Not to Cry Wolf: Distantly Supervised Multitask Learning in Critical Care

Schwab, Patrick ; Keller, Emanuela ; Muroi, Carl ; Mack, David J ; Strässle, Christian ; Karlen, Walter

Abstract: Patients in the intensive care unit (ICU) require constant and close supervision. To assist clinical staff in this task, hospitals use monitoring systems that trigger audiovisual alarms if their algorithms indicate that a patient's condition may be worsening. However, current monitoring systems are extremely sensitive to movement artefacts and technical errors. As a result, they typically trigger hundreds to thousands of false alarms per patient per day - drowning the important alarms in noise and adding to the exhaustion of clinical staff. In this setting, data is abundantly available, but obtaining trustworthy annotations by experts is laborious and expensive. We frame the problem of false alarm reduction from multivariate time series as a machine-learning task and address it with a novel multitask network architecture that utilises distant supervision through multiple related auxiliary tasks in order to reduce the number of expensive labels required for training. We show that our approach leads to significant improvements over several state-of-the-art baselines on real-world ICU data and provide new insights on the importance of task selection and architectural choices in distantly supervised multitask learning.

Not to Cry Wolf:
Distantly Supervised Multitask Learning in Critical Care

Patrick Schwab [1] **Emanuela Keller** [2] **Carl Muroi** [2] **David J. Mack** [2] **Christian Strässle** [2] **Walter Karlen** [1]

Abstract

Patients in the intensive care unit (ICU) require constant and close supervision. To assist clinical staff in this task, hospitals use monitoring systems that trigger audiovisual alarms if their algorithms indicate that a patient's condition may be worsening. However, current monitoring systems are extremely sensitive to movement artefacts and technical errors. As a result, they typically trigger hundreds to thousands of false alarms per patient per day - drowning the important alarms in noise and adding to the exhaustion of clinical staff. In this setting, data is abundantly available, but obtaining trustworthy annotations by experts is laborious and expensive. We frame this problem as a machine-learning task and address it with a novel multitask network architecture for semi-supervised learning that utilises distant supervision through hundreds of related auxiliary tasks. We show, in an extensive experimental evaluation, that our approach could reduce the burden of false alarms by 41%, 49% and 64% at a specificity greater than 95% using just 25, 50 and 100 labelled alarms, respectively - significantly outperforming several state-of-the-art baselines.

1. Introduction

False alarms are an enormous mental burden for clinical staff and are extremely dangerous to patients, as alarm fatigue and desensitisation may lead to clinically important alarms being missed (Drew et al., 2014). Reportedly, several hundreds of deaths a year are associated with false alarms in patient monitoring in the United States alone (Cvach, 2012).

[1]Institute of Robotics and Intelligent Systems, ETH Zurich, Switzerland [2]Neurocritical Care Unit, Department of Neurosurgery, University Hospital Zurich, Switzerland. Correspondence to: Patrick Schwab <patrick.schwab@hest.ethz.ch>.
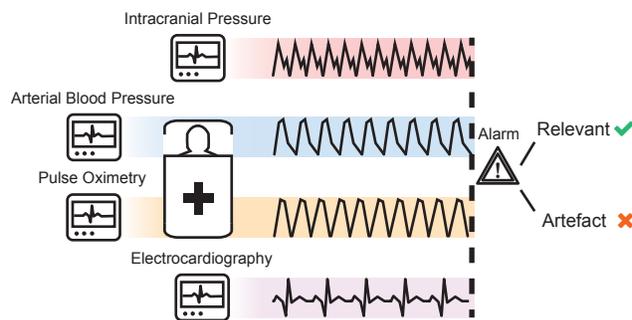
Figure 1. Schematic overview of the described problem setting in critical care. Before an alarm is brought to the attention of clinical staff, an alarm classification algorithm could analyse a recent window of the full set of available data streams in order to identify whether the alarm was likely caused by an artefact or technical error and may therefore be reported with a lower degree of urgency.

An intelligent alarm classification system could potentially reduce the burden of a large subset of those false alarms by assessing which alarms were likely caused by either an artefact or technical error and reporting those alarms with a lower degree of urgency (Figure 1). Roadblocks that have so far prevented the adoption of machine learning for this task are the heterogeneity of monitoring systems, the requirement for an extremely high specificity, to avoid suppressing important alarms, and the prohibitively high cost associated with obtaining a representative set of clinically validated labels for each of the manifold alarm types and monitoring system configurations in use at hospitals.

We present a semi-supervised approach to false alarm classification that automatically identifies and incorporates a large amount of distantly supervised auxiliary tasks in order to significantly reduce the number of expensive labels required for training. We demonstrate, on real-world ICU data, that our approach is able to correctly classify alarms originating from artefacts and technical errors better than several state-of-the-art methods for semi-supervised learning when using just 25, 50 and 100 labelled samples. Besides their importance for clinical practice, our results highlight the power of distant multitask supervision as a simple and effective tool for learning when unlabelled data are readily available, and shed new light on semi-supervised learning beyond low-resolution image benchmark datasets.

Contributions. We subdivide this work along the following distinct contributions:

 (i) We introduce a novel neural architecture built on the idea of utilising distant supervision through multiple auxiliary tasks in order to better harness unlabelled data (DSMT-Nets).

 (ii) We present a methodology for selecting a large set of related auxiliary tasks in time series data and a training procedure for DSMT-Nets that counteracts adverse gradient interactions between auxiliary tasks and the main task.

 (iii) We perform extensive quantitative experiments on a real-world ICU dataset consisting of almost 14,000 alarms in order to evaluate the relative classification performance and label efficiency of DSMT-Nets compared to several state-of-the-art methods.

2. Related Work

Background. Driven by widespread efforts to automate patient monitoring, there has been a recent surge in works applying machine learning to the vast amounts of data generated in ICUs. One notable driver is the MIMIC (Saeed et al., 2011) dataset that has made ICU data accessible to a large number of researchers. Related works have, for example, explored the use of multivariate ICU data for tasks such as mortality modelling (Ghassemi et al., 2014), illness assessment and forecasting (Ghassemi et al., 2015), diagnostic support (Lipton et al., 2016a), patient state prediction (Cheng et al., 2017) and learning weaning policies for mechanical ventilation (Prasad et al., 2017). Applying machine-learning approaches to clinical and physiological data is challenging, because it is heterogenous, noisy, confounded, sparse and of high temporal resolution over long periods of time. These properties are in stark contrast to many of the benchmark datasets that machine-learning approaches are typically developed and evaluated on. Several works therefore deal with adapting existing machine-learning approaches to the idiosyncrasies of clinical and physiological data, such as missingness (Lipton et al., 2016b; Che et al., 2016), long-term temporal dependencies (Choi et al., 2016), noise (Schwab et al., 2017), heterogeneity and sparsity (Lasko et al., 2013). We build on several of these innovations in this work.

Alarm Fatigue. The PhysioNet 2015 challenge on false alarm reduction in electrocardiography (ECG) monitoring (Clifford et al., 2015) was one of the most notable efforts to date to address the issue of false alarms in physiological monitoring. Within the challenge, researchers introduced several highly effective approaches to reducing the false alarm rate of arrhythmia alerts in ECG signals. However, clinicians in the ICU do not just monitor for arrhythmias, but many adverse events at once using a multitude of different monitoring systems. Typically, these monitoring systems operate in isolation on a single biosignal and trigger their own distinct sets of alarms. Previous research has shown that there is an opportunity to use data from other biosignals to identify false alarms in related waveforms (Aboukhalil et al., 2008). We therefore believe that a comprehensive technical solution to alarm fatigue requires a more holistic approach that accounts for the monitoring setup as a whole, rather than targeting specific systems or alarms in isolation.

Distant Supervision and Multitask Learning. Multitask learning has a rich history in healthcare applications and has, for example, been used for risk prediction in neonatal intensive care (Saria et al., 2010), drug discovery (Ramsundar et al., 2015) and prediction of Clostridium difficile (Wiens et al., 2016). A way of leveraging multitask learning to improve label-efficiency is to learn jointly from complementary unsupervised auxiliary tasks along with the supervised main task. Existing literature refers to the concept of applying indirect supervision through auxiliary tasks, be it for label-efficiency or additional predictive performance, as weak supervision (Oquab et al., 2015), distant supervision (Deriu et al., 2016) or self-supervision (Doersch & Zisserman, 2017). In particular, (Doersch & Zisserman, 2017) presented a comprehensive study of distantly supervised multitask learning in computer vision with up to four hand-engineered auxiliary tasks. Using auxiliary tasks in addition to a main task has also been shown to be a promising approach in reinforcement (Jaderberg et al., 2016) and adversarial learning (Salimans et al., 2016). Recently, (Laine & Aila, 2017) proposed to use outputs from the same model at different points in training and with varying amounts of regularisation as additional unsupervised targets for the main task. In contrast to existing works, we present the first approach to distantly supervised multitask learning that automatically identifies a large set of related data-dependant auxiliary tasks in heterogenous time series data and scales to hundreds of auxiliary tasks in a single end-to-end trained neural network (NN).

Semi-supervised Learning. Beside distant supervision, other state-of-the-art approaches to semi-supervised learning in NNs include, broadly, methods based on (i) reconstruction objectives, such as Variational Auto-Encoders (VAEs) (Kingma et al., 2014) and Ladder Networks (Rasmus et al., 2015), and (ii) adversarial learning (Springenberg, 2015; Dai et al., 2017; Li et al., 2017). However, with standard benchmarks consisting primarily of low-resolution image datasets, it is yet unclear to what degree these method's results generalise to heterogenous, long-term and high-resolution time series datasets, as commonly encountered, among others, in healthcare applications.
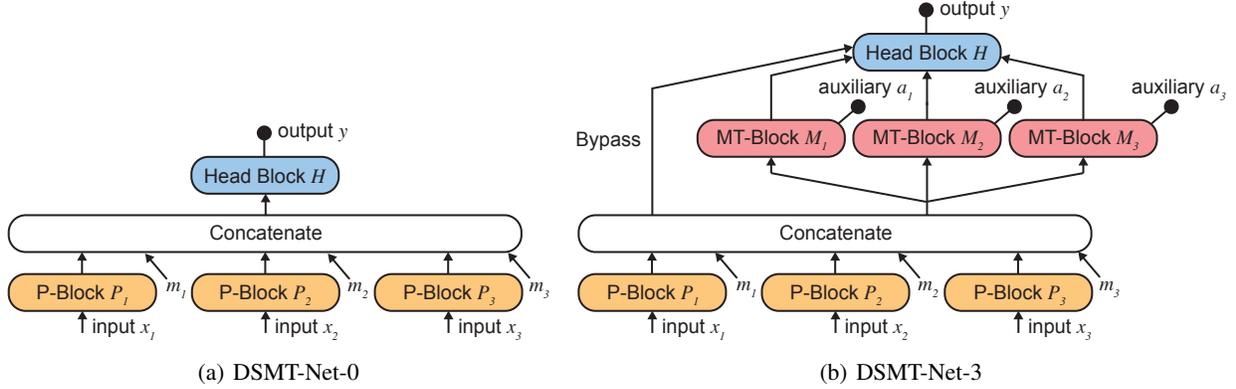
Figure 2. The architectures of two DSMT-Nets: (a) without (DSMT-Net-0) and (b) with three (DSMT-Net-3) auxiliary tasks. The number of horizontally aligned multitask blocks M_j (MT-Blocks; red) is variable. Each multitask block hosts its own auxiliary task a_j. An additional bypass connection gives the head block H (blue) direct access to the concatenated hidden states of the perception blocks P_i (P-Blocks; yellow). Each perception block operates on its own input data stream x_i. The model incorporates binary missing indicators m_i for each perception block to handle situations where input data streams are missing.

3. Distantly Supervised Multitask Networks

Distantly Supervised Multitask Networks (DSMT-Nets) are end-to-end trained NNs that process n heterogenous input data streams x_i in order to solve a multitask learning problem with one main task and k auxiliary tasks designed to augment the main task. Conceptually, a DSMT-Net consists of the following components: One perception block P_i with $i \in [1, n]$ for each of the n input data streams x_i, a variable number k of multitask blocks M_j with $j \in [1, k]$, and a single head block H (Figure 2b). Each of these block types is itself a NN with its own parameters and arbitrary architectures and hyperparameters. The role of the perception blocks P_i is to extract a hidden feature representation $h_{p,i}$ from their respective input data streams x_i:

$$h_{p,i} = P_i(x_i) \tag{1}$$

We separate the perception blocks by input stream x_i in order to be able to model a dynamic set of potentially missing input data streams. To allow our model to learn missingness patterns, we follow (Lipton et al., 2016b) and accompany each perception block with a missing indicator m_i that is set to 0 if the data stream x_i is present and 1 if it is missing. We additionally perform zero-imputation on the missing perception blocks' features $h_{p,i}$. We then concatenate the features $h_{p,i}$ extracted from the perception blocks and the corresponding missing indicators m_i into a joint feature representation P_c over all input data streams as follows:

$$P_c = \text{concatenate}(h_{p,1}, m_1, ..., h_{p,n}, m_n) \tag{2}$$

The joint feature representation P_c combines the information from all feature representations of the input data streams and serves as input to the higher level multitask blocks and the head block. The main role of multitask blocks M_j is to host auxiliary tasks a_j. All multitask blocks are aligned in parallel in order to minimise the distance gradients have to propagate through both to the joint feature representation P_c and from the head block. As output, each multitask block produces a hidden high-level feature representation $h_{m,j}$:

$$h_{m,j} = M_j(P_c) \tag{3}$$

Compared to the straightforward approach of directly appending the auxiliary tasks to the head block H, the positioning of multitask blocks below the head block achieves separation of concerns. In DSMT-Nets, the head block focuses on learning a hidden feature representation that is suitable for the main task rather than being forced to learn a joint feature representation that performs well on multiple (at times, possibly competing) tasks.

The head block H computes the final model output y and further processes the hidden feature representations $h_{m,j}$ of the multitask blocks via a combinator function (see equation (5)). In addition to the hidden feature representations of the multitask blocks, the head block retains direct access to P_c via a bypass connection. We motivate the inclusion of a bypass connection with the desire to learn hidden feature representations in multitask blocks that add information over P_c (He et al., 2016). Mathematically, we formulate the head block H as follows:

$$y = H(\text{combine}_{\text{MLP}}(P_c, h_{m,1}, ..., h_{m,k})) \tag{4}$$

We note that the DSMT-Net architecture without any multitask blocks corresponds to a naïve supervised NN over a mixture of expert networks for each input data stream x_i (DSMT-Net-0; Figure 2a).

Combinator Function. In DSMT-Nets, the combinator function integrates $m + 1$ data flows from the m multitask blocks' hidden representations as well as the joint feature representation P_c. We propose a combinator function (combine$_{\text{MLP}}$) that consists of a single hidden-layer multi-layer perceptron (MLP) with a dimensionality twice as big as a single multitask block's feature representation. As input, the MLP receives the concatenation of all the feature representations to be integrated:

$$\text{combine}_{\text{MLP}} = \text{MLP}(\text{concatenate}(P_c, h_1, ..., h_m)) \quad (5)$$

3.1. Selection of Auxiliary Tasks

One of the most important questions in distantly supervised learning is how to identify suitable auxiliary tasks. A common choice of auxiliary task to enable un- and semi-supervised learning is reconstruction over the feature and/or hidden representation space. Several modern semi-supervised methods take this approach (Vincent et al., 2008; Kingma & Welling, 2014; Kingma et al., 2014; Rasmus et al., 2015). Reconstruction is a convenient choice of auxiliary task because it is generically applicable to any input data, NN architecture and predictive task. However, given recent empirical successes in studies that used distant supervision with specifically engineered auxiliary tasks (Oquab et al., 2015; Deriu et al., 2016; Doersch & Zisserman, 2017), we reason that (i) more "related" tasks might be a better choice of auxiliary task for semi-supervised learning than reconstruction (Ben-David & Schuller, 2003) and that (ii) using multiple diverse auxiliary tasks might be more effective than just one (Baxter, 2000). Since a predictive feature for a main task is also a good auxiliary task for learning shared predictive representations (Ando & Zhang, 2005), we follow a simple two-step feature selection methodology (Christ et al., 2016) to automatically identify a large set of auxiliary tasks that are closely related to the main task:

1. We extract features from a large pool of manually-designed features from each input time series. Due to the large wealth of research in manual feature engineering, there exist vast repositories of such features for many data modalities, e.g. (Christ et al., 2016). For time series, examples of such features would be, e.g., the autocorrelation at different lag levels or the power spectral density over a specific frequency range.
2. We statistically test the extracted features for their importance related to the main task in order to rank the features by their estimated predictive potential. A suitable statistical test is, for example, a hypothesis test for correlation between the labels y_{true} and the extracted features using Kendall's τ (Kendall, 1945).

Using this approach, we are able to identify a large, ranked list of predictive features suitable for use as target labels for auxiliary tasks a_j in DSMT-Nets. There are two approaches to choosing a subset of those features as auxiliary targets: (i) in order of feature importance or (ii) randomly out of the set of relevant features. The main difference between the two approaches is that random selection has a higher expected task diversity as similar tasks are likely to also rank similarly in terms of importance. There are arguments both for (more information per task) and against (harder to learn shared feature representation) higher task diversity. We therefore evaluate both approaches in our experiments.

3.2. Training Distantly Supervised Multitask Networks

A key problem when training NNs on multiple tasks simultaneously using stochastic gradient descent (SGD) is that gradients from the different tasks can interfere adversely (Teh et al., 2017; Doersch & Zisserman, 2017). We therefore completely disentangle the training of the unsupervised and supervised tasks in DSMT-Nets. Instead of training the auxiliary tasks jointly with the main task, we alternate between optimising DSMT-Nets for the auxiliary tasks and the main task in each epoch, starting with the auxiliary tasks. At the computational cost of an additional pass during training, the two-step training procedure prevents any potential adverse intra-step gradient interactions between the two classes of tasks[1]. To ensure a similar training speed for both the main and auxiliary tasks, we weight the auxiliary tasks such that the total learning rate for the unsupervised and supervised step are approximately the same, i.e. a weight of $\frac{1}{k}$ for each auxiliary task when there are k auxiliary tasks. A similar training schedule, where generator and discriminator networks are trained one after another in each iteration, has been proposed to train generative adversarial networks (GANs) (Goodfellow et al., 2014).

4. Experiments

We performed extensive quantitative experiments[2] on real-world ICU data using a multitude of different hyperparameter settings in order to answer the following questions:

(1) *How do DSMT-Nets perform in terms of predictive performance and label efficiency in multivariate false alarm detection relative to state-of-the-art methods for semi-supervised learning?*
(2) *What is the relationship between the number of aux-*

---

[1] We note that intra-step gradient interactions are, however, still possible between auxiliary tasks.

[2] The source code, model binaries and training logs for all experiments are available online at `https://github.com/(omittedfordoubleblindreview)`.

iliary tasks, predictive performance and label efficiency?

To answer question (1), we systematically evaluated DSMT-Nets and several baseline models in terms of their area under the receiver operator curve (AUROC) using varying amounts of manually classified labels $(12, 25, 50, 100, 500, 1244)$[3] and varying amounts of auxiliary tasks $(6, 12, 25, 50, 100)$ in the DSMT-Nets. The comparison between the models' performances when using different levels of labels allows us to judge the label efficiency of the compared models, i.e. how high of a level of predictive performance they are able to reach with a limited amount of labels. By also changing the amount of auxiliary tasks used in the models, we are additionally able to assess the relationship of the number of auxiliary tasks with label efficiency and predictive performance (question (2)).

(3) *What is the importance of the architectural separation of auxiliary tasks and the main task and the two-step training procedure in DSMT-Nets?*

To answer question (3), we performed an ablation study using the DSMT-Nets with 100 auxiliary tasks (DSMT-Net-100) using varying amounts of manually classified labels as base models. We then trained the same models without the two-step training procedure (- two step train). In addition, we evaluated the performance of a deep Highway Network (Srivastava et al., 2015) with the same 100 auxiliary tasks distributed sequentially among layers (DSMT-Net-100D) to compare multitask learning in depth against width. Lastly, we also evaluated a multitask network where the auxiliary tasks are placed directly on the head block (Naïve Multitask Network). Through this process, we aimed to determine the relative importance of the individual design choices introduced in section 3.

(4) *Is there value in selecting a specific set of related auxiliary tasks for distantly supervised multitask learning over random selection?*

To answer question (4), we compared the predictive performance of DSMT-Nets using a random selection of all the significant features (as determined by our feature selection methodology) to that of DSMT-Nets that use a selection in order of feature importance. We do so with DSMT-Nets with 6 (DSMT-Net-6R) and 100 (DSMT-Net-100R) auxiliary tasks to additionally assess whether the importance of auxiliary task selection is sensitive to the number of auxiliary tasks.

In total, we trained 2730 distinct model configurations in order to gain a better understanding of the empirical strengths and weaknesses of DSMT-Nets.

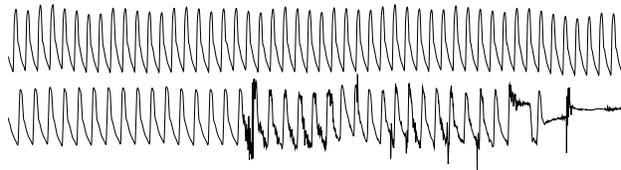[3]The label subsets are chosen at random without stratification.

Figure 3. Two clear examples of arterial blood pressure signals without (top) and with pronounced artefacts (bottom). Note the high frequency noise and atypical shape in the artefact sample.

4.1. Dataset

We collected biosignal monitoring data over the 8 month time frame from January to August 2017 from consenting patients admitted to the Neurocritical Care Unit at the University of Zurich, Switzerland. The data includes continuous, evenly-sampled waveforms obtained by electrocardiography (ECG; 200 Hz), arterial blood pressure (ART; 100 Hz), pulse oximetry (SpO$_2$; 100 Hz) and intracranial pressure (ICP; 100 Hz) measurements. For this study, we did not collect or make use of any personal, demographic or clinical data, such as prior diagnoses, treatments or electronic health records. To obtain a ground truth assessment of alarms, we provided clinical staff with a user interface and instructions[4] for annotating alarms that they believed were caused by artefacts or a technical error (Figure 3). Because of technical limitations in exporting data from our ICU database, we selected the subset of 20 monitoring days of 14 patients with the highest amount of manually labelled alarms for further analysis. On those 20 monitoring days, the vital sign monitors triggered a total of 13,938 alarms, yielding an average rate of 696.9 alarms per patient per day in the analysed dataset. This number is in line with those reported in previous works (Cvach, 2012). Of all alarms, 46.99% were caused by an alarm-generating algorithm operating on the ART waveform, 33.10% on SpO$_2$, 12.02% on ICP and 7.89% on either ECG or an ECG-derived signal, such as the heart rate.

Annotations. Out of the whole set of alarms, 1,777 (12.75%) alarms were manually labelled by clinical staff during the observed period. Because we used multiple annotators that were not calibrated to each other's assessments, we additionally conducted a review over all 1,777 annotations in order to ensure the internal consistency of the set of annotations as a whole. In this review round, we found that a total of 603 (33.93%) annotations were inconsistent. We subsequently assigned corrected labels to these alarms. Since label quality is paramount for (semi-)supervised learning and model validation, we suggest at least one label review round using the majority vote of a committee of labellers with clear instructions in order to maintain a sufficient degree of label consistency. Recent large-scale labelling efforts in physiological monitoring of

[4]We provide the detailed instructions and full qualitative samples in the supplementary material.

arrhythmias (Clifford et al., 2017) suggest that even more review rounds might be necessary to obtain a gold standard set of labels. In our final label set, 976 (45.08%) out of all annotated alarms are labelled as most likely being caused by an artefact or technical error. We note that a data collection effort of this scale is extremely expensive and therefore economically infeasible for most hospitals, motivating our search for a more label-efficient approach.

4.2. Evaluation Setup

As input data, we extracted a 40 second window of the time frame immediately before an alarm was triggered from each available vital signal. We considered a signal stream to be missing for a given alarm setting if the last recorded measurement of that type happened longer than 10 seconds ago. To reduce the computational resources required for our extensive experiments, we resampled the original data to $\frac{1}{16}^{th}$. In our initial evaluations, we did not see significant performance changes when using a higher sampling rate or a longer context window. Additionally, we standardised the extracted windows of each stream to the range of $[-1, 1]$ using the maximum and minimum values encountered in that window.

Baselines. To ensure a fair reference, we used the DSMT-Nets base architecture without any horizontal blocks and auxiliary tasks as the supervised baseline (DSMT-Net-0, Figure 2a). Because the supervised baselines have no auxiliary tasks, we trained them in a purely supervised manner on the labelled alarms only.

As a feature selection baseline, we used the automated feature extraction and selection approach from (Christ et al., 2016) to identify a large number (up to 875) of relevant time series features from the multivariate input data. Note that we followed this process separately for each distinct amount of labels in order to avoid information leakage. We then fed those features to a random forest (RF) classifier consisting of 4096 trees to produce predictions (Feature RF). As mentioned in section 3.2, we used the same feature selection approach to identify suitable auxiliary tasks for DSMT-Nets. The Feature RF baseline therefore serves as a reference for directly using the identified significant features to make a prediction.

For comparison to the state-of-the-art in reconstruction-based semi-supervised learning, we evaluated Ladder Networks (Rasmus et al., 2015) on the same dataset. We replaced the DSMT-Net components on top of the joint feature representation P_c with a Ladder Network in order to use a comparable architecture that is also able to model missingness and heterogenous input streams.

For comparison to the state-of-the-art in semi-supervised adversarial learning, we trained GANs using a semi-supervised objective function and feature matching (Salimans et al., 2016) on the same dataset. This type of GAN has been shown to be highly efficacious at semi-supervised learning in low-resolution image datasets (Salimans et al., 2016). We trained the generator networks to generate a context window of multiple high-resolution time series as input to a DSMT-Net discriminator without any auxiliary tasks. In terms of architecture, the generator networks used strided upsampling convolutions.

Hyperparameters. To ensure a fair comparison, we used a systematic approach to hyperparameter selection for each evaluated NN model. We trained each model 35 times with a random choice of the three variable hyperparameters bound to the same ranges ($1 - 3$ hidden layers, $16 - 32$ units/filters per hidden layer, $25\% - 85\%$ dropout). We reset the random seed to the same value for each model in order to make the search deterministic across training runs, i.e. all the models were evaluated on exactly the same set of hyperparameter values[5]. Note that this setup does not guarantee optimality for any model, however, with respect to the evaluated hyperparameters, it guarantees the models were evaluated fairly and given the same amount of scrutiny. To train the NN models, we used a learning rate of 0.001 for the first ten epochs and 0.0001 afterwards to optimise the binary cross-entropy for the main classification output and the mean squared error for all auxiliary tasks. We additionally used early stopping on the validation set with a patience of 13 epochs. For the extra hyperparameters in Ladder Networks, we set the noise level to be fixed at 0.2 at every layer, the denoising loss weight to 100 for the first hidden layer and to 0.1 for every following hidden layer. For the GAN models, we used a base learning rate of 0.0003 for the discriminator and a slightly increased learning rate of 0.003 for the generator to counteract the faster convergence of the discriminators. We trained GANs with an early stopping patience on the classification loss of 650 steps for a minimum of 2500 steps. To choose the extra hyperparameters of GANs and Ladder Networks, we followed the original author's published configurations (Rasmus et al., 2015; Salimans et al., 2016) and adjusted them slightly to ensure they converged.

Architectures. We used the conceptual architecture from Figure 2 as a base architecture for the DSMT-Nets. As perception blocks, we employed ResNets (He et al., 2016) with 1-dimensional convolutions over the time axis for each input data stream. As head block and multitask blocks, we used Highway Networks (Srivastava et al., 2015). The head block hosted a sigmoid binary output y that indicated whether or not the proposed alarm was likely caused by an artefact. In addition, we used batch normalisation in the DSMT-Net blocks.

---

[5]We report the exact values in the supplementary material.

Table 1. Comparison of the maximum AUROC value across the 35 distinct models we trained using different sets of hyperparameters and varying amounts of labels. We report the AUROC of the best encountered model as calculated on the held-out test set of 533 alarms. The best results in each column are highlighted in bold.

AUROC with # of Labels	12	25	50	100	500	1244
Feature RF	0.567	0.574	0.628	0.822	**0.942**	**0.955**
Supervised baseline	0.751	0.753	0.806	0.873	0.941	0.942
Naïve Multitask Network	0.791	0.804	0.828	0.887	0.941	0.940
Ladder Network	0.791	0.772	0.800	0.842	0.863	0.868
Feature Matching GAN	**0.846**	0.834	0.834	0.865	0.911	0.898
DSMT-Net-6	0.763	0.839	0.866	0.897	0.924	0.934
DSMT-Net-12	0.739	**0.872**	0.891	0.890	0.928	0.933
DSMT-Net-25	0.761	0.870	0.886	0.898	0.924	0.929
DSMT-Net-50	0.722	0.847	**0.901**	0.906	0.926	0.936
DSMT-Net-100	0.720	0.831	0.893	0.907	0.934	0.934
- two step train	0.733	0.798	0.785	0.814	0.849	0.898
DSMT-Net-6R	0.805	0.851	0.884	**0.909**	0.921	0.938
DSMT-Net-100R	0.790	0.860	0.883	**0.909**	0.918	0.932
DSMT-Net-100D	0.587	0.611	0.722	0.610	0.624	0.702

Metrics. For each approach, we report the AUROC of the best model encountered over all 35 hyperparameter runs.

Dataset Split. We applied a random split stratified by alarm classification to the whole set of annotated alarms to separate the available data into a training (70%, 1244 alarms) and test set (30%, 533 alarms). For some of our models, we only used a subset of all training labels in order to evaluate the relative label efficiency of different approaches.

5. Results and Discussion

We report the detailed results of our experiments in Table 1 and discuss them in the following paragraphs.

Predictive Performance. Overall, we find that the label limit after which the purely supervised approaches consistently outperform the semi-supervised approaches in the presented setting is between 100 and 500 labels. The strongest approach when using all 1244 available labels and the 500 label subset is the purely supervised Feature RF baseline. Out of all compared methods, DSMT-Nets are the most label-efficient approach when using 25, 50 and 100 labels. However, the Feature Matching GAN outperforms the DSMT-Nets when using just 12 labels. In our experimental setting, the best DSMT-Nets yield significant improvements in AUROC over both reconstruction-based as well as adversarial state-of-the-art approaches to semi-supervised learning on low-resolution image benchmarks. The relative improvements in AUROC amount to 13.0%, 12.6% and 8.0% over Ladder Networks and 9.4%, 10.4% and 5.6% over Feature Matching GANs at 25, 50 and 100 labels, respectively. We note that even Naïve Multitask Networks, that do not make use of any of the adaptions introduced by DSMT-Nets, with the exception of two cases outperform

both Ladder Networks and Feature Matching GANs - suggesting that distant supervision is indeed a highly efficacious approach to semi-supervised learning in this domain.

Interestingly, most of the evaluated semi-supervised approaches, with the exception of Naïve Multitask Networks, are (often by a large margin) outperformed by their purely supervised counterparts at lower amounts of labels than one would expect. Indeed, both Feature Matching GANs as well as Ladder Networks are eclipsed by the supervised baseline at just 100 labels. This suggests that either: (i) Feature Matching GANs and Ladder Networks require a higher degree of hyperparameter optimisation than the other evaluated approaches or (ii) the strengths of these approaches in the domain of low-resolution images do not (without adaptions) generalise to high-resolution time series. These are novel findings given that most other recent evaluations of state-of-the-art methods in semi-supervised learning have been confined solely to the low-resolution image domain. We believe that, in the future, more systematic replication studies, such as the one presented in this work, are necessary to evaluate the degree to which new methods generalise beyond benchmark datasets that often do not cover many practically important data modalities, such as time series data, and idiosyncrasies, such as missingness, heterogeneity, sparsity and noise.

In terms of sensitivity and specificity, our best models would be able to reduce the number of false alarms that are brought to the attention of clinical staff with the same degree of urgency as true alarms with sensitivity of 22.97% (Feature Matching GAN), 40.99% (DSMT-Net-12), 48.76% (DSMT-Net-50), 63.60% (DSMT-Net-100R), 66.43% (Feature RF) and 76.68% (Feature RF) using, respectively, 12, 25, 50, 100, 500 and 1244 labelled training samples at a specificity greater than 95%.

Number of Auxiliary Tasks. Empirically, it appears that, in our experimental setup, more auxiliary tasks yield slightly better performances once sufficient amounts of labels are available. We reason that, because the head block is trained on labelled samples only, a bigger number of annotations is necessary to effectively orchestrate the extra information provided by a larger number of multitask blocks.

Importance of Adaptions. We find that using DSMT-Net trained with auxiliary tasks distributed in depth (DSMT-Net-100D) perform worse than our proposed architecture - demonstrating that parallel alignment of multitask blocks is the superior architectural design choice. Similarly, DSMT-Net-100 variants without the two step training procedure (- two step train) consistently fail to reach the semi-supervised performance of their counterparts with the two step training procedure enabled (DSMT-Net-100) for more than 12 labels. This shows that disentangling the training of the auxiliary and the main task plays an integral role in the strong semi-supervised performance of DSMT-Nets and further reinforces prior reports that adverse gradient interactions are a key challenge for multitask learning in NNs (Teh et al., 2017; Doersch & Zisserman, 2017).

Task Selection. We find that random selection in most cases outperforms selection in order of feature importance when comparing the DSMT-Net-6 and DSMT-Net-6R variants. We believe this is the result of increased task diversity when selecting at random from the relevant auxiliary tasks, as similar features will rank close to each other in terms of feature importance. The fact that this effect is less pronounced between the same models with more auxiliary tasks (DSMT-Net-100R and DSMT-Net-100) supports this theory, as a larger set of tasks will automatically have a higher diversity due to the limited amount of highly similar features, thus decreasing the importance of accounting for diversity in the selection methodology.

6. Limitations

False alarms in the ICU are not solely a technical problem (Cvach, 2012; Drew et al., 2014). Organisational and processual aspects must also be considered to comprehensibly address this grave issue in clinical care (Drew et al., 2014). One such aspect is the question of how to best manage those alarms that have been flagged as false by an alarm classification system. We reason that, due to the inherent possibility of suppressing a true alarm, a sensible approach would be to report those errors with a lower degree of urgency, i.e. with a less pronounced sound, rather than completely suppressing them (Cvach, 2012). Another limitation of this work is that we only considered the detection of alarms that are caused by either artefacts or technical errors. Those alarms that are technically correct, but clinically require no intervention, are another important source

of false alarms (Drew et al., 2014) that we did not analyse in this work. Identifying clinically false alarms is significantly harder than those caused by artefacts and technical errors, as clinical reasoning requires deep knowledge of a patient's high-level physiological state as well as a significant amount of domain knowledge.

Lastly, while the presented distantly supervised approach to semi-supervised learning is highly efficacious on our dataset, its applicability to other datasets hinges on being able to determine multiple related auxiliary tasks. We only evaluated distantly supervised multitask learning on time series data, where a large number of suitable auxiliary tasks are readily available through automated feature extraction and selection (Christ et al., 2016). We hypothesise that it might not be trivial to find large repositories of auxiliary tasks suitable for distant supervision for all data types. However, previous works in using distant supervision in natural language processing (Deriu et al., 2016) and computer vision (Oquab et al., 2015; Doersch & Zisserman, 2017) suggest that large sets of auxiliary tasks suitable for distant supervision are also readily available for many other data modalities. Finally, our experiments yield insights into the importance of auxiliary task selection in DSMT-Nets but further theoretical analysis is necessary to understand exactly what types of auxiliary task are useful to what degree in distantly supervised multitask learning.

7. Conclusion

We present a semi-supervised approach to detecting false alarms in the ICU from a dynamic set of multiple heterogenous biosignal monitors. Using a real-world critical care dataset, we demonstrate that a fully supervised model would be able to reduce the number of false alarms that are brought to the attention of clinical staff with the same degree of urgency as true alarms by 76.68% at a specificity greater than 95% on 1244 annotated samples. However, for most hospitals, collecting such large amounts of annotations for each of their monitoring setups would be prohibitively expensive. We therefore introduce a novel approach (DSMT-Nets) to semi-supervised learning that leverages distant supervision through up to hundreds of auxiliary tasks to reduce the number of expensive labels required for training. To facilitate the training of DSMT-Nets, we present a methodology for selecting a large set of auxiliary tasks as well as an optimised training procedure to counter adverse gradient interactions between tasks. We show that DSMT-Nets realise 86.63% of the performance of the fully supervised model while using just 100 labels, less than a tenth of the fully supervised model. Finally, our thorough experimental evaluation yields new insights into the importance of task selection and architectural choices in distantly supervised multitask learning.

Acknowledgements

References

Aboukhalil, Anton, Nielsen, Larry, Saeed, Mohammed, Mark, Roger G, and Clifford, Gari D. Reducing false alarm rates for critical arrhythmias using the arterial blood pressure waveform. *Journal of Biomedical Informatics*, 41(3):442–451, 2008.

Ando, Rie Kubota and Zhang, Tong. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(Nov):1817–1853, 2005.

Baxter, Jonathan. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12(149-198):3, 2000.

Ben-David, Shai and Schuller, Reba. Exploiting task relatedness for multiple task learning. *Lecture Notes in Computer Science*, pp. 567–580, 2003.

Che, Zhengping, Purushotham, Sanjay, Cho, Kyunghyun, Sontag, David, and Liu, Yan. Recurrent neural networks for multivariate time series with missing values. *arXiv preprint arXiv:1606.01865*, 2016.

Cheng, Li-Fang, Darnell, Gregory, Chivers, Corey, Draugelis, Michael E, Li, Kai, and Engelhardt, Barbara E. Sparse multi-output Gaussian processes for medical time series prediction. *arXiv preprint arXiv:1703.09112*, 2017.

Choi, Edward, Bahadori, Mohammad Taha, Schuetz, Andy, Stewart, Walter F, and Sun, Jimeng. Doctor AI: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*, pp. 301–318, 2016.

Christ, Maximilian, Kempa-Liehr, Andreas W, and Feindt, Michael. Distributed and parallel time series feature extraction for industrial big data applications. *arXiv preprint arXiv:1610.07717*, 2016.

Clifford, G, Liu, Chengyu, Moody, Benjamin, Lehman, L, Silva, Ikaro, Li, Qiao, Johnson, A, and Mark, Roger G. AF classification from a short single lead ECG recording: The Physionet Computing in Cardiology Challenge 2017. *Computing in Cardiology*, 44, 2017.

Clifford, Gari D, Silva, Ikaro, Moody, Benjamin, Li, Qiao, Kella, Danesh, Shahin, Abdullah, Kooistra, Tristan, Perry, Diane, and Mark, Roger G. The PhysioNet/Computing in Cardiology Challenge 2015: Reducing false arrhythmia alarms in the ICU. In *Computing in Cardiology*, pp. 273–276. IEEE, 2015.

Cvach, Maria. Monitor alarm fatigue: An integrative review. *Biomedical Instrumentation & Technology*, 46(4): 268–277, 2012.

Dai, Zihang, Yang, Zhilin, Yang, Fan, Cohen, William W, and Salakhutdinov, Ruslan. Good Semi-supervised Learning that Requires a Bad GAN. *Advances in Neural Information Processing Systems*, 2017.

Deriu, Jan, Gonzenbach, Maurice, Uzdilli, Fatih, Lucchi, Aurelien, De Luca, Valeria, and Jaggi, Martin. SwissCheese at SemEval-2016 Task 4: Sentiment classification using an ensemble of convolutional neural networks with distant supervision. *Proceedings of SemEval*, pp. 1124–1128, 2016.

Doersch, Carl and Zisserman, Andrew. Multi-task self-supervised visual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2051–2060, 2017.

Drew, Barbara J, Harris, Patricia, Zègre-Hemsey, Jessica K, Mammone, Tina, Schindler, Daniel, Salas-Boni, Rebeca, Bai, Yong, Tinoco, Adelita, Ding, Quan, and Hu, Xiao. Insights into the problem of alarm fatigue with physiologic monitor devices: A comprehensive observational study of consecutive intensive care unit patients. *PloS one*, 9(10):e110274, 2014.

Ghassemi, Marzyeh, Naumann, Tristan, Doshi-Velez, Finale, Brimmer, Nicole, Joshi, Rohit, Rumshisky, Anna, and Szolovits, Peter. Unfolding physiological state: Mortality modelling in intensive care units. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 75–84. ACM, 2014.

Ghassemi, Marzyeh, Pimentel, Marco AF, Naumann, Tristan, Brennan, Thomas, Clifton, David A, Szolovits, Peter, and Feng, Mengling. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 446–453, 2015.

Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, and Bengio, Yoshua. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jaderberg, Max, Mnih, Volodymyr, Czarnecki, Wojciech Marian, Schaul, Tom, Leibo, Joel Z, Silver, David, and Kavukcuoglu, Koray. Reinforcement learning with unsupervised auxiliary tasks. *arXiv preprint arXiv:1611.05397*, 2016.

Kendall, Maurice G. The treatment of ties in ranking problems. *Biometrika*, pp. 239–251, 1945.

Kingma, Diederik P and Welling, Max. Auto-encoding variational bayes. *International Conference on Learning Representations*, 2014.

Kingma, Diederik P, Mohamed, Shakir, Rezende, Danilo Jimenez, and Welling, Max. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pp. 3581–3589, 2014.

Laine, Samuli and Aila, Timo. Temporal ensembling for semi-supervised learning. *International Conference on Learning Representations*, 2017.

Lasko, Thomas A, Denny, Joshua C, and Levy, Mia A. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PloS one*, 8(6):e66341, 2013.

Li, Chongxuan, Xu, Kun, Zhu, Jun, and Zhang, Bo. Triple generative adversarial nets. *Advances in Neural Information Processing Systems*, 2017.

Lipton, Zachary C, Kale, David C, Elkan, Charles, and Wetzell, Randall. Learning to diagnose with LSTM recurrent neural networks. *International Conference on Learning Representations*, 2016a.

Lipton, Zachary C, Kale, David C, and Wetzel, Randall. Directly modeling missing data in sequences with RNNs: Improved classification of clinical time series. In *Machine Learning for Healthcare Conference*, pp. 253–270, 2016b.

Oquab, Maxime, Bottou, Léon, Laptev, Ivan, and Sivic, Josef. Is object localization for free? Weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 685–694, 2015.

Prasad, Niranjani, Cheng, Li-Fang, Chivers, Corey, Draugelis, Michael, and Engelhardt, Barbara E. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. *arXiv preprint arXiv:1704.06300*, 2017.

Ramsundar, Bharath, Kearnes, Steven, Riley, Patrick, Webster, Dale, Konerding, David, and Pande, Vijay. Massively multitask networks for drug discovery. *arXiv preprint arXiv:1502.02072*, 2015.

Rasmus, Antti, Berglund, Mathias, Honkala, Mikko, Valpola, Harri, and Raiko, Tapani. Semi-supervised learning with ladder networks. In *Advances in Neural Information Processing Systems*, pp. 3546–3554, 2015.

Saeed, Mohammed, Villarroel, Mauricio, Reisner, Andrew T, Clifford, Gari D, Lehman, Li-Wei, Moody, George, Heldt, Thomas, Kyaw, Tin H, Moody, Benjamin, and Mark, Roger G. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database. *Critical Care Medicine*, 39(5):952, 2011.

Salimans, Tim, Goodfellow, Ian, Zaremba, Wojciech, Cheung, Vicki, Radford, Alec, and Chen, Xi. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016.

Saria, Suchi, Rajani, Anand K, Gould, Jeffrey, Koller, Daphne, and Penn, Anna A. Integration of early physiological responses predicts later illness severity in preterm infants. *Science translational medicine*, 2(48):48ra65–48ra65, 2010.

Schwab, Patrick, Scebba, Gaetano C., Zhang, Jia, Delai, Marco, and Karlen, Walter. Beat by Beat: Classifying Cardiac Arrhythmias with Recurrent Neural Networks. In *Computing in Cardiology*, 2017.

Springenberg, Jost Tobias. Unsupervised and semi-supervised learning with categorical generative adversarial networks. *arXiv preprint arXiv:1511.06390*, 2015.

Srivastava, Rupesh K, Greff, Klaus, and Schmidhuber, Jürgen. Training very deep networks. In *Advances in Neural Information Processing Systems*, pp. 2377–2385, 2015.

Teh, Yee, Bapst, Victor, Pascanu, Razvan, Heess, Nicolas, Quan, John, Kirkpatrick, James, Czarnecki, Wojciech M, and Hadsell, Raia. Distral: Robust multitask reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 4497–4507, 2017.

Vincent, Pascal, Larochelle, Hugo, Bengio, Yoshua, and Manzagol, Pierre-Antoine. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 1096–1103. ACM, 2008.

Wiens, Jenna, Guttag, John, and Horvitz, Eric. Patient risk stratification with time-varying parameters: A multitask learning approach. *The Journal of Machine Learning Research*, 17(1):2797–2819, 2016.

Supplementary Material for:
"Not to Cry Wolf:
Distantly Supervised Multitask Learning in Critical Care"

Patrick Schwab [1] **Emanuela Keller** [2] **Carl Muroi** [2] **David J. Mack** [2] **Christian Strässle** [2] **Walter Karlen** [1]

1. Source Code, Model Binaries and Training Logs

The source code, model binaries and training logs for all experiments are available online at `https://github.com/(omittedfordoubleblindreview)` (GPL-License).

2. Instructions for Annotators

We instructed our annotators to label a given alarm context window as caused by an artefact iff:

1. The signal that caused the alarm is not being recorded, as verified by visibility on the monitor.

2. The alarm-generating signal curve has an atypical shape.

3. Numerical values derived from the alarm-generating signal are not physiologically plausible.

Figures 1 and 2 depict qualitative examples of context windows that have been labelled as caused by an artefact.

[1]Department of Health Sciences and Technology, ETH Zurich, Switzerland [2]Neurocritical Care Unit, Department of Neurosurgery, University Hospital Zurich, Switzerland. Correspondence to: Patrick Schwab <patrick.schwab@hest.ethz.ch>.

Table 1. Comparison of the standard deviation of AUROC values across the 35 distinct models we trained using different sets of hyper-parameters and varying amounts of labels. We report the AUROC of the best encountered model as calculated on the held-out test set of 533 alarms. The worst result in each column is highlighted in bold. A higher variation in AUROC across hyperparameter choices and training runs may indicate higher sensitivity to hyperparameters in the evaluated range and/or lacking robustness of training in the presented setting. Most notably, we find that disentangling training of the auxiliary and the main task in DSMT-Nets generally improves training stability.

AUROC with # of Labels	12	25	50	100	500	1244
Feature RF	-	-	-	-	-	-
Supervised baseline	0.055	0.046	0.045	0.026	0.008	0.007
Naïve Multitask Network	0.061	0.057	0.048	0.054	0.049	0.041
Ladder Network	0.067	0.074	0.069	0.076	**0.066**	**0.076**
Feature Matching GAN	0.050	0.051	0.051	0.037	0.020	0.027
DSMT-Net-6	0.059	0.058	0.056	0.070	0.040	0.041
DSMT-Net-12	0.058	0.064	0.072	0.074	0.040	0.037
DSMT-Net-25	0.066	0.062	0.068	0.076	0.038	0.043
DSMT-Net-50	0.059	0.071	0.066	0.060	0.042	0.044
DSMT-Net-100	0.060	0.060	**0.075**	0.048	0.032	0.039
- two step train	**0.070**	**0.076**	0.065	**0.078**	0.058	0.061
DSMT-Net-6R	0.058	0.051	0.061	0.062	0.039	0.035
DSMT-Net-100R	0.070	0.062	0.054	0.047	0.034	0.048
DSMT-Net-100D	0.019	0.023	0.038	0.021	0.030	0.038

Table 2. Comparison of the minimum AUROC value across the 35 distinct models we trained using different sets of hyperparameters and varying amounts of labels. We report the AUROC of the best encountered model as calculated on the held-out test set of 533 alarms. The best results in each column are highlighted in bold. The difference between the maximum and minimum value indicates the range of values covered over the 35 hyperparameter settings.

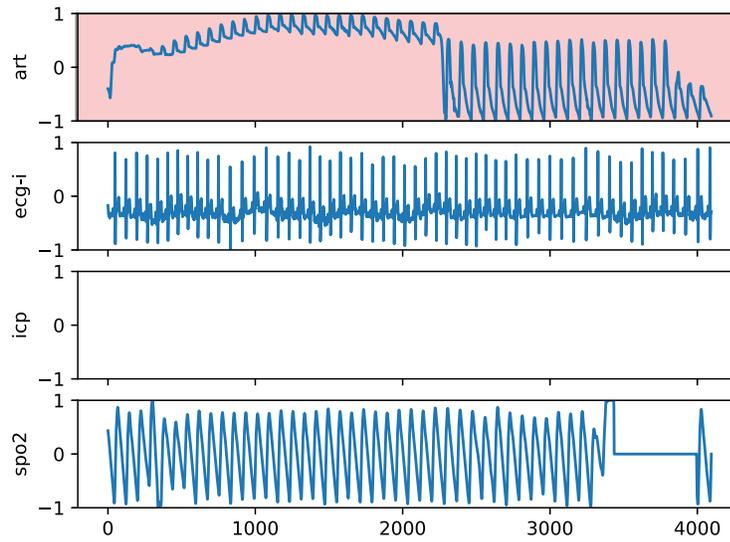| AUROC with # of Labels	12	25	50	100	500	1244
Feature RF	-	-	-	-	-	-
Supervised baseline	0.501	0.547	0.568	**0.763**	**0.907**	**0.911**
Naïve Multitask Network	0.516	0.577	0.613	0.648	0.693	0.732
Ladder Network	0.506	0.516	0.538	0.512	0.594	0.560
Feature Matching GAN	**0.629**	**0.628**	0.646	0.719	0.817	0.757
DSMT-Net-6	0.514	0.557	0.588	0.604	0.760	0.752
DSMT-Net-12	0.507	0.540	0.579	0.630	0.753	0.791
DSMT-Net-25	0.501	0.603	0.535	0.570	0.774	0.779
DSMT-Net-50	0.506	0.557	0.649	0.682	0.768	0.770
DSMT-Net-100	0.507	0.552	0.600	0.691	0.797	0.774
- two step train	0.502	0.500	0.539	0.525	0.645	0.685
DSMT-Net-6R	0.515	0.624	0.630	0.635	0.760	0.805
DSMT-Net-100R	0.506	0.601	**0.660**	0.686	0.771	0.771
DSMT-Net-100D	0.500	0.500	0.500	0.500	0.500	0.500

Figure 1. A qualitative example of an alarm caused by an artefact, as encountered in the ICU dataset. Depicted are the amplitudes (y-axis, standardised) over time (x-axis, in hundredths of a second) of the arterial blood pressure (art), electrocardiography (ecg-i), intracranial pressure (icp) and pulse oximetry (spo2) signals immediately before the alarm was triggered. An empty box indicates a missing signal. In this case, the alarm was triggered by the arterial blood pressure monitor (red). Note that there also appears to be an artefact in the pulse oximetry signal that might have triggered another independent alarm concurrently.
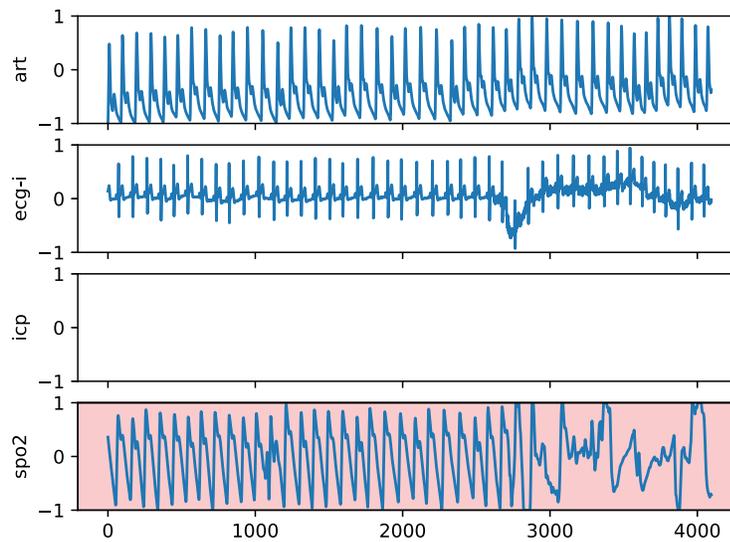

Figure 2. A qualitative example of an alarm caused by an artefact, as encountered in the ICU dataset. Depicted are the amplitudes (y-axis, standardised) over time (x-axis, in hundredths of a second) of the arterial blood pressure (art), electrocardiography (ecg-i), intracranial pressure (icp) and pulse oximetry (spo2) signals immediately before the alarm was triggered. An empty box indicates a missing signal. In this case, the alarm was triggered by the pulse oximetry monitor (red).

Table 3. The exact hyperparameter values used for each model for each of the 35 distinct training runs. We chose the values using a uniformly random selection within the ranges specified in the main paper. The number of hidden units per layer and the number of hidden layers were rounded to the nearest integer in our experiments.

Run	Dropout	Number of hidden units / layer	Number of hidden layers
1	0.5256	18.3015	1.4562
2	0.2926	26.3799	1.6650
3	0.3888	29.7946	1.4185
4	0.4633	29.1221	1.7195
5	0.3619	27.6884	1.7030
6	0.5049	26.5369	2.7647
7	0.7134	26.2866	1.4111
8	0.4486	23.7360	2.4363
9	0.2939	24.0741	1.1734
10	0.5652	21.2195	1.2685
11	0.3688	18.8924	2.5907
12	0.7542	20.2902	2.7300
13	0.2614	27.6143	1.5102
14	0.3820	24.7860	2.1281
15	0.3452	25.3250	2.9806
16	0.7308	30.3649	1.4315
17	0.6195	22.6811	1.7044
18	0.6170	21.3986	2.7229
19	0.7451	27.8114	2.2333
20	0.3469	22.9611	1.4900
21	0.5168	16.2036	2.9124
22	0.4098	20.5713	2.4480
23	0.3012	24.5169	1.3481
24	0.4475	17.3175	2.8138
25	0.2660	27.0517	1.2606
26	0.4830	21.8282	2.9766
27	0.7799	18.0746	2.1824
28	0.3712	24.3822	2.1989
29	0.5958	25.3871	2.8844
30	0.2649	30.3633	2.6249
31	0.6065	20.6158	1.9874
32	0.4623	16.1852	1.3220
33	0.2592	24.9682	1.8996
34	0.6531	26.4506	2.3409
35	0.7825	28.5137	2.9273