# Protecting User Data Privacy with Adversarial Perturbations

Ziqi Wang, Brian Wang, Mani B. Srivastava (University of California, Los Angeles)

NESL Networked & Embedded Systems Laboratory
UCLA

## Introduction

- The increased availability of on-body sensors gives researchers access to rich time-series data, many of which are related to human health conditions.
- Advanced data-driven machine learning models can make rich health-related inferences to help improve human well-being. Data-driven approaches benefit from widely accessible large datasets.

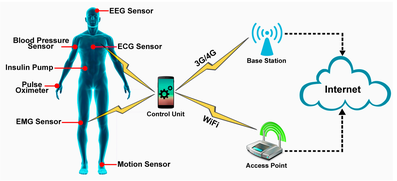




Figure Source: https://eecs.ceas.uc.edu/MNCL/images/slideshow/body-sensor.jpg https://md2k.org/

**Figure 1:** Data-driven algorithms require large datasets to achieve accurate health-related inferences

- Clinical datasets often contain sensitive information about their contributors. For example, health conditions, symptoms, and genetic disease. **Sharing of such data is strictly restricted because of privacy concerns.**
- One potential solution to this problem is Synthetic Generation, i.e., generate a synthetic dataset using the original dataset. (See Figure 2 left).
- Two requirements for synthetic data: **Utility** (the data remains useful for making health inferences) and **Privacy** (the dataset obfuscates the identities of the users and can be shared publicly).

## Hide-and-Seek Privacy Challenge

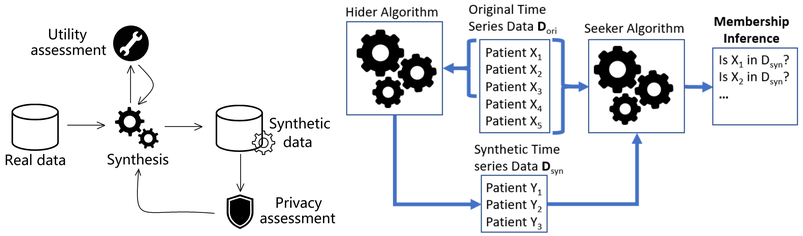






**Figure 2:** Synthetic generation (Left) and the NeurIPS 2020 Hide-and-seek Challenge setup

- In this work, we propose our solution to the NeurIPS 2020 Hide-and-Seek Privacy Challenge [2], in which there are two tracks: hiders and seekers.
- Hiders generate synthetic data from the AmsterdamUMCdb dataset, which contains data entries (multimodal time series) from ~ 20000 ICU patients.
- This competition [2] formulates privacy as robustness to membership inference attacks, where the goal of the seeker (attacker) is to infer whether or not a given data entry is used to generate the synthetic dataset.

- In terms of identity, ideal synthesizing process (hiders) should create new users that never exist. In the challenge, a subset of users in the raw dataset is used by the hiders for synthetic generation.
- If seekers can be successfully identified this subset of users with a membership inference attack, then there is a link between identities of these users and the synthetic users, indicating a privacy risk.
- Our proposed algorithm is a hider aiming to generate data with privacy and utility.

## Proposed Method

- We propose using adversarial perturbations to protect user privacy inspired by [3].
- A trained feature extraction neural network generates a feature embedding for each user's time series. This embedding is trained to represent the user's identity.
- A small perturbation (namely adversarial perturbations) in the raw time series may cause its embedding to change drastically because of the discontinuity of the function approximated by the neural network. The added noise is trained to "drag" the feature embedding of the current user towards the embedding of a different user, so that the network is confused about their identities.

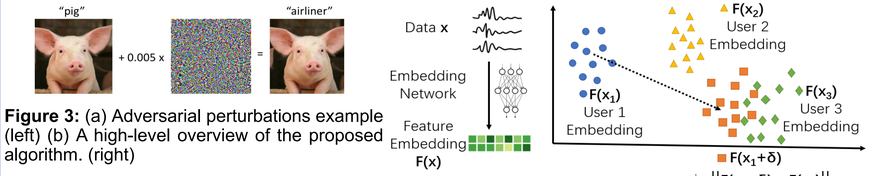
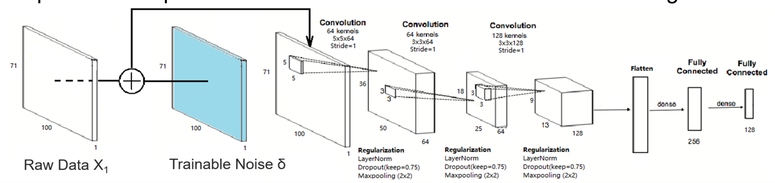



**Figure 3:** (a) Adversarial perturbations example (left) (b) A high-level overview of the proposed algorithm. (right)

- The proposed algorithm can be described in three steps:
- **Step 1:** Train a Convolutional Neural Network (CNN) feature extractor using a Siamese structure. The feature extractor works to recognize the user and maximize the embedding distances between different users using contrastive loss.
- **Step 2:** Define a CNN with exactly the same structure as previously, except that at the input layer, we add a layer of the same dimension to inject additive noise. Then we copy and fix the trained CNN weights. The only learnable parameter is the added noise $\boldsymbol{\delta}$ (See the Figure below).
- **Step 3:** For each user (for example, data entry $\boldsymbol{x}_1$ in Figure 3(b)), select a data entry $\boldsymbol{x}_3$ from another user, whose embedding $\boldsymbol{F}(\boldsymbol{x}_3)$ is distant from $\boldsymbol{F}(\boldsymbol{x}_1)$. Then we train the noise $\boldsymbol{\delta}$ using gradient descent to move the current embedding $\boldsymbol{F}(\boldsymbol{x}_1 + \boldsymbol{\delta})$ towards embedding $\boldsymbol{F}(\boldsymbol{x}_3)$. Then $\boldsymbol{x}_1 + \boldsymbol{\delta}_{opt}$ is returned as the generated data entry. This process is repeated for all the users to scramble their embeddings.




## Baselines and Metrics

- We compare our proposed algorithm against three baselines:
- **Add Noise** adds an i.i.d. Gaussian noise to every value in the dataset with zero mean and a standard deviation as a hyper-parameter.
- **TimeGAN** [4] seeks to use Generative Adversarial Networks to capture temporal dynamics of the time series and generate new data.
- **Genetic** is inspired by [1]. This approach uses the genetic algorithm, mutating user data by adding noise. In this baseline we train several models checking the utility of mutated data for the natural selection process.

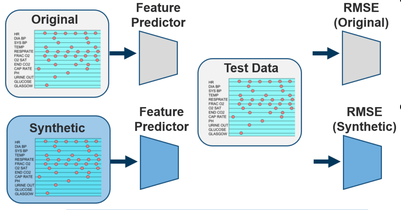



**Figure 5:** Synthetic data utility evaluation

- Each data entry has two dimensions: features and time. For utility, we train autoregressive NN predictors for a masked feature or a masked timestep.
- The test error (RMSE) of such predictors are demoted as **Feature Prediction Error** and **One-Step-Ahead Prediction Error**. The performance of models trained on synthetic data should be comparable to that trained with real data (Figure 5).
- In terms of privacy, we use the maximum re-identification (Re- ID) score across all nine seekers participating in this challenge as the metric. The Re-ID score is the percentage of the correct re-identifications made by a seeker.

## Evaluation Results

- We evaluate hiders from two aspects: utility and privacy (see Table 1).
- In the feature prediction test, only our proposed adversarial perturbation method passed all tests compared to the baselines. In terms of the one-step-ahead prediction, all the four methods pass.
- The last column, Re-ID score, shows the performance of privacy protection where a lower score is better. Our proposed method outperforms all the three baselines in front of 9 different seekers.

| Data Generator | Feature Prediction | One-step-ahead Prediction | Re-ID Score |
|---|---|---|---|
| Add Noise | 8/10 Pass | Pass | 0.5734 |
| TimeGAN | 6/10 Pass | Pass | 0.5047 |
| Genetic | 8/10 Pass | Pass | 0.6315 |
| Adversarial(Ours) | **10/10 Pass** | Pass | **0.5037** |

**Table 1: Evaluation Results**

References:
[1] Moustafa Alzantot, Bharathan Balaji, and Mani Srivastava. 2018. Did you hear that? adversarial examples against automatic speech recognition. arXiv preprint arXiv:1801.00554 (2018).
[2] James Jordon, et. al, 2020. Hide-and-Seek Privacy Challenge. arXiv preprint arXiv:2007.12087 (2020).
[3] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. 2020. Fawkes: Protecting privacy against unauthorized deep learning models. In 29th USENIX Security Symposium. 1589–1604.
[4] Jinsung Yoon, Daniel Jarrett, and Mihaela van der Schaar. 2019. Time-series generative adversarial networks. (2019).



Contact: wangzq312@g.ucla.edu wangbri1@g.ucla.edu (Equal Contribution)

GitHub: https://github.com/nesl/hide-and-seek-challenge

NSF NIH DOT SRC CONIX