




Privacy at Risk: Exploiting Similarities in Health Data for Identity Inference

Lucas Lange , Tobias Schreieder , Victor Christen , and Erhard Rahm 

Leipzig University & ScaDS.AI Dresden/Leipzig, Germany
{lange,christen,rahm}@informatik.uni-leipzig.de
{fp83rusi}@studserv.uni-leipzig.de

Abstract. Smartwatches enable the efficient collection of health data that can be used for research and comprehensive analysis to improve the health of individuals. In addition to the analysis capabilities, ensuring privacy when handling health data is a critical concern as the collection and analysis of such data become pervasive. Since health data contains sensitive information, it should be handled with responsibility and is therefore often treated anonymously. However, also the data itself can be exploited to reveal information and break anonymity. We propose a novel similarity-based re-identification attack on time-series health data and thereby unveil a significant vulnerability. Despite privacy measures that remove identifying information, our attack demonstrates that a brief amount of various sensor data from a target individual is adequate to possibly identify them within a database of other samples, solely based on sensor-level similarities. In our example scenario, where data owners leverage health data from smartwatches, findings show that we are able to correctly link the target data in two out of three cases. User privacy is thus already inherently threatened by the data itself and even when removing personal information.

Keywords: Data privacy · Re-identification attack · Similarity attack · Time-series health data · Smartwatch health data · Stress detection

1 Introduction

A general development in the Internet of Things (IoT) is the use of smart devices for tracking and recording personal health data. While suitable devices are becoming more and more widespread and collected data is growing rapidly, the issue of data privacy is only becoming more important and awareness in users is increasing. In their study, Ernst and Ernst [6] found that the perceived privacy risk has a direct influence on device adoption and could ultimately prove a deciding factor for interested users. Aside from general privacy concerns for personal data, the perceived risk directly correlates with the trust in a data owner's privacy promise. This is because once a user's data is collected, the responsibility for its privacy protection is completely transferred into the hands of the data owner. Therefore, they should have knowledge of threats and defences.

In such scenarios, de-identification presently constitutes a common privacy mechanism, i.e., the prevention that the personal identity of users can be revealed. For example, a large distributor of smartwatches states in their privacy policy and pledge, that they “... may share non-personal information that is aggregated or de-identified so that it cannot reasonably be used to identify an individual” [7, 8]. De-identification may hide user identities at first glance, but do not remove the inherent characteristics encoded in an individual’s health data and may thus not enough to provide real privacy against identity inference [5].

We demonstrate this risk with our similarity-based re-identification attack, that solely uses brief time-series health data of a target user for re-linking their de-identified data samples inside a data set back to them. Our attack utilizes Dynamic Time Warping (DTW) [11] for comparing time-series and exploits common features of the provided multi-modal sensor data. DTW delivers the distance between two samples, which however can be inverted to get our similarity measure. Figure 1 shows a matrix of such DTW similarities between 15 subjects, with the diagonal line comparing each subject to itself. The matrix shows how similarity varies

constantly, but the maximal difference in values is only 0.07 in distance. However, even these small differences offer the potential to distinguish the original individuals. We find our approach to be effective in breaking de-identification, especially in our example scenario, where data owners collect and leverage health data from smartwatches, which is why we emphasize stricter privacy measures.

Our contributions are:

- We propose a novel re-identification attack approach based on similarities from DTW alignments for multi-modal time-series data collected by smartwatches.
- We are also the first to evaluate data-specific optimization strategies that exploit the multi-modal and biological features of the underlying health data.
- Our results unravel and underscore the inherent re-identification threats that might be present in personal health data collected from smart devices.
- Our findings have practical relevance in smartwatch data collection scenarios which are currently being widely implemented using de-identification.

In Sect. 2 we briefly review the background relevant to our attack before focusing on existing related work in the domain in Sect. 3. Section 4 describes our attack and also introduces the example smartwatch scenario for our experiments. These experiments and their general outcomes are then outlined in Sect. 5. The following Sect. 6 is more centered around discussing the implications of our results and is divided into answering research questions and limitations regarding our approach. Finally, we provide both a concise summary of our findings and an outlook into future work in the conclusive Sect. 7.

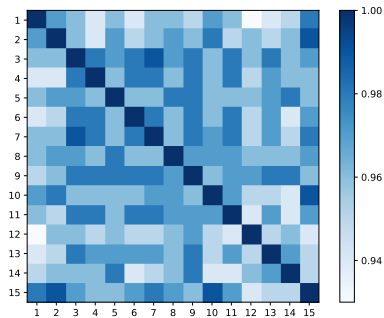


Fig. 1. Matrix of DTW alignments between 15 subjects, where higher scores indicate stronger similarity.

2 Background

In this section, we explain important concepts for better understanding this work.

2.1 Re-identification and Identity Inference

When de-identification is used as an anonymization technique, individual entries in a data set are stripped from their identifying personal information. This might include (user-)names, locations, affiliations, or other relevant metadata. In this way they are protected from harm, when their possibly sensitive data is released to the public, or at least this is what should be achieved.

However, there are many cases in which such privacy measures are broken and individuals are re-identified, especially in the context of health data [5]. In these cases, we see adversaries aiming to infer the identity behind a record from a data set or finding records related to a target individual, which opposes the concept of de-identification. The general goal of these attacks can be summarized under the term identity inference.

Another type of attacks are membership inference attacks introduced by Shokri et al. [17]. But these only aim to resolve whether a target is present in the data set at all, while identity inference goes one step further and tries to recognize the actual samples in the data set that belong to a target.

2.2 Dynamic Time Warping

DTW is a set of algorithms used to measure the similarity between temporal sequences based on alignment [11]. It aligns given time-series samples by minimizing the difference between corresponding elements, accommodating temporal distortions. To achieve this, the technique considers local temporal dependencies and enables flexible matching by warping or stretching one sequence to fit the other as closely as possible. The constructed alignment matrix quantifies the distance between each pair of elements to determine the overall correspondence.

3 Related Work

Several works propose similarity-based attacks based on the encoded data. The methods utilize the preservation of similarity in the original data space and the encoded data space.

In the domain of authentication, biometric images such as fingerprints are used as keys to log in to systems or applications. The original images are encoded to templates using e.g., Bloom filter, neural networks, etc. [15]. Due to the preservation of the similarity between original images and templates, similarity-based attacks aim to construct an image where the encoded template is similar to the target template. Therefore, similarity-based attack methods [3, 21] compare a fake template with a target template and iteratively optimize the construction process to obtain a new image being used to generate a new template.

Similarity-based attacks also exist in the context of privacy-preserving record linkage (PPRL). PPRL aims to identify duplicate records between two or more databases containing sensitive information. Therefore, the data owners encode the plain text data to encodings compared to determine duplicates [19]. Due to the preservation of similarities, the proposed attacks [2, 20] construct a graph consisting of records as nodes and similarities as edges using a publicly available plaintext database and the encoded one. The attack utilizes the similarity graphs to determine a mapping between nodes representing encoded and plaintext records based on similar graph features such as indegree/outdegree, PageRank, etc.

Due to the increasing relevance of sensors in manufacturing processes, mobility and life sciences, a tremendous amount of sensor data is collected and analysed. Especially, mobile devices such as mobile phones and smartwatches are equipped with various types of sensors. However, the collected data also bears the risk of endangering the privacy of users. For instance, accelerometer data can be used to predict the location of metro riders [12].

Recent work [14] proposed a re-identification attack using accelerometry data from 353 participants being recorded for 190,078 hours (70 days with at least 8 hours per day) resulting in 51.3 billion data points. The attack aims to determine the trace from an anonymized database regarding an available trace where the user is known. The attack computes similarities between the anonymized and known time series. Therefore, the traces are split into smaller segments to build meaningful features using a neural network. The network consists of convolutional layers and gated recurrent units to address the time aspect. Moreover, the base model classifies resulting features if the segment from the known user corresponds to the anonymized one. The authors suggest various aggregation strategies to determine the similarity between traces based on the segment similarities. In contrast to our approach, we do not utilize a supervised feature extraction and classification model where the performance depends on training data being rarely available. Our proposed method can be used for each available individual because we do not split the data into training and test data sets. Moreover, we consider various sensor data types and thus not only focus on accelerometry data. In contrast to our evaluation, the work only considers the true matching rate which does not allow a more differentiated view for the attacker.

4 Attack Description

This section covers a description of our approach from a methodological and experimental viewpoint. We focus on conveying the general idea and how we designed the actual attack implementations in our experiments in order to enable a thorough evaluation for the involved factors.

4.1 Threat Model

Our attack aims at showing the possible threat from just a short sample of sensor data from a target that contains enough information to identify them inside a

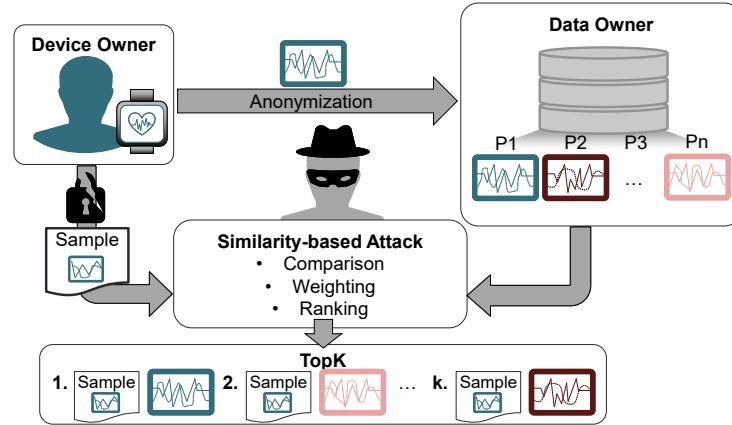


Fig. 2. The attack scenario consists of a device owner (target), a data owner and an attacker. The device owner sent the collected data via his device to the data owner who anonymizes and stores the data for analysis capabilities. The attacker aims to determine the corresponding time series data maintained by the data owner utilizing a sample from a known device owner.

database of other but similar samples based solely on the similarities in the given health data. Our threat model is illustrated in Fig. 2 and can be described by a scenario in which health data collected from e.g. smartwatches is handled by a data owner. The data owner here is different from the smartwatch or device owner and could be a company, institution, or person. The data owner wants to use this user data for improving their product or providing smart health features like stress detection through training machine learning models. To ensure privacy for device owners their incoming data is anonymized by removing any identifying information like name or location.

However, our proposed attack allows an adversary to still deduce the original identity of such data samples, thereby destroying any privacy guarantees. The only prerequisite is a short existing sample of data from the target to enable the similarity comparison. This could be obtained from intercepting data before anonymization, other more direct hacking methods, or simply from data repositories which were collected before anonymization promises were given. The assumed attacker could be an insider on the data owner side or personally targeting a user to gain information on them. It could as well be the data owner itself, who wants to get back his lost information from the given privacy promises.

By successfully performing our similarity attack on the time-series signals recorded from the target's device, the attacker can correlate the data samples back to them. One objective could be to collect more data stemming from the target. In any case, user privacy would be broken irreparably, making the anonymization defacto useless in terms of real provided security. In our described scenario it does not matter if user data is stored or only processed before deletion, since the attack can also be executed directly on any arriving data sample.

4.2 Data Model

Our attack can be performed on time-series health data in general but we also want to show how an informed attacker might use background knowledge to their advantage. For our data basis we thus decided on the WESAD data set introduced by Schmidt et al. [16]. It consists of 15 healthy subjects (12 male, 3 female), each of whom features a health data recording of about 36 minutes in length, which was acquired during a lab study. During this period data was continuously and simultaneously gathered from a wrist- and a chest-worn device, which both offer multiple modalities as time-series data. At each point in time the subjects' current status is assigned as one of three affective states: neutral, stress, or amusement.

Therefore, the usual focus of this data set is on stress detection from wearable devices based on their collected time-series sensor data. Since stress detection is a common feature that data owners provide as analysis for their users, we find it a fitting basis for our attack. We furthermore restrict the sensor modalities to the ones available from smartwatch devices, since these would be the most realistic and common devices when it comes to personal health data in everyday life. There is also already existing work focused on maintaining privacy when building machine learning models based on such sensitive personal health data [13]. With our attack, we can now deliver further reasoning in favor of using such privacy-preserving techniques. Our pre-processing steps for the data set is adapted from the methods described by Gil-Martin et al. [10].

4.3 Attacker Model

The attacker model definition refers mainly to the practical design and capabilities of the attacker in our described threat model from Sect. 4.1. We want the attacker to be closely related to the envisioned threat model given in Fig. 2, while also taking into account the existing possibilities and limitations imposed by the format of relevant time-series data as seen in Sect. 4.2. The attack preparation procedure described in the following can be directly adapted to other data sets but here sometimes refers to the WESAD data set for better explanation of specific details in the approach. A fundamental assumption is that all used time-series data undergoes the same pre-processing to at least ensure common frequency rates and windowing of features. Dividing between device owner (target) and data owner is simply achieved by selected one of the 15 subjects from the WESAD data set as a target, while the rest is taken as other samples present in the attacked collection of the data owner.

Now, the main difficulty is the separation of target data into a sample already owned by the attacker, called attacker set, and other data stored by the data owner, which should then be matched by the similarity attack. To achieve this, we cut out a snippet of predefined length from the target data to constitute the attacker set. The cut out part is removed from the time-series and the resulting shortened version is then added to the other samples in the data owner set. The attacker has information on his attacker set in the form of knowing that the

possessed sample belongs to the target. The collection of samples on the data owner side is anonymized and stripped from other metadata, making only the time-series data itself available to the attacker. It is important to be clear, that the attacker might as well be the data owner himself given this scenario.

Repeating this process for each of the 15 subjects in the data set leads to 15 single targets. We therefore target each subject once and create an attacker set snipped from the middle of their time-series, as well as, a data owner collection accordingly. The attacker set is the brief known sample of the attacker, while the data owner collection might contain other samples but also the rest of the target data not used as the attacker set. All data samples consist of time-series for each of the six provided sensor modalities by the smartwatch device of the WESAD data set, which are: blood volume pulse (BVP), electrodermal activity (EDA), body temperature (TEMP), and three-axis acceleration (ACC). We only name four instead of six sensors, but the ACC is actually recording each axis individually leading to an x, y, and z value at each time point and resulting in three separate sensor time-series. Since these three series are inherently correlated, we combine their individual similarity values into one sensor similarity result.

The attacker performs the similarity attack by calculating the DTW alignment scores between two given samples for each of their six sensors. The attack process always takes the attacker set as first reference and then picks one of the samples from the data owner collection. The attacker thereby assigns similarity scores to every candidate from the collection to then decide which sample was the original target. Now having a dictionary of scores for each sample and each sensor, there are multiple methods for proceeding with the attack. For example, the sensors can be combined and weighted according to their importance in the similarity scoring. Our evaluation of such options is described in the following Sect. 5.

5 Experiments

In this part, we detail the different settings for our experiments and evaluate them. We decide on a top-down stacking approach for our experimental setup, in which we examine individual parts of the parameters separately. For this purpose, the other influencing factors are averaged and only the aspect of interest is put into focus. In this way we can break down the complexity in such a way that we first choose the best general decisions and can then use these for the more detailed parameter investigations. Each experiment and this process are further described in their respective subsections.

5.1 Environment

On the software side we employ Python 3.9 as our programming language. For our DTW needs we utilize the open-source implementation from the dtw-python library [11]. For the experiments, the hardware configuration comprises machines with 64GB of RAM and eight cores of an AMD EPYC 7551P CPU each, facilitating efficient computations.

5.2 Evaluation Metrics

To allow a meaningful evaluation, we first have to decide on representative metrics relevant to our goals. In our attack scenario, the goal is to rank a set of data samples relative to their similarity to our existing target data. In this ranking scheme, the included de-identified samples of the target should be ranked as high as possible for our attack to be successful. We thus find the Precision@k ($P@k$) as the most relevant measure of success.

In general, the $P@k$ metric gives the proportion of relevant items among the top-k retrieved items and a higher $P@k$, therefore, indicates more relevant results in the top-k list. In our case, we want the correct target sample to be as highly ranked as possible and we thus only measure its inclusion or absence in the top-k list. The most relevant k-value to us is $k=1$ since it represents the likelihood of having our target data in the top-ranked spot, allowing direct re-identification. To get a better picture, we also include $k=3$, $k=5$, and the k-value at which $P@k$ reaches its 1.0 maximum ($\max@k$). The $\max@k$ metric can be important, because e.g. a maximum precision at $k=10$ tells us that our target is always included among the first 10 results, making the last 5 ranks obsolete and removable from our search. When deciding on the best-performing version of our attack, we mainly look at the most threatening case of $k=1$. If the results are equal, we then increase the k-value step by step until an unambiguous decision is possible.

The WESAD data set includes 15 subjects for evaluation. To perform a full sweep, we take each subject as a target once and perform a full attack evaluation. A full attack includes taking a target sample from the data and calculating the similarity scores of all subjects before ranking them. Subsequently, we derive the $P@k$ results regarding the 15 provided ranks. Finally, all 15 of these single-target attack results are then averaged based on their $P@k$ results to generate the overall precision values.

5.3 Ranking Methods

We first have to decide on a ranking method for our DTW alignment or similarity scores. Since we get individual scores at the sensor level, we need to combine them to decide on a final score for ranking the data samples in terms of their overall similarity. Our first option takes on the keyword *score* and is the simplest aggregation form, in which we calculate the mean overall sensor similarity scores and then rank the data samples based on this averaged result. Our second proposal is denoted as *rank* and takes into account, that some sensors might match really well, while others might not align at all. We thus rank each sensor individually, assigning one rank to the similarity score of each sensor based on the other tested samples. The individual ranks are then averaged into an overall ranking, that is used for the final re-ranking of the data samples. So, if a sensor score is low for all subjects but other results are better, the lower overall score could substantially draw down the averaged result for the *score* method, while *rank* might be more forgiving and instead focused on the outperforming similarity scores provided by the other sensors. In any cases, where multiple subjects share

the same rank, we refer to methods by Berrendorf et al. [1] to first apply a realistic ranking and if necessary follow with a pessimistic approach.

The results for our ranking methods are shown in Table 1, where we assume averaged results regarding the other experiment parameters to allow this completely separate comparison. Even though the *rank* method is slightly better at higher k-values, our preferred ranking method for further experiments is *score* based on its best-performing result at $k = 1$. We also included the mean result of both methods, which promotes the individual advantageous characteristics, delivering a good compromise over all k-values. None of the two methods is able to lower the $\max@k$ value from the worst-case $k=15$, which is the minimum since we rank 15 subjects.

Table 1. Ranking methods compared on Precision@k (P@k) with $\max@k$ giving the k-value, where P@k reached its 1.0 maximum.

P@k	Methods		
	rank	score	mean
k=1	0.178	0.188	0.183
k=3	0.439	0.428	0.433
k=5	0.601	0.588	0.594
$\max@k$	k=15	k=15	k=15

5.4 Classes

When we talk about classes, we are referring to the different labelled classes present in the WESAD data set. Each data point in the provided time-series samples is assigned a label out of neutral, stress, and amusement. This classification might be relevant to how good a person can be recognized since a subject’s recorded modalities can differ depending on his affective state [9]. The neutral is probably the most common state in daily life, while stress and amusement can certainly also appear regularly. Because the target data available to the attacker could be from any of the three types, we evaluate all of them individually to find the most dangerous in terms of privacy. In the end it would however be unrealistic to assume a single specific type of data in the attacker’s possession, which is why we instead focus on a mean value. Taking the standard mean would in turn miss the natural prevalence of neutral data, which is why we consider a weighted mean over the classes to be the most representative and effective for our data model. We thus introduce a weighted mean with weights of 0.53 for neutral, 0.3

Table 2. The three labelled classes from the WESAD data set compared on Precision@k (P@k) with $\max@k$ giving the k-value, where P@k reached its 1.0 maximum.

P@k	Classes			mean	weighted mean
	neutral	stress	amusement		
k=1	0.226	0.165	0.172	0.188	0.199
k=3	0.503	0.373	0.406	0.427	0.448
k=5	0.674	0.504	0.586	0.588	0.608
$\max@k$	k=15	k=15	k=15	k=15	k=15

for stress and 0.17 for amusement. These weights are calculated based on the average appearance of the classes in the data set.

Table 2 gives the results for our class experiments, where we take the score ranking method based on Sect. 5.3 and assume averaged results regarding the other experimental parameters. Conforming to expectations, the most prevalent neutral class is also the most effective for finding similarities, simply based on the available data for comparison. This trend is however not continued when comparing the stress and amusement classes, which on average constitute 30% and 17%, respectively. Here, the less appearing amusement class shows to be better for identifying the target, even though it provides just about half of the data compared to stress. In turn, the best similarity scores are not only dependent on available amounts of data but rather also heavily factor in the class-specific peculiarities in sensor readings, which show to be better for distinguishing between subjects in some cases.

Finally, however, it is unclear what kind of class the data samples of the attacker and also data owner collection might have. Therefore, for our proposed attack, it is most realistic to assume the weighted mean, which is based on the existing distribution of the classes in the data set. This result is also preferable to the simple mean, since individual classes can be over- or underrepresented, which could distort the actual threat depending on the underlying data.

5.5 Sensor Combinations

Our multi-modal sensor data consisting of ACC, BVP, EDA, and TEMP data, allows for generating an individual similarity scoring for the time series of each modality. Our attack might perform better or worse given different types of sensor data, since some could exhibit more recognizable differences between subjects, which would make them more relevant to our similarity comparison. Consequently we can choose from different possible combinations and for finding the best sensor types regarding our attack we evaluate each possibility. For now, when we combine sensors, we just calculate the equally weighted average regarding their individual scores.

As before, we take the earlier experimental results as our basis for this evaluation and now present the possible combinations along their outcomes in Table 3. On their own, BVP and ACC already seem to exhibit strong characteristics for our similarity comparison. The best results over all k -values are however obtained, when combining them with each other (BVP+ACC). Combinations with the other sensors can be beneficial but not to the same extent. At our desired level of $k=1$, the second best combination uses all four sensors, which then falls behind other variations at $k=3$ and $k=5$. The additionally gained dimensions from involving more than two sensor-level dependencies when comparing two data samples of the target, are however never enough to match BVP+ACC. This also means, that the other sensors can be omitted and are not needed for ranking in our attack scenario. We can also see first-time reductions in the $\max@k$ metric, where we now reach the maximum precision at already $k=14$ instead of the worst-case of $k=15$. We will therefore use BVP+ACC for further experiments.

Table 3. All possible combinations of the six sensor modalities from the WESAD data set compared on Precision@k (P@k) with max@k giving the k-value, where P@k reached its 1.0 maximum.

Sensor combinations	P@k			
	k=1	k=3	k=5	max@k
BVP	0.190	0.499	0.677	k=15
EDA	0.104	0.252	0.457	k=15
ACC	0.223	0.510	0.668	k=15
TEMP	0.127	0.357	0.456	k=15
BVP+EDA	0.179	0.448	0.626	k=15
BVP+ACC	0.375	0.643	0.754	k=14
BVP+TEMP	0.228	0.439	0.559	k=15
EDA+ACC	0.135	0.382	0.575	k=15
EDA+TEMP	0.107	0.337	0.520	k=15
ACC+TEMP	0.224	0.382	0.600	k=15
BVP+EDA+ACC	0.232	0.547	0.705	k=15
BVP+EDA+TEMP	0.162	0.446	0.603	k=15
BVP+ACC+TEMP	0.237	0.541	0.675	k=15
EDA+ACC+TEMP	0.187	0.409	0.586	k=15
BVP+EDA+ACC+TEMP	0.270	0.523	0.659	k=15

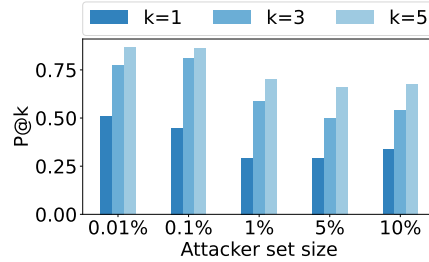
5.6 Attacker Set Sizes

An adversary might have different amounts of target data available for comparison in his similarity attack. We therefore test multiple sizes of attacker sets to evaluate varying data needs and corresponding threat levels. The sizes are chosen relative to the total target data, with our largest portion being 10%, i.e. an average of 3.6 minutes, and the smallest being 0.01%, i.e. 210ms which corresponds to exactly one single window from our time-series pre-processing.

The results for our tested attacker set sizes are accumulated in Fig. 3, which also assume the choices from previous experiments. The largest set of 10% first seems on par when reducing to 5% or 1%, but we can then quickly derive an ongoing trend towards shorter lengths increasing our attack performance. This might seem counterintuitive, but shorter samples allow for capturing even the tiniest overlaps of alignments found in the compared times-series data, whereas longer samples consequently need to match longer sequences correctly to achieve the same level of alignment. It thus seems to be most effective to promote shorter samples for similarity comparison, with our shortest attacker set of 0.01% size performing the best at k=1, which is about 210ms in length and corresponds

P@k	Attacker set sizes				
	0.01%	0.1%	1%	5%	10%
k=1	0.510	0.448	0.291	0.291	0.337
k=3	0.776	0.810	0.590	0.497	0.542
k=5	0.867	0.862	0.700	0.662	0.677
max@k	k=11	k=11	k=13	k=14	k=14

(a) Tabular result overview.



(b) Graphical result illustration.

Fig. 3. (a) Different attacker set sizes compared on Precision@k (P@k) with max@k giving the k-value, where P@k reached its 1.0 maximum. (b) Illustration of the relation between set size and attack success (P@k) for different k-values.

to one time-series window after pre-processing. It is therefore also the smallest set possible regarding our pre-processing and our attacker set size of choice. We additionally note a further change in the max@k metric, which is found to at k=11 instead of the earlier best of k=14 from Sect. 5.5.

5.7 Optimized Sensor Weighting

Our naive DTW results are based on all findings from our previous experiments with an attacker set size of just 0.01%, i.e. 210ms, and are already a threat to privacy in our scenario. Based on the knowledge of the underlying data, we can however further improve our attack by tailoring the approach to the available sensor modalities. To simulate a more sophisticated attacker that is closer to a worst-case scenario, we use a grid search approach to find the optimal sensor weightings based on our task and data. These weightings allow us to focus on the alignment scores calculated from the most relevant signals for re-identification while omitting or undervaluing the less important ones. When combining the DTW alignment scores into our similarity ranking we then factor in each signal-specific score depending on its found optimal weighting. This difference to the method applied in Sect. 5.5 allows us to factor in less relevant sensors with reduced impact so that they may provide some useful insights without harming the more important sensor alignments.

The radar charts regarding the results for our optimized weightings are given in Fig. 4, where we show the found combinations for each class at the three k-values $k=\{1,3,5\}$. In some cases there were multiple sensor weightings that produced the same attack results for a given k-value, which makes them equally desirable. The frequency of selected weighting values corresponds to the transparency of the area in the heat maps, where more transparent areas show values that saw less selection. Due to the large number of possibilities, we limit the results to illustrations at this point and instead refer to Appendix A for a complete tabular rundown of the weightings for each class and k-value.

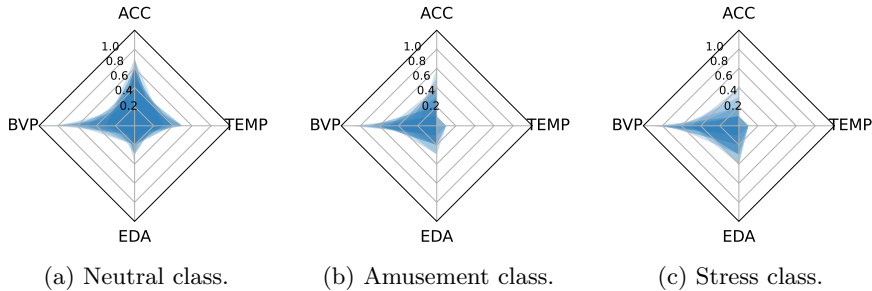


Fig. 4. Illustrations of the best found sensor weightings for each class at $k=\{1,3,5\}$. At each k -value, we included all combinations that delivered the same optimal results. A detailed overview on all combinations is given in Appendix A.

In general the classes show some preferences when it comes to the most relevant sensors for our similarity attack. We first see that our weightings favor the ACC and BVP signals, as seen before in Sect. 5.5. The amusement and stress classes more often focus on BVP over ACC, while the neutral class rather tends to have a more equal distribution. We see some inclusion of EDA in just a few stress cases and almost no inclusion of TEMP in amusement or stress. The neutral class on the other hand includes TEMP in multiple combinations.

Based on these elaborate weighting results, our more tailored solution is compared to our naive results in Table 4. With optimized sensor weighting, we are able to improve the $P@k$ for $k=1$ by 19.6% from 51.0% to 70.6%, leading to a drastically higher probability of directly identifying the target. At $k=3$ and $k=5$, our new method still outperforms by 13.7% and 9.7%, respectively. Another substantial improvement is visible for our $\max@k$ evaluation, where we now find the perfect score at $k=6$ already, which almost halves the previous result from our naive approach.

Table 4. Our naive DTW and sensor-weighted approaches compared on Precision@k ($P@k$) with $\max@k$ giving the k -value, where $P@k$ reached its 1.0 maximum.

P@k	Approaches	
	naive	weighted
k=1	0.510	0.706
k=3	0.776	0.913
k=5	0.867	0.964
$\max@k$	k=11	k=6

6 Discussion

In the first part, we structure our discussion by raising and addressing the main research questions stemming from our experiments and their results. In the second part, we then focus on the limitations of our approach and experimental design.

6.1 Derived Research Questions

The following collected research questions might arise from the presented results or also from the general attack description. We try to briefly answer them, giving more context to the outcomes detailed in this work.

RQ1: How severe is the actual threat level achieved in our example scenario?

To put our results into perspective, we first compare them to the probability of successfully random guessing the correct target to be in to the top-k results. This probability is given by $p = k/N$, where k is the top-k-value for the P@k and N is the number of possible ranks, which here equals the number of possible subjects ($N=15$). Therefore, the chances of ranking the subjects by random guessing and putting our target at the top spot ($k=1$) would be a probability of $p = 1/15 \approx 0.067$. The full comparison of our results to random

guessing is plotted in Fig. 5, where we see that our sensor-weighted approach performs more than $10\times$ better than random guessing at $k=1$, over $4\times$ better at $k=3$, and almost $3\times$ better at $k=5$. We are thus able to directly identify the correct target data in two out of three cases. These substantial advantages over the random chances put the privacy loss through our attack in context since random guessing results would instead translate to perfect privacy.

Another factor we consider devastating to privacy is our ability of simply reducing the candidate space through our attacks. When our attack achieves the maximum of $P@k=1.0$ at a given k-value, this result translates to the guaranteed inclusion of the target data in these top-k results. Consequently, we can omit all other candidate samples except for the top-k ranked ones from consideration. By that, we drastically reduce the search space and thereby also the difficulty of re-identification from e.g. additionally linked external information. For our best sensor-weighted attack we would be able to safely reduce the relevant subjects to a subset of 6 out of 15 candidates. Now, external information on e.g. existing heart disease or other medical conditions could further help in identifying the target and would be easier than before, because there are less samples to consider.

These conclusions can of course only be drawn for our example scenario, while in general we have to point out the limitations which we consider in Sect. 6.2.

RQ2: How much is the attack tailored to this exact task? Can it be adapted to other data sets? Our attack is generally transferable to other similar cases based on the description and process provided in Sect. 4. This is especially true for the naive approach, while the sensor-weighted approach on the other hand might not always work depending on the new data. Still, many tasks allow similar exploits with knowledge about the underlying data characteristics.

Our achieved results can also find use as a pre-trained attack model with already optimized parameters through our experiments. Even the provided sen-

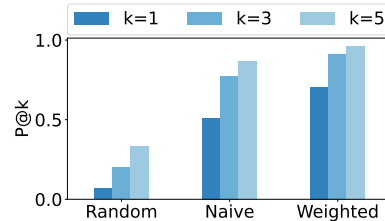


Fig. 5. Our attack results for the naive and the sensor-weighted methods compared against the probability of correctly ranking the target in the top-k results by random guessing.

weighting can be directly applied on newly collected data, giving a strong foundation for attacks. With our investigation, we generally want to convey the idea that similarity attacks can have a threat to time-series health data use cases and also show through our weighting approach, that some modalities are more threatened than others.

RQ3: Are there also beneficial use cases for such similarity searches?

Yes, one such use case would be linking similar patients or subjects for quickly improving personalized health applications. Such tasks could include of course stress detection, but also tuning a person’s medication dosage. By finding similar subjects in an existing data basis, a newly added individual might profit from the existing data when adjusting these applications to their personal but similar needs and health situation. We thus also see a favorable outcome from our attack.

RQ4: What are possible defence mechanisms that are more suitable than simple de-identification?

To defend against our attack, it is not enough to remove the identifying metadata for a data sample, since the similarity is calculated with respect to the data itself. Instead it would be necessary to hide these similarities between the data sample and the acquired target sample by directly breaking or hiding these links.

One option would be to only allow the collection of a data sample when enough closely similar samples of other subjects are already present in the data, making it harder to distinguish between them and reducing sensitivity. This would be comparable to solutions like k-anonymity [18]. However, this would depend heavily on the available data and could possibly remove large amounts of threatened but needed samples from such data sets.

Another solution would be to directly alter the data in such a way, that the differences between them are hidden well enough already, which would give the option of making any collected data more private. A widespread standard is the application of differential privacy, a theoretical privacy guarantee introduced by Dwork [4]. Differential privacy induces selective amounts of random noise into the data that ultimately makes each sample indistinguishable from each other.

6.2 Limitations

We are mainly limited by the available data for testing our approach. A deciding factor for this limitation are the data owners of smart device health data, which are presently for the most part the responsible companies themselves. Consequently, we are mostly dependent on study data from the public domain, where participant counts are way lower than users of e.g. smartwatches in daily life.

Like others, the publicly available WESAD data set thus only includes a low count of 15 subjects. It however offers a wider range of sensor modalities than other slightly larger data sets, which is closer to what real devices are able to collect and more realistic than just focusing on a single signal. Regarding this duality, there are no other bigger or less specialized data sets that provide the same qualities for our task, making a throughout evaluation of similarity attacks difficult. We of course expect diminishing success for uncovering the correct

identity, when possible candidate samples increase inside the data collection. It is however unclear, to what extent we really would be able to retrieve a target's data in such cases and the attack might still pose a critical privacy risk.

Another factor that is out of scope due to low data amounts is rigorous testing regarding scalability. With more samples to compare, the computing time for our DTW alignments could increase dramatically. Also, the comparison of longer time-series might show problematic because of the quadratic increases in computing needs caused by the underlying $n \times n$ Matrix of the DTW implementation.

For now, our results prove the possible severity of the problem and motivate further studies on hopefully larger data sets in the future.

7 Conclusion

With the increasing popularity of smartwatches in today's market, we also see an ever-increasing amount of collected personal health data flowing to their respective companies. Although these data collections usually support the improvement and creation of smart health services for users, they can also threaten their privacy. Our proposed attack highlights one potential privacy risk involved when working with personal health data from IoT and smart devices. Cases of identity inference are of special relevance where we usually find de-identified data, like in our example scenario on smartwatch data. Further, our optimizations exploiting the multi-modal and biological nature of such data show how a knowledgeable adversary is able to steeply increase the chances of finding a target individual. Data-driven optimization based on the underlying characteristics therefore shows to be a promising approach in similarity attacks.

For reducing the main limiting factor, future work should be primarily focused on acquiring more data for evaluating the scalability and relevance of our attack on larger data sets. Synthesizing artificial data through generative adversarial networks could prove advantageous in both directions by increasing available data for evaluation but also for pre-setting our sensor weightings on such data.

Availability. Reference code for all experiments is available from our repository at <https://github.com/tobiasschreieder/smartwatch-dtw-attack>.

Ethical Principles. All health data originated from public sources provided for research purposes and was solely used within the limited scope of this work.

Acknowledgments. We thank Nils Wenzlitschke for his contributions to our pre-processing implementation. The authors acknowledge the financial support by the Federal Ministry of Education and Research of Germany and by the Sächsische Staatsministerium für Wissenschaft Kultur und Tourismus in the program Center of Excellence for AI-research "Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig", project identification: ScaDS.AI. Computations for this work were done (in part) using resources of the Leipzig University Computing Centre.

References

1. Berrendorf, M., Faerman, E., Vermue, L., Tresp, V.: On the ambiguity of rank-based evaluation of entity alignment or link prediction methods. arXiv preprint arXiv:2002.06914 (2020)
2. Cullane, C., Rubinstein, B.I., Teague, V.: Vulnerabilities in the use of similarity tables in combination with pseudonymisation to preserve data privacy in the UK Office for National Statistics’ privacy-preserving record linkage. arXiv preprint arXiv:1712.00871 (2017)
3. Dong, X., Jin, Z., Jin, A.T.B.: A genetic algorithm enabled similarity-based attack on cancellable biometrics. In: 10th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2019, IEEE (2019)
4. Dwork, C.: Differential privacy. In: ICALP 2006, pp. 1–12, Springer (2006)
5. El Emam, K., Jonker, E., Arbuckle, L., Malin, B.: A Systematic Review of Re-Identification Attacks on Health Data. PLoS ONE **6**(12), e28071 (2011)
6. Ernst, C.P., Ernst, A.: The Influence of Privacy Risk on Smartwatch Usage. AMCIS 2016 Proceedings (2016)
7. Fitbit Health Solutions: Research pledge (2023), URL <https://healthsolutions.fitbit.com/research-pledge/>, accessed on July 7, 2023
8. Fitbit International Limited, Fitbit LLC: Fitbit privacy policy (2023), URL <https://www.fitbit.com/global/us/legal/privacy-policy>, accessed on July 7, 2023
9. Giannakakis, G., Grigoriadis, D., Giannakaki, K., Simantiraki, O., Roniotis, A., Tsiknakis, M.: Review on psychological stress detection using biosignals. IEEE Transactions on Affective Computing **13**(01) (2022)
10. Gil-Martin, M., San-Segundo, R., Mateos, A., Ferreiros-Lopez, J.: Human stress detection with wearable sensors using convolutional neural networks. IEEE Aerospace and Electronic Systems Magazine **37**(1), 60–70 (2022)
11. Giorgino, T.: Computing and visualizing dynamic time warping alignments in R: The dtw package. Journal of Statistical Software **31**(7), 1–24 (2009)
12. Hua, J., Shen, Z., Zhong, S.: We can track you if you take the metro: Tracking metro riders using accelerometers on smartphones. IEEE Trans. Inf. Forensics Secur. **12**(2), 286–297 (2017)
13. Lange, L., Degenkolb, B., Rahm, E.: Privacy-preserving stress detection using smartwatch health data. In: 4. Interdisciplinary Privacy & Security at Large Workshop, INFORMATIK 2023 (2023)
14. Saleheen, N., Ullah, M.A., Chakraborty, S., Ones, D.S., Srivastava, M., Kumar, S.: Wristprint: Characterizing user re-identification risks from wrist-worn accelerometry data. In: Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, p. 2807–2823, CCS ’21 (2021)
15. Sandhya, M., Prasad, M.V.N.K.: Biometric Template Protection: A Systematic Literature Review of Approaches and Modalities, pp. 323–370. Springer International Publishing, Cham (2017)
16. Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., Van Laerhoven, K.: Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. In: ACM ICMI ’18, pp. 400–408 (2018)

17. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: S&P 2017, IEEE (2017)
18. Sweeney, L.: k-anonymity: A model for protecting privacy. *INT J UNCERTAIN FUZZ* **10**(05), 557–570 (2002)
19. Vatsalan, D., Sehili, Z., Christen, P., Rahm, E.: Privacy-Preserving Record Linkage for Big Data: Current Approaches and Research Challenges. In: *Handbook of Big Data Technologies*, pp. 851–895, Springer (2017)
20. Vidanage, A., Christen, P., Ranbaduge, T., Schnell, R.: A graph matching attack on privacy-preserving record linkage. In: *International Conference on Information and Knowledge Management*, pp. 1485–1494, ACM (2020)
21. Yang, Z., Leng, L., Zhang, B., Li, M., Chu, J.: Two novel style-transfer palmprint reconstruction attacks. *Appl. Intell.* **53**(6), 6354–6371 (2023)

A Sensor Weightings

Table 5. Optimal sensor weight combinations for the three classes and relevant k-values.

Sensor weights					Sensor weights					Sensor weights					
k-value	ACC	BVP	EDA	TEMP	k-value	ACC	BVP	EDA	TEMP	k-value	ACC	BVP	EDA	TEMP	
k=1	0.4	0.3	0.2	0.1	k=1	0.6	0.2	0.2	0.0	k=1	0.1	0.4	0.4	0.1	
	0.4	0.1	0.2	0.3		0.5	0.3	0.2	0.0		0.4	0.5	0.1	0.0	
	0.5	0.1	0.2	0.2		0.4	0.4	0.2	0.0		0.1	0.5	0.3	0.1	
	0.6	0.1	0.2	0.1		0.4	0.5	0.1	0.0		0.2	0.5	0.3	0.0	
	0.5	0.2	0.1	0.2		0.2	0.7	0.1	0.0		0.3	0.6	0.1	0.0	
	0.5	0.2	0.2	0.1		k=3	0.1	0.6	0.2		0.1	0.2	0.7	0.1	0.0
	0.6	0.3	0.1	0.0			0.3	0.4	0.3		0.0	0.1	0.8	0.1	0.0
	0.7	0.0	0.0	0.3			0.3	0.4	0.3		0.0	0.0	0.8	0.2	0.0
	0.4	0.0	0.1	0.5			0.3	0.5	0.2		0.0	0.1	0.5	0.3	0.1
	0.5	0.0	0.1	0.4			0.1	0.5	0.3		0.1	0.0	0.5	0.4	0.1
k=3	0.6	0.0	0.1	0.3	k=5	0.1	0.6	0.2	0.1	k=3	0.1	0.6	0.2	0.1	
	0.7	0.0	0.1	0.2		0.2	0.6	0.2	0.0		0.0	0.7	0.3	0.0	
	0.4	0.0	0.2	0.4		0.2	0.7	0.1	0.0		0.0	0.8	0.1	0.1	
	0.5	0.0	0.2	0.3		0.0	0.8	0.1	0.1		0.0	0.8	0.2	0.0	
	0.6	0.0	0.2	0.2		0.1	0.8	0.1	0.0		0.1	0.9	0.0	0.0	
	0.5	0.0	0.3	0.2		0.1	0.9	0.0	0.0		0.0	0.9	0.1	0.0	
	0.4	0.1	0.0	0.5		0.0	0.9	0.1	0.0		k=5	0.0	0.9	0.1	0.0
	0.5	0.1	0.0	0.4		(b) Amusement class.	(c) Stress class.								
	0.6	0.1	0.0	0.3											
	0.7	0.1	0.0	0.2											
0.3	0.1	0.1	0.5												
0.4	0.1	0.1	0.4												
0.5	0.1	0.1	0.3												
0.6	0.1	0.1	0.2												
0.3	0.1	0.2	0.4												
0.4	0.1	0.2	0.3												
0.5	0.1	0.2	0.2												
0.6	0.1	0.2	0.1												
0.4	0.1	0.3	0.2												
0.5	0.1	0.3	0.1												
0.5	0.1	0.4	0.0												
0.3	0.2	0.0	0.5												
0.4	0.2	0.0	0.4												
0.5	0.2	0.0	0.3												
0.6	0.2	0.0	0.2												
0.4	0.2	0.1	0.3												
0.5	0.2	0.1	0.2												
0.6	0.2	0.1	0.1												
0.5	0.2	0.2	0.1												
0.6	0.2	0.2	0.0												
0.5	0.2	0.3	0.0												
0.5	0.3	0.0	0.2												
0.7	0.3	0.0	0.0												
0.5	0.3	0.1	0.1												
0.6	0.3	0.1	0.0												
0.5	0.4	0.1	0.0												
0.4	0.4	0.2	0.0												
0.4	0.5	0.1	0.0												
0.3	0.5	0.2	0.0												
0.3	0.6	0.1	0.0												
0.2	0.6	0.2	0.0												
0.2	0.7	0.1	0.0												
0.2	0.8	0.0	0.0												
0.1	0.8	0.1	0.0												
0.1	0.9	0.0	0.0												

(a) Neutral class.