A. Architecture

In this section, we expand on the details of our architecture. The full architecture of the encoder is depicted in Figure 5. For the non-linear projector and classifier, we used an inner layer with the same dimension as the representation size. Thus for all tasks, we used an inner dimension of 64. Because we deal with time-series, we used causal dilated convolutions to not break temporal ordering. We built our pipeline using `tensorflow 2.3` and `keras-tcn 3.3`.

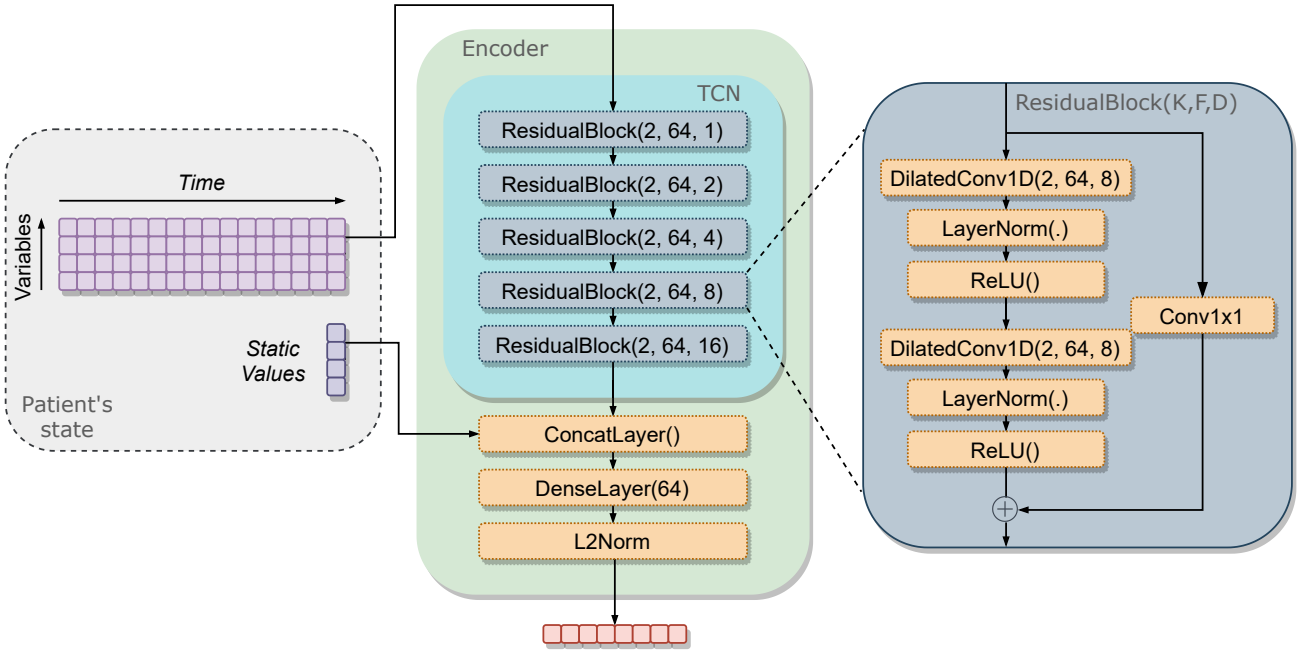

Figure 5. Encoder architecture we used for all methods. In the figure, K, F, and D represent respectively, the kernel size, number of filters, and dilation rate. We use a similar TCN block than the original paper (Bai et al., 2018) with the exception that we use layer normalization. We use a fully-connected layer to incorporate static features in the representation. Finally, we normalize this representation to the unit sphere as in He et al. (2020)

B. Data augmentation

In this section, we further expand on the data augmentations used in all contrastive methods. To choose each function's hyperparameters we performed a random search on the validation performance for both MIMIC-III Benchmark and Physionet 2019 for the regular CL method. As a result of the random search, we chose the following parameters.

1. **History Crop**: We apply a crop with a probability of 0.5 and minimum size of 50% of the initial sequence.
2. **History Cutout**: We apply time cutout of 8 steps with a probability of 0.8.
3. **Channel Dropout**: We mask out each channel randomly with a probability of 0.2
4. **Gaussian Noise**: We add random Gaussian noise to each variable independently with a standard deviation of 0.1

Also, we verify that composing augmentations (Chen et al., 2020a) improves performances. We find, as in Cheng et al. (2020) and (Kiyasseh et al., 2020), that composing temporal and spatial augmentations yields the best performances as shown in Figure 6. It obtains lower performance than composing all transformations, which achieves an AUPRC of 35.5 on validation for the 5 same seeds. Therefore, we applied these four augmentations sequentially to both branches of the pipeline for all contrastive methods.

C. Data sets

In this section we expand further on the datasets we performed experiments on.

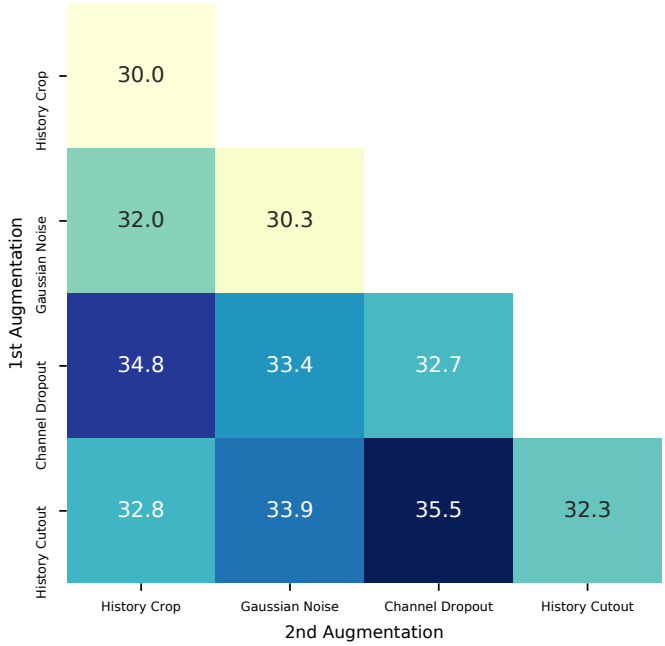

Figure 6. Comparison of performance between different choices of augmentation. Result are reported on validation set AUPRC for the decompensation task and are averaged over 5 seeds.

C.1. MIMIC-III Benchmark

As shown in Table 6, MIMIC-III Benchmark provides 17 measurements in addition to the time since admission. After one-hot encoding of the categorical features, we obtain an input dimension of 42.

In Table 5, we detail the splitting and prevalence of the dataset. We observe that, compared to Physionet 2019, the length of patient stays are significantly greater. Moreover, we also observe that decompensation is a highly unbalanced task.

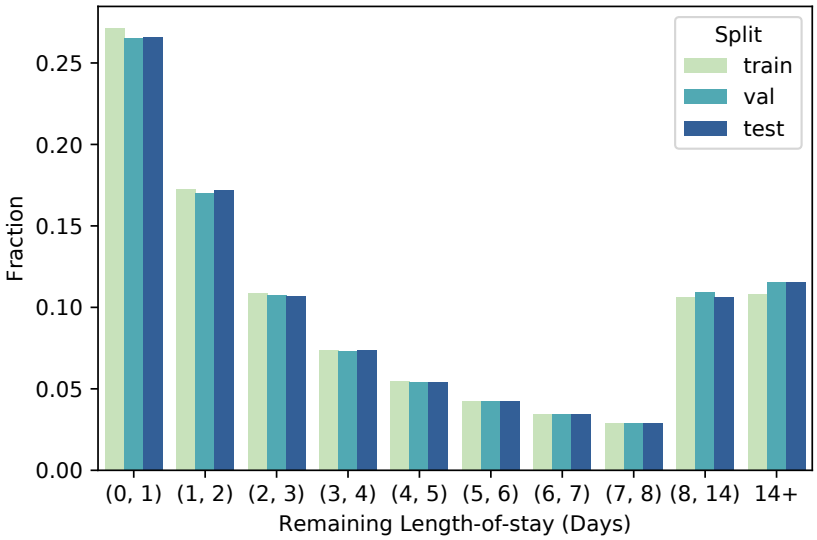

Figure 7. Prevalence of each temporal bin used in the *Length-of-stay* task. We used the same bins as (Harutyunyan et al., 2019).

Table 5. Number of patients and samples in the full data-set as well as for individual predictive tasks.

MIMIC-III	**Number of patients**		
	Train	Test	Val
	29250	6281	6371
Length of stay	**Number of samples**		
	Train	Test	Val
	2,586,619	563,742	572,032
Decompensation	**Number of samples**		
	Train	Test	Val
	2,377,738	523,200	530,638
	Number of positives		
	Train	Test	Val
	49,260	9,683	11,752

Table 6. Measurements recorded and re-sampled hourly in the MIMIC-III benchmark dataset. BP: Blood pressure, MAP: Mean arterial pressure, FiO_2: Fraction of inspired oxygen. GCS: Glasgow Coma Scale. SpO_2: Pulse oxygen saturation.

Measurement	**Type**
Time since admission	Continuous
Height	Static (Continuous)
Capillary refill rate	Categorical
GCS eye opening	Categorical
GCS motor response	Categorical
GCS verbal response	Categorical
GCS total	Categorical
Diastolic BP	Continuous
FiO_2	Continuous
Glucose	Continuous
Heart Rate	Continuous
MAP	Continuous
SpO_2	Continuous
Respiratory rate	Continuous
Systolic BP	Continuous
Temperature	Continuous
Weight	Continuous
pH	Continuous

C.2. Physionet 2019

As shown in Table 7 Physionet 2019 provides 40 measurements. As all categorical features are binary, the final input dimension is 40 as well. In Table 8, we detail the splitting and prevalence of the dataset. We, once again, highlight the very low prevalence of positive labels.

Table 7. Measurement recorded and re-sampled hourly for Physionet 2019. BP: Blood pressure, MAP: Mean arterial pressure, FiO_2: Fraction of inspired oxygen. $PaCO_2$: Partial pressure of carbon dioxide from arterial blood. SaO_2: Oxygen saturation from arterial blood. SpO_2: Pulse oxygen saturation.

Measurement	Type	Measurement	Type
Time since admission (ICU)	Continuous	SaO_2	Continuous
Age	Static (Continuous)	Aspartate transaminase	Continuous
Gender	Static (Categorical)	Blood urea nitrogen	Continuous
Hospital Admission Time	Static (Continuous)	Alkaline phosphatase	Continuous
ICU Unit 1	Static (Categorical)	Calcium	Continuous
ICU Unit 2	Static (Categorical)	Chloride	Continuous
Heart rate	Continuous	Creatinine	Continuous
SpO_2	Continuous	Bilirubin direct	Continuous
Temperature	Continuous	Total bilirubin	Continuous
Systolic BP	Continuous	Serum glucose	Continuous
MAP	Continuous	Lactic acid	Continuous
Diastolic BP	Continuous	Troponin I	Continuous
Respiratory rate	Continuous	Hematocrit	Continuous
End tidal carbon dioxide	Continuous	Hemoglobin	Continuous
Excess Bicarbonate	Continuous	Partial Thromboplastin time	Continuous
Bicarbonate	Continuous	Leukocyte count	Continuous
FiO_2	Continuous	Fibrinogen	Continuous
$PaCO_2$	Continuous	Platelets	Continuous

Table 8. Description of Physionet 2019 statistics by patient and sample.

Physionet 2019	**Number of patients**		
	Train	Test	Val
	25,813	8,066	6,454
Sepsis onset	**Number of samples**		
	Train	Test	Val
	992,732	312,078	247,283
	Number of positives		
	Train	Test	Val
	17,891	5,550	4,475

D. Hyperparameter selection

We tuned all existing hyperparameters over validation performances. For MIMIC-III, we used AUPRC on *Decompensation* task as a reference. For Physionet 2019 we used the Utility metric from (Reyna et al., 2019).

D.1. Architecture parameters

The main hyperparameters of the TCN architecture are the kernel size and the number of filters. We tuned these parameters on the End-to-end model and then used them for all other methods.

Kernel Size	AUPRC (Validation set)
2	**37.0** $\pm$ 0.5
4	36.7 $\pm$ 0.4
8	35.7 $\pm$ 0.6

Number of filters	AUPRC (Validation set)
16	37.1 $\pm$ 0.4
32	37.0 $\pm$ 0.5
64	**37.3** $\pm$ 0.5
128	37.1 $\pm$ 0.5
256	36.8 $\pm$ 1.0
512	35.7 $\pm$ 2.0

Table 9. (a) Impact of the kernel size parameter on the validation AUPRC metric for end-to-end training on MIMIC-III decompensation task. Results are averaged over 5 seeds and number of filters was set to 32, (b) Impact of the number of filters on the validation AUPRC metric for end-to-end training on the MIMIC-III decompensation task. Results are averaged over 5 seeds and kernel size was set to 2

Kernel Size	Utility (Validation set)
2	**28.7** $\pm$ 0.7
4	28.0 $\pm$ 1.1
8	**29.1** $\pm$ **1.6**

Number of filters	Utility (Validation set)
16	27.8 $\pm$ 1.4
32	28.8 $\pm$ 0.6
64	**29.0** $\pm$ 0.8
128	28.8 $\pm$ 1.3
256	26.8 $\pm$ 3.1

Table 10. (a) Impact of the kernel size parameter on the validation Utility metric for end-to-end training on Physionet 2019. Results are averaged over 5 seeds and the number of filters was set to 32., (b) Impact of the number of filter on the validation Utility metric for end-to-end training on Physionet 2019. Results are averaged over 5 seeds and the kernel size was set to 2

D.2. Contrastive parameters

The two main contrastive parameters shared across methods are the momentum ρ and the temperature τ. For a fair comparison, we used the same values for these parameters based on the performance of regular CL as shown in Figure 8 and 9.

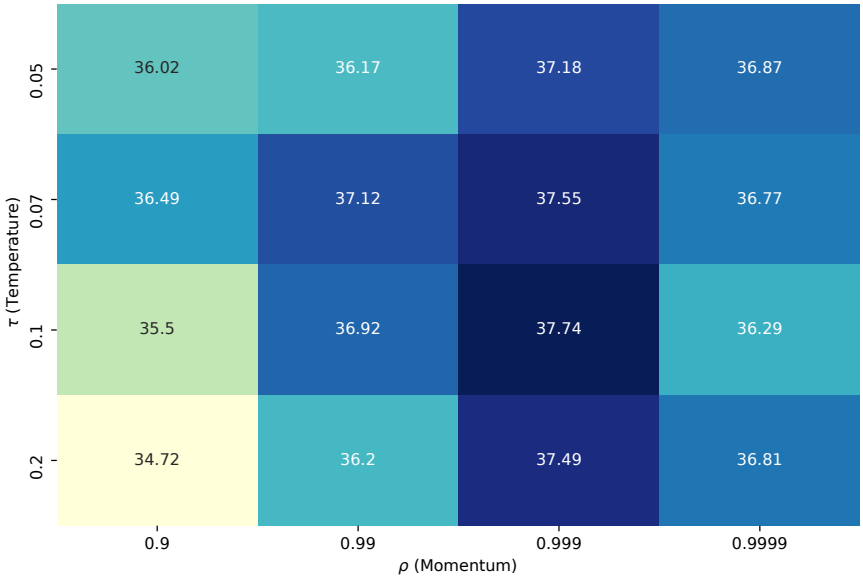

Figure 8. Grid search over τ (temperature) and ρ (momentum) for regular Contrastive Learning method on MIMIC-III. Here result are averaged over 5 runs. Reported metric is AUPRC on validation set for *Decompensation* task.

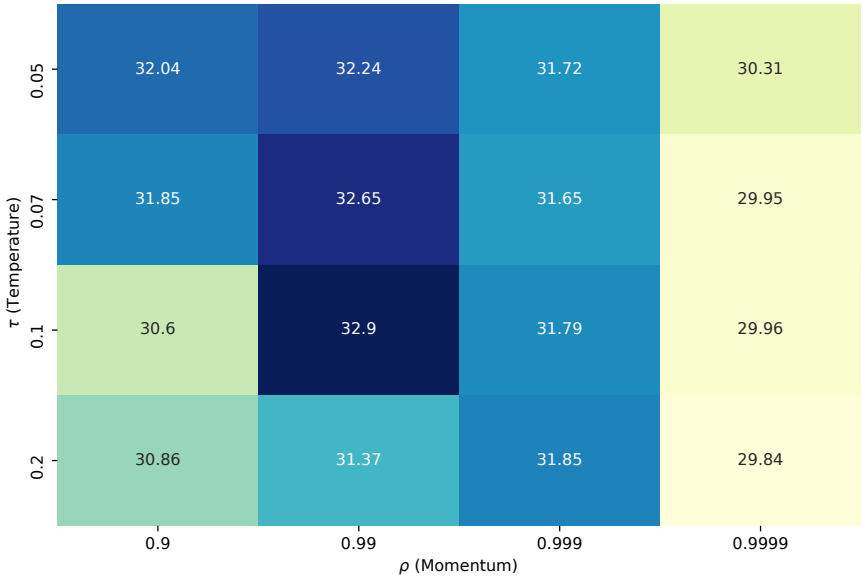

Figure 9. Grid search over τ (temperature) and ρ (momentum) for regular Contrastive Learning method on Physionet 2019. Here result are averaged over 5 runs. Reported metric is Utility on validation set for sepsis task.

D.3. Neighborhood parameters

As shown in Figures 13 and 14, we select the specific parameters to n_w with a grid search over 5 runs. If parameters yielding good performance are stable for MIMIC-III, we found that performance on Utility metric varied significantly for Physionet

2019. We believe a reason for that is the fact these metrics depend on a threshold for making a prediction. Thus, contrary to AUROC or AUPRC, in addition to evaluating the model performances, it also evaluates its calibration.

As showed in Figures 10, 11 and 12 we selected α for NCL(n_Y) on validation set performance. We observe that for all tasks, taking $\alpha = 0.9$ yields the best performance on the training task and in transfer learning.

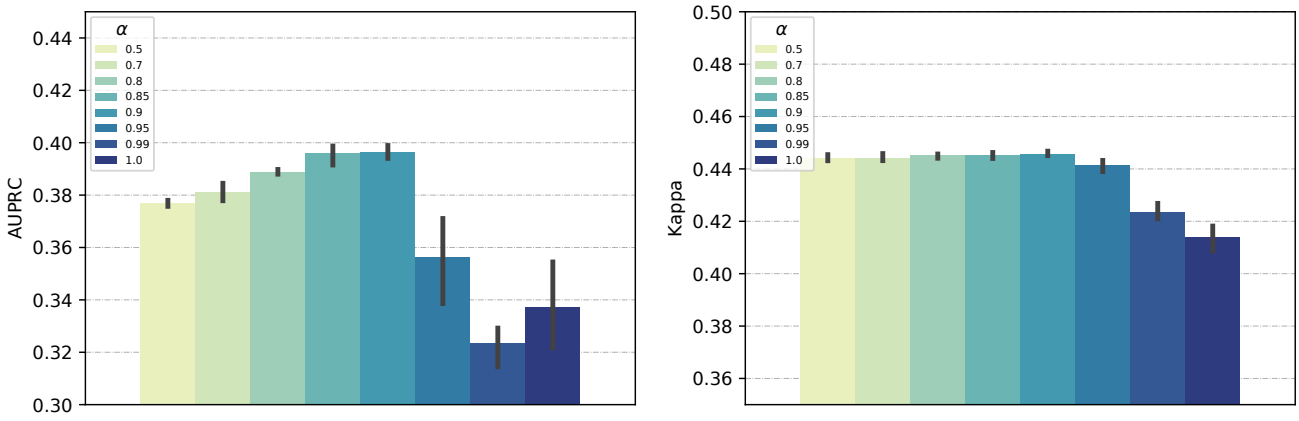

Figure 10. Parameter search over α for NCL(n_Y) method on MIMIC-III Benchmark when trained using *Decompensation* labels. Here result are averaged over 5 runs. Reported metric is AUPRC (for *Decompensation*) and Kappa (for *Length-of-stay*)on validation set.

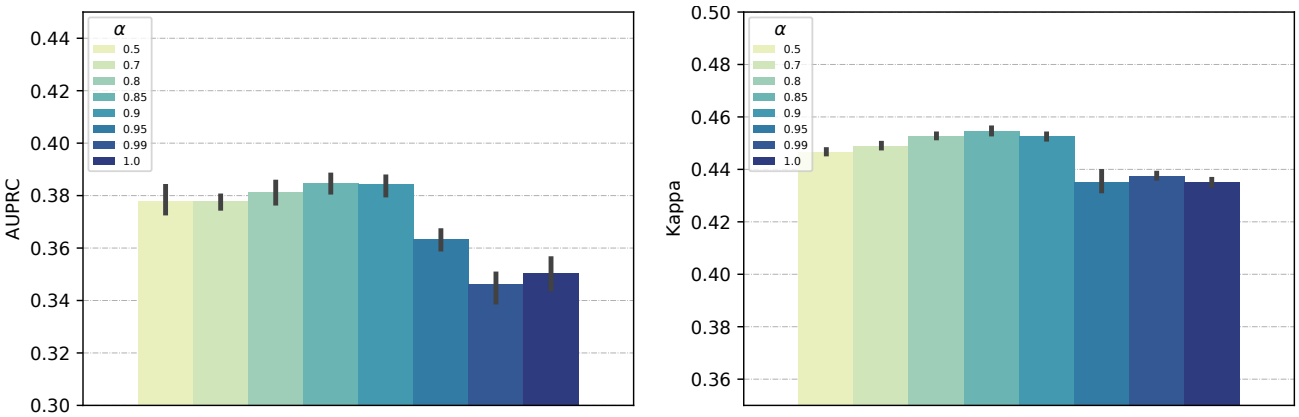

Figure 11. Parameter search over α for NCL(n_Y) method on MIMIC-III Benchmark when trained using *Length-of-stay* labels. Here result are averaged over 5 runs. Reported metric is AUPRC (for *Decompensation*) and Kappa (for *Length-of-stay*)on validation set.

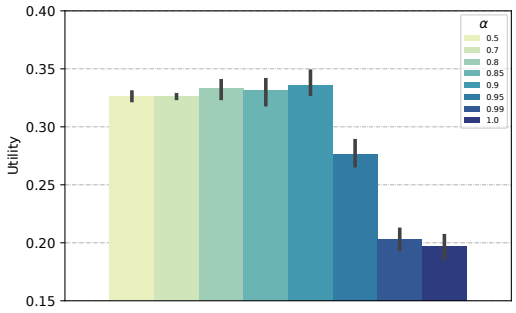

Figure 12. Parameter search over α for NCL(n_Y) method on Physionet 2019. Here result are averaged over 5 runs. Reported metric is Utility on validation set for sepsis task.

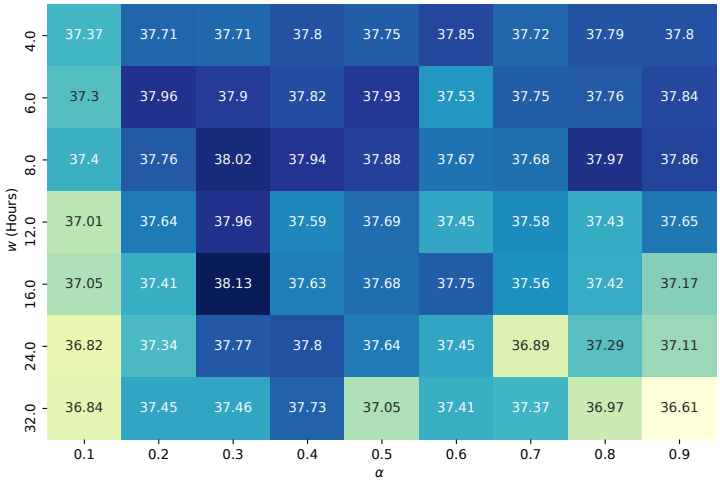

Figure 13. Grid search over w and α for NCL(n_w) method on MIMIC-III Benchmark. Here result are averaged over 5 runs. Reported metric is AUPRC on validation set for decompensation task.

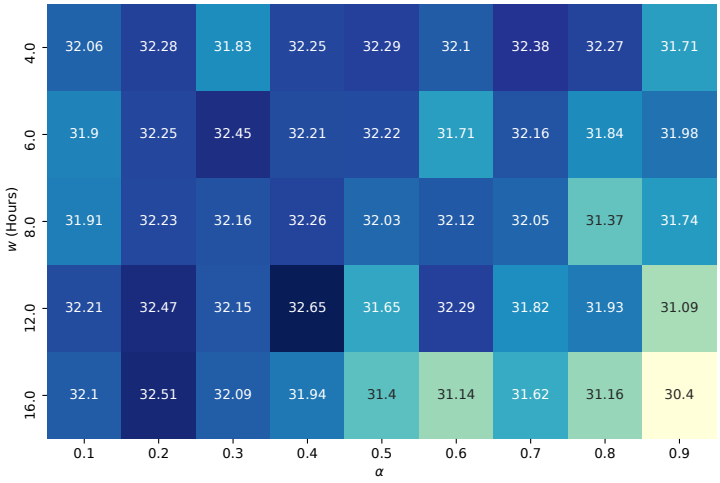

Figure 14. Grid search over w and α for NCL(n_w) method on Physionet 2019. Here result are averaged over 5 runs. Reported metric is Utility on validation set for sepsis task.

E. Supplementary Results

E.1. Semi-supervised neighborhood

We explored another neighborhood possibility for our supervised approach, as the intersection of N_w and N_Y, called NCL($n_{w\cap Y}$). We make two observations from the results in Table 11. First, regardless of the label used for training, results were similar on all tasks and competitive with end-to-end. Second, we managed to achieve a stable training even though $\alpha = 1.0$ by considering as positive the sample with the same label and close temporally. However, we under-perform compared to NCL(n_Y), suggesting that using the $\mathcal{L}^{\text{NCL}}$ objective is a better alternative in this case.

Table 11. Supplementary results on MIMIC-III when using other neighborhood function for supervised approach NCL($n_{w\cap Y}$). (D) and (L) stands for Decompensation and Length-of-Stay indicating which labels were used at training. For NCL($n_{w\cap Y}$) we used a trade-off parameter $\alpha = 1.0$

Task	Decompensation				Length-of-stay	
Metric	AUPRC		AUROC		Kappa	
Head	Linear	MLP	Linear	MLP	Linear	MLP
End-to-End	34.3 ± 1.1	34.2 ± 0.6	90.6 ± 0.3	90.6 ± 0.2	43.3 ± 0.2	**43.4** ± 0.2
NCL($n_{w\cap Y}$) (D) (Ours)	31.3 ± 0.7	34.4 ± 0.6	89.4 ± 0.3	90.7 ± 0.1	40.8 ± 0.3	43.2 ± 0.2
NCL($n_{w\cap Y}$) (L) (Ours)	31.2 ± 0.5	34.0 ± 0.4	89.2 ± 0.2	90.6 ± 0.1	40.6 ± 0.2	43.2 ± 0.2
NCL(n_Y) (D) (Ours)	**37.0** ± 0.6	**37.1** ± 0.7	90.3 ± 0.2	**90.9** ± 0.1	40.8 ± 0.3	43.3 ± 0.2
NCL(n_Y) (L) (Ours)	33.5 ± 1.0	36.0 ± 0.6	88.2 ± 0.5	90.5 ± 0.2	**43.7** ± 0.2	**43.8** ± 0.3

E.2. Fine-tuning representation

The preferred approach to compare representations is to use a frozen classifier. Contrary to fine-tuning, it preserves what was learned at the training step, allowing a fair comparison. However, if one is interested in the absolute performance on a downstream task, fine-tuning the representation encoder with the classification head usually yields better results. We show in Table 12 that for MIMIC-III, we observe this effect by further improving our unsupervised method. However, as shown in Table 13, on Physionet 2019 where our unsupervised method already improved over end-to-end training, performances are degraded. Moreover, we observe that the existing gap between classification heads disappears when fine-tuning. It highlights that fine-tuning shouldn't be used to compare representations learning approaches.

Table 12. Fine-tuning results for MIMIC-III. The results are averaged over the same 20 runs as frozen evaluation.

Task	Decompensation				Length-of-stay	
Metric	AUPRC		AUROC		Kappa	
Head	Linear	MLP	Linear	MLP	Linear	MLP
End-to-End	34.3 ± 1.1	34.2 ± 0.6	90.6 ± 0.3	90.6 ± 0.2	43.3 ± 0.2	43.4 ± 0.2
NCL(n_w) (Ours)	36.7 ± 0.5	36.6 ± 0.4	91.1 ± 0.1	91.3 ± 0.2	43.7 ± 0.3	43.9 ± 0.3
NCL(n_y) (D) (Ours)	36.7 ± 0.7	37.1 ± 0.7	90.7 ± 0.2	90.9 ± 0.1	44.0 ± 0.2	44.0 ± 0.3
NCL(n_y) (L) (Ours)	34.8 ± 1.1	34.7 ± 1.0	90.2 ± 0.2	90.3 ± 0.3	42.5 ± 0.3	42.8 ± 0.3

Table 13. Fine-tuning results for Physionet 2019. The results are averaged over the same 20 runs as frozen evaluation.

Task	Sepsis					
Metric	AUPRC		AUROC		Utility	
Head	Linear	MLP	Linear	MLP	Linear	MLP
End-to-End	7.6 ± 0.2	8.1 ± 0.4	78.9 ± 0.3	78.8 ± 0.4	27.9 ± 0.8	27.5 ± 1.0
NCL(n_w) (Ours)	8.8 ± 0.4	8.9 ± 0.4	80.6 ± 0.4	80.7 ± 0.3	30.2 ± 0.9	30.3 ± 0.9
NCL(n_y) (Ours)	8.9 ± 0.4	9.5 ± 0.4	80.6 ± 0.3	80.9 ± 0.2	30.5 ± 0.7	31.6 ± 0.7

E.3. Visualizing Aggregation Impact

In Figure 15, using T-SNE plots we highlight that by increasing α in $\mathcal{L}^{\text{NCL}}$, we gradually increase aggregation among neighbors. As conjectured, using only $\mathcal{L}^{\text{NA}}$ or $\mathcal{L}^{\text{ND}}$. yields either a collapsed representation or a patient-independent representation.

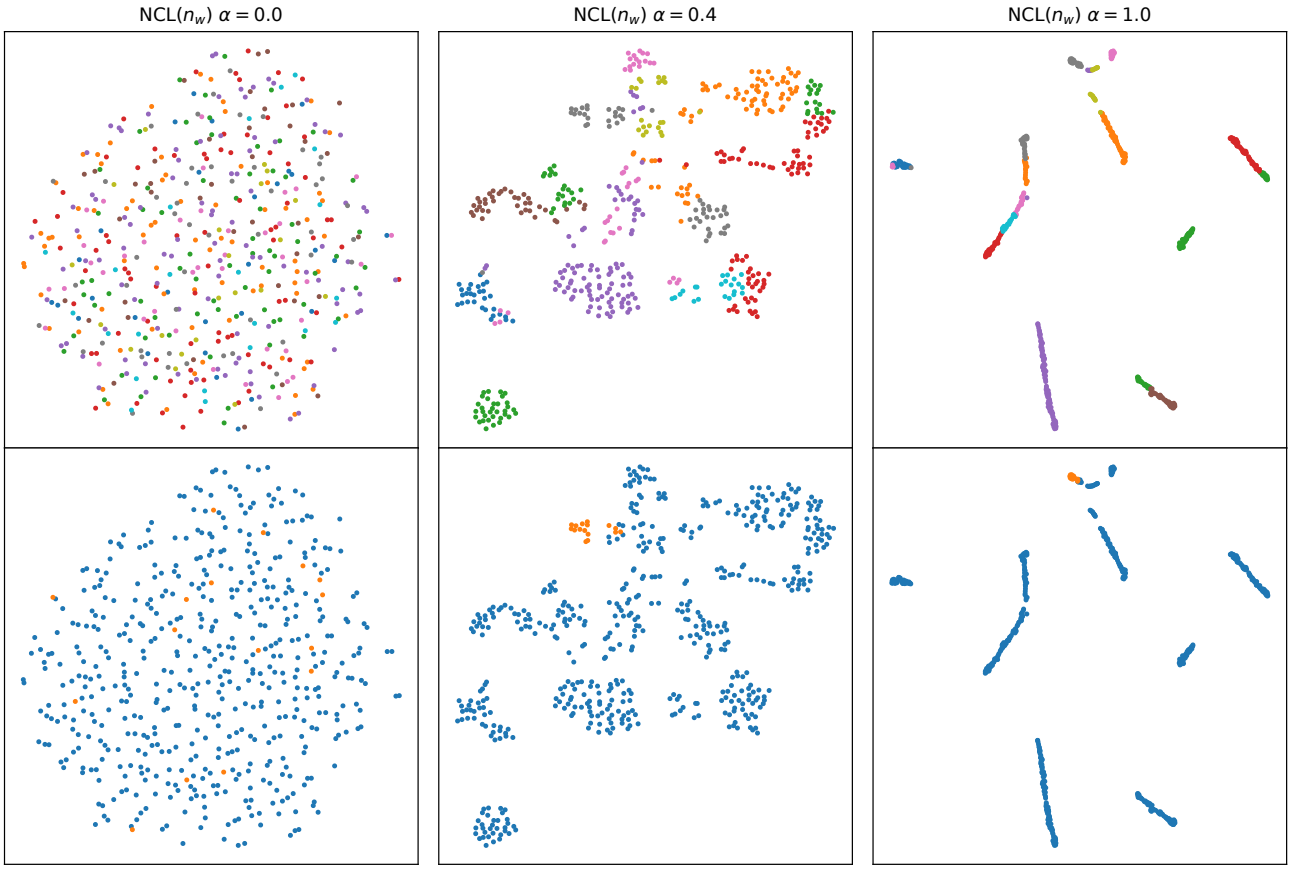

Figure 15. T-SNE plot (Mikolov et al., 2013) of learned representations on MIMIC-III Benchmark dataset for different values of α. (Top row) Labeled per patient. (Bottom row) Labeled with *Decompensation* task. We observe that as conjectured a trade-off in neighbors aggregation is obtained where taking an intermediate value for α.