to the average training loss, computed after each training run using the current model parameters θ. This ensures that τ adapts to each model instance and reflects the actual distribution of losses encountered during training.

ROC curve By varying the threshold τ and calculating the corresponding TPR and FPR values, we generate the Receiver Operating Characteristic (ROC) curve. The area under the ROC curve (AUROC) provides a quantitative measure of the model's susceptibility to membership inference. An AUROC value of 0.5 indicates that the training and held-out sets are indistinguishable, implying no vulnerability to the attack. Conversely, an AUROC value approaching 1 suggests that the sets are highly separable, indicating a significant privacy risk.

5. Data Augmentation

The central idea of data augmentation is to improve a model's generalization performance by re-training on a dataset augmented with synthesized examples that cover aspects of the data distribution that are not found in the training sample. In our work, we focus on data augmentation techniques that synthesize data in the space of neural embeddings. We compare mix-based data augmentation (Zhang et al., 2018; Cao et al., 2024) with zeroth-order optimization in embedding space (Nesterov and Spokoiny, 2015; Chen et al., 2017). We use data augmentation not only to reduce MSE on the held-out set but also to lower the TPR/FPR ratio of the attacker. In our experiments, new data is generated within each epoch and the model is trained on a mix of original training data and synthetic data, while the embedding layer is kept fixed.

ZOO Zeroth-order optimization methods are of interest for machine learning problems where only the zeroth-order oracle, i.e., the value of the objective function but no explicit gradient, is available. Our application of ZOO is inspired by a black-box attack on deep neural networks where adversarial images that lead to misclassification are found by approximating the gradient through a comparison of function values at random perturbations of input images (Chen et al., 2017) or language (Berger et al., 2021). In addition to being simple and scalable, ZOO methods can be adjusted to optimize non-differentiable

functions and are still provably convergent (Fu, 2006; Nesterov and Spokoiny, 2015).

The central idea of ZOO is to perform optimization in the space of neural embeddings by iteratively updating an embedding for S steps by evaluating an objective function g over randomly perturbed points. This procedure includes the following iterative steps:

1. Generate k random perturbations u_i drawn from a normal distribution $\mathcal{N}(0, \mathbb{I})$ and normalized to $\|u_i\| = 1$, $i = 1, \ldots, k$.
2. Given a data point $x = (e, y, m)$, consisting of an embedding matrix $e \in \mathbb{R}^{24\times n}$, a ground truth time series $y \in \mathbb{R}^{T\times|F|}$, and a mask $m \in \{0,1\}^{T\times|F|}$, compute perturbed samples x_i:
$$x_i^{\pm} = (e_s \pm \mu u_i, m, y)$$
3. Update the embedding matrix using:
$$e_{s+1} = e_s - \lambda \frac{1}{k} \sum_{i=1}^{k} \frac{g(x_i^+, X, \theta) - g(x_i^-, X, \theta)}{2\mu} u_i$$

- $\lambda \in \mathbb{R}$: Learning rate controlling the step size of updates.
- $\mu \in \mathbb{R}$: Perturbation width controlling the scale of sampling around e_s.

We use a parameterized objective function g_α that is designed to improve both utility and privacy of medical data. Utility is optimized by searching for embedding matrices that represent examples that are sufficiently different from the training examples, in order to allow a decrease in held-out loss after retraining. This is achieved by using a negative mMSE term (Equation 1) in the loss objective. Privacy protection is optimized by searching for embedding matrices that are similar to the training examples in order to distract the attacker. This is achieved by using a negative PL term (Equation 7) in the loss objective. The interpolation weight $\alpha \in [0, 1]$ balances these two competing goals: when $\alpha = 1$, we purely optimize for diversity (high MSE on synthetic samples encourages exploration); when $\alpha = 0$, we purely optimize for privacy (low PL makes synthetic samples indistinguishable from training data). In practice, intermediate values of α achieve the best utility-privacy tradeoff:

$$g_\alpha(x, X, \theta) = -\Big(\alpha\, \mathrm{mMSE}(x, \theta) + (1-\alpha)\, \mathrm{PL}(x, X, \theta)\Big) \tag{8}$$

ZOO-PCA In traditional ZOO, the perturbations are generated randomly from a standard normal distribution, meaning that each feature in the embedding matrix is perturbed independently. While this can help explore the parameter space, it may not always result in meaningful or efficient perturbations, especially when certain features are more important than others for the task at hand.

In ZOO-PCA, by performing Principal Component Analysis (PCA) first, we are essentially identifying the most significant directions in the data that explain its variance. These principal components represent the major patterns or features of variation in the data, rather than treating all features equally. By keeping only the components that explain a significant portion of the variance (based on the cumulative explained variance ratio threshold), ZOO-PCA guides perturbations toward the most relevant variance components of the data, rather than random or less important components.

MixUp The MixUp technique of Zhang et al. (2018) generates a new data point by interpolating between two randomly selected data points $x_1 = (e_1, y_1, m_1)$ and $x_2 = (e_2, y_2, m_2)$. With a $\lambda \sim \mathrm{Beta}(\beta, \beta)$, we receive:

$$x_{\mathrm{MixUp1,2}} = \begin{cases} \big(\lambda e_1 + (1-\lambda) e_2, y_1, m_1\big), & \text{if } \lambda > 0.5 \\ \big(\lambda e_1 + (1-\lambda) e_2, y_2, m_2\big), & \text{if } \lambda \leq 0.5 \end{cases} \tag{9}$$

The new data point has a new embedding matrix, and the mask and target of the dominant data point. Due to the high sparsity of medical time series this is a good way to have a working perturbation instead of mixing most of the time *zero values* with observations. With a small β ($\beta \ll 1$), the new data points resemble one example more closely, while with a large β ($\beta \gg 1$), the examples are pushed toward a mixture of both points.

6. Experimental Setup

In our experiments, we use the Medical Information Mart for Intensive Care III (MIMIC-III) data (Johnson et al., 2016) and the larger eICU dataset (Pollard et al., 2018). A detailed description of the data is given in Appendix A.

In all of our experiments, we go beyond the standard forecasting setup – predicting hours 24 to 48, based on input from hours 0 to 24 – by employing a

Table 1: Evaluation results for baseline and data augmentation methods for TSF on MIMIC-III and eICU. All results are reported on unseen test data. Generalization performance of models is reported as MSE on test data MSE(X_{test}). Effectiveness of a loss-based MIA is reported by TPR/FPR ratio using the average training loss as threshold. An approximate randomization test (Noreen, 1989) shows that all result differences are significant (with $p < 0.05$) except for the MSE between baseline and ZOO-PCA for the eICU dataset. Best results for data augmentation methods are shown in bold face. FPR@τ values show that ZOO-based methods and DP-SGD increase false positive rates, effectively distracting the attacker.

	MIMIC-III			eICU		
Method	MSE(X_{test})	$\frac{TPR}{FPR}$@τ	FPR@τ	MSE(X_{test})	$\frac{TPR}{FPR}$@τ	FPR@τ
Baseline	0.5053	3.5482	0.1870	0.5604	1.3206	0.5137
DP-SGD	0.7446	0.9887	0.6560	0.8228	0.9969	0.6676
ZOO	0.4945	1.9925	0.4148	0.5576	1.1716	0.6877
ZOO-PCA	0.5000	**1.4271**	0.6421	0.5607	**1.1299**	0.7712
MixUp	**0.4918**	3.2264	0.1936	**0.5519**	1.2927	0.5000

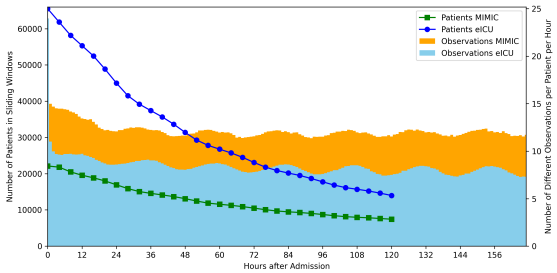

Figure 2: The bar plots (right axis) show the average number of different clinical variables recorded per patient per hour after admission (multiple measurements of the same variable within one hour are deleted during binning). These counts remain relatively stable over time for both MIMIC (orange) and eICU (blue). In contrast, the line plots (left axis) display the number of patients contributing to each 4-hour sliding window, which declines over time – more sharply in eICU – reflecting the decreasing number of long-staying patients.

sliding window technique that covers a wider range of temporal patterns. Specifically, for each patient, we shift both the observation and prediction windows in four-hour intervals throughout the ICU stay, enabling forecasts such as hours 40 to 64 based on inputs from hours 16 to 40. This strategy allows us to capture temporal dynamics more comprehensively and is illustrated in Figure 2. As a result, we do not restrict our analysis to the initial data points but include all windowed segments up to the first four days after admission, while the set splits stay on patient level. This significantly expands our datasets: from 21,573 to 342,213 samples in MIMIC-III (a 15.9-fold increase) and from 65,053 to 868,487 in eICU (a 13.4-fold increase).

For each data augmentation run, we generate 32,000 new synthetic examples. After the generation process, the model is re-trained on a balanced dataset composed of 50% original and 50% synthetic data. Due to memory constraints, the number of synthetic examples is limited such that it never exceeds 50% of the size of the training set. If this limit is reached, the oldest synthetic samples are discarded to make room for newly generated ones.

The updated model parameters are denoted by θ, and the augmented dataset X_{aug} comprises both the original training data X_{train} and the generated synthetic data. The mean squared error (MSE; see Equation 2) is evaluated on a held-out set using the current model parameters. The privacy metric (Equation 6) treats the training set as the member set, the held-out set as the non-member set, and uses the augmented set as the reference for computing the average loss τ.

For all training runs, the model is only updated if the following conditions are satisfied:

$$\begin{aligned}
&\text{Priv}(X_{\text{train}}, X_{\text{heldout}}, \tau, \theta) \leq (1 + \varepsilon_{\text{priv}})\,\text{Priv}_{\text{best}},\\
&\text{MSE}(X_{\text{heldout}}, \theta) \leq (1 + \varepsilon_{\text{MSE}})\,\text{MSE}_{\text{best}},\\
&\text{Priv}(X_{\text{train}}, X_{\text{heldout}}, \tau, \theta) + \beta\,\text{MSE}(X_{\text{heldout}}, \theta)\\
&\quad \leq \text{Priv}_{\text{best}} + \beta\,\text{MSE}_{\text{best}}.
\end{aligned}$$

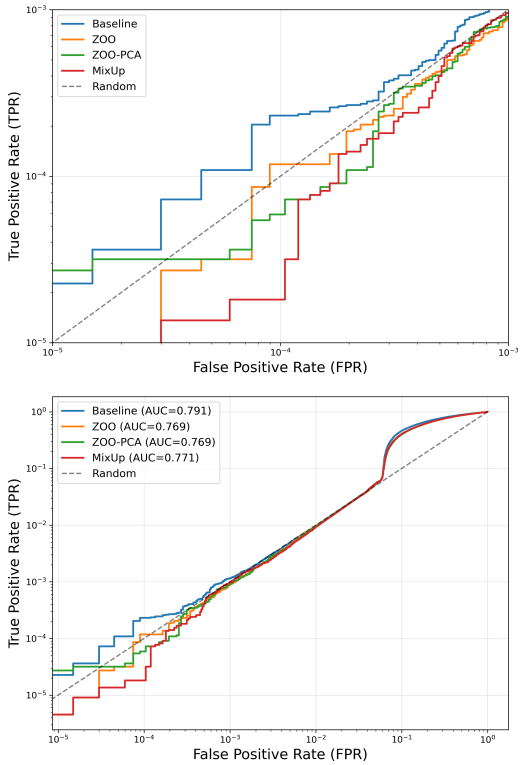

Figure 3: MIMIC-III: ROC curves (log-log scaling) for varying thresholds of loss-based MIAs on models trained with and without data augmentation. Upper plot magnifies the area for FPR $< 0.1\%$. The DP-SGD curve (not shown) is nearly indistinguishable from the diagonal, representing random guessing.

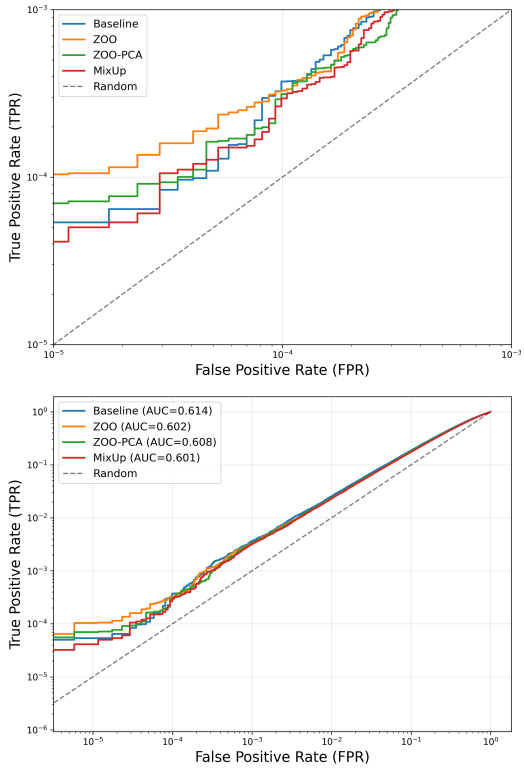

Figure 4: eICU: ROC curves (log-log scaling) for varying thresholds of loss-based MIAs on models trained with and without data augmentation. Upper plot magnifies the area for FPR $< 0.1\%$. The DP-SGD curve (not shown) is nearly indistinguishable from the diagonal, representing random guessing.

These criteria ensure that the model is not updated unless it preserves a balance between utility and privacy: neither the privacy risk nor the held-out MSE may increase beyond a small threshold, and the combined metric must improve. Thus, the optimization does not aim for absolute minima in either privacy or utility alone, but instead for an improved tradeoff between both. In our experiments, we set $\varepsilon_{\text{priv}} = \varepsilon_{\text{MSE}} = 0.5\%$ and $\beta = 3$. Complete meta-parameter settings are given in Appendix B.

7. Experimental Results

In our experiments, all results are reported on unseen test data. The evaluation metric used is MSE (Equation 2) on test data to measure the generalization performance. The effectiveness of the loss-based MIA described in Definition 1 is measured by the TPR/FPR ratio with τ set to the average training loss, and negative examples taken from the test set. Results are reported in Table 1 for the best tradeoff between privacy protection and utility on validation data for each method, respectively.

We present two complementary evaluations of MIA's effectiveness: Table 1 follows Yeom et al.'s (2018) fixed-threshold scenario where τ is set to the average training loss (the only information realistically available to an attacker), while Figures 3 and 4 show Carlini et al.'s (2022) recommended evaluation using ROC curves with varying thresholds, particularly focusing on the low FPR regime where privacy breaches are most critical.

As baseline we use a model that is trained to convergence on the training set without the use of synthetic data.

Table 2: Evaluation of DP-SGD with varying noise multipliers (σ) and fixed clipping norm $C = 2$ on MIMIC-III and eICU datasets.

	MIMIC-III		eICU	
σ	MSE(X_{test})	$\frac{TPR}{FPR}$@τ	MSE(X_{test})	$\frac{TPR}{FPR}$@τ
1.1	0.7446	0.9887	0.8228	0.9969
1.5	0.7655	0.9913	0.9324	0.9976
2.0	0.8403	0.9844	0.9938	0.9945
3.0	0.7952	0.9838	1.0610	0.9929
4.0	0.7788	0.9816	1.0302	0.9952
5.0	0.7900	0.9823	0.9931	0.9942
6.0	0.8044	0.9795	0.9835	0.9966
7.0	0.8150	0.9861	0.9774	0.9963
8.0	0.8320	0.9814	0.9762	0.9951

Furthermore, we compare to DP-SGD (Abadi et al., 2016), a standard method to foster differential privacy by clipping and adding Gaussian noise to the gradient updates. Following Abadi et al. (2016), the privacy accounting (moments accountant) provides meaningful ε-DP bounds primarily for $\sigma > 1$. We implemented DP-SGD with noise multipliers ranging from $\sigma = 1.1$ to $\sigma = 10$, gradient clipping norms $C \in \{1.1, 1.5, 2.0\}$, and increased the learning rate by 100 against the baseline to compensate for the added noise. While DP-SGD achieves strong privacy protection with TPR/FPR ratios approaching 1.0 at any noise levels, this comes at the cost of severe utility degradation. Table 2 shows that even at the lowest noise level ($\sigma = 1.1$), DP-SGD achieves strong privacy protection (TPR/FPR ≈ 1) but at significant cost to utility, with MSE degrading by 47% on MIMIC-III (0.5053) and 47% on eICU (0.5604) compared to baseline. This demonstrates that DP-SGD's noise injection mechanism is poorly suited to high-dimensional, sparse EHR forecasting tasks. While DP-SGD can provide theoretical guarantees during the augmentation phase, it is important to note that our overall system cannot provide end-to-end ε-DP guarantees. This is because our pretrained embedding layer was trained without DP. By the composition theorem (Dwork, 2006), $\varepsilon_{\text{total}} = \varepsilon_{\text{pretrain}} + \varepsilon_{\text{augmentation}}$, and since $\varepsilon_{\text{pretrain}}$ is unbounded, the partial privacy bound has limited practical value in our specific setup. For the results in Table 1, we used $\sigma = 1.1$ and $C = 2$ with a learning rate of 0.005.

The data augmentation methods are ZOO, ZOO-PCA, and MixUp as described in Section 5. The MSE(X_{test}) results in Table 1 show that all data augmentation techniques improve generalization performance on test data compared to the baseline, while best results are achieved for MixUp due to its ability to explore distant data regions while ZOO stays close to training points. This shows MixUp to be a strong technique to improve generalization by data augmentation. The TPR/FPR ratio results in Table 1 show that all data augmentation methods achieve a decrease in the effectiveness of the loss-based MIA, with best results obtained by the ZOO-PCA technique. In our experiments, we keep as many principal components as necessary to explain 70% of the variance. This threshold was found to be optimal by experiments on the held-out set testing 50% to 90% (see Appendix C). This shows that the setting of the ZOO algorithm that prefers synthesis of complex data in order to increase the FPR of the attacker, together with a focus on principal components in embedding space, yields the best tradeoff between generalization and privacy protection.

Figures 3 and 4 show the ROC curves for the loss-based MIA for varying thresholds τ. We see that especially at low FPR values, the TPR/FPR ratio is high for the baseline model, and substantially reduced for the models trained with data augmentation. The ZOO data augmentation method forces the ROC curve of the attacker close to the diagonal, which corresponds to an attacker based on random choice.

For a visual analysis of the privacy-utility tradeoff across different augmentation parameters, see Appendix F.

8. Discussion

Our study demonstrates the effectiveness of data augmentation techniques, particularly zeroth-order optimization, and its PCA-restricted variant, in optimizing both utility and privacy in medical time series forecasting. In fact we found that all data augmentation methods improve privacy against MIA at similar or smaller MSE values on test data, albeit not by huge amounts. In contrast, noise injection on gradients as done in DP-SGD requires noise levels that destroy the model's ability to learn useful patterns. The results presented in Table 1 highlight a significant reduction in the effectiveness of loss-based membership inference attacks through the strategic generation of synthetic data. The ROC curves in Figures 3 and 4 emphasize the privacy benefits of our approach. At low FPR values, where the risk

of privacy breaches is most critical, models trained with ZOO-based augmentation significantly outperform the baseline model. The near-diagonal ROC curve achieved by ZOO indicates that the attacker's performance is close to random guessing, effectively mitigating the risk of MIA.

The ZOO-PCA method, which perturbs embeddings along the directions of principal components, proved particularly effective in reducing the TPR/FPR ratio. We hypothesize that ZOO-PCA directs the augmentation process along the most significant data variations, whereas random perturbations could lead to embeddings that are outside the convex hull of the input embedding space. Since ZOO-PCA effectively performs sparse zeroth-order optimization, it enjoys favorable convergence rates (Balasubramanian and Ghadimi, 2022). Furthermore, directing the augmentation process along the most significant data variations can better confuse the attacker while maintaining or even improving model utility. In contrast, MixUp, a randomization-based approach, demonstrated strong generalization capabilities, achieving the lowest MSE on the held-out set. This highlights the complementary strengths of gradient-free optimization and randomization techniques in data augmentation.

9. Conclusion

Our study demonstrates that embedding-space data augmentation can effectively mitigate membership inference attacks while preserving predictive performance in clinical time series forecasting. ZOO-PCA achieves the best privacy-utility tradeoff, while MixUp excels at generalization, highlighting complementary strengths of different augmentation strategies. In future work we intend to explore hybrid approaches and investigate applicability to other deep learning architectures and privacy attack scenarios.

Acknowledgments

We acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG.

The first author was supported by the Helmholtz Association under the joint research school HIDSS4Health — Helmholtz Information and Data Science School for Health.

References

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 23rd ACM SIGSAC Conference on Computer and Communications Security (CCS)*, Vienna, Austria, 2016. URL https://dl.acm.org/doi/10.1145/2976749.2978318.

Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. *Found. Comput. Math.*, 22(1): 35–76, 2022. URL https://doi.org/10.1007/s10208-021-09499-8.

Nathaniel Berger, Stefan Riezler, Artem Sokolov, and Sebastian Ebert. Don't search for a search method – simple heuristics suffice for adversarial text attacks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Punta Cana, Dominican Republic, 2021. URL https://aclanthology.org/2021.emnlp-main.647.pdf.

T. Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *Annals of Statistics*, 49(5):2825–2850, 2021. URL https://doi.org/10.1214/21-AOS2058.

Chengtai Cao, Fan Zhou, Yurou Dai, Jianping Wang, and Kunpeng Zhang. A survey of mix-based data augmentation: Taxonomy, methods, applications, and explainability. *ACM Comput. Surv.*, 57(2), 2024. URL https://doi.org/10.1145/3696206.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, 2022. URL http://doi.org/10.1109/SP46214.2022.9833649.

Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (AISec)*, Dallas, TX, USA, 2017. URL https://doi.org/10.1145/3128572.3140448.

Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the Twenty-Second ACM Symposium on Principles of Database Systems (PODS)*, San Diego, California, 2003. URL https://doi.org/10.1145/773153.773173.

Cynthia Dwork. Differential privacy. In *Proceedings of the International Colloquium on Automata, Languages and Programming (ICALP)*, Venice, Italy, 2006. URL https://doi.org/10.1007/11787006_1.

Cynthia Dwork. Differential privacy: A survey of results. In *Proceedings of the International Conference on Theory and Applications of Models of Computation (TAMC)*, Xi'an, China, 2008. URL https://doi.org/10.1007/978-3-540-79228-4_1.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9 (3-4):211–407, 2014. URL https://doi.org/10.1561/0400000042.

Michael C. Fu. Gradient estimation. In S.G. Henderson and B.L. Nelson, editors, *Handbook in Operations Research and Management Science*, volume 13, pages 575–616. 2006.

Dominik Hintersdorf, Lukas Struppek, and Kristian Kersting. To trust or not to trust prediction scores for membership inference attacks. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*, 2022. URL https://doi.org/10.24963/ijcai.2022/422.

Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In *Proceedings of the 28th USENIX Conference on Security Symposium (SEC'19)*, Santa Clara, CA, USA, 2019. URL https://www.usenix.org/conference/usenixsecurity19/presentation/jayaraman.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035, 2016. URL https://doi.org/10.1038/sdata.2016.35.

Yigitcan Kaya and Tudor Dumitras. When does data augmentation help with membership inference attacks? In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, virtual, 2021. URL https://proceedings.mlr.press/v139/kaya21a.html.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 2015. URL https://doi.org/10.1007/s10208-015-9296-2.

Eric W. Noreen. *Computer Intensive Methods for Testing Hypotheses. An Introduction.* Wiley, 1989. URL https://www.wiley.com/en-us/Computer+Intensive+Methods+for+Testing+Hypotheses:+An+Introduction-p-9780471611363.

Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. The eICU collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5(180178), 2018. URL https://doi.org/10.1038/sdata.2018.178.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Herve Jegou. White-box vs black-box: Bayes optimal strategies for membership inference. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019. URL https://proceedings.mlr.press/v97/sablayrolles19a.html.

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)*, 2017. URL http://doi.org/10.1109/SP.2017.41.

Michael Staniek, Marius Fracarolli, Michael Hagmann, and Stefan Riezler. Early prediction of causes (not effects) in healthcare by long-term clinical time series forecasting. *Proceedings of Machine Learning Research: Machine Learning for Healthcare*, 252:1–29, 2024. URL https://arxiv.org/abs/2408.03816.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, Long

Beach, CA, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022. URL https://doi.org/10.1145/3548606.3560675.

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)*, 2018. URL http://doi.org/10.1109/CSF.2018.00027.

Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. How does data augmentation affect privacy in machine learning? In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, virtual, 2021. URL http://doi.org/10.1609/aaai.v35i12.17284.

Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-cost high-power membership inference attacks. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL https://proceedings.mlr.press/v235/zarifzadeh24a.html.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

Appendix A. Data

MIMIC-III: MIMIC-III was collected from the Beth Israel Deaconess Medical Center between 2001 and 2012 and contain over 40k patients. After filtering for patients with an ICU stay of at least 24 hours with reported gender and age of at least 18 years, our dataset contains 44,858 ICU stays with 56 million data points. We split the data into partitions for training (28,791), held-out (7,144), and test (8,878). For our study, we used 131 different clinical variables. The full list of extracted MIMIC-III features is given in Appendix D.

Converting the data to a dense one hour representation yields 89.08% missing data, changing per variable from under 15% (HR, RR, SBP, DBP, MBP, and O2 Saturation) to more than 90% for 101 variables, and exceeding 99% for 42 variables. On the other side we are losing 17.73% of the data points through the densification procedure where multiple measurements occur within the same hour and for the same variable. The high sparsity of the MIMIC-III data presents a significant challenge for time series forecasting, requiring careful handling of missing values and the development of robust models.

eICU: The eICU data was collected from over 200 US hospitals and comprise over 200,000 ICU stays. After filtering for patients with an ICU stay of at least 48 hours, reported gender and aged 18 years or older, we arrived at 77,704 ICU stays with 415 million data points. This set was partitioned in subsets for training (49,730), development (12,433), and testing (15,541). As shown in Appendix E, we extracted 98 clinical variables for our experiments. The measurements in the eICU data set are denser than in MIMIC-III since the number of observations per patient per hour is three times higher than for MIMIC-III and decreases at a slower rate with length of stay, also there are a lot of bed values. After our binning process, this reduces to one sixth of the measurements (16.86%) leading to still 89.85% missing data. The same six variables (HR, RR, SBP, DBP, MPB, and O2 Saturation) as before are quite complete.

Appendix B. Meta-parameters

The following meta-parameter settings were used in our experiments:

Table 3: Hyperparameter configurations

ZOO and ZOO-PCA parameters

λ	μ	k	`steps`
3000	300	3	10

Variants of ZOO/ZOO-PCA (α) and MixUp (β)

α	β
{0, 0.25, 0.5, 0.75, 1}	{0.2, 1, 5}

Acceptance Criteria

$\varepsilon_{\text{priv}}$	ε_{MSE}	β
0.5%	0.5%	3

DP-SGD parameter

Noise Multiplier	Clipping Norm
{1.1, 1.5, 2}	{1.1, 1.5, 2}
{2}	{3,4,5,6,7,8,16,50}

Appendix C. Evaluation of ZOO-PCA variance explanation

Table 4: Evaluation results for baseline and PCA ZOO data augmentations. The most fitting variance explanation parameter was selected via grid search on validation data from $\{50\%, 70\%, 90\%\}$. Compare to Table 1.

	MIMIC-III		eICU	
PCA ratio	MSE(X_{test})	$\frac{TPR}{FPR}@\tau$	MSE(X_{test})	$\frac{TPR}{FPR}@\tau$
Baseline	0.5053	3.5479	0.5604	1.3206
50%	0.5037	1.713	0.5602	1.1526
70%	0.5000	**1.4271**	0.5607	**1.1299**
90 %	**0.4994**	1.455	**0.5597**	1.1612

Appendix D. MIMIC-III features

Table 5: For MIMIC-III, 131 dynamic variables were extracted.

ALP	Epinephrine	LDH	Packed RBC
ALT	Famotidine	Lactate	Pantoprazole
AST	Fentanyl	Lactated Ringers	Phosphate
Albumin	FiO2	Levofloxacin	Piggyback
Albumin 25%	Fiber	Lorazepam	Piperacillin
Albumin 5%	Free Water	Lymphocytes	Platelet Count
Amiodarone	Fresh Frozen Plasma	Lymphocytes (Absolute)	Potassium
Anion Gap	Furosemide	MBP	Pre-admission Intake
BUN	GCS_eye	MCH	Pre-admission Output
Base Excess	GCS_motor	MCHC	Propofol
Basophils	GCS_verbal	MCV	RBC
Bicarbonate	GT Flush	Magnesium	RDW
Bilirubin (Direct)	Gastric	Magnesium Sulfate (Bolus)	RR
Bilirubin (Indirect)	Gastric Meds	Magnesium Sulphate	Residual
Bilirubin (Total)	Glucose (Blood)	Mechanically ventilated	SBP
CRR	Glucose (Serum)	Metoprolol	SG Urine
Calcium Free	Glucose (Whole Blood)	Midazolam	Sodium
Calcium Gluconate	HR	Milrinone	Solution
Calcium Total	Half Normal Saline	Monocytes	Sterile Water
Cefazolin	Hct	Morphine Sulfate	Stool
Chest Tube	Heparin	Neosynephrine	TPN
Chloride	Hgb	Neutrophils	Temperature
Colloid	Hydralazine	Nitroglycerine	Total CO2
Creatinine Blood	Hydromorphone	Nitroprusside	Ultrafiltrate
Creatinine Urine	INR	Norepinephrine	Urine
D5W	Insulin Humalog	Normal Saline	Vancomycin
DBP	Insulin NPH	O2 Saturation	Vasopressin
Dextrose Other	Insulin Regular	OR/PACU Crystalloid	WBC
Dobutamine	Insulin glargine	PCO2	Weight
Dopamine	Intubated	PO intake	pH Blood
EBL	Jackson-Pratt	PO2	pH Urine
Emesis	KCl	PT	
Eosinophils	KCl (Bolus)	PTT	

Appendix E. eICU features

Table 6: For eICU, 100 variables were extracted. The 35 variables on the right column are drug-related. Some of them seem redundant due to different hospitals but can not be merged because of different or not standardized concentrations.

ALP	Lactate	Amiodarone
ALT	Lymphocytes	Dobutamine dose
AST	MBP	Dobutamine ratio
Albumin	MCH	Dopamine dose
Anion Gap	MCHC	Dopamine ratio
BUN	MCV	Epinephrine dose
Base Deficit	MPV	Epinephrine ratio
Base Excess	Magnesium	Fentanyl 1
Basophils	Monocytes	Fentanyl 2
Bedside Glucose	Neutrophils	Fentanyl 3
Bicarbonate	O2 L/%	Furosemide
Bilirubin (Direct)	O2 Saturation	Heparin 1
Bilirubin (Total)	PT	Heparin 2
Bodyweight (kg)	PTT	Heparin 3
CO2 (Total)	PaCO2	Heparin vol
Calcium	PaO2	Insulin 1
Chloride	Phosphate	Insulin 2
Creatinine (Blood)	Platelets	Insulin 3
Creatinine (Urine)	Potassium	Midazolam 1
DBP	Protein (Total)	Midazolam 2
Eosinophils	RBC	Milrinone 1
EtCO2	RDW	Milrinone 2
FiO2	RR	Nitroglycerin 1
Fibrinogen	SBP	Nitroglycerin 2
GCS eye	Sodium	Nitroprusside
GCS motor	Stool	Norepinephrine 1
GCS verbal	Temperature	Norepinephrine 2
Glucose	Troponin - I	Norepinephrine ratio
HR	Urine	Pantoprazole
Hct	WBC	Propofol 1
Hgb	pH	Propofol 2
INR		Propofol 3
		Vasopressin 1
		Vasopressin 2
		Vasopressin 3

Appendix F. Privacy-Utility Tradeoff Visualization

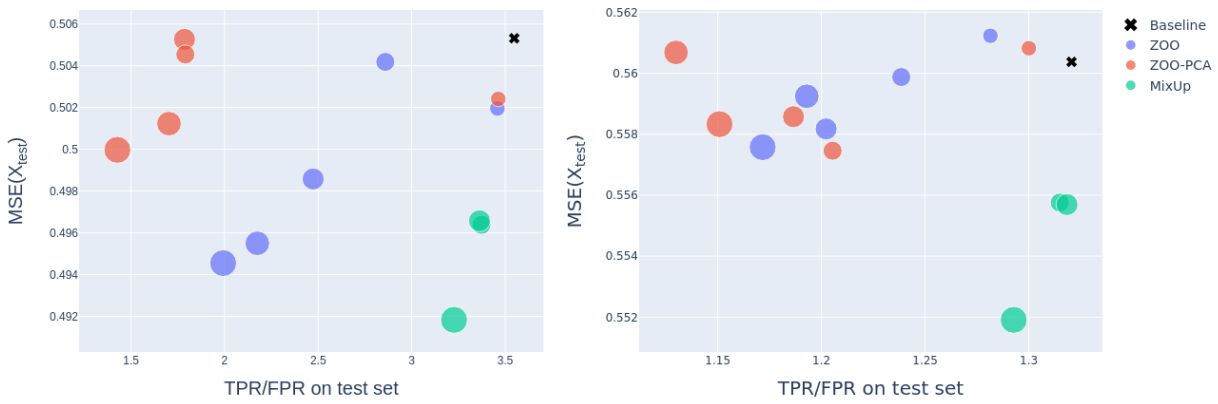

Figure 5: TPR/FPR ratio of MIAs against generalization performance on test data on MIMIC-III (top) resp. eICU (bottom). The size of the interpolation parameter $\alpha \in \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}$ for ZOO and ZOO-PCA, resp. $\beta \in \{0.2, 1, 5\}$ for MixUp in data augmentation is indicated by the size of the ball.

Figure 5 plots the TPR/FPR ratio of the loss-based MIA against generalization performance on test data on MIMIC-III. The focus on improvements in mMSE or on privacy protection is shown by the size of the interpolation parameter $\alpha \in \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}$. Best results are achieved with higher values of α.