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

The source code for this work is available online at `https://github.com/d909b/DSMT-Nets`.

2. Instructions for Annotators

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

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 S1 and S2 depict qualitative examples of context windows that have been labelled as caused by an artefact.

[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>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

Table S1. Comparison of the standard deviation of AUROC values across the 35 distinct models (vertical) that we trained using different sets of hyperparameters and varying amounts of labels (horizontal). We report the AUROC of the best encountered model as calculated on the 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 improves training stability in most cases.

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 S2. Comparison of the minimum AUROC value across the 35 distinct models (vertical) that we trained using different sets of hyperparameters and varying amounts of labels (horizontal). We report the AUROC of the best encountered model as calculated on the 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

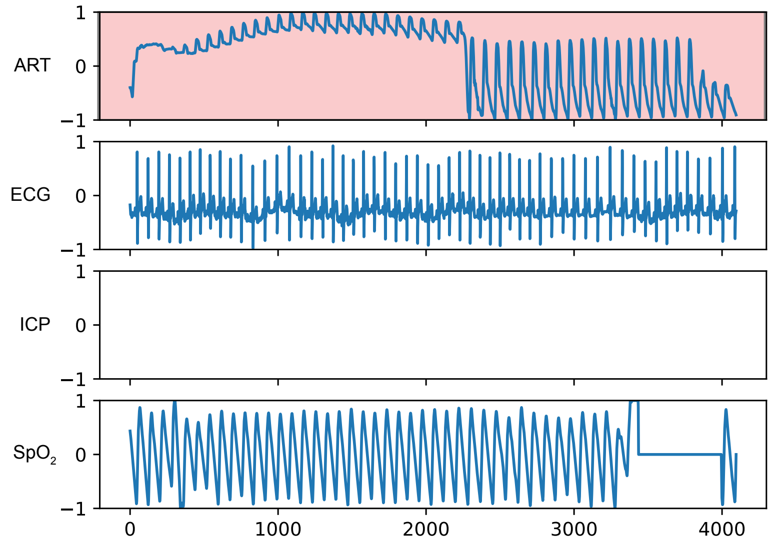

Figure S1. A qualitative example of an alarm caused by an artefact, as encountered in the ICU dataset. Depicted are the amplitudes (y-axis, standardised) over time (x-axis, in hundredths of a second) of the arterial blood pressure (ART), electrocardiography (ECG), intracranial pressure (ICP) and pulse oximetry (SpO_2) 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.

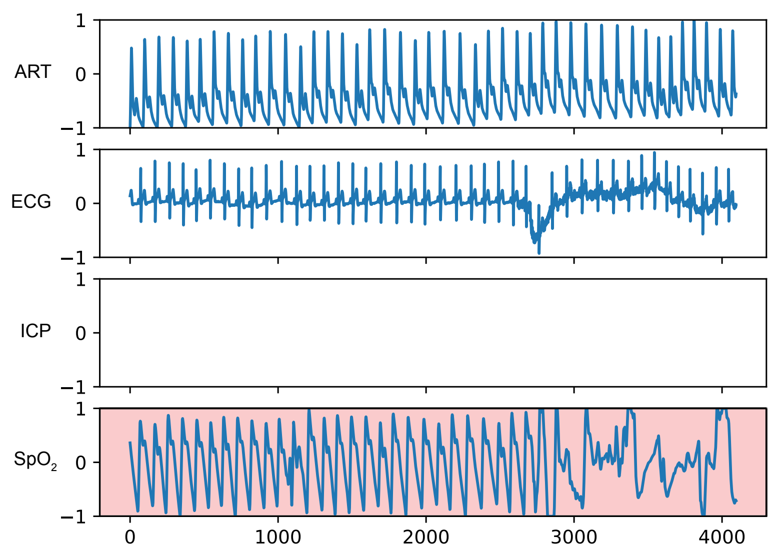

Figure S2. A qualitative example of an alarm caused by an artefact, as encountered in the ICU dataset. Depicted are the amplitudes (y-axis, standardised) over time (x-axis, in hundredths of a second) of the arterial blood pressure (ART), electrocardiography (ECG), intracranial pressure (ICP) and pulse oximetry (SpO_2) 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 S3. 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