Data Privatizer for Biometric Applications and Online Identity Management

Giuseppe Garofalo, Davy Preuveneers, Wouter Joosen

Data privatizer for biometric applications and online identity management

Giuseppe Garofalo, Davy Preuveneers, and Wouter Joosen

imec - DistriNet, KU Leuven; Celestijnenlaan 200A, B-3001 Heverlee, Belgium
{firstname.lastname}@cs.kuleuven.be

Abstract. Biometric data embeds information about the user which enables transparent and frictionless authentication. Despite being a more reliable alternative to traditional knowledge-based mechanisms, sharing the biometric template with third-parties raises privacy concerns for the user. Recent research has shown how biometric traces can be used to infer sensitive attributes like medical conditions or soft biometrics, e.g. age and gender. In this work, we investigate a novel methodology for private feature extraction in online biometric authentication. We aim to suppress soft biometrics, i.e. age and gender, while boosting the identification potential of the input trace. To this extent, we devise a min-max loss function which combines a siamese network for authentication and a predictor for private attribute inference. The multi-objective loss function harnesses the output of the predictor through adversarial optimization and gradient flipping to maximize the final gain. We empirically evaluate our model on gait data extracted from accelerometer and gyroscope sensors: our experiments show a drop from 73% to 52% accuracy for gender classification while loosing around 6% in the identity verification task. Our work demonstrates that a better trade-off between privacy and utility in biometric authentication is not only desirable but feasible.

1 Introduction

Biometrics have become a prevalent form of authentication. A broad spectrum of services, with their own unique security requirements, uses some form of biometric authentication, e.g. messaging applications or banking services. Biometrics are preferred over traditional knowledge-based systems, such as PINs and passwords, due to their ease of use, robustness and uniqueness. Moreover, the wide availability of mobile sensors allows for the deployment of near frictionless multi-modal systems.

Sensor based gait recognition is regarded as a promising approach towards unobtrusive user authentication [9, 17, 29, 30]. Despite being less robust than well-established biometrics, motion data takes advantage of body worn sensors that are widely implemented in modern devices and require little to no effort by the user. By enabling continuous user authentication, gait authentication is a natural candidate for multi-modal settings, e.g. by combining face recognition to walking data [17]. In this way, we not only improve accuracy, but also strengthen

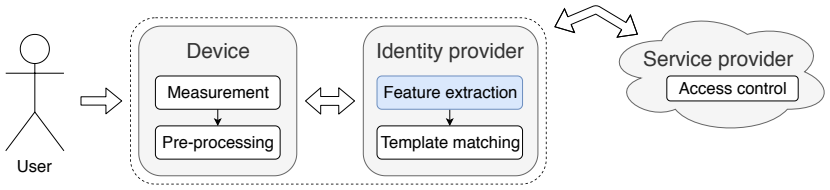


Fig. 1: Biometric authentication pipeline for online identity management. The user interacts with the identity provider and the service provider through his/her device.

our system against forging and spoofing attacks [9]. Moreover, gender recognition is a relevant topic to be addressed in gait recognition for future applications in healthcare [17, 30]. The ever-improving resilience of continuous authentication systems based on accelerometer and gyroscope measurements clashes with the lack of a comprehensive assessment in terms of sensitive data leakage, demanding for techniques to protect a user's privacy against sensitive inferences. In this work, we explore gait authentication and soft biometric recognition, i.e. gender and age, as a use case for our adversarial framework for privacy.

As shown in Figure 1, biometric authentication systems typically involve three entities [26]: a device equipped with sensors, a service provider that authorizes access, and an identity provider that verifies the identity. The authentication pipeline is composed of three steps. During step one, the user device collects the biometric signal. The latter is then cleaned and prepared for feature extraction, which is the second step in the pipeline. Consequently, the pre-processed signal is transformed into a set of relevant features that can be matched with a stored user template. For example, a face image may be turned into a vector of numeric features, while a gait trace could become a 2D image. Herein, the feature extraction and matching scheme are implemented by the identity provider. The final step consists of sending the output of template matching to the service provider, which grants access to its services based on proper access control policies. These three blocks can be incorporated as parts of the user's device or exist in isolation. In alternative, hybrid implementation are possible, e.g feature are extracted in the user's device while the templates are matched remotely. The latter scenario, i.e. online authentication, requires the user to send sensitive data over an unreliable network, exposing him/her to potential privacy leaks. Handling biometric data, including storing and processing templates, calls for additional security and privacy guarantees.

Misuse of biometric templates leads to severe privacy leakages for the end-user. Recent work has shown the presence of sensitive data in biometric traces, including medical conditions and soft biometrics [2, 18, 22]. If the user consented for his/her biometric template to be stored on a third-party server, he/she has to be aware of the potential disclosure of such sensitive data. For example, a *curious* service provider might want to learn more about its customers to advertise them with tailored products and increase its sales. Even in the unrealistic hypothesis that the user can blindly trust his/her recipient, adversaries might steal the user's

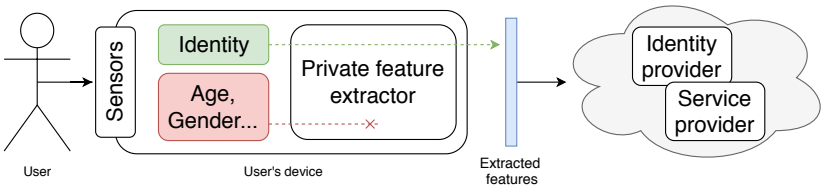


Fig. 2: Pipeline for private feature extraction in a biometric sharing scenario. The central element is the extractor, which is shared with the user.

template by impersonating the trusted party or attacking remote servers [32]. However, in many cases the third-party will need access to relevant information to keep its services alive. This calls for techniques which enable the sharing of the least amount of sensitive identifying information, e.g. *private biometric features*.

Recent history testifies that algorithms are prone to discriminate based on racial or sex attributes [4, 5], leading to discussions on how to mitigate bias in AI [13]. Typical sources of bias are the *training dataset*, directly reflecting unfair patterns in the external world, and a *flawed machine learning (ML) process*, not accounting for discrimination or, even worse, deliberately fueling inequalities. By suppressing highly-sensitive information like gender or age, we propose a novel representation of the biometric trace which discourages differences based on group membership.

Our work is also motivated by the General Data Protection Regulation (GDPR) [8], which tackles private data collection and processing problems. Article 25 puts the emphasis on the scope limitation by defining the *data minimization principle* which explicitly requires third-parties to limit their data collection to what is useful for their purposes, upon consent. However, storing data such as biometrics clashes with this principle because of its intrinsic re-purposable nature, which has been theorized and validated thoroughly in recent years. Hence, there is a need to design data reduction processes for sharing sensitive information in order to protect the users against unprompted attribute inferences. By sharing only what is needed for a predefined task, we can also address the discrimination problem, bringing fairness and transparency in the ML pipeline.

Figure 2 shows our pipeline for private biometric sharing. The *user's device* is in charge of collecting motion data and extracting features to be shared with an external authentication provider. However, the user typically lacks the necessary resources, both hardware and data, to train a feature extractor him/her-self. Thus, we embrace a data-driven approach to derive a private feature extractor on the server side. This approach exploits a publicly available dataset to train the model, allowing to derive only one feature extractor for all the users willing to authenticate. In particular, our adversarial framework is composed of three entities closely working together: a (private) feature extractor, an (adversarial) siamese identity verifier and a predictor for the private task. During training, the feature extractor will iteratively adapt to changes in the classifier, which in turn will challenge the extractor. This procedure models the mutual information

between the identity and the private attributes, guaranteeing protection against sensitive inferences. Eventually, the feature extractor is published by the identity/service provider and becomes available for local usage on the user's device, as depicted in Figure 2. By harnessing our feature extractor, neither third parties nor channel eavesdroppers can accurately infer the target classification attribute from the shared embedding.

We apply our proposed framework in a gait verification scenario using fixed inertial sensors. In our evaluation, we compare different privatized traces to assess the identifiability of the users while the new extracted features cannot be used to infer the user's gender or age, which are our private attributes. We emulate our adversary's ability to infer the sensitive attributes by means of transfer learning, i.e. training an unseen classifier for the private task on the private features. Our main contributions include:

1. Devising an adversarial framework exploiting a novel loss function to train a private feature extractor starting from variable-length gait traces.
2. Evaluating the privacy-utility trade-off w.r.t soft biometrics privacy in the gait authentication domain.
3. Using the biggest known inertial sensors dataset, which includes almost 500 users and 5 different activities.

The rest of the paper is organized as follows. We identify the gap and differences with related work in Section 2. We present our framework for privacy-preserving feature extraction in Section 3. The experimental protocol and results are presented in Section 4. Section 5 concludes our work.

2 Related Work

Mordini and Ashton [22] have performed an extensive study of medical pattern retainment in biometric templates: psychiatric conditions can be inferred from gait traces, chromosomal diseases can be accurately guessed from face images or fingerprints, while neurological pathologies have been associated to a broad range of behavioural biometrics. The same leakage potential holds true for electrocardiogram (ECG) signals [18], iris recognition [2] and other bio or behavio-metrics [6]. Similarly, soft biometrics like age, gender or race are linked to physiological or behavioural traits of the user. In a recent work, we proved the feasibility of age and gender estimation from gait traces in the frame of the *OU-ISIR Wearable Sensor-based Gait Challenge: Age and Gender* (GAG 2019) competition at the 12th IAPR International Conference on Biometrics[1] [28]. The goal of this competition was to improve the state-of-the-art in soft biometric prediction from accelerometer and gyroscope traces. Even without crucial information on sensors position, we were able to achieve ∼76% accuracy for gender classification and a mean absolute error of ∼6 years for age estimation, eventually obtaining the best result among all contestants. Our model is inspired

[1] http://www.am.sanken.osaka-u.ac.jp/GAG2019/

by [28] as follows: we harness temporal convolutional networks (TCNs) for feature extraction and few dense layers for soft biometric prediction. On top of the extracted features, we have built a siamese network for user verification and we have plugged a gradient reversal layer for attribute privatization.

Several works tackled the problem of discrimination in the ML pipeline [1,7]. Typically, they focus on the output of the decision function and how to make it independent from a particular group membership. In contrast to previous work, we address this problem indirectly, aiming to achieve **soft biometric privacy**. By suppressing information deemed to be private, we discourage discriminative attributes to influence the learning process, thereby representing a source of bias. For example, by minimizing the information about the gender in motion data, we encourage the building of a gender-agnostic gait verification system. In the soft biometric privacy landscape, our work is the first one focusing on time sequences and, specifically, gait authentication.

The approaches to protect user's privacy divide into context-free and context-aware techniques. Context-free techniques, like differential privacy (DP), model worst-case adversaries regardless of his/her real capabilities and discarding relevant contextual information, i.e. about the problem to be solved. DP provides strong privacy guarantees, delivering a shrinking in data usefulness. Context-aware strategies, on the other hand, incorporate the retainment of task-specific utility by selectively adding noise where it matters. This advantage comes at the expenses of a formal characterization of the relationship between public variables, i.e. what we aim to share, and private variables, i.e. what we aim to protect, which is rarely available in practice.

Data-driven optimization has been recently proposed as a mean to achieve context-aware privacy. By exploiting recent advances in adversarial optimization, it is possible to model the joint distribution between shared and private variables. Generative adversarial networks (GANs) have been recently proposed as an effective tool to achieve this goal [12]. They model a min-max game between a generator and a discriminator, where the former tries to fool the latter in an iterative learning process. This concept has been first adapted to the privacy domain by Huang et al. who define the generative adversarial privacy (GAP) framework [14]. Inspired by their work, we harness adversarial training to obtain a private feature extractor, representing our feature generator. This generator is used to obtain a compressed template representation for a specific, measurable and limited purpose, while also minimizing sensitive disclosure.

Morales et al. [21] recently proposed a method to reduce gender and race information in latent representations of face images. Their method is based on a modification of the triplet loss function, which is commonly employed in face verification scenarios [27]. Our work differs from theirs for two reasons: first, we exploit adversarial optimization to maximize the privacy-utility trade-off; second, our feature extractor is designed to model temporal dependencies in the input data, making it more suitable for gait samples than face images. Similarly, Mirjalili et al. proposed a framework to impart gender privacy to face images [19, 20]. Drawing from their work, we empirically evaluate the privacy

and generalizability of our approach by training several models, which simulate the ability of a malicious entity. As before, they focus on face recognition systems rather than temporal data.

Malekzadeh et al. [16] have considered motion data and gait authentication in a different min-max optimization scenario: perturbing identity while preserving task-specific utility. Their classification task is activity recognition, which has been extensively studied in the gait literature in addition to being arguably a private variable. Moreover, we shift the focus towards building a private extractor to be used by end-users instead of generating a privatized trace in the input domain. By compacting the trace in a latent space representation, we reduce the interpretability of the shared sample while minimizing the risks of sensitive inference. Similarly, Osia et al. [24] investigates the use of siamese networks for privatizing the user's identity while preserving gender classification accuracy. Besides the different learning goal, they focus on fine-tuning existing, pre-trained networks. In our framework, the minimization of sensitive attribute is embedded in the learning process itself. By simply applying fine-tuning to the last layers of the feature extractor, we would discourage the achievement of a better sub-optimal solution for the min-max optimization problem.

3 Private Feature Extraction Framework

We propose a novel framework for protecting sensitive variables when sharing biometric data in an online authentication scenario. In this section, we tackle 4 key aspects which characterize our framework: (i) the *main steps*, stakeholders, and threat model, (ii) the nominal *privacy-preserving loss function*, (iii) the designed *architecture*, i.e. the neural networks to approximate the nominal loss, (iv) and the architecture *min-max optimization strategy*.

3.1 High-level framework and threat model

Our framework faces the problem of sharing biometric data without exposing user's private information. This process requires the interaction between two entities: the user and the service provider. The user is willing to share what is needed to accomplish the main task but he is worried that certain information might leak along the way. Let us assume user A wants to be authenticated towards service B, then sharing the raw data will reveal attributes which A might consider private, e.g. A's gender. As discussed in Section 2, traditional techniques can be employed to solve this problem, however they come with several limitations such as expensive computations at the edge or having to trust external entities. Instead, we propose the use of a contained set of *private features* extracted from biometric data on the user side. These features are carefully optimized during the training phase, with a twofold purpose: (1) to preserve the information which identifies the user and (2) to suppress a specific private variable, e.g. the gender of the user.

Our proposed framework requires three actions by the involved actors: (i) the identity/service provider trains a private feature extractor for authentication purposes; (ii) the feature extractor is published, which allows the user to extract authentication features and assess its privacy guarantees; (iii) by following the authentication protocol, the authentication features are shared with the service provider which grants access to the system based on given access control policies. This three-step procedure protects the user against unprompted sensitive inferences: the service provider will not be able to improve its knowledge about the user w.r.t. to the selected private variables, thus we protect against *function creep* [31]. This is inherent to the local feature extraction step (ii), which suppresses private variables within the user's device. Moreover, sharing the private extractor enables any external entities, like the user, to assess and analyze the privacy of the model, which is only trained on *public data*. It is worth noticing that the authentication-party might want to train a ML model to authenticate the user based on the received features, i.e. step (iii), and this part of the model has to be kept private in order to guarantee users' privacy. In addition, the feature extractor could be trained by different parties than the third-party, but we rely on the realistic hypothesis that the features are especially crafted for the main task. Therefore, the service provider is better suited to design the feature extraction step. If we assume that the features are intended for different uses, then we can assign step (i) to another external, mediating entity without affecting the presented framework.

While our solution overcomes traditional techniques limitations, several challenges regarding privacy estimation arise. By delegating the training phase to a cloud-based service, we cut down the computational power which is requested to the user. Thus, only feature extraction of a pre-trained model is performed on the user side. In addition to energy consumption, by extracting the biometric representation locally, we avoid the need for an external mediator. Therefore, we free the user from trusting an external entity. However, unlike traditional techniques like DP, we can only provide empirical privacy guarantees of the shared representation. This is due to a different sharing scenario, which involves single temporal traces as opposed to large databases of many users.

We evaluate the privacy of the extracted representation by looking at the performance of a newly trained ML classifier. Having fixed the discernment of a discrete sensitive variable as the learning goal, and provided the extracted features as the model input, we derive an empirical definition of privacy: the better the classifier performs, the higher the sensitive leakage. In practice, the classifier mimics the capabilities of a *curious service provider* willing to obtain valuable information about its users. The provider has access to the public dataset used to train the feature extractor (step (i)), and it also knows the training details as well as the trained model weights. Thus, the third-party is able to use public data to train a classifier discerning the private variable from the extracted features. We test the classifier against a group of test users, which simulates the sharing of features during regular usage. The service provider acts as the most powerful adversary for knowledge and resources, so we indirectly test the

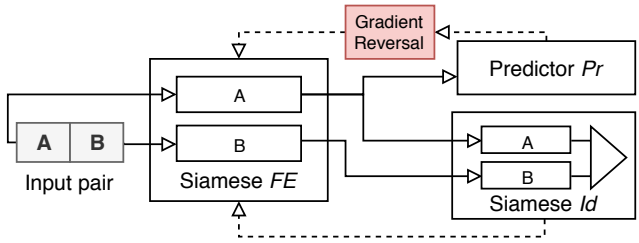


Fig. 3: Our framework with its three challenging entities: the siamese feature extractor (FE), the soft biometric predictor, and the siamese identity verifier (Id).

privacy of the extractor against any, less or equally powerful, external adversaries trying to infer sensitive information from the shared representation. In conclusion, if the service provider is not able to train a classifier which discerns the private variable from the shared representation, we consider the features to be *safe* against inferences. However, it has to be stressed that a more powerful estimator which is able to extract private information from the shared template might exist, which is a realistic assumption drawn from recent work [25].

3.2 Privacy-preserving training objective

We present here our nominal loss function for soft biometric privacy, i.e. the ideal training objective to be approximated via neural networks optimization. Given an input X, we search for the optimal feature extractor $FE(\cdot)$, which outputs the private embedding $Z = FE(X)$. We define $D(\cdot, \cdot)$ as a measure of the dependency between two variables, such that $D(Z, T)$ measures the dependency between the private embedding and the classification task T we aim to suppress, while $D(Z, I)$ describes the usefulness of the latent representation for authentication purposes, being I the identity of the user. The nominal loss to be minimized becomes

$$NL = \alpha * D(Z, I) - \beta * D(Z, T) \tag{1}$$

where α and β regulate the importance of each term, which purpose is to fine-tune the privacy-utility trade-off. As mentioned before, we make use of deep neural networks to approximate $D(\cdot, \cdot)$. Therefore, an estimation of this measure is embedded in the weights of the models after training. We estimate it by extracting features and analyzing what newly created ML models are capable of learning from these features.

3.3 Neural networks architecture

Our framework is composed of three competing blocks: a private extractor (Figure 4), an identity verifier (Figure 5b) and a task-dependent predictor (Figure 5a). As mentioned earlier, the classification pipeline is inspired by [28] with

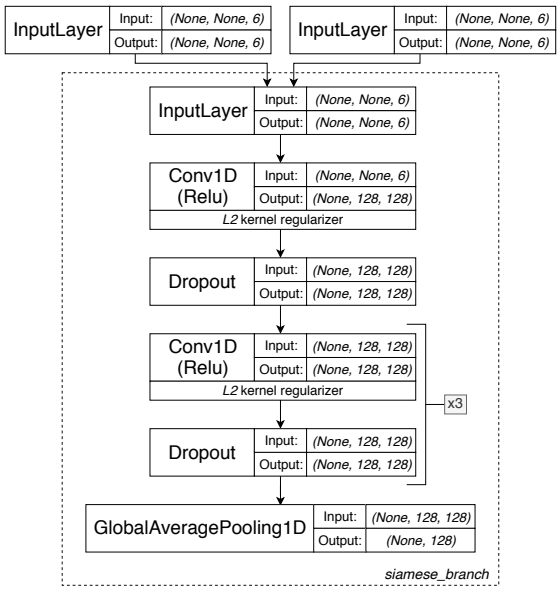
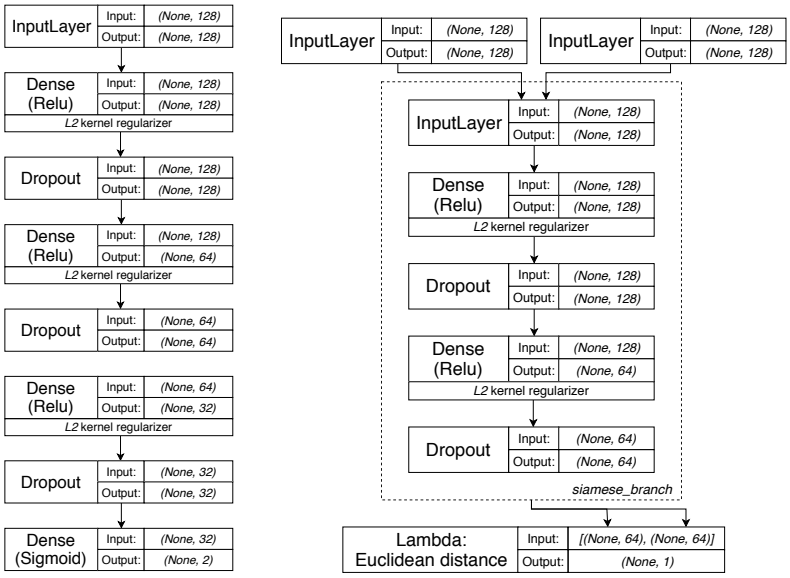


Fig. 4: Implementation of *Siamese FE*. The sub-model *siamese_branch* is shared among the two input layers and contains a quintessential element, namely *Conv1D*.

the addition of a siamese block for gait verification. An overview of the interacting components is presented in Figure 3.

The input of the model is 3-dimensional raw accelerometer and 3-dimensional raw gyroscope measurements recorded from inertial sensors. In our implementation, we account for variable-length input by simply stacking sensors measurements (6-dimensional measurements) without pruning the obtained sequences.

Input data is fed to a *siamese feature extractor FE*, which details are presented in Figure 4. Every layer is presented along with its input and output shapes, where *None* represents either a variable length trace or a variable number of samples, i.e. a batch. As first proposed by Bai et al. [3], we perform dynamic feature extraction by harnessing temporal convolutional networks (TCNs). This family of networks have been demonstrated to achieve state-of-the-art performance when dealing with temporal data, behaving equal to or better than recurrent neural networks. TCNs potential is mainly due to *dilated* convolutions, which address both complexity of the network and low-level spatial accuracy. Convolutional layers are intertwined with *dropout* layers to prevent over-fitting the training set, thus acting as regularizers by zeroing-out random filters which are re-activated when testing the model. Following best practices, $L2$ regularization is also introduced to penalize complex models through the loss function. Finally, a *global average layer* flattens the output of the extractor to obtain 128 features, which are the objective of the privatization of our optimization framework. They are then conveyed into two branches: a *soft biometric predictor* and a *siamese identity verifier*.



(a) (b)

Fig. 5: Implementations of *Predictor* (a) and *Siamese Id* (b). In (b), the sub-model *siamese_branch* is shared among the two input layers.

The first branch is the soft biometric predictor Pr, which details are depicted in Figure 5a. This model is composed of several fully connected layers with $ReLU$ activations. As before, dropout and regularization techniques are employed to improve generalizability. The model is designed once but tuned and optimized separately for gender and age. During optimization, we minimize *binary cross-entropy* to maximize our sensitive variable prediction accuracy.

The second branch is a siamese neural network for identity verification Id, which details are shown in Figure 5b. This model is composed of two stacked fully connected layers which are duplicated into two parallel branches sharing their weights (in Figure 5b, only one branch is shown). These branches converge into a distance-based function, the *contrastive loss*, which is typically used in a verification scenario to increase the similarity, i.e. decrease the distance in the latent space, of samples belonging to the same class while driving away dissimilar pairs. In order to obtain two feature vectors, the feature extractor is turned into a siamese model as follows: we duplicate the FE, obtaining two branches which share their weights and allowing to feed a pair of gait traces into the input layer.

By combining our three blocks, we end up with the following multi-objective loss function

$$G(\theta, \phi) = min_{\theta,\phi}(\alpha * Id(\phi) + \beta * Pr(\theta)) \tag{2}$$

where θ and ϕ are the weights of PR and Id, respectively, and α and β weight each term. By minimizing G, however, the predictor accuracy is maximized, thus we need to reverse this trend to protect the latent representation against inferences of the sensitive variable. First introduced in domain adaptation by Ganin et al. [10], we plug a *gradient reversal layer* between FE and Pr which minimizes the Pr accuracy when training FE. This leads to a min-max loss function which is the approximated realization of Equation 1.

3.4 Optimization of the networks

As common in adversarial training, the three blocks are optimized separately, via *strict alternation*. Figure 3 shows the interaction among the elements, which single iteration works as follows. Id and Pr are fed with features extracted from FE and trained for one epoch to minimize the *contrastive loss* and the *cross-entropy*, respectively. FE is later trained for a single epoch to optimize Equation 2 by freezing the weights of the other networks. By reversing the gradient of Pr, FE will adapt to its changes becoming more and more resilient to sensitive inferences. This challenge approximates the dependency measure D (cfr. Equation 1) which is key to achieve a satisfactory sub-optimal trade-off between privacy and utility.

By alternating the gradient updates among the three networks, we enforce dynamic adaptation to future updates. Suppose Pr is trained w.r.t. the true labels until a acceptable sub-optimal solution is found. FE can be trained to beat Pr, hiding the sensitive variable from the learnt representation. However, this holds true only for one sub-optimal solution of Pr and does not prevent inferences from future re-training. Instead, by letting Pr adapt to changes in FE, and vice-versa, both models converge to a more satisfactory solution in terms of privacy vs. utility. Hyper-parameters play an important role in network convergence, impeding Pr to win over FE; yet, we can only empirically estimate the best hyper-parameters for our task.

Our feature extraction strategy differs from traditional dimensionality reduction techniques because we actively trigger the reaction of a discriminator, exploiting its input to refine our final output. However, approaches like Principal Component Analysis (PCA) or noise addition at the bottleneck of the network could still be applied as a complementary, simple aid to achieve a better sub-optimal solution.

4 Evaluation

In this section, we evaluate the ability of our framework to hinder classification of gender from motion features while preserving the identity of the user. We present the experimental setup, followed by the evidence we found.

4.1 Experimental setup

We define our networks using Keras 2.2.4 with a Tensorflow backend, which runs on a machine with a 3.4GHz i5-7500 CPU, 16GB RAM, and an NVIDIA Titan V GPU. As in [28], we choose the OU-ISIR labelled gait action dataset [23] to train our model. Angular velocity and acceleration in the 3 dimensions are collected from sensors fixed on a belt, at a sampling rate of 100 Hz. Every trace is associated to age, gender and current activity of the user.

In order to generate training pairs, we undertake several steps. We first divide users into training and test sets following a 80%-20% proportion, after which we use the same ratio to split the training set into training and validation sets. Then we generate fixed length sequences through windowing with overlapping: a user is selected and his/her trace is divided into several traces of length 2.56s with 20% overlap, i.e. 0.5s at the end and 0.5s at the beginning of the trace. Finally, we create pairs to feed our model with through an iterative procedure in two steps: (1) for each user, the current trace is coupled with the subsequent one and a *similar pair label* is assigned to the pair; (2) the second trace of the previous pair is coupled with a random trace from a different user and a *dissimilar pair label* is assigned to this couple. Since the first branch of the siamese *FE* (i.e., branch A in Figure 3) is responsible for feeding *Pr*, a label with the gender of the first user is associated to the pair.

We empirically select the best hyper-parameter configuration for our networks. We select the Nadam optimizer to train the models, following the procedure explained in Section 3.4. Every model is trained on mini-batches within the set [5,25,50], while the number of epochs varies between 15 and 50. We fix α and β to 1. Intuitively, one can expect α and β to have a predictable impact on the privacy-utility trade-off when training the features extractor. Due to the adversarial nature of the training procedure, however, tuning these variables proved to be highly sensitive w.r.t. the given setting and selected hyper-parameters. We argue that this effect can be associated to the training of the predictors, which are independent from α and β in our implementation. Nevertheless, as underlined before, trivial noise addition and dimensionality reduction could be employed as a better, more stable alternative for privacy-utility trade-off tuning.

After training, the sensitive variable predictor and the identity verifier are re-trained on the privatized features. As suggested by previous work [24], the privacy of a sensitive variable can be evaluated via transfer learning, i.e. freezing our pre-trained feature extractor and training a soft biometric predictor from scratch. In order for a thorough evaluation of the generalizability of our approach, unseen classifiers have to be taken into consideration. Hence, we define two models: (1) a DNN-based predictor resembling the one we used in our adversarial framework, which ensures protection against our target model; (2) a Support Vector Machine (SVM) which is typically used downstream of DNN for feature extraction. The SVM is optimized by applying grid search, which exhaustively searches for the hyper-parameter combination with the best score. $L2$-normalization is also performed to maximize our accuracy metric, i.e. *f1-score*.

Table 1: Scores for baseline siamese verifier trained without predictor feedback.

Epochs	Verification accuracy	f1-score (SVM)
200 (early stopping)	90.93%±0.15	72.58%

Table 2: Re-train f1-score for the predictors and accuracy for the siamese identity verifier after adversarial privatization.

Epochs	Batch size	Verification accuracy	f1-score	f1-score (SVM)
15	25	82.14%±0.89%	51.15%±0.70%	60.97%
15	50	87.15%±0.38%	50.68%±0.57%	65.26%
25	25	84.47%±0.53%	50.20%±0.15%	63.28%
25	50	85.28%±0.48%	50.10%±0.00%	**52.99%**

We repeat each experiment 10 times, reporting the f1-score for the gender and the average verification accuracy for the verifier. All the results, besides the f1-score which results from a deterministic search, are presented with their standard deviation.

4.2 Experimental results

Table 1 shows the results for our baseline model: a siamese feature extractor for identity verification. After training, the extracted features are used to infer the gender of the user, resulting in a f1-score of 72.58% for the SVM in the best configuration. This underlines the retainment of soft biometric information in the authentication features, especially if we compare this figure to the state-of-the-art gender prediction accuracy presented in [28], i.e. 75.77%.

We compare our baseline with our proposed approach for feature privatization, which is summarized in Table 2. A high variability in the results can be observed, which is mainly due to the instability of the adversarial learning procedure. The *f1-score* shows how our privatization mechanism protects the features against possible re-training of our target predictor in each setting. However, we have to take into account generalization and we must be able to protect against unseen classifiers. SVM *f1-score* proves that we are able to achieve a nearly optimal result (50%) by carefully tuning our hyper-parameters, i.e. number of epochs and batch size. This comes at the expense of a nearly 6% loss in verification accuracy. For a batch size of 50, and increase in the number of epochs corresponds with a slight decline in verification accuracy. This drop indicates how letting the network train for a larger number of epochs improves the empirical privacy (see smaller f1-scores in Table 2) while decreasing the utility, even by just a tiny fraction.

We identify two main limitations for this work which are linked to our privacy evaluation and chosen dataset. First, we empirically evaluate the privacy of our

framework by re-training different models on the extracted features. Future work could derive a formal evaluation of the privacy guarantees from an information-theory point of view. Second, the OU-ISIR dataset provides us with sufficient data for our scopes but its data is collected in a constrained environment by sensors fixed on a belt. In a real world scenario, we deal with different orientations of mobile devices, and its sensors, carried by the users. A more realistic dataset is needed to properly evaluate and compare solutions in the gait domain.

As a future direction, we aim to tackle the linkability of templates across services while hiding different private variables for one user. Since we are not delivering a full-fledged biometric template protection (BTP) scheme, we do not directly address linkability of traces, assessing instead the retainment of private information for a specific use-case, i.e. gender classification. Hence, our framework alone does not fulfill the two requirements of the standard on biometric data protection ISO/IEC 24745 (2011) [15], i.e. *irreversibility* and *unlinkability*. However, by tackling the data minimization problem we aim to address problems which are complementary to BTP schemes: (1) we exclude unnecessary data from transmission and processing, possibly improving privacy and performance of crypto schemes, and (2) we help preventing or fighting back algorithmic bias by feeding algorithms with more neutral and task-specific data. We envision a hybrid system where BTP schemes and adversarial training maximize the utility for a specific task without compromising users privacy. Future directions include exploring age or race prediction to evaluate cross-task linkability, and analyzing the advantage of applying a biometric crypto scheme on top of our minimization framework for compliance with existing requirements for private and secure biometric data processing and management. To this extent, Barrero et al. [11] have proposed a metric to evaluate the local and global linkability of biometric templates.

5 Conclusion

In this work, we demonstrated the effectiveness of an adversarial learning technique towards privatization of biometric features from sequential data. We evaluated our approach on the gender estimation use-case, inspired by a recent work. Our evaluation supported our approach, showing a dip in the f1-score from 73% to 52.99% in the best case, which is very close to random guess (50%).

Further evaluation is needed to assess the effectiveness of our approach against different use-cases, but our results show that a solution to the long standing problem of data-minimization for biometrics is possible. Data-driven techniques have the potential to achieve the optimal trade-off between privacy and utility, something traditional techniques usually struggle with. We advocate for new tools for the user to manage his own identity and the amount of sensitive information which is shared with third-parties.

Acknowledgement. This research is partially funded by the Research Fund KU Leuven. Work for this paper was supported by the European Commission

through the H2020 project CyberSec4Europe (https://www.cybersec4europe.eu/) under grant No. 830929. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

References

1. Alvi, M., Zisserman, A., Nellaaker, C.: Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In: The European Conference on Computer Vision (ECCV) Workshops (September 2018)
2. American Academy of Ophthalmology: Evidence mounts that an eye scan may detect early alzheimer's disease (2018), https://www.aao.org/newsroom/news-releases/detail/evidence-eye-scan-may-detect-early-alzheimers, last accessed 14 May 2019
3. Bai, S., Kolter, J.Z., Koltun, V.: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv:1803.01271 (2018)
4. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency. Proceedings of Machine Learning Research, vol. 81, pp. 77–91. PMLR, New York, NY, USA (23–24 Feb 2018)
5. Cohn, J.: Googles algorithms discriminate against women and people of coloue (2019), http://theconversation.com/googles-algorithms-discriminate-against-women-and-people-of-colour-112516, last accessed 14 May 2019
6. Dantcheva, A., Elia, P., Ross, A.: What else does your biometric data reveal? a survey on soft biometrics. IEEE Transactions on Information Forensics and Security **11**(3), 441–467 (March 2016)
7. Das, A., Dantcheva, A., Bremond, F.: Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach. In: The European Conference on Computer Vision (ECCV) Workshops (September 2018)
8. European Parliament: Regulation (EU) 2016 of the European Parliament and of the Council, on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (2016)
9. Gafurov, D.: A survey of biometric gait recognition: Approaches, security and challenges. In: NIK conference (2007)
10. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., Lempitsky, V.: Domain-adversarial training of neural networks. Journal of Machine Learning Research **17**(59), 1–35 (2016)
11. Gomez-Barrero, M., Galbally, J., Rathgeb, C., Busch, C.: General framework to evaluate unlinkability in biometric template protection systems. IEEE Transactions on Information Forensics and Security **13**(6), 1406–1420 (June 2018)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems 27, pp. 2672–2680. Curran Associates, Inc. (2014)
13. Hao, K.: This is how AI bias really happensand why its so hard to fix (2019), https://www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/, last accessed 14 May 2019
14. Huang, C., Kairouz, P., Chen, X., Sankar, L., Rajagopal, R.: Context-aware generative adversarial privacy. Entropy **19**(12) (2017)

15. Information technology - Security techniques - Biometric information protection. Standard, International Organization for Standardization (2011)
16. Malekzadeh, M., Clegg, R.G., Cavallaro, A., Haddadi, H.: Mobile sensor data anonymization. In: Proceedings of the International Conference on Internet of Things Design and Implementation. pp. 49–58. IoTDI '19, ACM (2019)
17. Marsico, M.D., Mecca, A.: A survey on gait recognition via wearable sensors. ACM Comput. Surv. **52**(4), 86:1–86:39 (Aug 2019)
18. Matovu, R., Serwadda, A.: Your substance abuse disorder is an open secret! gleaning sensitive personal information from templates in an eeg-based authentication system. In: 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS). pp. 1–7 (Sep 2016)
19. Mirjalili, V., Raschka, S., Ross, A.: Gender privacy: An ensemble of semi adversarial networks for confounding arbitrary gender classifiers. In: 2018 IEEE 9th Int. Conf. on Biometrics Theory, Applications and Systems (BTAS). pp. 1–10 (Oct 2018)
20. Mirjalili, V., Raschka, S., Ross, A.: Flowsan: Privacy-enhancing semi-adversarial networks to confound arbitrary face-based gender classifiers. IEEE Access **7**, 99735–99745 (2019)
21. Morales, A., Fiérrez, J., Vera-Rodríguez, R.: Sensitivenets: Learning agnostic representations with application to face recognition. CoRR **abs/1902.00334** (2019)
22. Mordini, E., Ashton, H.: The Transparent Body: Medical Information, Physical Privacy and Respect for Body Integrity, pp. 257–283. Springer Netherlands (2012)
23. Ngo, T.T., Makihara, Y., Nagahara, H., Mukaigawa, Y., Yagi, Y.: Similar gait action recognition using an inertial sensor. Pattern Recognition **48**(4), 1289 – 1301 (2015)
24. Ossia, S.A., Shamsabadi, A.S., Taheri, A., Rabiee, H.R., Lane, N.D., Haddadi, H.: A hybrid deep learning architecture for privacy-preserving mobile analytics. CoRR **abs/1703.02952** (2017)
25. Pittaluga, F., Koppal, S., Chakrabarti, A.: Learning Privacy Preserving Encodings Through Adversarial Training. 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE (2019)
26. Rui, Z., Yan, Z.: A survey on biometric authentication: Toward secure and privacy-preserving identification. IEEE Access **7**, 5994–6009 (2019)
27. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 815–823 (June 2015)
28. Van hamme, T., Garofalo, G., Argones Rúa, E., Preuveneers, D., Joosen, W.: A systematic comparison of age and gender prediction on imu sensor-based gait traces. Sensors **19**(13) (2019)
29. Van hamme, T., Preuveneers, D., Joosen, W.: Improving resilience of behaviometric based continuous authentication with multiple accelerometers. In: Livraga, G., Zhu, S. (eds.) Data and Applications Security and Privacy XXXI. pp. 473–485 (2017)
30. Wan, C., Wang, L., Phoha, V.V.: A survey on gait recognition. ACM Comput. Surv. **51**(5), 89:1–89:35 (Aug 2018)
31. Winner, L.: Autonomous Technology: Technics-Out-of-Control as a Theme in Political Thought. MIT Press (1977)
32. Zeitz, C., Scheidat, T., Dittmann, J., Vielhauer, C., González-Agulla, E., Muras, E.O., García-Mateo, C., Alba-Castro, J.L.: Security issues of internet-based biometric authentication systems: risks of man-in-the-middle and biophishing on the example of biowebauth. In: Security, Forensics, Steganography, and Watermarking of Multimedia Contents. p. 68190R (2008)