MULTI-MODAL CONTRASTIVE LEARNING FOR ONLINE CLINICAL TIME-SERIES APPLICATIONS

Fabian Baldenweg, Manuel Burger, Gunnar Rätsch, Rita Kuznetsova
Department of Computer Science, ETH Zürich, Switzerland
{bafabian,burgerm,raetsch,mkuznetsova}@ethz.ch

ABSTRACT

Electronic Health Record (EHR) datasets from Intensive Care Units (ICU) contain a diverse set of data modalities. While prior works have successfully leveraged multiple modalities in supervised settings, we apply advanced self-supervised multi-modal contrastive learning techniques to ICU data, specifically focusing on clinical notes and time-series for clinically relevant online prediction tasks. We introduce a loss function *Multi-Modal Neighborhood Contrastive Loss (MM-NCL)*, a *soft* neighborhood function, and showcase the excellent linear probe and zero-shot performance of our approach.

1 INTRODUCTION

Electronic Health Record (EHR) data from Intensive Care Units (ICUs) has emerged as a valuable resource for predicting clinically relevant quantities in recent years (Hyland et al., 2020; Hüser et al., 2024; Yèche et al., 2022; Pace et al., 2022). However, the diverse nature of EHR data, encompassing different modalities such as clinical notes and time series, presents a challenge for effective utilization. The majority of models leveraging multiple modalities rely on supervised learning Husmann et al. (2022); Jain et al. (2023); Khadanga et al. (2019), necessitating separate training for each task, demanding substantial amounts of annotated data, and learning only task-specific modality interactions. There remains a gap for an architecture, which fuses modalities in a task-agnostic manner.

To address these challenges, there is growing interest in developing self-supervised approaches (van den Oord et al., 2019; Yèche et al., 2021) that can learn task-agnostic representations, thereby reducing or eliminating the dependency on annotated data. Encouragingly, contrastive learning has proven successful in creating such multi-modal representations for text and images without task-specific training (Radford et al., 2021; Wang et al., 2022; Li et al., 2023). Radford et al. (2021) even demonstrate strong zero-shot classification performance based on their multi-modal shared latent space.

Our Contribution: Motivated by initial exploration (Radford et al., 2021; King et al., 2023), we aim to apply multi-modal contrastive learning, specifically focusing on clinical notes and medical time-series, while aiming to improve performance for online prediction tasks (Yèche et al., 2021). We introduce a loss function titled *Multi-Modal Neighborhood Contrastive Learning (MM-NCL)* together with a novel *soft* neighborhood function. We showcase the strong linear probe and zero-shot performance of our approach on in-hospital mortality and, most importantly, decompensation tasks. To the best of our knowledge, our decompensation results represent the best successful benchmarked zero-shot performance on an online ICU prediction task.

2 RELATED WORK

Learning on Medical Time-Series A wide range of research has been conducted on supervised learning applied to ICU time series (Harutyunyan et al., 2019; Kuznetsova et al., 2023; Yèche et al., 2022; van de Water et al., 2024). Uni-modal contrastive methods (Yèche et al., 2021; Zhang et al., 2022; Weatherhead et al., 2022), on the other hand, are more closely related to this work. They often rely on *InfoNCE* (van den Oord et al., 2019) or a variant thereof to pull augmented views of the same data closer together and push different samples apart in the representation space Liu et al. (2023b).

Multi-modal Learning Multi-modal contrastive learning was popularized by CLIP (Radford et al., 2021) contrasting images and captions, leveraging it to train models for zero-shot classifica-

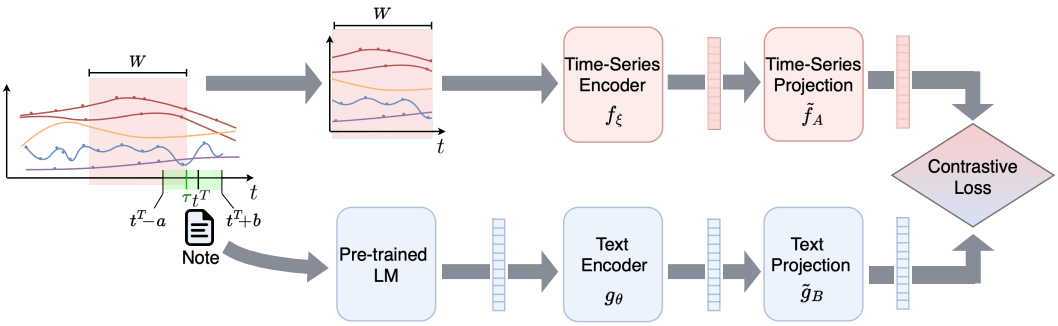


Figure 1: Training pipeline

tion, and laying the groundwork for state-of-the-art image generators (Rombach et al., 2022). Adaptations of CLIP have been proposed for video input (Liu et al., 2023a), multi-lingual text (Chen et al., 2022), and pre-trained uni-modal encoders (Li et al., 2023). In the medical domain, MedCLIP (Wang et al., 2022) adopts this framework and applies it to X-ray images and clinical notes. To overcome the scarcity of paired multi-modal samples they propose a soft assignment based on a text similarity scoring. Li & Gao (2022) used multi-modal contrastive learning to align medical time-series and hyperbolic embeddings of ICD codes, while MedFuse (Hayat et al., 2023) does so with images. King et al. (2023) recently explored multi-modal contrastive learning between clinical notes and medical time-series, however, their approach differs from ours in several aspects, their architecture is only suitable for offline tasks, and we show vastly better performance especially in the zero-shot setting.

3 METHODS

Notation We consider a patient dataset of multiple paired time series X^S and clinical note series X^T, where each pair represents a patient's stay in the ICU. The time series contains d_v hourly vital signs of patients in the ICU, while the clinical note series is sparse across time. $X_t^S \in \mathbb{R}^{d_v}$ is step t of X^S and $X^S[t_1 : t_2]$ for $t_1, t_2 \in \mathbb{N}, t_1 \leq t_2$ denotes the sub-sequence of X^S between steps t_1 (inclusive) and t_2 (exclusive). Let $X_{i,j}^T$ denote the j-th note in ICU stay i. A batch consists of K pairs of texts $X_{i,j}^T$ and time series $X_i^S, i \in \{1, 2, \dots K-1\}$. A sub-sequence of length w (window size) of each time-series near the creation time of the note is fed to the time-series encoder. Let $\textit{time}(X_{i,j}^T)$ be the creation time of note $X_{i,j}^T$. Then, for each note $X_{i,j}^T$, a target time τ_i is drawn uniformly at random from $[\textit{time}(X_{i,j}^T) - a, \textit{time}(X_{i,j}^T) + b]$ where a and b may depend on the type of notes. The final time-series $\tilde{X}_{i,j}^S, i \in \{1, \dots, K-1\}$ where $\tilde{X}_{i,j,\tau}^S = X_i^S[\tau_i - w : \tau_i]$ are fed to the model.

Model Architecture Our model (see Figure 1) consists of a time-series encoder $f_\xi(X^S) = \text{GRU}_\xi(X^S)$, a time-series projection $\tilde{f}_A(x) = W_A x$, a text encoder $g_\theta(X^T) = \textit{concat}(\text{MLP}_\theta(\text{LM}(X^T)), \text{LM}(X^T))$ and a text projection $\tilde{g}_B(x) = W_B x$. GRU_ξ is the last hidden state of a Gated Recurrent Unit (Cho et al., 2014) with parameters ξ. MLP_θ is a multi-layer perceptron (MLP) with parameters θ and LM is a pre-trained language model (Huang et al., 2020), which we use to compute a single representation vector per clinical note. *concat* concatenates two vectors. W_A, W_B are trainable matrices.

Loss Based on Yèche et al. (2021) we propose a *Multi-Modal Neighborhood Contrastive Loss (MM-NCL)* $\mathcal{L}_{MM-NCL}$. We also do experiments with our pipeline using the original loss from CLIP (Radford et al., 2021) (and refer to it using *MM-InfoNCE*).

Let B be the set of index tuples $\iota = (i, j, \tau)$ in a batch of size K, containing stay index, note index, and target time. Let $\nu > 0$ be a trainable temperature parameter and the following define the normalized embeddings ($norm(x) := x / \|x\|$) of the time-series $\mathbf{h}_\iota^S$ and the texts $\mathbf{h}_\iota^T$ passed to the contrastive objective:

$$\mathbf{h}_\iota^S := norm\big(\tilde{f}_A(f_\xi(X_{i,j,\tau}^S))\big) \in \mathbb{R}^c, \quad \mathbf{h}_\iota^T := norm\big(\tilde{g}_B(g_\theta(X_{i,j,\tau}^T))\big) \in \mathbb{R}^c \tag{1}$$

MM-NCL consists of two components: (**i**) The neighborhood aware loss $\mathcal{L}_A$ (Eqn. 2) and (**ii**) the neighborhood discriminative loss $\mathcal{L}_D$ (Eqn. 3). Eqn. 4 and 5 define a novel *soft* neighborhood function relating neighboring clinical notes and time-series windows w.r.t. their distance in time:

$$\mathcal{L}_A \coloneqq \sum_{l\in B}\sum_{m\in B} -\frac{N_{l,m}}{2K}\left(\log\frac{\exp\left(\mathbf{h}_l^S\cdot\mathbf{h}_m^T/\nu\right)}{\sum_{n\neq l}^{B}\exp\left(\mathbf{h}_l^S\cdot\mathbf{h}_n^T/\nu\right)}+\log\frac{\exp\left(\mathbf{h}_l^T\cdot\mathbf{h}_m^S/\nu\right)}{\sum_{n\neq l}^{B}\exp\left(\mathbf{h}_l^T\cdot\mathbf{h}_n^S/\nu\right)}\right) \tag{2}$$

$$\mathcal{L}_D \coloneqq \sum_{l\in B} -\frac{1}{2K}\left(\log\frac{\exp(\mathbf{h}_l^S\cdot\mathbf{h}_l^T/\nu)}{\sum_{m}^{B}\mathbb{1}_{l,m}^N\exp(\mathbf{h}_l^S\cdot\mathbf{h}_m^T/\nu)}+\log\frac{\exp(\mathbf{h}_l^T\cdot\mathbf{h}_l^S/\nu)}{\sum_{m}^{B}\mathbb{1}_{l,m}^N\exp(\mathbf{h}_l^T\cdot\mathbf{h}_m^S/\nu)}\right) \tag{3}$$

where

$$N_{l,m} \coloneqq \frac{\tilde{N}(l,m)}{\sum_{n\in B}\tilde{N}(l,m)}, \qquad \mathbb{1}_{l,m}^N \coloneqq \begin{cases}1, & \text{if } N_{l,m}\neq 0\\ 0, & \text{if } N_{l,m}=0\end{cases} \tag{4}$$

$$\tilde{N}(l,m) \coloneqq \begin{cases}\frac{\beta}{\beta+|\tau_m-\tau_l|}, & \text{if } i_l=i_m\wedge|j_l-j_m|\leq 1\\ 0, & \text{otherwise}\end{cases} \tag{5}$$

$\beta\in\mathbb{R}^{\geq 1}$ is a hyperparameter (defining the soft neighborhood decay w.r.t. temporal distance) and the final loss function is a linear combination with trade-off hyperparameter α:

$$\mathcal{L}_{MM-NCL} \coloneqq \alpha\mathcal{L}_A + (1-\alpha)\mathcal{L}_D, \qquad \alpha\in(0,1] \tag{6}$$

While Yèche et al. (2021) have proposed the separation of a contrastive loss into neighborhood *aware* and *distriminative* components, we have expanded their definition to the multi-modal setting and introduce a soft neighborhood function, where they have only considered the binary case.

4 Results and Discussion

Experimental Setup We use the time-series features, cohort selection, splits, and label definitions from MIMIC-III Benchmark (Harutyunyan et al., 2019) and extracted the clinical notes from the MIMIC-III dataset (Johnson et al., 2016). We benchmark in-hospital mortality and decompensation as defined by Harutyunyan et al. (2019). In-hospital mortality is an offline binary classification of predicting patient mortality after the first 48 hours of stay in the ICU. Decompensation is an hourly *online* binary classification task to predict the onset of death in the next 24 hours. More details in Appendix A and B.

Evaluation Note that for all results, while the contrastive pretraining considers multiple modalities, for inference only time-series data is passed to the model. This is different from some supervised baselines (Husmann et al., 2022; Khadanga et al., 2019) and is more suitable for an online deployment scenario, where measurements and lab results are naturally stored in databases, but clinical notes require a physician to analyse the respective data streams and write the note, before they would become visible to the system. We evaluate linear probes (Alain & Bengio, 2017) on the output of the frozen base time-series encoder $f_\xi(X^S)$. Further, we consider zero-shot classification by scoring the alignment of an embedded time-series window with class-specific text prompts (Radford et al., 2021). Time-series windows are classified by their similarity with positive (e.g. *"patient died"*) and negative prompt ensembles (e.g. *"patient survived"*, more examples in App. C). Let $\mathcal{P}_+^T$ and $\mathcal{P}_-^T$ be the set of positive and negative text prompts. Then the zero-shot probabilities $\hat{y}_{zs}$ for a time-series window X^S, encoded to $\mathbf{h^S}$ as in Eqn. 1, are (Eqn. 7):

$$\hat{y}_{zs} = \text{softmax}(\mathbf{h^S}\cdot\mathbf{p}_+^T,\ \mathbf{h^S}\cdot\mathbf{p}_-^T) \quad \text{where} \quad \mathbf{p}_{+/-}^T \coloneqq \frac{1}{|\mathcal{P}_{+/-}^T|}\sum_{p\in\mathcal{P}_{+/-}^T} norm\big(g_B(g_\theta(p))\big) \tag{7}$$

Results Table 1 shows comparisons to a supervised time-series baseline (Harutyunyan et al., 2019) and supervised multi-modal baselines (Khadanga et al., 2019; Husmann et al., 2022). Further, we compare to self-supervised results for online predictions (Yèche et al., 2021) and self-supervised multi-modal results (King et al., 2023). We strongly outperform prior work by King et al. (2023) in

Table 1: Model Performance Comparison. All values are in % and denoted as $mean \pm std$. Bold marks the best result in each section. *MM-Train.* and *-Infer.* mark if multiple modalities are used for training and inference. Missing values are not provided by the respective references.

Method	*MM Train.*	*MM Infer.*	**Mortality** *AuPRC*	**Mortality** *AuROC*	**Decompensation** *AuPRC*	**Decompensation** *AuROC*
Supervised						
Harutyunyan et al. (2019)	✗	✗	50.1 ± 1.3	86.1 ± 0.3	34.1 ± 0.5	90.7±0.2
Khadanga et al. (2019)	✓	✓	52.5±1.3	86.5±0.4	34.5±0.7	90.7±0.7
Husmann et al. (2022)	✓	✓	**52.7±1.0**	**87.1±0.6**	**39.7±0.6**	**92.2±0.2**
Self-Supervised Linear Probes						
Yèche et al. (2021)	✗	✗	-	-	31.2 ± 0.5	88.9 ± 0.3
King et al. (2023)	✓	✗	40.2 ± 5.3	82.8 ± 2.0	-	-
MM-NCL (ours)	✓	✗	**52.1 ± 0.5**	**85.9 ± 0.2**	**32.6 ± 1.2**	**90.2 ± 0.4**
Self-Supervised Zero-Shot						
King et al. (2023)	✓	✗	21.4 ± nan	70.9 ± nan	-	-
MM-InfoNCE (ours)	✓	✗	**48.3 ± 1.4**	**83.2 ± 0.63**	26.9 ± 2.2	**87.8 ± 0.2**
MM-NCL (ours)	✓	✗	45.1 ± 2.8	80.0 ± 2.4	**30.9 ± 0.7**	87.4 ± 0.7

both the probed and the zero-shot setting. Our probed results on mortality can even compete with a strong supervised multi-modal baseline by Husmann et al. (2022), while on decompensation we can slightly improve upon the results by Yèche et al. (2021).

In the zero-shot setting, we present the first results on an online patient prediction task (decompensation). Our loss function for multi-modal neighborhood contrastive learning in online settings achieves a zero-shot performance getting close to probed results. Additionally, we vastly outperform the only available prior result on multi-modal contrastive learning for time-series and clinical notes on mortality by King et al. (2023). The difference in performance on mortality for MM-InfoNCE and MM-NCL can be attributed to the nature of our proposed loss function focusing on more local, clinically relevant (Yèche et al., 2021), online patient state changes, while on the offline mortality prediction global alignment is favored.

Scarce label setting Figure 2 compares supervised, linear probe and zero-shot results on training sets with reduced labels. It clearly shows the superiority of zero-shot predictions in the scarce label regime.

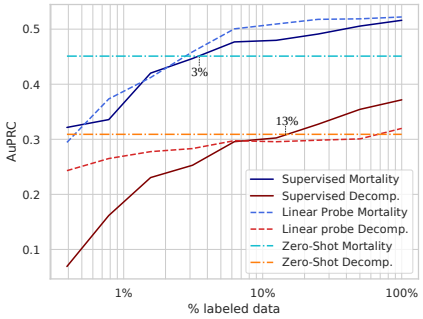
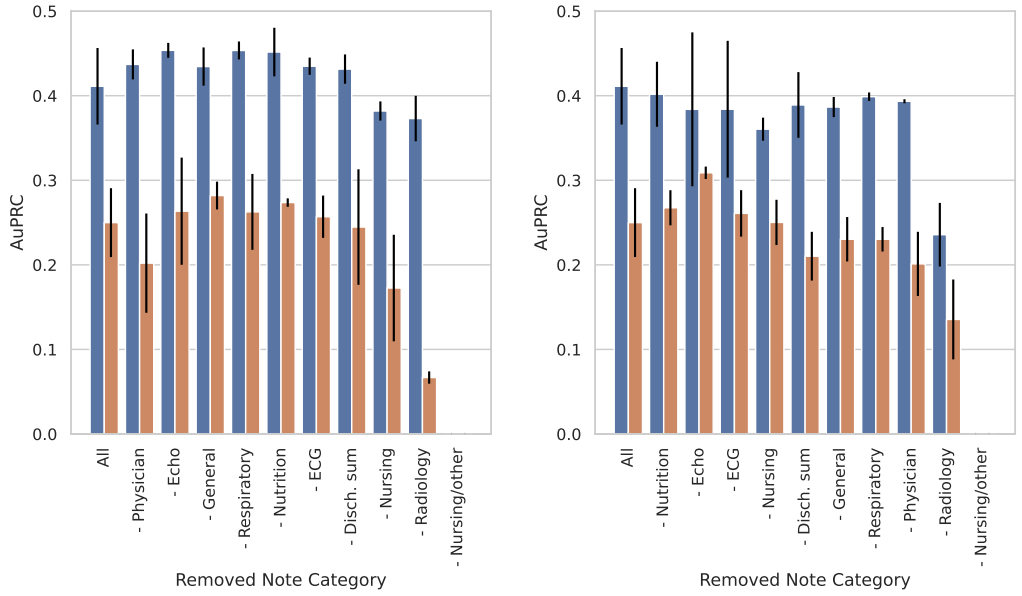


Figure 2: AuPRC when training with reduced labels, x-axis shows the percentage of of labels used from the full training set. All results were obtained using the same time-series architecture. We mark the percentage, where supervised outperforms zero-shot.

Ablation on Note Types For each task, we optimized the set of note types used during training by greedily removing note types from the training data. In each run, we early-stopped based on the validation AuPRC of the task we optimized for. In Figure 3, we observe a strong effect of note type selection on model performance. Notable differences between the tasks are that `Physician` notes seem to be a lot more important for decompensation and that `Nursing` notes seem to be more important for mortality. To be expected was that the last remaining categories are `Radiology` and `Nursing/other` notes, as they make up 65% of all considered notes. Also expected was that discharge summaries are more helpful for mortality as they highlight information relevant over the entire patient stay and tend to mention patient outcomes. Note that differences in the two plots for pre-training on `Nursing/other` only stem from optimizing the selection (including early stopping of the pre-training) for different tasks.



(a) Mortality

(b) Decompensation

Figure 3: Zero-shot AuPRC for mortality (blue) and decompensation (orange) for different sets of note types for *MM-NCL*. We greedily remove note types from the left to the right based on mortality (Fig. 3a) and decompensation (Fig. 3b) AuPRC. Removing the last remaining category (`Nursing/other` in both cases) leaves no training data for the text modality, so there is no result in the rightmost columns.

5 Conclusion

We proposed a new multi-modal contrastive loss function for clinical notes and time-series. Leveraging a soft neighborhood function we can train a multi-modal shared latent space, which exhibits strong performance under linear probing and facilitates unseen zero-shot classification performance in this application domain. Further research remains to validate our findings on other datasets and experiment with the inclusion of additional data modalities.

Acknowledgments

We would like to thank Hugo Yèche for insightful discussions and valuable advice. This project was supported by grant #2022-278 of the Strategic Focus Area "Personalized Health and Related Technologies (PHRT)" of the ETH Domain (Swiss Federal Institutes of Technology; to G.R.).

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2017. URL https://openreview.net/forum?id=ryF7rTqgl.

Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Fulong Ye, Qinghong Yang, and Ledell Wu. Altclip: Altering the language encoder in clip for extended language capabilities, 2022.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96, Jun 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0103-9. URL https://doi.org/10.1038/s41597-019-0103-9.

Nasir Hayat, Krzysztof J. Geras, and Farah E. Shamout. Medfuse: Multi-modal fusion with clinical time-series data and chest x-ray images, 2023.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission, 2020.

Matthias Hüser, Xinrui Lyu, Martin Faltys, Alizée Pace, Marine Hoche, Stephanie Hyland, Hugo Yèche, Manuel Burger, Tobias M Merz, and Gunnar Rätsch. A comprehensive ml-based respiratory monitoring system for physiological monitoring & resource planning in the icu. *medRxiv*, 2024. doi: 10.1101/2024.01.23.24301516. URL https://www.medrxiv.org/content/early/2024/01/23/2024.01.23.24301516.

Severin Husmann, Hugo Yèche, Gunnar Ratsch, and Rita Kuznetsova. On the importance of clinical notes in multi-modal learning for ehr data. In *NeurIPS 2022 Workshop on Learning from Time Series for Health*, 2022.

Stephanie L Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature medicine*, 26(3):364–373, 2020.

Samyak Jain, Manuel Burger, Gunnar Rätsch, and Rita Kuznetsova. Knowledge graph representations to enhance intensive care time-series predictions, 2023. URL https://arxiv.org/abs/2311.07180.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):160035, May 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.35. URL https://doi.org/10.1038/sdata.2016.35.

Swaraj Khadanga, Karan Aggarwal, Shafiq Joty, and Jaideep Srivastava. Using clinical notes with time series data for icu management. *arXiv preprint arXiv:1909.09702*, 2019.

Ryan King, Tianbao Yang, and Bobak Mortazavi. Multimodal pretraining of medical time series and notes, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Rita Kuznetsova, Alizée Pace, Manuel Burger, Hugo Yèche, and Gunnar Rätsch. On the importance of step-wise embeddings for heterogeneous clinical time-series. In Stefan Hegselmann, Antonio Parziale, Divya Shanmugam, Shengpu Tang, Mercy Nyamewaa Asiedu, Serina Chang, Tom Hartvigsen, and Harvineet Singh (eds.), *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pp. 268–291. PMLR, 10 Dec 2023. URL https://proceedings.mlr.press/v225/kuznetsova23a.html.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

Rui Li and Jing Gao. Multi-modal contrastive learning for healthcare data analytics. In *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*, pp. 120–127, 2022. doi: 10.1109/ICHI54592.2022.00029.

Ruyang Liu, Jingjia Huang, Ge Li, Jiashi Feng, Xinglong Wu, and Thomas H. Li. Revisiting temporal modeling for clip-based image-to-video knowledge transferring, 2023a.

Ziyu Liu, Azadeh Alavi, Minyi Li, and Xiang Zhang. Self-supervised contrastive learning for medical time series: A systematic review. *Sensors*, 23(9), 2023b. ISSN 1424-8220. doi: 10.3390/s23094221. URL https://www.mdpi.com/1424-8220/23/9/4221.

Alizée Pace, Alex Chan, and Mihaela van der Schaar. POETREE: Interpretable policy learning with adaptive decision trees. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=AJsI-ymaKn_.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

Vimal Thilak, Chen Huang, Omid Saremi, Laurent Dinh, Hanlin Goh, Preetum Nakkiran, Joshua M. Susskind, and Etai Littwin. Lidar: Sensing linear probing performance in joint embedding ssl architectures, 2023.

Robin van de Water, Hendrik Schmidt, Paul Elbers, Patrick Thoral, Bert Arnrich, and Patrick Rockenschaub. Yet another ICU benchmark: A flexible multi-center framework for clinical ML. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ox2ATRM90I.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text, 2022.

Addison Weatherhead, Robert Greer, Michael-Alice Moga, Mjaye Mazwi, Danny Eytan, Anna Goldenberg, and Sana Tonekaboni. Learning unsupervised representations for icu timeseries. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann (eds.), *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 152–168. PMLR, 07–08 Apr 2022. URL https://proceedings.mlr.press/v174/weatherhead22a.html.

Hugo Yèche, Gideon Dresdner, Francesco Locatello, Matthias Hüser, and Gunnar Rätsch. Neighborhood contrastive learning applied to online patient monitoring. In *International Conference on Machine Learning*, pp. 11964–11974. PMLR, 2021.

Hugo Yèche, Rita Kuznetsova, Marc Zimmermann, Matthias Hüser, Xinrui Lyu, Martin Faltys, and Gunnar Rätsch. Hirid-icu-benchmark – a comprehensive machine learning benchmark on high-resolution icu data, 2022.

Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency, 2022.

A DATA

A.1 TASKS

We benchmark two mortality-related tasks defined by Harutyunyan et al. (2019) on the MIMIC-III dataset (Johnson et al., 2016). Harutyunyan et al. (2019) provide cohort selection and patient splits for training. Mortality information is directly extracted by Harutyunyan et al. (2019) from the MIMIC-III patient metadata.

In-Hospital Mortality Given an ICU stay and its time-series X^S the binary classification task is to predict, whether the patient died in the ICU or was discharged alive based on the first 48 hours of stay ($X^S[0:48]$).

Decompensation Given an ICU stay the *online* binary classification task is to predict the onset of death at every hour of the patient's stay until death or discharged alive.

A.2 TIME-SERIES

We prepare time-series data from MIMIC-III (Johnson et al., 2016) as published by Harutyunyan et al. (2019). Missing values are forwarded imputed if prior measurements are available. Data is standard-scaled and the remaining missing values after forward imputation are zero-imputed, which corresponds to a population mean imputation.

A.3 CLINICAL NOTES

We consider the same set of notes as Jain et al. (2023) and provide their description and details on the clinical notes published in the MIMIC-III (Johnson et al., 2016) dataset in the `NOTEEVENTS` table.

In total, there are about 2 million individual text notes of 10 categories (`Discharge summary`, `ECG`, `Echo`, `General`, `Nursing`, `Nursing/other`, `Nutrition`, `Physician`, `Radiology` and `Respiratory`). The median length of such a note is 1090 characters and we observe a median of about 14 individual notes per patient (with 7 at the first quartile and 30 at the third quartile).

Each note is associated with a specific timestamp (`CHARTDATE` and `CHARTTIME`) during a single admission (`HADM_ID`) of a given patient (`SUBJECT_ID`). However, for a given patient admission we do not observe a note at every single time point on our resampled grid used during training. Some time points might have no clinical note associated with them, whereas others might have multiple, and they thus build an irregularly sampled time series of textual descriptions of the patient state.

B TRAINING DETAILS

B.1 HYPERPARAMETERS

Table 2: Hyperparameter ranges, chosen ones are in bold

Parameter Name	**Values**
GRU hidden dimension	[128, **256**]
GRU depth	[1,**2**,3]
GRU dropout	[**0.1**, 0.2, 0.3]
Text Encoder number of hidden dimensions	[**1**]
Text Encoder MLP hidden dimension	[**4096**]
Loss parameter α	[0.1, **0.3**, 0.5, 0.8, 0.95, 1.0]
Loss parameter β	[**2**, 4, 8, 16, 24, 48, 96]
Window size	[8, **16**, 24, 48]
Batch size	**512** Patient Stays
Learning Rate Adam (Kingma & Ba, 2017)	$5e^{-4}$

We tuned the hyperparameters on the validation performances in several grid searches, refining iteratively over time. Table 2 shows an overview. We use PyTorch (Paszke et al., 2019) for train-

Table 3: Zero-shot performance for different window sizes w on the MIMIC-III Benchmark tasks. All values are in % and denoted as $mean \pm std$. All hyperparameters except for window size are kept fixed.

Window Size	**Mortality**		**Decompensation**	
hours	*AuPRC*	*AuROC*	*AuPRC*	*AuROC*
8	42.3 ± 1.3	80.6 ± 1.1	27.2 ± 0.5	87.1 ± 1.0
16	45.0 ± 3.2	81.2 ± 0.7	29.5 ± 0.5	87.5 ± 0.8
24	43.6 ± 2.8	79.0 ± 2.7	28.8 ± 2.1	88.1 ± 0.6
48	42.3 ± 0.2	80.1 ± 0.5	21.9 ± 3.9	86.8 ± 1.2

Table 4: Zero-shot performance for different window sizes w on the MIMIC-III Benchmark tasks. All values are in % and denoted as $mean \pm std$.

Class	**Mortality**	**Decompensation**
Positive Prompts	patient deceased passed away patient died died deceased expired condition: expired care withdrawn	Discharge Condition: Expired Expired died dnr
Negative Prompts	survived stable discharged	stable stable condition discharged today

ing the models. We trained all models with a single `NVIDIA RTX2080Ti` and an Intel `Xeon E5-2630v4` CPU.

For the target time selection hyperparameters we chose $b = 3$ and $a = 10$ for discharge summaries, $a = 30$ for radiology notes, and $a = 3$ for everything else. Those are set in time-steps, which translates to hours on the MIMIC-III Benchmark (Harutyunyan et al., 2019).

B.2 Model Selection

We pretrain the model with *MM-NCL* for 30 epochs, which has been tuned for best aggregated zero-shot task performance on the validation set. Future work should look into incorporating more efficient early-stopping methods such as *LiDAR* (Thilak et al., 2023), which enables early stopping online in a more efficient way.

C Evaluation

Ablation on window sizes Table 3 shows the zero-shot performance of our model with different window sizes w.

Model Prompts

We present a collection of prompts used in the zero-shot classification prompt ensembles in Table 4. The prompts have been selected based on an inspection of the notes found in the MIMIC-III dataset (Johnson et al., 2016) conditioned on the class labels.