A Privacy Filter Framework for Internet of Robotic Things Applications

Zahir Alsulaimawi
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR 97331
alsulaiz@oregonstate.edu

***Abstract*—Traditionally robots have been stand-alone systems. In recent years, however, they have increasingly been connected to external knowledge resources through the Internet of Things (IoT). These robots are thus becoming part of IoT and can realistically allocate Internet of Robotic Things (IoRT) technologies. IoRT can facilitate Human-Robot Interaction (HRI) at functional (commanding and programming) and social levels, as well as a means for remote-interaction. IoRT-HRI can cause privacy issues for humans, in part because robots can collect data using IoT and move in the real world, partly because robots can learn to read human social cues and adapt or correct their behavior accordingly. In this paper, we address the topic of privacy-preserving for IoRT-HRI applications. The objective is to design a data release framework called a Privacy Filter (PF) that can prevent an adversary from private mining information from the released data while keeping utility data. In the experiments, we test our framework on two accessible datasets: MNIST (handwritten digits) and UCI-HAR (activity recognition from motion). Our experimental results on these datasets show that PF is highly effective in removing private information from the dataset while allowing utility data to be mined effectively.**

***Index Terms*—Privacy-preserving, probabilistic model, IoRT, variational mutual information, deep learning.**

I. Introduction

The Internet of Things (IoT) is a network of Internet-enabled devices that can sense, communicate, and react to changes in their environment. There are various highly integrated goals so far pushing the IoT and robotics communities; the first focuses on supporting information services for remote sensing, tracking, and widespread monitoring; the latter in work production, interaction, and independent behavior. For this reason, it is increasingly demanded that the creation of an Internet of Robotic Things (IoRT) that combines the outcomes of the two communities will bring significant added value. [1] [2]. Human-Robot Interaction (HRI) is a field of study assigned to the understanding, design, and evaluation of automated systems for use by or with humans. HRI could then be implemented into a robot used for assisted living in a care home, senior houses, etc.. The arrival of IoT and, in particular, IoRT implies that HRI should now consider the countless complex interaction scenarios that stem from new interconnected systems such as robots, sensors, and humans. In the era of IoRT-HRI, customers are expected to be continuously monitored by a large number of robots, and the leak of raw robot data to an adversarial entity can significantly undermine the privacy of customers, primarily due to the intimate nature of the data [3].

There has been a recent rise in high profile cyber-attacks able to shut down or corrupt the actions of IoT devices that control equipment or interact in some other way with the physical world. The information leaks remind those IoRT applications, without a proper defense mechanism, they can be subject to similar attacks; the privacy concern needs to be adequately addressed before the deployment of IoRT applications. The majority of solution techniques to secure privacy are based on cryptography via the use of secret keys [4] [5]. However, in IoT applications, it is challenging to design a secure, efficient, and reliable mechanism for key exchange/distribution among the vast network of heterogeneous devices [6].

We introduce a novel practical framework, the Privacy Filter (PF), that removes sensitive information to improve privacy protection without significantly decreasing utility in IoRT-HRI applications. To address this issue, we consider a formulation to learn an autoencoder as PF, through adversarial training against adversaries classifiers that is simultaneously training to succeed at recovering the private information from disclosed data. The autoencoder, in turn, trains to prevent these inferences, while also maintaining non-private data from disclosing to gain some utility. Our work is based on a proposed concept called the ”privacy funnel” [7] to represent the trade-off between data utility and data privacy. In this work, we introduce a simple but efficient variational approach that relies on adversarial training. Our empirical results show that our approach is stable and outperforms in terms of both accuracy and mutual information estimation.

Main Contributions: The contributions of this paper are as follows.

- We present a novel supervised learning algorithm of privacy-preserving for privacy funnel using deep convolution neural networks.
- We introduce a general framework to capture the problems of privacy-utility trade-off, which is given in terms of mutual information. We explicitly formulate the issue in a novel way that we feel deserves further study.
- Our approach helps in understanding and interpreting the relation between deep networks and information theory, and we hope it contributes to a foundation for such knowledge.
- This paper presents the privacy-preserving issue as a Bayesian Network (BN) model.
- We demonstrate through experimental analysis using real-word datasets that our proposed solutions, for a privacy-preserving, outperform in terms of accuracy.

II. Related Work

The IoT is a novel paradigm in the scenario of modern wireless telecommunications that incorporates billions of devices that are owned by different organizations and people who

DOI 10.1109/SPW50608.2020.00059

are deploying and using them for their purposes. However, it raises serious concerns over individual privacy in the new environment of smart things [8]. Anonymization of data is the process of removing specific information that may lead to personal identification so that the persons/objects described in this data remain anonymous. Numerous attempts have been made to provide anonymization and image clarity mechanisms for IoT applications, especially for images and videos [9]. Encryption is based on complex algorithms called ciphers. The primary purpose of any encryption method is to keep sensitive information secret from others by processing readable data into a long series of random or pseudo-random ciphers. Many IoT devices currently use the encryption protocol to protect their dada, e.g., the health-care industry [10].

Within the private preserving framework, there exist privacy-preserving data mining (PPDM) techniques in the database community [11] [12] [13] whose goal is to prevent association of any instance in a database to a person. In addition to PPDM, many privacy-preserving machine learning (PPML) techniques [14],[15], [16], [17], [18], [19] have been proposed to deal with data beyond those in the traditional databases. Most existing PPML literature focuses on ensuring that private data can not be mined and make no premise about the non-private data. On the other hand, our work assumes pre-specified sets of private and non-private data. Such a formulation not only makes the proposed data sanitization more effective, but also provides a a flexible trade-off between privacy and the ability to mine non-private information from the sanitized data.

Within the private preserving framework, there exist privacy-preserving data mining (PPDM) techniques in the database community [11] [12]. In [20], describes many definitions and models for privacy-preserving computation, and compares several different approaches. The most prominent, however, is given by differential privacy (DP) [21]. This guarantee requires that algorithms operating on data sets consisting of nearly the same individuals should yield similar results with a high probability. In this case, the dependency of the algorithm's output on a particular individual is low and does not leak information. To protect against record linkage attacks, Samarati and Sweeney [22] proposed k-anonymity. This privacy model states that every record in the published data table must be indistinguishable from at least $k-1$ other records over the as quasi-identifiers attributes.

Several studies investigate feature selection as a tool for obtaining privacy for sensitive data [23] [24]. The idea is not to release the entire information in the data, but only selected features. Unlike these studies in [25] consider the privacy of information that can be extracted from a single feature vector by zeroing out feature components in the approximate null space of the linear operator. The framework proposed in [26] transforms the data in a way that the covariance between the data and the desired information is increased, while the covariance between the data and confidential information is decreased.

Privacy-preserving has been addressed from an information theoretic viewpoint in [7] [27] [28] [29] [30] where both utility and privacy are measured in terms of information-theoretic quantities.

III. Preliminaries

In this section, we first give some background knowledge of the cross-entropy (CE) loss function. Then, we introduce some concepts of information theory, that are used in the design of our approach.

A. Cross-entropy Loss Function

Cross-entropy (CE) loss function measures the performance of a model whose output is a probability value between 0 and 1. It is defined as, $CE(\hat{y},y) = \mathbb{E}_y[-\log(\hat{y})]$ where $\hat{y} \in \{0,1\}$ is the predicted probability (a.k.a. data), and $y \in \{0,1\}$ is the class label (a.k.a. model). CE is also called negative log-likelihood.

B. Information Theory

We adopt the same notation for information theoretic quantities used in [31]. Specifically, the symbol H will denote entropy, I mutual information, and KL Kullback-Leibler divergence. We briefly recall those concepts that we will use in this paper.

- The entropy of a discrete random variable X with probability mass function P is a
$$H(X) = \mathbb{E}_X[-\log P(X)]$$
- Let $P(X)$ and $Q(X)$ be two probability distributions over the same alphabet. The KL divergence $KL(P||Q)$ is a measure of their discrepancy. It can be defined as
$$KL(P||Q) = \mathbb{E}_P\big[\log\frac{P(X)}{Q(X)}\big]$$
- The mutual information $I(X;Y)$ of two random variables X and Y is a measure of the amount of information that one random variable contains about the other, satisfying $I(X;Y) \geq 0$, with equality if and only if, X and Y are mutually independent. It is defined as the KL divergence between the joint distribution $P(X,Y)$ and the independent distribution $P(X)P(Y)$ generated by the marginal ones
$$\begin{aligned} I(X;Y) &= KL(P(X,Y)||P(X)P(Y)) \\ &= \mathbb{E}_{X,Y}\big[\log\frac{P(X,Y)}{P(X)P(Y)}\big] \\ &= H(X) - H(X|Y) = H(Y) - H(Y|X) \end{aligned}$$

IV. Setting

A. Problem Statement

Assuming data $X \in \mathbb{R}^n$ is Random Variable (RV) of continuous high-dimensional raw data, and it has private label $S = (s_1, s_2, ..., s_m)$ where s_i represents the ith private task for an adversary, e.g., gender or race. The private label can be can be discrete, continuous, and/or high-dimensional vector. Let $P_{\hat{X}|X}$ PF, which is a probabilistic privacy mapping converting

X into $\hat{X} \in \mathbb{R}^n$, a disclosed data. In a privacy preserving data release, the goal is to find a probabilistic mapping $P_{\hat{X}|X}$ such that releasing $\hat{X}$ will not violate the privacy of individuals. Without privacy in mind, we could think of this as features transformation. This framework is specified by joint probability function $P_{X,\hat{X},S} = P_X P_{\hat{X}|X} P_{S|\hat{X},X}$. For privacy-preserving we want S to be independent of X for a given $\hat{X}$. So that the joint probability of our approach can be factorized into $P_{X,\hat{X},S} = P_X P_{\hat{X}|X} P_{S|\hat{X}}$, where P_X is a raw data, $P_{\hat{X}|X}$ is PF inference, and $P_{S|\hat{X}}$ is an adversary inference, which form a Bayesian network as shown in Figure 1.

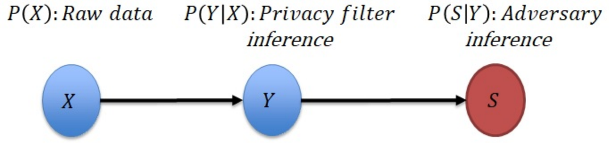

$P(X)$: *Raw data*
$P(Y|X)$: *Privacy filter inference*
$P(S|Y)$: *Adversary inference*
X
Y
S

Fig. 1. An illustration of Bayesian network

Our approach depends on privacy funnel concept to obtain privacy-preserving mapping $P_{\hat{X}|X}$. Briefly, the privacy funnel assumes that the original data X is transformed into $\hat{X}$ before disclosing and the log loss is used for both privacy and utility metric. Then the problem can be modeled as finding a mapping $X \rightarrow \hat{X}$ that maximizes mutual information between X and $\hat{X}$ subject to a constraint that the mutual information between $\hat{X}$ and private data S is smaller than a predefined threshold ε.

B. Threat Model

There are several situations where IoRT-HRI privacy-preserving is an issue. We introduce the problem through the practical scenario shown in Figure 1. In a remote health monitoring system, patients are continuously monitored by robots in their residential space, and the system object is to be able to detect indicators or symptoms of medical conditions based on sensor measurements. The robot collects the data and sends them to the specialists through the Local Area Network (LAN) or the internet. While the robot provides information about patients' medical conditions, it may also convey sensitive information that they do not want to share. For example, motion sensor data might disclose the weight or gender of a user or enable their re-identification. Also, these robots present a video record and chat interface with the addition of a mobile base so the remote operator can look around and even drive from place to place. It seems clear that robot in a remote health monitoring system, cause concerns about privacy [32] [33] [34]. The proposed PF can be potentially applied to the robot's sensors to filter out information about private data while guaranteeing that the disclosed data, can be used to detect medical events with high accuracy.

V. THE PROPOSED APPROACH

Let us consider we pass X through a probabilistic mapping $P_{\hat{X}|X}$ to reveal $\hat{X}$ to the public. The purpose of this mapping is to make $\hat{X}$ informative about X and uninformative about S. In other words, we want to design PF $P_{\hat{X}|X}$ to maximize the amount of information $I(X;\hat{X})$ that the user discloses about the public information, X, while minimizing the collateral information about the private variable S measured by $I(S;\hat{X})$. The trade-off between disclosure and privacy in the design of PF is represented by the following optimization

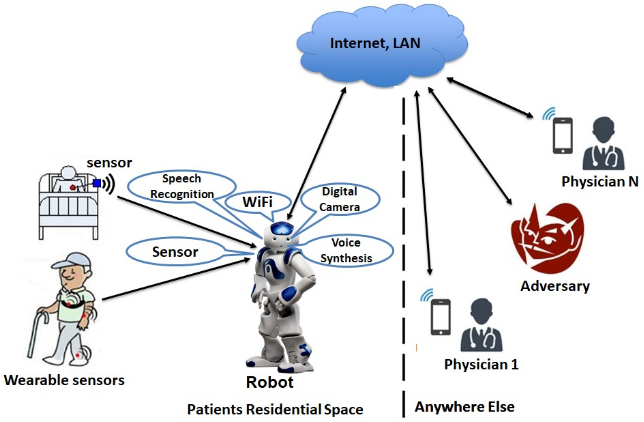


Fig. 2. Attack model: Remote health monitoring system

$$\begin{aligned} P_{\hat{X}|X} = \underset{P_{\hat{X}|X} \in \mathbb{P}}{\arg\min} \quad & I(S;\hat{X}) \\ \text{s.t.} \quad & I(X,\hat{X}) \geq R \end{aligned} \tag{1}$$

where R is the disclosure level, and $\mathbb{P}$ is the set of all possible probabilistic mapping for PF. The constraint in (1) can be written as $H(\hat{X}) - H(\hat{X}|X) \geq R$. So that (1) can be rewritten as

$$\begin{aligned} P_{\hat{X}|X} = \underset{P_{\hat{X}|X} \in \mathbb{P}}{\arg\min} \quad & I(S;\hat{X}) \\ \text{s.t.} \quad & H(\hat{X}|X) \leq D \end{aligned} \tag{2}$$

where $D = H(\hat{X}) - R$. By introducing a Lagrange multiplier $\beta > 0$, we can express (2) as the variational minimization problem of finding

$$P_{\hat{X}|X} = \underset{P_{\hat{X}|X} \in \mathbb{P}}{\arg\min} \quad (I(S;\hat{X}) + \beta H(\hat{X}|X)) \tag{3}$$

As we cannot practically search over all possible probabilistic mapping $\mathbb{P}$, we consider a transform $T_\theta(X) : X \rightarrow \hat{X}$, where θ is the parameter set, is a type of Artificial Neural Network (ANN) to approximate the required $P_{\hat{X}|X}$ and look for the optimal parameter set through training. The network optimizer finds the optimal parameter set θ^* by searching the space of all the possible parameter set, Θ, as

$$\theta^* = \underset{\theta \in \Theta}{\arg\min}(I(S;\hat{X}) + \beta H(\hat{X}|X)) \tag{4}$$

First term of (4): Let's find a variational lower bound of mutual information between $\hat{X}$ and S

$$I(S;\hat{X}) = H(S) - H(S|\hat{X}) \tag{5}$$

$$I(S;\hat{X}) = H(S) + \mathbb{E}_{\hat{X}}\mathbb{E}_{S|\hat{X}}[\log P(S|\hat{X})] \tag{6}$$

In practice, the mutual information term $I(S;\hat{X})$ is hard to minimize directly as it requires access to the posterior $P(S|\hat{X}) = \frac{P(S,\hat{X})}{\int_S P(S,\hat{X})ds}$. The marginalization over S to calculate $P(\hat{X})$ in the denominator is typically intractable because this integral is unavailable in closed form. Fortunately, we can obtain a lower bound of $I(S;\hat{X})$ by defining a parametric probability distribution $Q_\phi(S|\hat{X})$ to approximate $P(S|\hat{X})$. We define $Q_\phi(S|\hat{X})$ as an ANN having weights and biases both are represented by ϕ.

$$I(S;\hat{X}) = H(S) + \mathbb{E}_{\hat{X}}\mathbb{E}_{S|\hat{X}}\left[\log \frac{Q_\phi(S|\hat{X})P(S|\hat{X})}{Q_\phi(S|\hat{X})}\right] \tag{7}$$

$$\begin{aligned} I(S;\hat{X}) = H(S) + \mathbb{E}_{\hat{X}}\mathbb{E}_{S|\hat{X}}[\log Q_\phi(S|\hat{X})] \\ + \mathbb{E}_{\hat{X}}\mathbb{E}_{S|\hat{X}}\left[\log \frac{P(S|\hat{X})}{Q_\phi(S|\hat{X})}\right] \end{aligned} \tag{8}$$

$$\begin{aligned} I(S;\hat{X}) = H(S) + \mathbb{E}_{S,\hat{X}}[\log Q_\phi(S|\hat{X})] \\ + \mathbb{E}_{\hat{X}} KL[P(S|\hat{X})||Q_\phi(S|\hat{X})] \end{aligned} \tag{9}$$

The KL divergence is a non-negative value that indicates how close two probability distributions are, therefore the lower bound to hold is

$$I(S;\hat{X}) \geq H(S) + \mathbb{E}_{S,\hat{X}}[\log Q_\phi(S|\hat{X})] \tag{10}$$

If $P(S|\hat{X}) = Q_\phi(S|\hat{X})$, the KL divergence is zero and the bound is tight. So, with the constant $H(S)$ term dropped, we can write this lower bound alternatively in the following way

$$I(S;\hat{X}) = \max_{\phi\in\Phi} \mathbb{E}_{S,\hat{X}}[\log Q_\phi(S|\hat{X})] \tag{11}$$

The max problem in equation (11) is the objective function of the adversary.

Second term of (4): The conditional entropy of $\hat{X}$ given X can be written as

$$H(\hat{X}|X) = \mathbb{E}_{\hat{X},X}[-\log Q_\theta(\hat{X}|X)] \tag{12}$$

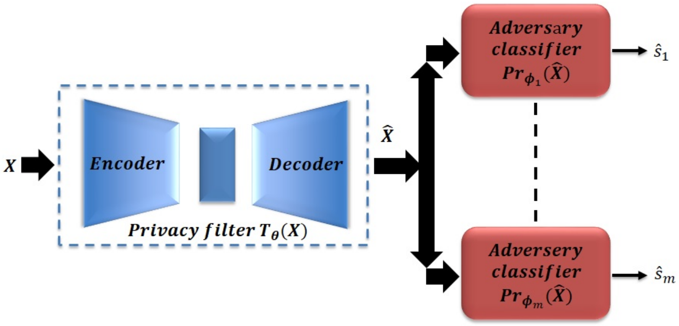
sub (11) and (12) in (4) we can find the multi-objective loss function of our approach as

$$\begin{aligned} \theta^* = \arg\min_{\theta\in\Theta}\max_{\phi\in\Phi} \mathbb{E}_{S,\hat{X}}[\log Q_\phi(S|\hat{X})] \\ + \beta\mathbb{E}_{\hat{X},X}[-\log Q_\theta(\hat{X}|X)] \end{aligned} \tag{13}$$

We obtain θ^* using backpropagation with stochastic gradient descent (SGD) and the multi-objective loss function. Our minimax formulation in (13) is similar to a Generative Adversarial Network (GAN) [35] objective function. It can be interpreted as PF wants to minimize the privacy loss, while the adversary is trying to maximize privacy loss. This optimization problem can be practically addressed via the training of two neural networks: PF as an autoencoder, and an adversary $Q_\phi(S|T_\theta(X))$ as classifier. To make the notations simple, we define PF as $T_\theta(X)$, which is equal to $\hat{X}$, and adversary classifier $Q_\phi(S|\hat{X})$ as $Pr_\phi(T_\theta(X))$. The equation (13) can be rewritten with help of the CE loss function as

$$\min_\theta \; (\beta CE(X, T_\theta(X)) - \textstyle\sum_{i=1}^{m} \min_{\phi_i} CE(s_i, Pr_{\phi_i}(T_\theta(X))) \tag{14}$$

which is the objective function of our approach. The objective function is close to to a adversary task for small $\beta \ll 1$ and for large $\beta \gg 1$ is close to utility task. The solution to (14) will refer to as PF and is define as optimal filter for privacy-utility tradeoffs in term of autoencoder as PF and adversary classifiers. The architecture of PF framework is illustrated in Figure 3. The algorithmic approach that we use to solve the optimization in (14) are detailed in algorithm 1.



Adversary classifier $Pr_{\phi_1}(\hat{X})$
$\hat{s}_1$
X
Encoder
Decoder
Privacy filter $T_\theta(X)$
$\hat{X}$
Adversery classifier $Pr_{\phi_m}(\hat{X})$
$\hat{s}_m$

Fig. 3. Schematic diagram of the proposed PF framework. Training alternates between optimizing the weights of adversaries classifiers keeping PF fixed and vice-versa.

VI. EXPERIMENTS

To evaluate our work, we use two datasets: MNIST (handwritten digits) dataset [36] and UCI Human Activity Recognition (UCI-HAR) dataset [37]. The networks were trained using the Pytorch deep learning platform using the Adam optimizer with a learning rate of 0.0001. We set β equal to 1 and 0.5 for MNIST and UCI-HARA respectively and $b = 64$ for both datasets. We use $k = 2$ for MNIST while $k = 3$ for UCI-HAR.

To evaluate a trained PF, we implement utility and adversary classifiers as ANNs that are trained separately using the disclosed training instances. Presumably, these classifiers act as ideal classifiers for detecting utility and private information. Therefore, the performance of PF can be characterized by the area under the ROC curves (AUC) resulting from these utility and adversary classifiers. A better PF would have a larger AUC for the Utility classifier and a smaller AUC for the adversary classifier.

MNIST dataset The original MNIST dataset is a handwritten digit dataset consisting of 60,000 training examples and 10,000 testing examples. Each sample is a 28 × 28 grayscale image. We create a new synthetic dataset where each synthetic image is a two-digit image (ranging from 70 to

Algorithm 1 PF training procedure for privacy funnel.

```
Require: Require: $\alpha$, learning rate. $b$, the bach size. $k$, a hyperparameter to be used for updating $\phi_{(1,..,m)}$ in each iteration. $\beta$, Lagrange multiplier.
$T_\theta(X), Pr_{\phi_{(1,..,m)}}(T_\theta(X)) \leftarrow$ Random initialization
while $\theta$ has not converged do
    for $k$ steps do
        Sample $\{x^i, s^i\}_{i=1}^{b}$ a bach from the real data.
        $\{\hat{x}^i\}_{i=1}^{b} \leftarrow T_\theta(\{x^i\}_{i=1}^{b})$
        Perform SGD-updates for $\phi_{(1,..,m)}$
        for $j = 1 : m$ do
            $g_{\phi_j} \leftarrow \nabla_{\phi_j} \frac{1}{b} \sum_{i=1}^{b} CE(s_j^i, Pr_{\phi_j}(\hat{x}^i))$
            $\phi_j \leftarrow \phi_j$ - $\alpha$ . AdamOptimizer$(\phi_j, g_{\phi_j})$
        end for
    end for
    Sample $\{x^i, s^i\}_{i=1}^{b}$ a bach from the real data.
    $\{\hat{x}^i\}_{i=1}^{b} \leftarrow T_\theta(\{x^i\}_{i=1}^{b})$
    Perform SGD-updates for $\theta$
    $g_\theta \leftarrow \nabla_\theta \frac{1}{b} \sum_{i=1}^{b} \left\{ \beta CE(x^i, \hat{x}^i)) - \sum_{j=1}^{v} CE(s_j^i, Pr_{\phi_j}(\hat{x}^i)) \right\}$
    $\theta \leftarrow \theta$ - $\alpha$ . AdamOptimizer$(\theta, g_\theta)$
end while
```

89) generated by concatenating two handwritten images into one with 56×56 pixels. We use 40,000 synthetic images for training, and 5,000 were used for testing. A private data is defined as the two-digit number in the synthetic image that is greater than or equal to 80, i.e., we want to hide the first digit so that an adversary can not guess if it is eight or seven. For the testing phase, the utility information is defined, whether as the two-digit number in the image is odd. In Figure 4, we visualize outputs for various images from our learned PF. On the top are original images and in the bottom reconstructed images, where the first digit is private information that we want to hide. The perturbation of the first digit is cause misclassification for an adversary classifier and does not determine if it is a seven or an eight. To evaluate the effectiveness of the proposed PF to retain the data features that allow for accurate classification, Figure 5 shows the ROC curves for the utility and adversary classifiers trained based on the output of a trained PF. As seen, the AUC is close to 1 for the utility classifier and near 0.5 for the adversary classifier. That indicates PF produces sanitized features that allow the utility data to be mined effectively. In contrast, private data cannot be inferred, i.e., the adversary classifier performs like a random guess.

UCI-HAR Dataset: The data was collected from 30 subjects aged between 19 and 48 years old performing one of six standard activities (Walking, Walking Upstairs, Walking Downstairs, Sitting Standing, Laying) while wearing a waist-mounted smartphone that recorded the movement data. The result was a 561 element vector of features and 10929 instances. We split the dataset into 70% for training and 30% for testing. The utility part is activity recognition, and the

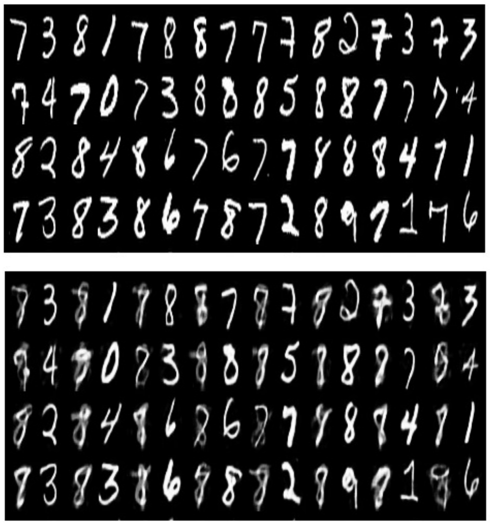


Fig. 4. MNIST experiment: The first four rows show the original images, and the remaining rows visualize outputs for original images from our learned PF.

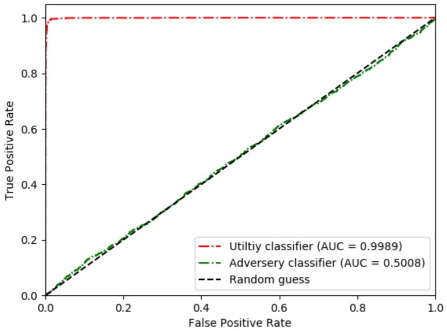


Fig. 5. MNIST experiment: ROC curves for utility and adversary classifiers.

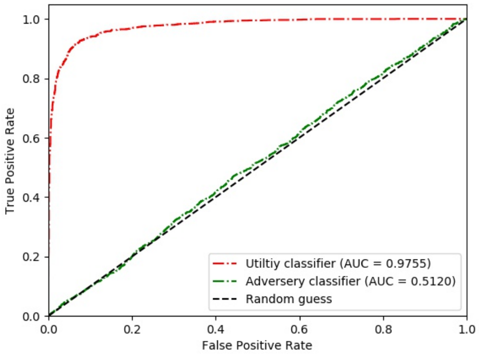


Fig. 6. UCI-HAR experiment: ROC curves for utility and adversary classifiers.

sensitive information is the identities of the users, i.e., $m = 30$. We average utility and adversary accuracy for all classifiers output to get one plot. As seen in Figure. 6, the ROC for the utility classifiers is quite good, while the ROC for the adversary classifiers is slightly better than that of a random guess.

VII. Conclusion

In this paper, we have designed, implemented, and evaluated a novel PF framework that is resilient against adversarial attacks in IoRT-HRI applications. Correctly, PF is assessed in the context problem encountered by users who want to disclose some data to gain utility in real-time while preserving their private information. Individually, we consider the setting in which the data is continuous and high-dimensional, and private labels can be high dimensional vector. The experimental results on two datasets MNIST and UCI-HAR show that PF framework is highly effective and achieves the highest classification accuracy.

References

[1] Partha Pratim Ray, "Internet of robotic things: Concept, technologies, and challenges," *IEEE Access*, vol. 4, pp. 9489–9500, 2016.

[2] Shancang Li, Li Da Xu, and Shanshan Zhao, "The internet of things: a survey," *Information Systems Frontiers*, vol. 17, no. 2, pp. 243–259, 2015.

[3] Matthew Rueben and William D Smart, "Privacy in human-robot interaction: Survey and future work," in *We Robot 2016: the Fifth Annual Conference on Legal and Policy Issues relating to Robotics. University of Miami School of Law, 2016, Discussant: Ashkan Soltani, Independent Researcher*, 2016.

[4] Jonathan Katz, Alfred J Menezes, Paul C Van Oorschot, and Scott A Vanstone, *Handbook of applied cryptography*, CRC press, 1996.

[5] Gilles Brassard et al., *Modern cryptology: A tutorial*, vol. 325, Springer, 1988.

[6] Ahmad-Reza Sadeghi, Christian Wachsmann, and Michael Waidner, "Security and privacy challenges in industrial internet of things," in *2015 52nd ACM/EDAC/IEEE Design Automation Conference (DAC)*. IEEE, 2015, pp. 1–6.

[7] Ali Makhdoumi, Salman Salamatian, Nadia Fawaz, and Muriel Médard, "From the information bottleneck to the privacy funnel," in *2014 IEEE Information Theory Workshop (ITW 2014)*. IEEE, 2014, pp. 501–505.

[8] Yuchen Yang, Longfei Wu, Guisheng Yin, Lijie Li, and Hongbin Zhao, "A survey on security and privacy issues in internet-of-things," *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1250–1258, 2017.

[9] Bin Zhou, Jian Pei, and WoShun Luk, "A brief survey on anonymization techniques for privacy preserving publishing of social network data," *ACM Sigkdd Explorations Newsletter*, vol. 10, no. 2, pp. 12–22, 2008.

[10] Harsh Kupwade Patil and Ravi Seshadri, "Big data security and privacy issues in healthcare," in *2014 IEEE international congress on big data*. IEEE, 2014, pp. 762–765.

[11] Stan Matwin, "Privacy-preserving data mining techniques: survey and challenges," in *Discrimination and Privacy in the Information Society*, pp. 209–221. Springer, 2013.

[12] Ricardo Mendes and João P Vilela, "Privacy-preserving data mining: Methods, metrics, and applications," *IEEE Access*, vol. 5, pp. 10562–10582, 2017.

[13] Aleksandra Korolova, "Privacy violations using microtargeted ads: A case study," in *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*. IEEE, 2010, pp. 474–482.

[14] Ismini Psychoula, Erinc Merdivan, Deepika Singh, Liming Chen, Feng Chen, Sten Hanke, Johannes Kropf, Andreas Holzinger, and Matthieu Geist, "A deep learning approach for privacy preservation in assisted living," *arXiv preprint arXiv:1802.09359*, 2018.

[15] Jianxin Zhao, Richard Mortier, Jon Crowcroft, and Liang Wang, "Privacy-preserving machine learning based data analytics on edge devices," 2018.

[16] Anand D Sarwate and Kamalika Chaudhuri, "Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data," *IEEE signal processing magazine*, vol. 30, no. 5, pp. 86–94, 2013.

[17] Benjamin IP Rubinstein, Peter L Bartlett, Ling Huang, and Nina Taft, "Learning in a large function space: Privacy-preserving mechanisms for svm learning," *arXiv preprint arXiv:0911.5708*, 2009.

[18] Kamalika Chaudhuri, Anand D Sarwate, and Kaushik Sinha, "A near-optimal algorithm for differentially-private principal components," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 2905–2943, 2013.

[19] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2016, pp. 308–318.

[20] Isabel Wagner and David Eckhoff, "Technical privacy metrics: a systematic survey," *ACM Computing Surveys (CSUR)*, vol. 51, no. 3, pp. 57, 2018.

[21] Cynthia Dwork, Aaron Roth, et al., "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[22] Pierangela Samarati and Latanya Sweeney, "Generalizing data to provide anonymity when disclosing information," in *PODS*. Citeseer, 1998, vol. 98, p. 188.

[23] Madhushri Banerjee and Sumit Chakravarty, "Privacy preserving feature selection for distributed data using virtual dimension," in *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2011, pp. 2281–2284.

[24] Yasser Jafer, Stan Matwin, and Marina Sokolova, "A framework for a privacy-aware feature selection evaluation measure," in *2015 13th Annual Conference on Privacy, Security and Trust (PST)*. IEEE, 2015, pp. 62–69.

[25] Ke Xu, Tongyi Cao, Swair Shah, Crystal Maung, and Haim Schweitzer, "Cleaning the null space: A privacy mechanism for predictors," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[26] Miro Enev, Jaeyeon Jung, Liefeng Bo, Xiaofeng Ren, and Tadayoshi Kohno, "Sensorsift: balancing sensor data privacy and utility in automated face understanding," in *Proceedings of the 28th Annual Computer Security Applications Conference*. ACM, 2012, pp. 149–158.

[27] Yuksel Ozan Basciftci, Ye Wang, and Prakash Ishwar, "On privacy-utility tradeoffs for constrained data release mechanisms," in *2016 Information Theory and Applications Workshop (ITA)*. IEEE, 2016, pp. 1–6.

[28] Ye Wang, Yuksel Ozan Basciftci, and Prakash Ishwar, "Privacy-utility tradeoffs under constrained data release mechanisms," *arXiv preprint arXiv:1710.09295*, 2017.

[29] Flávio du Pin Calmon and Nadia Fawaz, "Privacy against statistical inference," in *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2012, pp. 1401–1408.

[30] Naftali Tishby, Fernando C Pereira, and William Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[31] Thomas M Cover and Joy A Thomas, *Elements of information theory*, John Wiley & Sons, 2012.

[32] Christoph Lutz and Aurelia Tamò, "Privacy and healthcare robots–an ant analysis," in *We Robot 2016: the Fifth Annual Conference on Legal and Policy Issues relating to Robotics. University of Miami School of Law, 2016, Discussant: Matt Beane, University of California Santa Barbara*, 2016.

[33] M Ryan Calo, "12 robots and privacy," *Robot ethics: The ethical and social implications of robotics*, p. 187, 2011.

[34] Dag Sverre Syrdal, Michael L Walters, Nuno Otero, Kheng Lee Koay, and Kerstin Dautenhahn, "He knows when you are sleeping-privacy and the personal robot companion," in *Proc. Workshop Human Implications of Human-Robot Interaction, Association for the Advancement of Artificial Intelligence (AAAI'07)*, 2007, pp. 28–33.

[35] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[36] Yann LeCun, "The mnist database of handwritten digits," *http://yann.lecun. com/exdb/mnist/*, 1998.

[37] Olivier Martin, Irene Kotsia, Benoit Macq, and Ioannis Pitas, "The enterface'05 audio-visual emotion database," in *22nd International Conference on Data Engineering Workshops (ICDEW'06)*. IEEE, 2006, pp. 8–8.