

Generating Synthetic Health Sensor Data for Privacy-Preserving Wearable Stress Detection

Lucas Lange^{1,*†} , Nils Wenzlitschke¹ and Erhard Rahm¹ 

¹ Leipzig University & ScaDS.AI Dresden/Leipzig, Augustusplatz 10, 04109 Leipzig

* Correspondence: lange@informatik.uni-leipzig.de

† Current address: Room P416, Augustusplatz 10, 04109 Leipzig.

Abstract: Smartwatch health sensor data is increasingly utilized in smart health applications and patient monitoring, including stress detection. However, such medical data often comprises sensitive personal information and is resource-intensive to acquire for research purposes. In response to this challenge, we introduce the privacy-aware synthetization of multi-sensor smartwatch health readings related to moments of stress. Our method involves the generation of synthetic sequence data through Generative Adversarial Networks (GANs), coupled with the implementation of Differential Privacy (DP) safeguards for protecting patient information during model training. To ensure the integrity of our synthetic data, we employ a range of quality assessments and monitor the plausibility between synthetic and original data. To test the usefulness, we create private machine learning models on a commonly used, albeit small, stress detection dataset, exploring strategies for enhancing the existing data foundation with our synthetic data. Through our GAN-based augmentation methods, we observe improvements in model performance, both in non-private (0.45% F1) and private (11.90–15.48% F1) training scenarios. We underline the potential of differentially private synthetic data in optimizing utility-privacy trade-offs, especially with limited availability of real training samples.

Keywords: generative adversarial network; stress recognition; privacy-preserving machine learning; differential privacy; smartwatch; time series; physiological sensor data; synthetic data; smart health

1. Introduction

Healthcare applications see an ever-growing need for high quality medical data in abundant amounts. In particular, the uprising of smart health services can provide valuable insights into individual health conditions and personalized remedy recommendations. For example, solutions for detecting stress from physiological measurements of wearable devices receive attention from academic [1–3] and industrial [4–6] communities alike.

However, a major problem lies in the sensitive nature of the analyzed data. Each entry in a medical dataset may contain extensive information about an individual’s health status, making them highly sensitive even under anonymization techniques [7–10]. These dangers ultimately result in many complex privacy requirements, that make it difficult to collect and access large enough amounts of patient data for any real-world research purposes [11].

Regarding patient privacy, training machine learning models under the constraints of Differential Privacy (DP) [12] provides a robust and verifiable privacy guarantee. This approach ensures the secure handling of sensitive data and effectively mitigates the risk of potential attacks when these models are deployed in operational settings.

To address the limitations related to data availability, one effective strategy is the synthesis of data points, often achieved through techniques like Generative Adversarial Networks (GANs) [13]. GANs enable the development of models that capture the statistical distribution of a given dataset and subsequently leverage this knowledge to generate new synthetic data samples that adhere to the same foundational principles. In addition, we can directly integrate the privacy assurances of DP into our GAN training process, enabling the direct creation of a privacy-preserving generation model. This ensures that the synthetically generated images offer and maintain privacy guarantees [14].

In this work, we train both non-private GAN and private DP-GAN models for generating new time-series data needed for smartwatch stress detection. Existing datasets for

stress detection are small and can benefit from augmentation, especially when considering difficulties in private training of detection models using DP [15]. We present and evaluate multiple strategies for incorporating both non-private and private synthetic data to enhance the utility-privacy trade-off introduced by DP. Through this augmentation, our aim is to optimize the performance of privacy-preserving models in a scenario where we are constrained by limited amounts of real data.

Our contributions are:

- We achieve data generation models based on GANs that produce synthetic multi-modal time-series sequences corresponding to available smartwatch health sensors. Each data point presents a moment of stress or non-stress and is labelled accordingly.
- Our models can generate realistic data that is close to the original distribution, allowing us to effectively expand or replace publicly available, albeit limited, data collections for stress detection while keeping their characteristics and offering privacy guarantees.
- With our solutions for training stress detection models with synthetic data, we are able to improve on state-of-the-art results. Our private synthetic data generators for training DP-conform classifiers, help us in applying DP with much better utility-privacy trade-offs and lead to higher performance than before. We give a quick overview regarding the improvements over related work in Table 1.

In Section 2 we briefly review relevant basic knowledge and concepts before focusing on existing related work regarding synthetic health data and stress detection in Section 3. Section 4 presents an overview of our methodology, describes our experiments, and gives reference on the environment for our implementations. The outcome of our experiments are then detailed and evaluated in Section 5. The following Section 6 is centered around discussing the implications of our results and determining the actual best strategies from different perspectives, as well as, their possible limitations. In the final Section 7, we provide both a concise summary of our key findings and an outlook into future work.

Table 1. Performance results of relevant related work evaluated on the WESAD dataset for modalities collected from wrist devices regarding the binary (stress vs. non-stress) classification task. We compare accuracy (%) and F1-score (%) and include the achieved ϵ -guarantee regarding DP.

Reference	Model	Data	Accuracy	F1-score	Privacy Budget ϵ
[16]	RF	WESAD	87.12	84.11	∞
[17]	LDA	WESAD	87.40	N/A	∞
[18]	CNN	WESAD	92.70	92.55	∞
[15]	TSCT	WESAD	91.89	91.61	∞
Ours	CNN-LSTM	CGAN + WESAD	92.98	93.01	∞
[15]	DP-TSCT	WESAD	78.88	76.14	10
Ours	CNN	DP-CGAN	88.08	88.04	10
[15]	DP-TSCT	WESAD	78.16	71.26	1
Ours	CNN	DP-CGAN	85.46	85.36	1
[15]	DP-TSCT	WESAD	71.15	68.71	0.1
Ours	CNN-LSTM	DP-CGAN	84.16	84.19	0.1

2. Background

The following section introduces some of the fundamental concepts used in this work.

2.1. Stress Detection from Physiological Measurements

There are numerous reactions of the human body when answering situations of stress or amusement. Giannakakis et al. [1] give a comprehensive list of studies and give measurable biosignals related to stress as a separation into two categories: physiological (EEG, ECG, EDA, EMG) and physical measures (respiratory rate, speech, skin temperature, pupil size, eye activity). Some of the found correlations are: e.g., TEMP: non-stress high / stress low; EDA: non-stress low / stress high; or BVP, which has a higher frequency in the stress state than in the non-stress state. A key factor in stress detection systems is the availability and processing of these sensor readings, which also leads to the questions of which sensors we are able to measure using wearables and how relevant each sensor might be in classifying stress correctly. In wrist-worn wearable devices commonly used for stress-related research purposes like the Empatica E4¹, we find three-way acceleration (ACC), electrodermal activity (EDA) also known as galvanic skin response (GSR), skin temperature (TEMP) and blood volume pressure (BVP), which also doubles as an indicator of heart rate (HR). Especially EDA is known a key instrument for identifying moments of stress, while the electroencephalogram (EEG) also gives strong indications but has less availability in continuous wrist readings [1].

2.2. Generative Adversarial Network

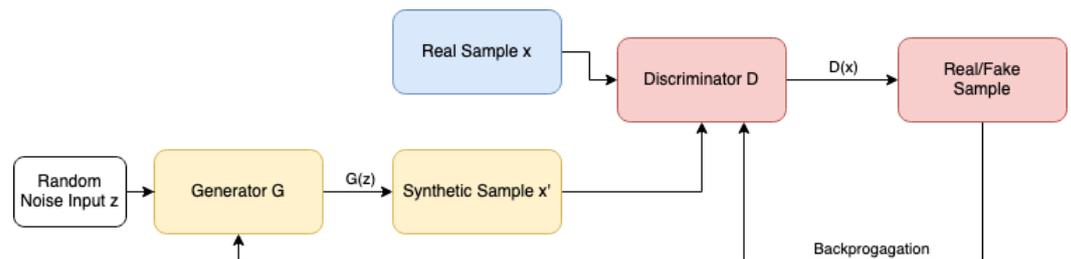


Figure 1. Overview of the fundamental GAN structure: The generator G produces a synthetic sample x' based on a random noise input z . The synthetic samples x' , along with real samples x , are the input to the discriminator D , that classifies each sample either as real or fake. Based on this classification, the loss is calculated to update the generator and discriminator via backpropagation.

The Generative Adversarial Network (GAN) was introduced by Goodfellow et al. [13] and is a class of machine learning frameworks that simultaneously train two neural networks. On the one hand, a generative model G that captures the data distribution, and, on the other hand, a discriminative model D that calculates the probability if a sample originates from the dataset rather than G . Figure 1 shows the architecture of the originally proposed GAN model. The generator G aims to create realistic sample data. Then, the discriminator D receives the synthetically created samples and real samples to classify each sample as either real or fake. The generator does not have direct access to the real sample, instead the interaction with the discriminator is the only way it learns. The discriminator receives the error signal based on the ground truth of knowing whether the sample originated from the real dataset or from the generator. The same error signal is then passed through the discriminator to train the generator, resulting in the generation of better synthetic samples. As a result G is trained to maximize D 's probability of making a mistake and training takes place in the form of a min-max game, where G 's error should be minimized and D 's error maximized.

¹ More information here: <http://www.empatica.com/research/e4/>

2.3. Differential Privacy

The concept of DP [12] is a mathematical definition of privacy that ensures that the inclusion or exclusion of any individual's data in a dataset does not significantly change the output of statistical queries on that dataset. The framework adds noise to the queries to hide individual influences while trying to preserve the utility of the dataset for analysis.

Formally, an algorithm A training on a set S is called (ϵ, δ) -differentially-private, if for all datasets D and D' that differ by exactly one record:

$$\Pr[A(D) \in S] \leq e^\epsilon \Pr[A(D') \in S] + \delta \quad (1)$$

The ϵ -parameter measures the level of privacy protection provided by the mechanism, also called privacy budget. It thus determines the amount of random noise that is added to algorithms on the dataset. Smaller values of ϵ result in better privacy, but also in less utility through more added noise. Guarantees of $\epsilon \leq 1$ are commonly considered as strong [19–21].

2.4. Differentially Private Stochastic Gradient Descent

Differentially Private Stochastic Gradient Descent (DP-SGD) [22] is a variant of the stochastic gradient descent optimization algorithm that provides differential privacy guarantees. DP-SGD works by adding controlled noise to the gradients computed on each mini-batch of data during the training process. The amount of noise added is controlled by a privacy parameter ϵ , which determines the strength of privacy protection. Tuning the privacy parameter can be challenging, and careful selection and calibration of the noise level are required to achieve the desired privacy-utility trade-off.

3. Related Work

This section gives reviews other literature in associated fields of research.

3.1. Synthetic Data for Stress Detection

Ehrhart et al. [23] successfully introduced a GAN approach for a very similar use case, however without considering privacy. In their study, they collected Empatica E4 wristband sensor data from 35 individuals in baseline neutral situations and when inducing stress through air horn sounds. They then trained a CGAN architecture to generate realistic EDA and TEMP signal data. This data is then used to augment the existing data basis and improve their stress detection results. Due to data protection laws, we are not able to use their dataset for our private approach and we are instead limited to using the publicly available but smaller WESAD dataset [16] with 15 participants, which was also collected using the E4 wristband. In contrast to Ehrhart et al. [23], we focus on generating the full range of the available six sensor modalities (ACC $[x,y,z]$, EDA, TEMP, BVP), while they only focused on two of them in their GAN model. We build on their valuable research by using data available to the public, including more sensor modalities, and furthermore by giving a new perspective on approaches for privacy preservation in stress detection from such data.

3.2. Privacy of Synthetic Data

The relevance of privacy methods might seem contradictory at first, since the approach of using synthetic data instead of real data itself already seems to hide the original information found in the data source. Contrary to this intuition, we find that synthetic data can still provide exploitable information on the dataset it is meant to resemble, which is especially true for data generated by GANs [24]. This contradiction is less surprising on second thought, since the goal of synthetic data is to closely follow the distribution of real data, there has to be some inherent information on its distributional qualities hidden inside the synthetic fakes. Another factor making GAN data vulnerable is the general nature of machine learning, where models tend to overly memorize their training data and, as all models, GANs will have the same difficulties escaping this paradigm [20]. Xie et al. [14] give a solution to these privacy concerns in the form of their DP-GAN model, which

is a general GAN model, where the generator is trained using the widespread DP-SGD algorithm to attain private models that guarantee DP. Thanks to this modular design, the DP-GAN approach can be applied to different underlying GAN models and for any given data, like a possible DP-CGAN architecture presented by Torkzadehmahani et al. [25].

3.3. Stress Detection on the WESAD Dataset

There are multiple recent works in smartwatch stress detection, that are evaluated on the WESAD dataset introduced by Schmidt et al. [16], which is a common choice inside the research field. We list the relevant results from both in Table 1 but filter them to only include models based on wrist-based wearable devices, that classify samples into *stress* and *non-stress*. The Convolutional Neural Network (CNN) model [18] delivers the best performance in the non-private setting at $\epsilon = \infty$, outperforming amongst others the Random Forest (RF) [16] and Linear Discriminant Analysis (LDA) [17] solutions. The Time-Series Classification Transformer (TSCT) approach [15] also stays slightly behind, but on the other hand, showed to be the only related work employing DP for this task. Taking these numbers as our reference for the utility-privacy trade-off suggests, that we should expect a substantial draw-down in performance when aiming for any of these privacy guarantees. However, when comparing our best results using synthetic data in Table 1, we improve on both the non-private and private setting. The utility-privacy trade-off improves significantly, especially at $\epsilon = 0.1$, which is a very strict guarantee.

4. Methodology

In this part, we detail the different methods and settings for our experiments. A general overview is given in Figure 2, while each this process and each presented part are further described in the following.

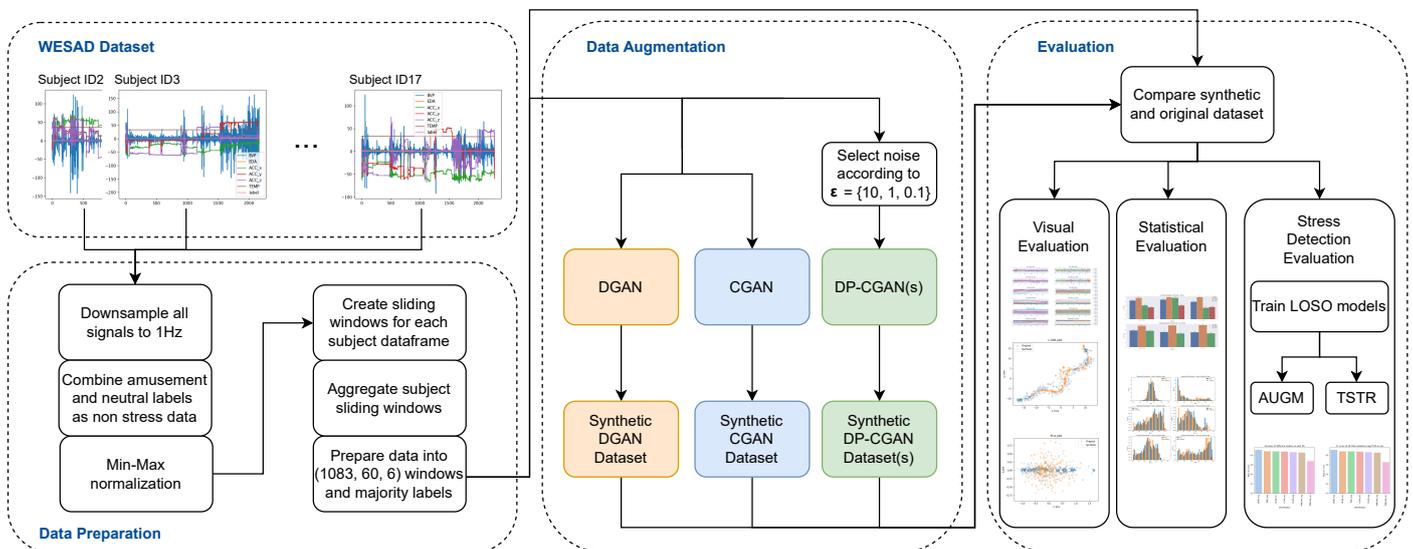


Figure 2. Our experimental methods are illustrated by the given workflow. In the first step, we load and pre-process the WESAD dataset. We then train different GAN models for our data augmentation purposes. Each resulting model generates synthetic data which is evaluated on data quality and finally, compared on their ability to improve our stress detection models.

4.1. Environment

On the software side we employ Python 3.8 as our programming language and utilize the Tensorflow framework for our machine learning models. The accompanying Tensorflow Privacy library provides the relevant DP-SGD training implementations. Our hardware configuration for the experiments comprises machines with 32GB of RAM and an NVIDIA GeForce RTX 2080 Ti graphics card. We further set the random seed to 42.

4.2. Dataset Description

We study our proposed approach on the publicly available multimodal WESAD dataset [16], which is a commonly used dataset for stress detection. The dataset consists of 15 healthy subjects (12 male, 3 female) with about 36 minutes of recorded health data each, 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. We only consider signals available from a wrist-worn wearable like smartwatches—i.e. the Empatica E4 device used in the dataset. The wristband offer the following six modalities at different sampling frequencies: blood volume pulse (BVP), electrodermal activity (EDA), body temperature (TEMP), and three-axis acceleration (ACC[x,y,z]). The three relevant affective states recorded are neutral, stress, and amusement. We focus on binary classification as stress vs. non-stress, which combines the neutral and amusement classes, and where we ultimately find about 30% stress and 70% non-stress data.

4.3. Data Preparation

Transforming time-series signal data to match the expected input format requires several pre-processing steps and is a crucial step in achieving good models. For our approach, we adopt the process of Gil-Martin et al. [18] in many points. We however change some key transformations to better accommodate the data to our setting and stop at 60s windows since we want fed them into our GANs instead of their CNN model. Our process can be divided into four general steps.

First, since the Empatica E4 signal modalities are recorded at different sampling rates due to technical implementations, they need to be resampled to a unified sampling rate. We further need to align these sampled data points to ensure that for each point of time in the time-series, there is a corresponding entry for all signals. To achieve this, signal data is downsampled to a consistent sampling rate of 1Hz using the Fourier method. Despite the reduction in original data points, most of the crucial non-stress/stress dynamics are still captured after the Fourier transformation process, while model training is greatly accelerated by reducing the number of samples per second.

In a second step, we adjust the labels by combining *neutral* and *amusement* into the common *non-stress* label. In addition to this data, we only keep the *stress* part of the dataset. This is reduction in labels is mainly due to the fact, that we want to enhance binary stress detection, that only distinguishes between moments of stress and moments without stress. However, only keeping neutral data would underestimate the importance of differentiating the amusement phase to the stress phase, since there is an overlap in signal characteristics, such as BVP or ACC, for amusement and stress [16]. After the first and this relabeling step, we get a intermediate result of 23,186 non-stress and 9,966 stress labelled seconds.

Thirdly, we normalize the signals using a min-max normalization in the range of [0,1] to eliminate the differences in scale among the modalities while still capturing their relationships. In addition, the normalization has a great impact on the subsequent training process, as it helps the model to converge faster, thus shortening the time to learn an optimal weight distribution.

Given that the dataset consists of about 36 minutes sessions per subject, in our fourth and final step, we divide these long sessions into smaller time frames to pose as input windows for our models. We transform each into 60s long windows but additionally, as described by Dzieżyc et al. [26], we introduce a sliding window effect of 30s. This means in instead of neatly dividing into 60s windows, we instead create a 60s window after every 30 seconds of data stream. These additional intermediate windows fill the gap between clean aligned 60s windows by overlapping with the previous window by 30s and next window by 30s, providing more contextual information by capturing the correlated time-series between individual windows. Additionally, sliding windows increase the amount of data points available for subsequent training. To assign a label for a window, we determine the

majority class in the given 60s frame. Finally, we concatenate the 60-second windows and their associated class labels from all subjects into a final training dataset.

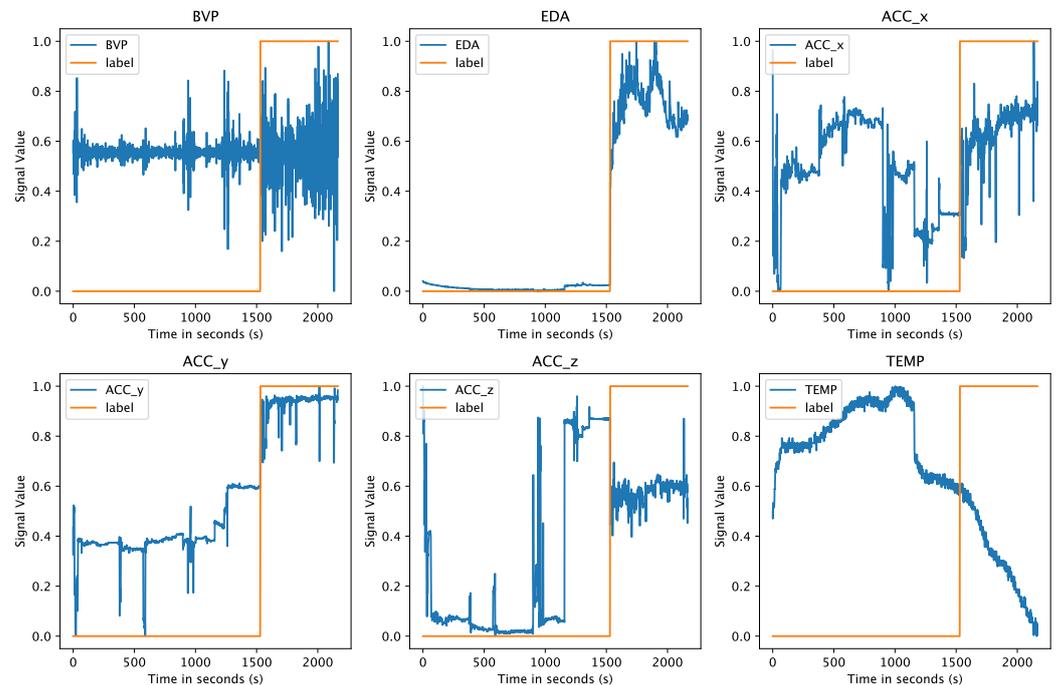


Figure 3. Individual signal modalities plotted for Subject ID4 after resampling, relabeling, and normalizing the data. The orange line shows the label, which equals 0 for non-stress and 1 for stress.

An example of pre-processed data is given in Figure 3, where we show the graphs for Subject ID4 from the WESAD dataset after the first three processing steps. The orange line represents the associated label for each signal plot and is given as 0 for non-stress and 1 for stress. We can already spot certain differences between the two states in relation to the signal curves simply when looking at the given plots.

4.4. Generative Models

After transforming our signal data to a suitable and consistent input format, it is important to determine the proper model architecture the given data characteristics. Compared to the original GAN architecture [13], we face three main challenges:

1. *Time-series data:* Instead of singular and individual input samples, we find continuous time-dependent data recorded over a specific time interval. Further each data point is correlated to the rest of the sequence before and after it.
2. *Multi-modal signal data:* For each point in time we find not one sample but a sample for all our six signal modalities each. This multimodality is further complicated by the fact that the modalities correlate to each other and to their assigned class label.
3. *Class labels:* Each sample also has a corresponding class label as stress or non-stress. This is solvable with standard GANs by training a separate GAN for each class, like when using the Time-series GAN (TimeGAN) [27]. However, with such individual models some correlation between label and signal data might be lost.

Based on these data characteristics and resulting challenges, we have selected the following three GAN architectures that address these criteria in different ways.

4.4.1. Conditional GAN

The Conditional GAN (CGAN) architecture was first introduced by Mirza and Osindero [28]. Here, both the generator and the discriminator receive additional auxiliary input information such as a class label with each sample. This means, in addition to solely

generating synthetic samples, the CGAN is able to learn and output the corresponding labels for synthetic samples, effectively allowing the synthetization of labelled multimodal data. For our time-series CGAN variant, we mainly adopt the architecture and approach from the related work by Ehrhart et al. [23], which is based on the LSTM-CGAN [29] but expanded by a diversity term to stabilize training and a FCN discriminator model with convolutional layers. We instead rely on a LSTM discriminator by stacking two LSTM layers, that performs better in our scenario [30]. As hyperparameters, we choose the diversity term $\lambda = 8$ and employ an Adam [31] optimizer with a learning rate of $2e - 4$. We further pick 64 for the batch size and train for 1,600 epochs. We derived these values from hyperparameter tuning.

4.4.2. DoppelGANger GAN

The other architecture considered is the DoppelGANger GAN (DGAN) by Lin et al. [32]. Like the CGAN, the DGAN uses LSTMs to capture relationships inside the time-series data. Thanks to a new architectural element, the DGAN is able to include multiple generators in its training process. The goal is to decouple the conditional generation part from the time-series generation. They thus include separate generators for auxiliary metadata, like labels, and continuous measurements. In the same vein, they use an auxiliary discriminator in addition to the standard discriminator, which exclusively judges the correctness of metadata outputs. To address mode collapse problems, they further introduce a third generator, which again treats the min and max of signal values as metadata. By combining these techniques, Lin et al. [32] try to incorporate the relationships between the many different attributes. This approach also offers the advantage that a trained model can be further refined, and by flexibly changing the metadata, can generate synthetic data for a different use case. In terms of hyperparameters, we choose a learning rate of $1e - 3$ and train for 10,000 epochs with the number of training samples as batch size.

4.4.3. DP-CGAN

Our private DP-GAN architecture of choice is the DP-CGAN, which was already used by Torkzadehmahani et al. [25] without our focus on time-series. Through the multiple generators and discriminator parts, the DGAN has a harder time in complying to private training, which is why we stay with the CGAN for private training that performed well in initial tests. To incorporate our task into the architecture, we take the CGAN part from Ehrhart et al. [23] and make it private using DP-SGD. More specifically we use the DP-Adam optimizer, which is an Adam variant of DP-SGD. For privatizing the CGAN architecture, we draw on the DP-GAN ideas by both Xie et al. [14] and Liu et al. [33]. Both approaches introduce the concept of securing a DP guarantee for GANs via applying noise to the gradients through the optimizer during training. During GAN training, the generator only reacts to the feedback received from the discriminator, while discriminator is the part that access real data for calculating the loss function [33]. From this, we can determine that just the discriminator needs to implement noise injection when seeing real samples to hide their influence. Thus, just the discriminator needs to switch to the DP optimizer and the generator can keep its standard training procedure. The hyperparameters of DP-CGAN training are described in Section 4.8, where we focus on the necessary information for implementing the private training.

4.5. Synthetic Data Quality Evaluation

Under the term of data quality, we unite the visual and statistical evaluation methods for our synthetic data. We use the following four strategies to get a good understanding of the achieved diversity and fidelity provided by our GANs:

1. *Principal Component Analysis (PCA)* [34]. As a statistical technique for simplifying and visualizing a dataset, PCA converts many correlated statistical variables into principal components to reduce the dimensional space. Generally, PCA is able to identify the principal components that identify the data while preserving its coarser structure.

2. *t-Distributed Stochastic Neighbor Embedding (t-SNE)* [35]. Another method for visualizing high-dimensional data is using t-SNE. Each data point is assigned a position in a two-dimensional space. This reduces the dimension while maintaining significant variance. Unlike PCA, it is less qualified at preserving the location of distant points, but can better represent the equality between nearby points.
3. *Signal correlation and distribution*. To validate the relationship between signal modalities and also to the respective labels, we analyse the strength of correlation found inside the data. A successful GAN model should be able to output synthetic data with a similar correlation as its original training data. Even though correlation does not imply causation, the correlation between labels and signals can be essential to train classification models. We further take a look at the actual distribution of signal values, to see if the GANs are able to replicate these statistics.
4. *Classifier Two-Sample Test (C2ST)*. To evaluate whether the generated data is overall comparable to real WESAD data, we employ a C2ST mostly as described by Lopez-Paz and Oquab [36]. The C2ST uses a classification model that is trained on a portion of both real and synthetic data, with the task of differentiating between the two classes. Afterwards the model is fed with a test set that again consists of real and synthetic samples in equal amounts. Now, if the synthetic data is close to the real data, the classifier would have a hard time in correctly labelling the different samples, leaving it with a low accuracy result. In an optimal case, the classifier would label all given test samples as real and thus only achieve 0.5 of accuracy. This test method allows us to see if the generated data is indistinguishable from real data for a trained classifier. For our C2ST model we decided on a Naive Bayes approach.

4.6. Use Case Specific Evaluation

We test the usefulness of our generated data in an actual stress detection task for classifying stress and non-stress data. The task is based on the WESAD dataset and follows a evaluation scheme using Leave One Subject Out (LOSO) cross-validation. In standard machine learning evaluation, we would split the subjects from the WESAD dataset in distinct train and test sets. In this scenario, we would only test on the selected subjects and these would also be excluded from training. In the LOSO format, we instead train 15 different models, one for each subject in the WESAD dataset. A training run uses 14 of the 15 subjects from the WESAD dataset as training data and the 15th subject as the test set for evaluation. Thereby, when cycling through the whole dataset using this strategy, every subject constitutes the test set once and is included in the training for the 14 other runs. This allows us to evaluate the classification results for each subject. For the final result, all 15 test set results are averaged into one score, simulating an evaluation over all subjects. This process is also performed by the related work presented in Table 1.

To evaluate our synthetic data, we generate time series sequences per GAN model with the size of an average subject of roughly 36 minutes in the WESAD dataset. We also conform to the same distribution of stress and non-stress with about 70% and 30%, respectively. By this we want to generate comparable subject data that allows us to realistically augment or replace the original WESAD dataset with synthetic data. We can then evaluate the influence of additional subjects on the classification. The synthetic subjects are included in each training round of the LOSO evaluation but the test sets are only based on the original 15 subjects to get comparable and consistent results. The GANs are also part of the to the LOSO procedure, which means that the subject which currently provides the test set is omitted from their training. Finally, each full LOSO evaluation run is performed 10 times to better account for randomness and fluctuations from GAN data, classifier training, and DP noise. The results are then again averaged into one final score.

For as evaluation metric we use the F1-score over accuracy, since it combines both precision and recall and shows the balance between these metrics. The F1-score gives their harmonic mean and is particularly useful for unbalanced datasets, such as the WESAD

dataset with its minority label distribution for stress. Precision is defined as $Prec = \frac{TP}{TP+FP}$, while recall is $Rec = \frac{TP}{TP+FN}$ and F1-score is then given as: $F1 = 2 * \frac{Prec*Rec}{Prec+Rec}$.

To improve the current state-of-the-art classification results using our synthetic data, we test the following two strategies in both non-private and private training scenarios:

1. *Train Synthetic Test Real (TSTR)*. The TSTR framework is commonly used in the synthetic data domain and means that the classification model is trained on just the synthetic data and then evaluated on the real data for testing. We implement this concept by generating synthetic subject data in differing amounts, i.e. number of subjects. We decide to first use the same size as the WESAD set of 15 subjects to simulate a synthetic replacement of the dataset. We then evaluate a larger synthetic set of 100 subjects. Complying to the LOSO method, the model is trained using the respective GAN model leaving out the test subject on which it is then tested. The average over all subject results is then compared to the original WESAD LOSO result. Private TSTR models can use our already privatized DP-CGAN data in normal training.
2. *Synthetic Data Augmentation (AUGM)*. The AUGM strategy focuses on enlarging the original WESAD dataset with synthetic data. For each LOSO run of a WESAD subject, we combine the respective original training data and our LOSO-conform GAN data in differing amounts. As before in TSTR, we consider 15 and 100 synthetic subjects. Testing is also performed in the LOSO format. With this setup we evaluate if adding more subjects, even though synthetic and of the same nature, helps the classification. Private training in this scenario takes the privatized DP-CGAN data but also has to consider the not yet private original WESAD data it is combined with. Therefore, the private AUGM models still undergo a DP-SGD training process to guarantee DP.

4.7. Stress Classifiers

Following, we present the tested classifier architectures and their needed pre-processing.

4.7.1. Pre-processing for Classification

After already pre-processing our WESAD data for GAN training as described in Section 4.3, we now need the aforementioned further processing steps from Gil-Martin et al. [18] to get our training data into the correct shape for posing as inputs to our classification models. The 60s long windows from Section 4.3 are present in both the WESAD and synthetically generated data. The only difference between the two types of data is that we use a 30s sliding window to the original WESAD data as we did before for GAN training.

We transform each subwindow into a frequency dependent representation via Fast Fourier Transformation (FFT). FFT is an algorithm that efficiently calculates the Fourier transform to convert a time-dependent signal into the relevant frequency components that form the original signal. This means that these subwindows are transformed into frequency spectra. But before applying the FFT, we first divide the 60s windows into further subwindows of different lengths depending on the signal type and apply a sliding window of 0.25 seconds. The different lengths of subwindows results from the unique frequency spectrum characteristics of each signal type. Depending on the frequency range, we adjust the subwindow length to obtain a constant spectrum shape of 210 frequency points.

Table 2. Subwindow length per signal depending on its frequency range and the resulting number of inputs for the classification model, as described by Gil-Martin et al. [18].

Signal	Frequency range	Subwindow length	# Inputs
ACC (x,y,z)	0–30 Hz	7s	210
BVP	0–7 Hz	30s	210
EDA	0–7 Hz	30s	210
TEMP	0–6 Hz	35s	210

Gil-Martin et al. [18] provide each signal’s frequency range and give the corresponding subwindow length as shown in 2. The subwindow lengths are chosen to always result in

the desired 210 data points when multiplied by the frequency range upper bound, which will be the input size for the classification models. An important intermediate step for our GAN-generated data to avoid possible errors in dealing with missing frequencies in the higher ranges, is to in some cases pad the FFT subwindows with additional zeroes to reach the desired 210 points. The frequency spectra are then averaged along all subwindows inside a 60s window to finally obtain a single frequency spectrum per 60s window.

4.7.2. Time-Series Classification Transformer

As our first classification model, we pick the Time-Series Classification Transformer (TSCT) from Lange et al. [15] that delivers the only comparison for related work in privacy-preserving stress detection, which is also described in Section 3. The model is however unable to reach the best state-of-the-art results for the non-private setting. In their work, the authors argue that the transformer model could drastically benefit from more training samples, like our synthetic data. In our implementation we use class weights and train for 110 epochs with a batch size of 50 using the Adam optimizer at a $1e - 3$ learning rate.

4.7.3. Convolutional Neural Network

The Convolutional Neural Network (CNN) is the currently best performing model in the non-private setting presented by Gil-Martin et al. [18]. For our approach, we also include their model in our evaluations to see if it keeps the top spot. We mostly keep the setup of the TSCT in terms of hyperparameters but train the CNN for just 10 epochs.

4.7.4. Hybrid Convolutional Neural Network

As the final architecture, we consider a hybrid LSTM-CNN model, for which we take the same CNN architecture but add two LSTM layer of sizes 128 and 64 between the convolutional part and the dense layers. Through these additions, we want to combine the advantages of the state-of-the-art CNN and the ability to recognize spatial correlations in the time series from the LSTM. For hyperparameters, we keep the same setup as for the standard CNN but increase the training time to 20 epochs.

4.8. Private Training

In this section we go over necessary steps and parameters to follow our privacy implementation. We first focus on the training of our private DP-CGANs and then follow with the private training of our classification models.

We want to evaluate three DP guarantees that represent different levels of privacy. The first has the budget of $\epsilon = 10$ and is a more relaxed although still private setting. The second and third options are significantly stricter in their guarantees, with a budget of $\epsilon = 1$ and $\epsilon = 0.1$. The budget of $\epsilon = 1$ is already considered strong in literature [19–21], making the setting of $\epsilon = 0.1$ a very strict guarantee. Giving less privacy budget leads to higher induced noise during training and therefore a higher utility loss. We want to test all three values to see how the models react to the different amounts of randomness and privacy.

4.8.1. For Generative Models

We already described our private DP-CGAN models in Section 4.4 and now offer further details on how we choose the hyperparameters relevant to their private training. The induced noise at every training step needs to be calculated depending on the wanted DP-guarantee and under consideration of the training setup. We switch to a learning rate of $1e - 3$, set the epochs to 420 and take a batch size of 8, which is also our number of minibatches. Next we determine the number of samples in the training dataset, which for us is the number of windows. Since we apply a 30s sliding window over the 60s windows of data, when preparing the WESAD dataset for our GANs, we technically double our

training data. We thus see $n \leq 1,000$ as the (maximum)² number of windows for each DP-CGAN after leaving a test subject out for LOSO training. The number of unique windows, on the other hand, stays at $n \leq 496$, since the overlapping windows from sliding do not include new unique data points but instead just re-sample the already included points from the original 60s windows. Thus, the original data points are only duplicated into the created intermediate sliding windows, meaning they are not unique anymore. To resolve this issue, we calculate the noise using the unique training set size of $n \leq 496$. We however, take $2 \times$ the number of epochs, which translates to seeing each unique data point twice during training and accounts for our increased sampling probability for each data point. We subsequently choose $\delta = 1e - 3$ according to $\delta \ll \frac{1}{n}$ [12] and use a norm clip of $C = 1.0$.

4.8.2. For Classification Models

When training our three different classification models in the privacy-preserving setting, we only need to apply DP when including original WESAD data, since the DP-CGANs already produce private synthetic data. In these cases, we mostly keep the same hyperparameters for training as before. We however exchange the Adam for the DP-Adam optimizer with the same learning rate from the Tensorflow Privacy library, which is an Adam version of DP-SGD. Regarding the DP noise, we calculate the needed amount conforming to the wanted guarantee before training. We already know the number of epochs and the batch size, which we also set for the microbatches. We however also have to consider other relevant parameters. The needed noise depends on the number of training samples, which for us is the number of windows. Since we do not use the 30s sliding windows when training classifiers on original WESAD data, all windows are unique. We find (at most) $n \leq 496$ remaining windows, when omitting a test subject for LOSO training. This leads to $\delta = 1e - 3$ according to $\delta \ll \frac{1}{n}$ [12]. We finally choose a norm clip of $C = 1.0$.

5. Results

In this section, we present the achieved results for our different evaluation criteria.

5.1. Synthetic Data Quality Results

This section summarizes the results of our analysis regarding the ability of our generated data to simulate original data. We give visual and statistical evaluations.

5.1.1. Two-dimensional Visualization

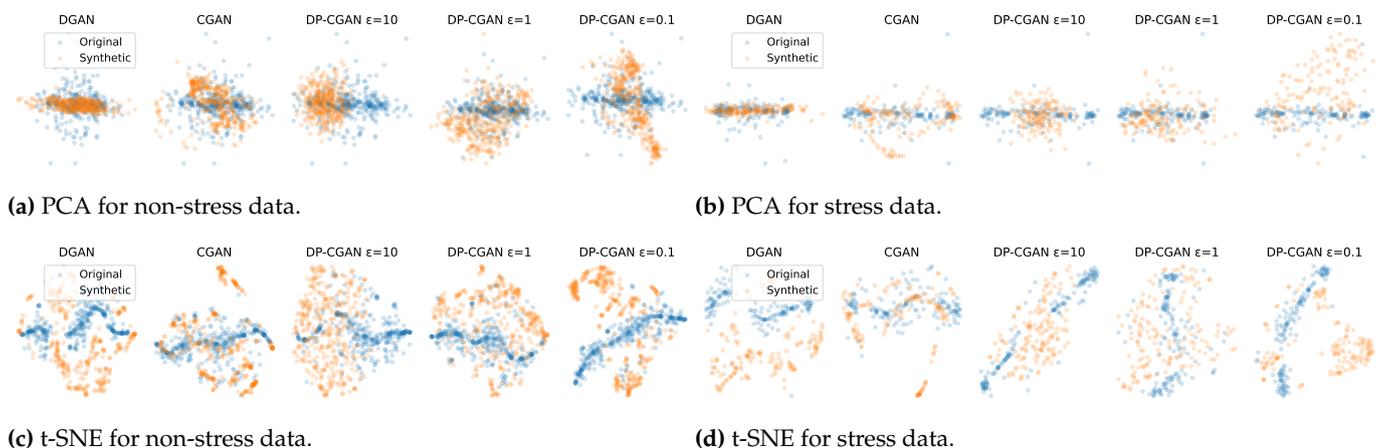


Figure 4. Visualization of synthetic data from our GANs using PCA and t-SNE to cluster data points against original WESAD data. Generated data is more realistic, when it fits the original data points.

² Since subjects offer differing numbers of training windows, the total amount for each LOSO run depends on the current test subject. We give the maximum as an indicator, but consider the actual counts for each LOSO run in the experiments. The ranges are: $n = [494, 496]$ without and $n = [995, 1000]$ with 30s sliding windows.

In Figure 4, we use PCA and t-SNE to visualize the multi-modal signal profiles in a lower two-dimensional space. We give separate diagrams for each model and also differentiate between non-stress and stress data. PCA and the t-SNE visualizations both show how well the diversity of the original data distribution has been mimicked and whether synthetic data point clusters form outside of it or miss the outliers of the real data.

Except for missing some smaller outlier clusters, the CGAN and DP-CGAN at $\varepsilon = 1$ visually seem to give a good representation of the original allocation. The CGAN shows to have a slight advantage in t-SNE as seen in Figures 4c and 4d, where the DP-CGAN ($\varepsilon = 1$) gives a straighter line cluster and thereby misses the bordering zones of the point cloud.

The other GANs generally also show some clusters that mostly stay within the original data. However, they tend to show more and stricter separation from the original points. They also miss clusters and form bigger clusters than original data in some locations. The DGAN shows an especially strict separation to the original cluster for the t-SNE stress data in Figure 4d, which induces problems when training with both data and might not correctly represent the original data.

5.1.2. Signal correlation and distribution

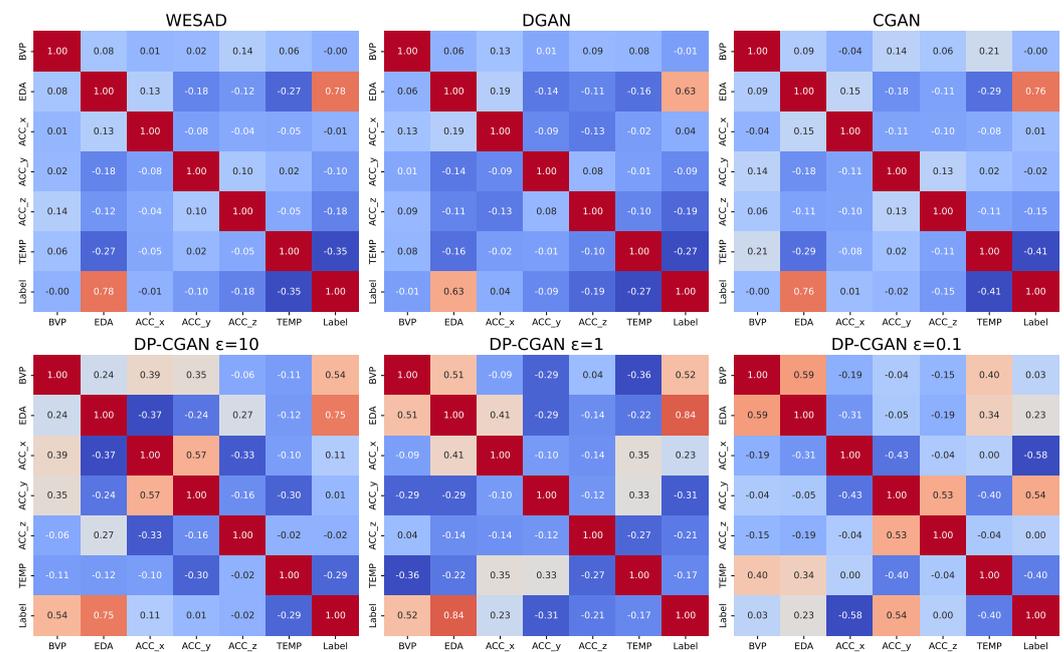


Figure 5. Matrices showing the correlation between the available signals. We compare real WESAD data and data from each of our GANs. A higher value signifies a stronger correlation.

Looking at the signal correlation matrices presented in Figure 5, the main focus is to find correlations between the labelling of non-stress and stress with any of the signals. For the WESAD dataset we mainly see a strong positive correlation from EDA and some already significantly lower but still visible negative correlation from TEMP. For the GANs it is important to stay fairly close to this label correlation ratio for allowing a good stress classification on their data. We can see that both EDA and TEMP ratios are caught well by the DGAN and CGAN data. This is also true for the rest of the correlation matrix with the CGAN being slightly more precise overall.

In lower row we see the DP-CGAN results, where the GANs at $\varepsilon = 10$ and $\varepsilon = 1$ are able to keep the highest correlation for EDA. We, however, also observe a clear over-correlation of BVP and also between multiple other signals, when comparing to WESAD data. Thus, the overall quality is already reduced. Finally comparing to the DP-CGAN at $\varepsilon = 0.1$, we see that the model transitions away from EDA to instead focus on ACC and

TEMP. The correlations between other signals are comparable to $\varepsilon = 10$ and $\varepsilon = 1$, but with losing the EDA ratio, the GAN at $\varepsilon = 0.1$ loses its grip on the main correlation attributes.

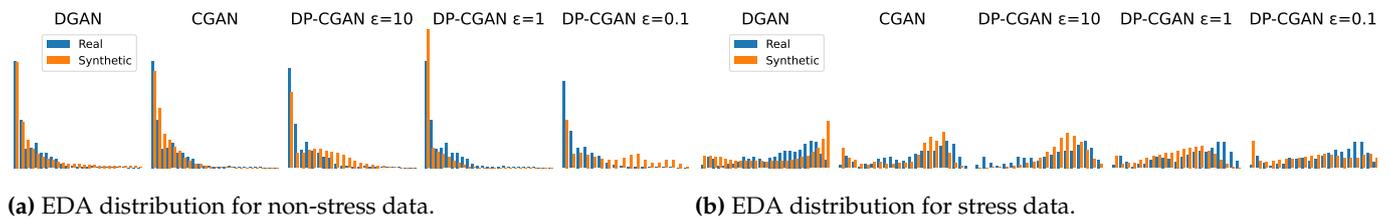


Figure 6. Histograms showing the distribution density of EDA signal values comparing between original and generated data. The y-axis gives the density as $y = [0, 12]$ and the x-axis the normalized signal value as $x = [0, 1]$. Plots for all signal modalities are located in Figure A1 of Appendix A.

We now take a closer look at the distribution density histogram of EDA signal data in the GAN datasets compared to the original WESAD dataset in Figure 6. We picked EDA as our sample because of its strong correlation to the stress labelling and therefore significance for the classification task. The evaluation results for all modalities is available in Figure A1 of Appendix A. Comparing the distribution density in non-stress data, we can see how EDA in WESAD is mostly represented with very low values because of a large peak at zero and a clustering on the left end of the x-axis ($x = [0.0, 0.3]$). While the DGAN and CGAN show similar precision with only smaller deviations from the original distribution, we can see the DP-CGANs struggle with adhering to it in different ways. The DP-CGANs at $\varepsilon = 10$ and $\varepsilon = 0.1$ tend to overvalue EDA leading to a skewing of the distribution to the right on the x-axis. The DP-CGAN $\varepsilon = 1$, however, shows the opposite direction and greatly underrepresents EDA by shifting further to the left and showing an extremely high density at $x = 0$ that neglects the other values.

When comparing EDA distribution for stress, we instead observe a variety of values and a cluster located on right half of the x-axis ($x = [0.6, 0.8]$). Here, the CGAN clearly wins by delivering a good representation over all values. The DGAN, on the other hand, shows a too high distribution on the highest signal values ($x = [0.9, 1.0]$). The private GANs at $\varepsilon = 10$ and $\varepsilon = 1$ generally show a good representation which is only slightly shifted to favor lower values than in the original data. The DP-CGAN at $\varepsilon = 0.1$ goes a bit too far in this direction by keeping a high density at $x = 0$, leading to the worst representation of the general direction of higher EDA values for the original stress data.

5.1.3. Indistinguishability

Table 3. Results of the classifier two-sample test (C2ST), where a low accuracy closer to 0.5 is better. We also include the results on unseen WESAD test data, which constitutes an empirical lower bound.

	WESAD (unseen)	DGAN	CGAN	DP- CGAN $\varepsilon = 10$	DP- CGAN $\varepsilon = 1$	DP- CGAN $\varepsilon = 0.1$
Both	0.59	0.93	0.61	0.93	0.77	0.75
Stress	0.72	0.94	0.77	1.00	0.90	0.85
Non-stress	0.70	0.90	0.71	0.99	0.83	0.91

The results of our C2ST for indistinguishability are given in Table 3. Next to the generated data from our GAN models, we also include a test result on original WESAD data that was not seen by the classifier, i.e. is different than the WESAD data we hand the classifier as real data. Creating synthetic data that comes close to original WESAD data would be the optimal case and thus the performance of our classifier in detecting such data as fake is the empirical lower bound achievable for our GANs. With this in mind, we can see that the CGAN not only has the best results but also comes close the unseen WESAD data results, showing that the CGAN data is almost indistinguishable from real

data. For the DP-CGANs, we see mixed success, where the classifier performs especially well in identifying our synthetic stress data but is fooled more by the non-stress data from the GANs at $\epsilon = 1$ and $\epsilon = 0.1$. DP-GAN data at $\epsilon = 10$ and DGAN data both seem to be an easy task for the classifier, which is able to clearly separate them from original data.

5.2. Stress Detection Use Case Results

Table 4. Summarization of our stress classification results in a comparison of our strategies using synthetic data with the the results using just original WESAD data. We include different counts of generated subjects, privacy budgets, and classification models. Each setting is compared on F1-score (%) as our utility metric.

Strategy	Dataset(s)	# Subjects	Privacy Budget ϵ	TSCT	CNN	CNN-LSTM
Original	WESAD	15	∞	80.65	88.00	86.48
TSTR	DGAN	15	∞	80.60	85.89	85.33
TSTR	CGAN	15	∞	87.04	88.50	90.24
TSTR	DGAN	100	∞	73.90	84.46	79.31
TSTR	CGAN	100	∞	86.97	87.96	91.33
AUGM	DGAN + WESAD	15+15	∞	82.86	88.45	90.67
AUGM	CGAN + WESAD	15+15	∞	88.00	91.13	90.83
AUGM	DGAN + WESAD	100+15	∞	86.94	87.28	88.14
AUGM	CGAN + WESAD	100+15	∞	90.67	91.40	93.01
Original	WESAD	15	10	59.81	46.21	73.18
TSTR	DP-CGAN	15	10	87.55	88.04	84.84
TSTR	DP-CGAN	100	10	85.28	86.41	85.19
AUGM	DP-CGAN + WESAD	15+15	10	64.24	73.66	71.70
AUGM	DP-CGAN + WESAD	100+15	10	71.96	73.50	69.59
Original	WESAD	15	1	58.31	26.82	71.82
TSTR	DP-CGAN	15	1	82.90	85.36	78.07
TSTR	DP-CGAN	100	1	83.75	77.43	83.94
AUGM	DP-CGAN + WESAD	15+15	1	68.55	75.76	71.70
AUGM	DP-CGAN + WESAD	100+15	1	50.06	62.03	71.75
Original	WESAD	15	0.1	58.81	28.32	71.70
TSTR	DP-CGAN	15	0.1	76.27	81.35	76.53
TSTR	DP-CGAN	100	0.1	76.54	83.00	84.19
AUGM	DP-CGAN + WESAD	15+15	0.1	68.99	73.89	71.70
AUGM	DP-CGAN + WESAD	100+15	0.1	35.05	61.99	71.70

We evaluate the usefulness of our synthetic data in a practical scenario of stress detection on the WESAD dataset. To enhance existing methods we introduce synthetic GAN data into the training using our AUGM and TSTR settings as described in Section 4.6. In Table 4 we give a full summarizing view of our results for both settings and taking into account different amounts of synthetic data, as well as, differing privacy levels.

On the WESAD dataset our models perform good but not exceptionally well regarding related work presented in Section 3, which could be due to the in some aspects differing data preparation we employed to train our GANs. The subsequently generated data inherently has the same processing and we thus also used it for the WESAD dataset to better combine

the data in our stress detection evaluation. It seems like the stress classification models disagree with the GAN models to some extent in terms of how the data should be processed. This is especially true for the TSCT model, which stays behind the CNN and CNN-LSTM by a good portion. We can see however, that the introduction of GAN data instead brings back the advantage of our pre-processing strategy, leading to stronger classification results on all privacy levels.

Another general trend are the CGAN outperforming the DGAN data, which is in line with the data quality results in Section 5.1. We further see, that an increased number of synthetic subjects is not always better in performance, since datasets of 15 generated subjects and 100 subjects place closely together and exchange the crown between settings.

Comparing the AUGM and TSTR settings, we can see a clear favorite in the non-private setting at $\epsilon = \infty$. Here, the AUGM strategy using both original WESAD and GAN data, clearly outperforms our TSTR datasets with solely synthetic data. We achieve our best result of about 93% using AUGM with 100 subjects of the CGAN and using a CNN-LSTM model. The TSTR results still tell a story though. From the non-private TSTR we can see the high quality of our synthetic data because we can already reach 91.33% without adding original WESAD data.

We observe a paradigm shift in the private settings of $\epsilon = \{10, 1, 0.1\}$, where TSTR strategy using DP-CGANs reigns supreme over the AUGM approach. The main difference lies in the training setup, where TSTR induces the needed noise already in the DP-CGAN training process. The WESAD-based methods instead (also) have to rely on noise when training the classifier, which shows to be at a substantial disadvantage. While the CNN-LSTM holds good results over all privacy levels with just the WESAD dataset, the TSCT and CNN fail miserably. The AUGM method is able to lift their performance but stays significantly behind the TSTR results. TSTR takes the lead with results of 88.04% and 85.36% at $\epsilon = 10$ and $\epsilon = 1$, respectively. In both cases we use 15 synthetic subjects and a CNN model. This changes for $\epsilon = 0.1$, where we achieve 84.19% using 100 subjects and a CNN-LSTM. The utility-privacy trade-off of our DP approach compared to the best non-private performance of 93.01% is $\Delta F1 = \{-4.97\%, -7.65\%, -8.82\%$ for $\epsilon = \{10, 1, 0.1\}$, which can be considered a low utility loss especially for our stricter privacy budgets.

6. Discussion

The CGAN wins over the DGAN in our usefulness evaluation regarding an actual stress detection task conducted in Section 5.2. In non-private classification, we are however still unable to match state-of-the-art results listed in Table 1 with just our synthetic CGAN data. In contrast, we are able to surpass them slightly by +0.45% at 93.01% F1-score when combining synthetic and original data in our AUGM setup using an CNN-LSTM. The TSCT model generally tends to underperform, while the performance of the CNN and CNN-LSTM models fluctuates, with each model outperforming the other depending on the specific setting. Our private classification models, that work best when only using synthetic data from DP-CGANs in the TSTR setting, show a favorable utility-privacy trade-off by keeping high performance over all privacy levels. With an F1-score of 84.19% at $\epsilon = 0.1$ our most private model still delivers usable performance with a loss of just -8.82% compared to the best non-private model, while also offering a very strict privacy guarantee. Compared to other private models from related work presented in Table 1, we are able to give substantial improvement in utility ranging from +11.90% at $\epsilon = 10$, to +14.10% at $\epsilon = 1$, to +15.48% at $\epsilon = 0.1$ regarding F1-score. Related work on private stress detection further indicates a large number of failing models due to increasing noise when training with strict DP budgets [15]. We do not find any bad models when using our strategies supported by GAN data, making synthetic data a feasible solution to this problem. Our overall results in the privacy-preserving domain indicate that creating private synthetic data using DP-GANs before actual training of a stress classifier, is more effective than later applying DP in its training. Using just already privatized synthetic data shows to be favorable because GANs seem to work better with the induced DP noise than the classification model itself.

Table 5. LOSO results for Subject ID14 and ID17 from the WESAD dataset. We compare the achieved F1-scores (%) based on the original WESAD data and on the best synthetically-enhanced models. Full coverage of all subject results is found in Table A1 of Appendix A.

WESAD Subject	WESAD	CGAN	CGAN + WESAD	DP-CGAN $\epsilon = 10$	DP-CGAN $\epsilon = 1$	DP-CGAN $\epsilon = 0.1$
ID14	54.46	74.88	77.22	69.44	61.00	57.22
ID17	53.57	91.39	88.61	65.18	43.04	83.33

Until now, we only consider the overall average performance from our LOSO evaluation runs, it is however also interesting to take a closer look at the actual per subject results. In this way, we can identify if our synthetic data just boosts the already well-recognized subjects or also enables better results for the otherwise poorly classified and thereby underrepresented subjects. In our results on the original WESAD data, we see that Subject ID14 and ID17 from the WESAD dataset are the hardest to classify correctly. In Table 5 we therefore give a concise overview of the results for the LOSO runs with Subject ID14 and ID17 as our test sets. We include the F1-scores delivered by our best synthetically-enhanced models at each privacy level and compare them to the best result from original WESAD data, as found in Table 4. We can see, that our added synthetic data mostly allows for a better generalization and improves the classification on difficult subjects. Even our DP-CGANs at $\epsilon = 10$ and $\epsilon = 0.1$, that are subject to a utility loss from DP, display increased scores. The other DP-CGAN at $\epsilon = 10$ however, struggles on Subject ID14. A complete rundown of each subject-based result for the selected models is given in Table A1 of Appendix A. The key insights from the full overview are that our GANs mostly facilitate enhancements in challenging subjects. However, especially non-private GANs somewhat equalize the performance across all subjects, which also leads to a decrease in performance in less challenging subjects. In contrast, private DP-CGANs tend to exhibit considerable differences between subjects, excelling in some while falling short in others. The observed inconsistency is linked to the DP-CGANs' struggle to correctly learn the full distribution, a challenge exacerbated by the noise introduced through DP. Such inconsistencies may pose a potential constraint on the actual performance of our DP-CGANs on specific subjects.

While improving the classification task is our main objective, we also consider the quality of our synthetic data in Section 5.1. The CGAN shows to generate the best data for our use case that is comparable to the original dataset in all data quality tests, while also performing best in classification. The DGAN achieves good results for most tested qualities but stays slightly behind the CGAN in all features and performs especially weak in our indistinguishability test. We notice more and more reduced data quality from increasing DP guarantees in our DP-CGANs but still see huge improvements in utility for our private classification. Taking these advantages and drawbacks into account, the CGAN might be able to create a feasible dataset as a realistic WESAD extension or substitute. The DP-CGANs on the other hand, show their advantages only in classification but considering their added privacy attributes, the resulting data quality trade-off could still be tolerable depending on what the synthetic data is used for. The data shows to still be feasible for our use case of stress detection. For usage in applications outside of stress classification, e.g. other analyses in clinical or similar critical settings, however, the DP-CGAN data might already be too inaccurate.

Beyond the aforementioned points, our synthetic data approach, to a certain extent, inherits the limitations found in the original dataset it was trained on. Consequently, we encounter the same challenges that are inherent in the WESAD data. These include a small number of subjects, an uneven distribution of gender and age, and the specific characteristics of the study itself, such as the particular method used to trigger stress moments. With such small datasets, training GANs carries the risk of overfitting. However, we have mitigated this risk through the use of LOSO cross-validation. Further, as demonstrated in Table 5, our GANs have proven capable of enhancing performance on subjects who are

underrepresented in earlier classification models. Nevertheless, questions remain regarding the generalizability of our stress classifiers to e.g. subjects with other stressor profiles and the extent to which our GANs can help overcome all shortcomings of the WESAD dataset.

7. Conclusion

We present an approach for generating synthetic health sensor data to improve stress detection in wrist-worn wearables, applicable in both non-private and private training scenarios. Our models generate multi-modal time-series sequences based on original data, encompassing both stress and non-stress periods. This allows for the substitution or augmentation of the original dataset when e.g. implementing machine learning algorithms. Given the significant privacy concerns associated with personal health data, our DP-compliant GAN models facilitate the creation of privatized data at various privacy levels, enabling privacy-aware usage. While our non-private classification results show only slight improvements over current state-of-the-art methods, our approach to include private synthetic data generation effectively manages the utility-privacy trade-offs inherent in DP training for privacy-preserving stress detection. We significantly improve upon the results found in related work, maintaining usable performance levels while ensuring privacy through strict DP budgets. However, the generalizability of our classifiers to subject data with differing stressors, and the potential enhancement of these capabilities through our synthetic data, remain uncertain without additional public data for evaluation.

Therefore, looking ahead into future work, it would be intriguing to explore how far private synthetic data can be applied in different contexts, especially with abundant availability of training data. In particular, assessing whether the private DP-CGAN data remains viable beyond our use case is worth considering. This could pave the way for replacing original data with fully private alternatives, while still maintaining utility.

Author Contributions: L.L. conducted the conceptualization and writing process. L.L. and N.W. contributed to the methodology, which E.R. supported. N.W. and L.L. implemented the experiment code used in the paper. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: <https://ubicomp.eti.uni-siegen.de/home/datasets/icmi18/> [16]. Implementations for the experiments in this work can be found here: <https://github.com/luckyos-code/Privacy-Preserving-Smartwatch-Health-Data-Generation-Using-DP-GANs>.

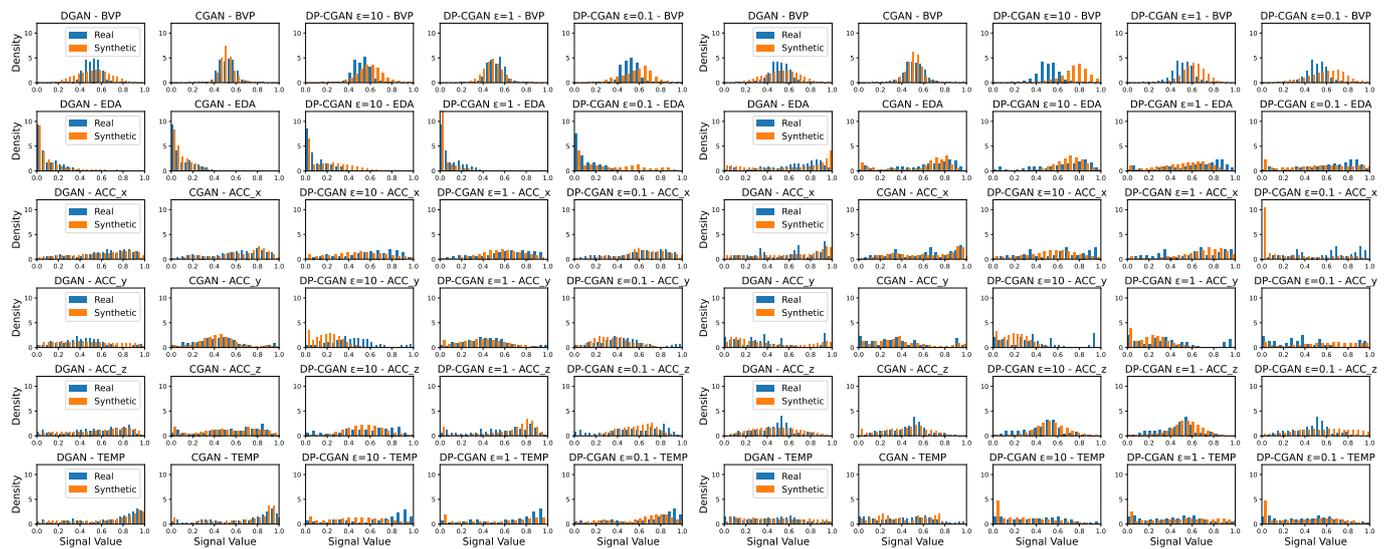
Acknowledgments: We thank Maximilian Ehrhart and Bernd Resch for providing their insights into training CGANs. We thank Victor Christen for his comments on earlier drafts. The authors acknowledge the financial support by the Federal Ministry of Education and Research of Germany and by the Sächsische 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.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Expanded Results

This appendix contains additional information and covers a wider result presentations that complements the more focused results reported in the paper.

Appendix A.1. Signal Distribution Plots



(a) Density histograms showing the distribution of each signal in non-stress data.

(b) Density histograms showing the distribution of each signal in stress data.

Figure A1. Overview of histograms giving the distribution density of signal values, while comparing generated and original data. Covers the omitted signals from Figure 6, which solely focused on EDA.

Appendix A.2. Per Subject LOSO Classification Results

Table A1. Averaged LOSO results broken down per subject and measured by F1-score (%). We compare the achieved scores based on the original WESAD data and on the best synthetically-enhanced models. This extends the before presented extract of the results in Table 5.

WESAD Subject	WESAD	CGAN	CGAN + WESAD	DP-CGAN $\epsilon = 10$	DP-CGAN $\epsilon = 1$	DP-CGAN $\epsilon = 0.1$
ID2	91.76	95.59	92.35	88.24	93.67	96.47
ID3	74.04	70.00	77.65	65.39	61.86	66.57
ID4	98.14	93.59	100.00	80.71	100.00	80.29
ID5	97.15	96.29	97.43	100.00	100.00	86.57
ID6	93.43	98.29	97.43	100.00	97.14	99.71
ID7	90.12	92.29	91.43	97.14	84.26	86.86
ID8	94.85	96.29	96.57	89.14	90.54	88.86
ID9	97.14	97.71	98.86	100.00	97.14	90.96
ID10	99.03	96.80	95.83	100.00	100.00	88.49
ID11	79.84	83.77	90.43	73.58	79.89	78.89
ID13	99.72	93.89	96.39	99.44	85.81	99.72
ID14	54.46	74.88	77.22	69.44	61.00	57.22
ID15	100.00	100.00	100.00	97.22	94.82	86.94
ID16	96.73	89.17	95.00	95.13	91.18	71.94
ID17	53.57	91.39	88.61	65.18	43.04	83.33
Average	88.00	91.33	93.01	88.04	85.36	84.19

References

1. Giannakakis, G.; Grigoriadis, D.; Giannakaki, K.; Simantiraki, O.; Roniotis, A.; Tsiknakis, M. Review on psychological stress detection using biosignals. *IEEE Transactions on Affective Computing* **2019**, *13*, 440–460.
2. Schmidt, P.; Reiss, A.; Dürichen, R.; Van Laerhoven, K. Wearable-based affect recognition—A review. *Sensors* **2019**, *19*, 4079.
3. Panicker, S.S.; Gayathri, P. A survey of machine learning techniques in physiology based mental stress detection systems. *Biocybernetics and Biomedical Engineering* **2019**, *39*, 444–469.
4. Perez, E.; Abdel-Ghaffar, S.; (Google/Fitbit). How we trained Fitbit’s Body Response feature to detect stress, 2023. Accessed on October 17, 2023.
5. Garmin Technology. Stress Tracking, 2023. Accessed on October 17, 2023.
6. Samsung Electronics. Measure your stress level with Samsung Health, 2023. Accessed on October 17, 2023.
7. Narayanan, A.; Shmatikov, V. Robust de-anonymization of large sparse datasets. In Proceedings of the 2008 IEEE Symposium on Security and Privacy (sp 2008). IEEE, 2008, pp. 111–125.
8. Lange, L.; Schreieder, T.; Christen, V.; Rahm, E. Privacy at Risk: Exploiting Similarities in Health Data for Identity Inference. *arXiv preprint arXiv:2308.08310* **2023**.
9. Perez, A.J.; Zeadally, S. Privacy issues and solutions for consumer wearables. *It Professional* **2017**, *20*, 46–56.
10. Jafarlou, S.; Rahmani, A.M.; Dutt, N.; Mousavi, S.R. ECG Biosignal Deidentification Using Conditional Generative Adversarial Networks. In Proceedings of the 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2022, pp. 1366–1370.
11. Kokosi, T.; Harron, K. Synthetic data in medical research. *BMJ medicine* **2022**, *1*.
12. Dwork, C. Differential privacy. In Proceedings of the International colloquium on automata, languages, and programming. Springer, 2006, pp. 1–12.
13. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems* **2014**, *27*.
14. Xie, L.; Lin, K.; Wang, S.; Wang, F.; Zhou, J. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739* **2018**.
15. Lange, L.; Degenkolb, B.; Rahm, E. Privacy-Preserving Stress Detection Using Smartwatch Health Data. *4. Interdisciplinary Privacy & Security at Large Workshop, INFORMATIK 2023* **2023**.
16. Schmidt, P.; Reiss, A.; Duerichen, R.; Marberger, C.; Van Laerhoven, K. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In Proceedings of the Proceedings of the 20th ACM international conference on multimodal interaction, 2018, pp. 400–408.
17. Siirtola, P. Continuous stress detection using the sensors of commercial smartwatch. In Proceedings of the Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers, 2019, pp. 1198–1201.
18. 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* **2022**, *37*, 60–70.
19. Nasr, M.; Songi, S.; Thakurta, A.; Papernot, N.; Carlin, N. Adversary instantiation: Lower bounds for differentially private machine learning. In Proceedings of the 2021 IEEE Symposium on security and privacy (SP). IEEE, 2021, pp. 866–882.
20. Carlini, N.; Liu, C.; Erlingsson, Ú.; Kos, J.; Song, D. The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks. In Proceedings of the 28th USENIX Security Symposium (USENIX Security 19), 2019, pp. 267–284.
21. Lange, L.; Schneider, M.; Christen, P.; Rahm, E. Privacy in Practice: Private COVID-19 Detection in X-Ray Images. In Proceedings of the 20th International Conference on Security and Cryptography (SECRYPT 2023). SciTePress, 2023, pp. 624–633. <https://doi.org/10.5220/0012048100003555>.
22. Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H.B.; Mironov, I.; Talwar, K.; Zhang, L. Deep learning with differential privacy. In Proceedings of the Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, 2016, pp. 308–318.
23. Ehrhart, M.; Resch, B.; Havas, C.; Niederseer, D. A Conditional GAN for Generating Time Series Data for Stress Detection in Wearable Physiological Sensor Data. *Sensors* **2022**, *22*, 5969.
24. Stadler, T.; Oprisanu, B.; Troncoso, C. Synthetic Data – Anonymisation Groundhog Day. In Proceedings of the 31st USENIX Security Symposium (USENIX Security 22), 2022, pp. 1451–1468.
25. Torkzadehmahani, R.; Kairouz, P.; Paten, B. DP-CGAN: Differentially Private Synthetic Data and Label Generation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 2019; pp. 98–104. <https://doi.org/10.1109/CVPRW.2019.00018>.
26. Dzieżyc, M.; Gjoreski, M.; Kazienko, P.; Saganowski, S.; Gams, M. Can we ditch feature engineering? end-to-end deep learning for affect recognition from physiological sensor data. *Sensors* **2020**, *20*, 6535.
27. Yoon, J.; Jarrett, D.; Van der Schaar, M. Time-series generative adversarial networks. *Advances in neural information processing systems* **2019**, *32*.
28. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* **2014**.
29. Esteban, C.; Hyland, S.L.; Ratsch, G. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633* **2017**.

30. Wenzlitschke, N. Privacy-Preserving Smartwatch Health Data Generation For Stress Detection Using GANs. Master's thesis, University Leipzig, Leipzig, 2023.
31. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**.
32. Lin, Z.; Jain, A.; Wang, C.; Fanti, G.; Sekar, V. Using gans for sharing networked time series data: Challenges, initial promise, and open questions. In Proceedings of the Proceedings of the ACM Internet Measurement Conference, 2020, pp. 464–483.
33. Liu, Y.; Peng, J.; James, J.; Wu, Y. PPGAN: Privacy-preserving generative adversarial network. In Proceedings of the 2019 IEEE 25Th international conference on parallel and distributed systems (ICPADS). IEEE, 2019, pp. 985–989.
34. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemometrics and intelligent laboratory systems* **1987**, *2*, 37–52.
35. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **2008**, *9*.
36. Lopez-Paz, D.; Oquab, M. Revisiting classifier two-sample tests. *arXiv preprint arXiv:1610.06545* **2016**.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.