On the Importance of Clinical Notes in Multi-modal Learning for EHR Data

Severin Husmann **Hugo Yèche** **Gunnar Rätsch** *

Rita Kuznetsova *

{shusmann,hyeche,raetsch,mkuznetsova}@ethz.ch
Department of Computer Science, ETH Zürich

Abstract

Understanding deep learning model behavior is critical to accepting machine learning-based decision support systems in the medical community. Previous research has shown that jointly using clinical notes with electronic health record (EHR) data improved predictive performance for patient monitoring in the intensive care unit (ICU). In this work, we explore the underlying reasons for these improvements. While relying on a basic attention-based model to allow for interpretability, we first confirm that performance significantly improves over state-of-the-art EHR data models when combining EHR data and clinical notes. We then provide an analysis showing improvements arise almost exclusively from a subset of notes containing broader context on patient state rather than clinician notes. We believe such findings highlight deep learning models for EHR data to be more limited by partially-descriptive data than by modeling choice, motivating a more data-centric approach in the field.

1 Introduction

Recently, researchers have explored using deep neural networks for clinical decision support, particularly for Intensive Care Unit (ICU) prediction tasks. Most existing literature focused on using either available electronic health records [1, 2, 3, 4, 5, 6, 7] (EHR) or clinical notes [8, 9, 10, 11] individually. However, more recent research shows that combining both data types in a cross-modal fashion improves performance [12, 13] on a popular suite of MIMIC-III [14] benchmark tasks [15]. Ultimately, the goal of such improvement in the field can help better plan the resource usage in ICU [12, 16]. Though, as for any real-world application, acceptance by clinicians will require an understanding of the behavior of these new cross-modal models. Other studies provided analyses for uni-modal models using EHR data [4] or clinical notes [8]. To the best of our knowledge, similar work investigating the cross-modal interactions when combining both modalities to monitor patients in ICU does not exist.

We aim to fill that gap by exploring the underlying reasons for performance improvements obtained from jointly using EHR data and clinical notes. Like Horn et al. [4], which analyzed the performances of EHR-only models, we rely on a transformer-based [17] model, as attention weights provide a direct explanation of the deep learning model behavior. Due to the multi-modal nature of the data, we adopt a multi-modal extension [18] to this original transformer architecture. In this work, first, we

*Co-supervised

Workshop on Learning from Time Series for Health, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

confirm that the most basic cross-modal transformer model combining EHR data and clinical notes performs significantly better than state-of-the-art uni-modal approaches on all classification tasks from the MIMIC-III benchmark [15]. Second, by performing a two-step analysis, we show that the performance gain compared to EHR-only models arises exclusively from a subset of nurses' notes and radiology reports. These notes provide additional context on patients' states. Moreover, we also show that using clinicians' notes, which summarize EHR contents rather than containing additional information, does not provide any performance gain. Thereby, we refute the hypothesis that current architectures poorly represent EHR information. We believe our findings motivate more data-centric approaches in the field, focusing on leveraging more descriptive data to model patients' states in the ICU.

2 Related work

Deep Learning for ICU patient monitoring. As previously mentioned, we can categorize existing research using deep learning to monitor patients' evolution in the ICU according to the modality used. First, the most common line of work relies on the EHR, viewed as a multivariate time series. A large group of related literature tackles this problem with various architectures in both supervised [1, 2, 3, 4, 5, 6, 7] and unsupervised fashion [19, 20]. Other researchers focus exclusively on clinical notes [8, 9, 10, 11]. At last, motivating our work, more recent research explored fusing both modalities [12, 13, 21, 22], yielding significant improvement over previous uni-modal approaches. Unfortunately, in this last body of work, the authors work with independent representations for each modality making it difficult to retrieve cross-modal interaction, thus limiting interpretability.

Interpretability in deep learning. Widely used methods in deep learning interpretability rely on saliency-map-based methods like LIME [23], GRAD-CAM [24], and SHAP [25]. Though, these methods are limited [26] and are not necessarily compatible with sequence data. Thus, we base our analysis on attention-based methods, a more common approach for sequence modeling. Various papers have investigated the use of attention weights as a potential post hoc interpretability mechanism [27, 28, 29, 30]. In addition, other literature further developed methods beyond attention weights, such as attention flow and rollout [31]. In the case of attention flow, [32] proved that attention flows are layer-wise Shapley values. Attention rollout is often employed in computer vision to provide attention heatmaps over the input [33, 34]. In the context of ICU data, similarly to us, Horn et al. [4] also based their uni-modal model behavior analysis on attention weights.

3 Method

In this section, we introduce our interpretable model to handle both clinical notes and EHR data. We also describe hyperparameter choice and data pre-processing to ensure a better analysis of clinical notes' importance.

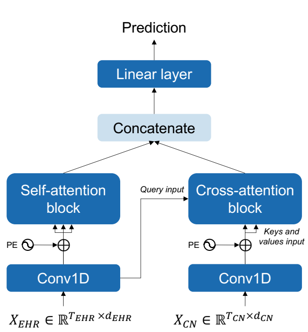

Figure 1: Proposed cross-modal architecture

Architecture. We consider two modalities $\mathbf{X}_{EHR} \in \mathbb{R}^{T_{EHR} \times d_{EHR}}$ and $\mathbf{X}_{CN} \in \mathbb{R}^{T_{CN} \times d_{CN}}$, where T_{EHR} and T_{CN} denote the length of the EHR data and clinical notes during the given patient's stay, and d_{EHR} and d_{CN} the input feature dimensionality. To obtain $\mathbf{X}_{CN}$ we use the pre-trained Clinical BERT model [35] as a text feature extractor, $d_{CN} = 769$. Note that d_{CN} is greater than the original Clinical BERT output dimension by one. This is because clinical notes are not regularly spaced, which cannot be accounted for by positional encoding. Thus, we add the entry time of notes as the last feature. The dimensionality of EHR data is $d_{EHR} = 42$ after one-hot encoding of categorical variables. Following the original paper [18], we pass the input modalities through a 1D convolutional layer to embed the text series and the time series into the same latent space. Then we add positional encoding to each modality. Since we make predictions at $\mathbf{X}_{EHR}$ frequency, we always obtain queries from $\mathbf{X}_{EHR}$ and keys and values from $\mathbf{X}_{EHR}$ or $\mathbf{X}_{CN}$ in the self-attention and cross-attention layers, respectively. Figure 1 shows the employed architecture.

Tasks definition. Previous literature jointly using EHR data and clinical notes [12, 13] relied on MIMIC-III data [36]. Like other existing datasets, MIMIC-III consists of physiological measurements and information about laboratory tests. However, contrary to others, in MIMIC-III, a significant number of patients have clinical notes written by various medical personnel allowing for a multi-modal approach. Thus, for our experiments, we follow previous research by analyzing clinical notes' importance on Harutyunyan et al. [15]'s benchmark tasks. In particular, we focus our work on three tasks. First, *decompensation*, an hourly binary prediction task on whether the patient dies in the next 24 hours. Second, *in-hospital mortality* (IHM), a similar task with a single binary prediction task after 48 hours of observed ICU data on whether the patient dies in the ICU. Finally, *phenotyping*, a multi-class classification problem where one predicts which acute care conditions were present during a patient's stay out of 25 care conditions. The benchmark contains another task, hourly predicting the remaining length of stay. We do not consider this task due to its lower clinical relevance and the poor performance of all existing deep learning models on the task. We further discuss the data and its processing in Appendix A.

Hyperparameter tuning. Increasing the number of heads or layers in transformers usually improves performance but also makes the models' analysis more complex. Since we already observed significant improvement over state-of-the-art uni-modal models using a single layer and head, we fixed these two hyperparameters to 1 in all experiments. Doing so allowed us to preserve a more explainable model at the cost of a slight loss in performance. All other hyperparameters were selected using a grid search on the validation performance of *decompensation*. For the convolutional layer, we use 64 filters with a size equal to the input space. In the transformer, we use a latent space of size 64 everywhere and a dropout rate of 0.2. We trained all models with a batch size of 16 patient stays, a learning rate of $1e^{-5}$, and Adam optimizer [37].

Data processing. As it is common for MIMIC-III benchmark data [15], we used forward imputation and standard scaling as pre-processing steps. More importantly, we mask out the last note in all tasks to avoid label leaks through direct death mentions. This is a simple way to reduce information leaking. Though, it is possible to tackle this more in-depth by filtering samples using keyword search as [11]. As for the choice of hyperparameters, masking the last note might reduce absolute performance compared to previous research not following the same procedure [12, 13], but ensure a fairer analysis of the role of the clinical notes.

4 Experiments

In this section, we first show that even with hard constraints on model complexity, jointly using EHR data and clinical notes outperforms existing uni-modal models on the considered tasks. We then perform an ablation study on note types and analyze cross-modal attention weights highlighting the importance of either "nurse" or "radiology" notes for cross-modal model decision-making. All results reported are from the test set and obtained over 5 random seeds.

Overall performance. As shown in Table 1, while only employing a single cross-attention layer with a single head and not attending to the last note, using a joint model surpasses previous literature's performance utilizing either EHR data [15] or clinical notes [12]. This observation confirms prior findings on the addition brought by clinical notes in patient monitoring [12, 13] and justifies the need to understand the underlying reason for such improvement. Interestingly, we also observe that our model performs on par with the best cross-modal model on both *decompensation* and *phenotyping* tasks despite our constraints.

Note type ablation. To understand which note types are important to the decision-making in our cross-modal model, we carry out ablation studies where we cumulatively add note types by increasing or by decreasing frequency (see Appendix B). Figure 2 shows the ablation studies for the *decompensation* and *IHM* tasks by increasing frequency. We draw two conclusions. On the one hand, adding doctors' ("Physicians") notes does not improve performance. This suggests that current representations obtained from the EHR data alone are sufficient. On the other hand, adding nurses' notes and radiology reports for *IHM* significantly improves AUPRC on both tasks. Similar results can be found for *phenotyping* in Appendix B. These findings show that, when using clinical notes, deep learning models benefit from additional textual information rather than a possibly better representation of similar information in doctors' notes summarizing the patient's state.

Table 1: Performances on MIMIC-III Benchmark [15] for uni-modal *(top, middle)* and cross-modal models *(bottom)*. Reported errors, if provided, are 95% confidence intervals on the mean.

Task	Decompensation		IHM		Phenotyping	
Metric	AUPRC	AUROC	AUPRC	AUROC	Macro-AUC	Micro-AUC
EHR-Only LSTM [15]	33.3 ± 1.0	90.6 ± 0.3	51.5 ± 5.5	86.2 ± 2.2	77.6 ± 0.4	82.5 ± 0.3
EHR-Only Transformer (**Ours**)	31.7 ± 0.4	90.3 ± 0.2	49.8 ± 0.9	85.5 ± 0.2	74.1 ± 0.1	79.9 ± 0.1
Text-only LSTM [12]	8.1 ± 0.7	79.3 ± 0.7	30.3 ± 1.3	79.3 ± 0.4	-	-
Text-only Transformer (**Ours**)	19.2 ± 0.6	83.3 ± 0.3	34.1 ± 0.6	77.9 ± 0.5	77.8 ± 0.2	82.7 ± 0.2
Cross-Modal LSTM [12]	34.5 ± 0.7	90.7 ± 0.7	52.5 ± 1.3	86.5 ± 0.4	-	-
Cross-Modal Ensemble [13]	40.4 ± n.a.	92.0 ± n.a.	58.2 ± n.a.	88.6 ± n.a.	82.9 ± n.a.	87.0 ± n.a.
Cross-Modal Transformer (**Ours**)	39.7 ± 0.6	92.2 ± 0.2	52.7 ± 1.0	87.1 ± 0.6	82.6 ± 0.1	86.1 ± 0.1

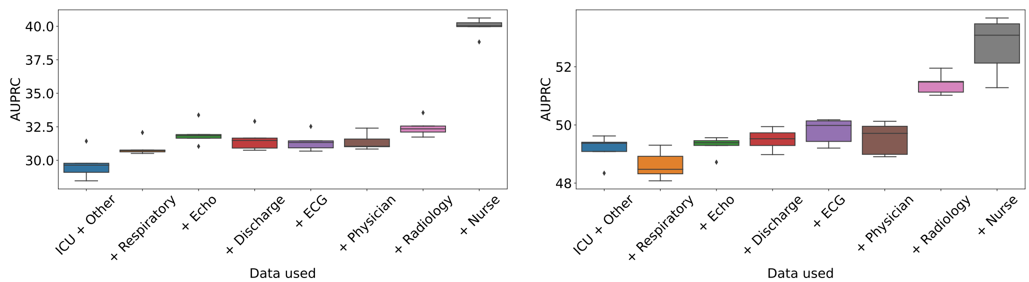

Figure 2: Area under the precision-recall curve performance of the cross-modal transformer when adding different note types by increasing frequency. *(left)* Decompensation. *(right)* In-hospital mortality.

Cross-modal attention weights analysis. For all tasks, we observed that notes among the two most frequent types, nurses' and radiology notes, lead to the greatest improvements. Thus, one could argue that the note frequency explains the increase in performance over the note type itself. To answer this question, we perform an analysis of cross-modal attention weights. As shown with an illustrative example in Figure 3, attention weights are very salient, making each EHR timestep attend to only a subset of notes. In addition, also shown in Figure 3, highly attended notes, such as the 7th note in the example, also impact the model's final prediction. Indeed, when it obtains this note at hour 45, the cross-modal model's predictions diverge from the EHR-only model ones. This high saliency of attention weights refutes the hypothesis that the nurses' and radiology notes are the most relevant types due to their high frequency. Indeed, in practice, only a subset of these notes are attended to independently of the notes type frequency. We provide additional examples to support this claim in Appendix C.

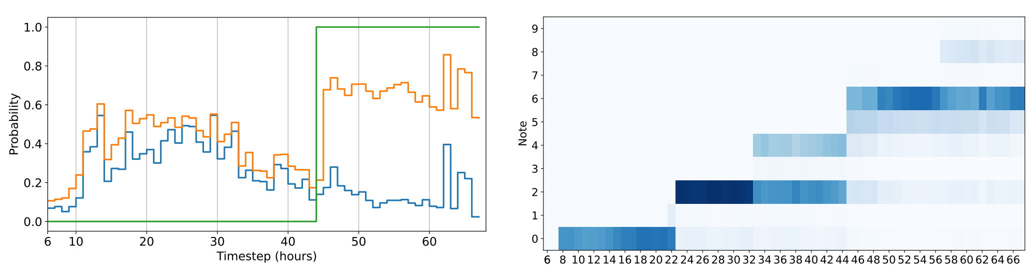

Figure 3: Illustrative example of model's behaviour for decompensation. (*left*) Predictions for EHR-only (blue) and cross-modal (orange) models over time compared to ground truth labels (green). (*right*) Cross-attention over time for the cross-modal model where EHR timesteps (x-axis) attends to clinical notes (y-axis). Both figures represent patient 11555.

5 Conclusion

In this paper, we aim to understand the underlying reason behind the significant improvement brought by jointly using EHR data and clinical notes in a multi-modal fashion in ICU-related tasks. To answer this question, we analyze the performance impact of the available note types in MIMIC-III and the cross-modal attention weights in our model. Based on our experimental results, we conclude that improvements yielded by adding clinical notes come from notes containing additional context on patient states, such as nurses' notes and radiology reports. We also find that doctors' notes, which often summarize EHR content, do not improve performance, suggesting that representations learned from EHR data alone preserve the input information. We believe these findings show how EHR data alone is only a narrow view of patient states in the ICU, motivating more data-centric approaches in the field.

References

[1] Joseph Futoma, Sanjay Hariharan, and Katherine Heller. Learning to detect sepsis with a multitask gaussian process rnn classifier. In *International conference on machine learning*, pages 1174–1182. PMLR, 2017.

[2] Nenad Tomašev, Xavier Glorot, Jack W Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572(7767):116–119, 2019.

[3] Patrick Schwab, Arash Mehrjou, Sonali Parbhoo, Leo Anthony Celi, Jürgen Hetzel, Markus Hofer, Bernhard Schölkopf, and Stefan Bauer. Real-time prediction of covid-19 related mortality using electronic health records. *arXiv preprint arXiv:2008.13412*, 2020.

[4] Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten Borgwardt. Set functions for time series. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4353–4363. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/horn20a.html`.

[5] Xiang Li, Xiao Xu, Fei Xie, Xian Xu, Yuyao Sun, Xiaoshuang Liu, Xiaoyu Jia, Yanni Kang, Lixin Xie, Fei Wang, et al. A time-phased machine learning model for real-time prediction of sepsis in critical care. *Critical Care Medicine*, 48(10):e884–e888, 2020.

[6] 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.

[7] Hugo Yèche, Rita Kuznetsova, Marc Zimmermann, Matthias Hüser, Xinrui Lyu, Martin Faltys, and Gunnar Ratsch. HiRID-ICU-benchmark — a comprehensive machine learning benchmark on high-resolution ICU data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL `https://openreview.net/forum?id=SnC9rUeqiqd`.

[8] Paulina Grnarova, Florian Schmidt, Stephanie L Hyland, and Carsten Eickhoff. Neural document embeddings for intensive care patient mortality prediction. *arXiv preprint arXiv:1612.00467*, 2016.

[9] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission, 2019.

[10] Jinyue Feng, Chantal Shaib, and Frank Rudzicz. Explainable clinical decision support from text. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 1478–1489, 2020.

[11] Betty van Aken, Jens-Michalis Papaioannou, Manuel Mayrdorfer, Klemens Budde, Felix A. Gers, and Alexander Löser. Clinical outcome prediction from admission notes using self-supervised knowledge integration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 881–893. Association for Computational Linguistics, 2021. URL `https://www.aclweb.org/anthology/2021.eacl-main.75/`.

[12] 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.

[13] John Chen, Ian Berlot-Attwell, Safwan Hossain, Xindi Wang, and Frank Rudzicz. Exploring text specific and blackbox fairness algorithms in multimodal clinical nlp. *arXiv preprint arXiv:2011.09625*, 2020.

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

[15] Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, and A. G. Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6, 2019.

[16] David W. Bates, Suchi Saria, Lucila Ohno-Machado, Anand Shah, and Gabriel Escobar. Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Affairs*, 33(7):1123–1131, 2014. ISSN 0278-2715. doi: 10.1377/hlthaff.2014.0041.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[18] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1656. URL `https://aclanthology.org/P19-1656`.

[19] Xinrui Lyu, Matthias Hueser, Stephanie L Hyland, George Zerveas, and Gunnar Rätsch. Improving clinical predictions through unsupervised time series representation learning. *arXiv preprint arXiv:1812.00490*, 2018.

[20] 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*, pages 11964–11974. PMLR, 2021.

[21] Mengqi Jin, Mohammad Taha Bahadori, Aaron Colak, Parminder Bhatia, Busra Celikkaya, Ram Bhakta, Selvan Senthivel, Mohammed Khalilia, Daniel Navarro, Borui Zhang, et al. Improving hospital mortality prediction with medical named entities and multimodal learning. *arXiv preprint arXiv:1811.12276*, 2018.

[22] Haiyang Yang, Li Kuang, and FengQiang Xia. Multimodal temporal-clinical note network for mortality prediction. *Journal of Biomedical Semantics*, 12(1):1–14, 2021.

[23] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL `https://doi.org/10.1145/2939672.2939778`.

[24] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[25] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[26] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

[27] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4828. URL `https://aclanthology.org/W19-4828`.

[28] Jesse Vig. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-3007. URL `https://www.aclweb.org/anthology/P19-3007`.

[29] Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.574. URL `https://aclanthology.org/2020.emnlp-main.574`.

[30] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020. doi: 10.1162/tacl_a_00349. URL `https://aclanthology.org/2020.tacl-1.54`.

[31] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.385. URL https://aclanthology.org/2020.acl-main.385.

[32] Kawin Ethayarajh and Dan Jurafsky. Attention flows are shapley value explanations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 49–54, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.8. URL https://aclanthology.org/2021.acl-short.8.

[33] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[34] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 14200–14213. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/76ba9f564ebbc35b1014ac498fafadd0-Paper.pdf.

[35] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1909. URL https://aclanthology.org/W19-1909.

[36] Alistair E. W. Johnson, Tom J. Pollard, and Roger G. Mark. Reproducibility in critical care: a mortality prediction case study. In Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 361–376. PMLR, 18–19 Aug 2017. URL https://proceedings.mlr.press/v68/johnson17a.html.

[37] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

A Dataset

A.1 Pre-processing

We first run the pre-processing step provided by Harutyunyan et al. [15] for EHR data. Next, we extract all patients with notes. Echo, electrocardiogram (ECG), and discharge summary notes have no *CHARTTIME*, the exact time at which the note is entered in the system. However, these notes might contain valuable medical information. Thus, we set their *CHARTTIME* to the end of the day of their *CHARTDATE*, which contains only the day on which the note was entered. It ensures no temporal leak of information. Following previous work [12], we then exclude patients without any notes. From the benchmark task with 6281 ICU stays in the test set, we exclude 27 patients and end up with 6254 ICU stays. For some of these stays, only phenotyping task labels exist, hence, for the decompensation task, we end up with 6215 labeled ICU stays in the test set. We follow the same approach for the training set, validation set, and test set.

A.2 Statistics

For further details, we provide in Table 2 the test set statistics for the *decompensation* and *IHM* tasks. In addition, Table 3 shows the note type counts in the training and test set. We ran the ablation studies according to the note type frequency in the training set.

Table 2: Number of negative and positive labels for the IHM task and the decompensation task test sets.

(a) Decompensation task test set

Label	Number of labels
0	512862
1	9427
Total	522289

(b) IHM task test set

Label	Number of labels
0	2860
1	365
Total	3225

Table 3: Number of notes per note type in the test and training cohort.

(a) Note type distribution in test set

Note type	Number of notes
Nursing	66206
Radiology	39744
ECG	16031
Physician	14261
Discharge summary	6809
Echo	3762
Respiratory	3036
Nutrition	885
General	812
Rehab Services	562
Social Work	263
Case Management	102
Pharmacy	13
Consult	12

(b) Note type distribution in training set

Note type	Number of notes
Nursing	316015
Radiology	183490
Physician	74089
ECG	73227
Discharge summary	31561
Echo	17298
Respiratory	16247
Nutrition	4777
General	4376
Rehab Services	2744
Social Work	1400
Case Management	535
Pharmacy	50
Consult	43

B Ablation Studies

In this section, we provide further results regarding notes type ablations.

Table 4: Note type ablation study results when adding clinical notes by increasing frequency.

Task	Decompensation		IHM		Phenotyping	
	AUPRC	AUROC	AUPRC	AUROC	AUROC	
Model					Macro	Micro
ICU + Other	29.7 ± 1.1	90.0 ± 0.1	49.2 ± 0.5	85.4 ± 0.2	74.2 ± 0.3	80.0 ± 0.2
+ Respiratory	30.9 ± 0.6	90.1 ± 0.1	48.6 ± 0.5	85.1 ± 0.2	74.2 ± 0.2	80.0 ± 0.1
+ Echo	32.0 ± 0.9	90.2 ± 0.1	49.3 ± 0.3	84.9 ± 0.2	74.8 ± 0.1	80.4 ± 0.1
+ Discharge	31.5 ± 0.9	90.1 ± 0.1	49.5 ± 0.4	84.9 ± 0.2	76.0 ± 0.1	81.3 ± 0.1
+ ECG	31.4 ± 0.7	90.1 ± 0.1	49.8 ± 0.4	85.8 ± 0.2	78.0 ± 0.2	82.8 ± 0.1
+ Physician	31.4 ± 0.6	90.0 ± 0.2	49.5 ± 0.6	85.7 ± 0.2	78.6 ± 0.2	83.2 ± 0.1
+ Radiology	32.5 ± 0.7	90.6 ± 0.2	51.4 ± 0.4	86.4 ± 0.4	81.4 ± 0.2	85.2 ± 0.1
+ Nurse	39.9 ± 0.7	92.3 ± 0.2	52.7 ± 1.0	87.1 ± 0.6	82.6 ± 0.1	86.1 ± 0.1

B.1 Increasing Frequency

We first summarize results from the ablation by increasing frequency from Figure 2 in Table 4. We find, as previously mentioned, that for all tasks, nurses' or radiology notes contribute to improved performance. For the phenotyping task, we see that also other note types help. It may be because these types contain information for specific illnesses not necessarily present in the other notes, thus also from some additional context on patients' states. Nonetheless, the main contribution to performance remains from radiology reports.

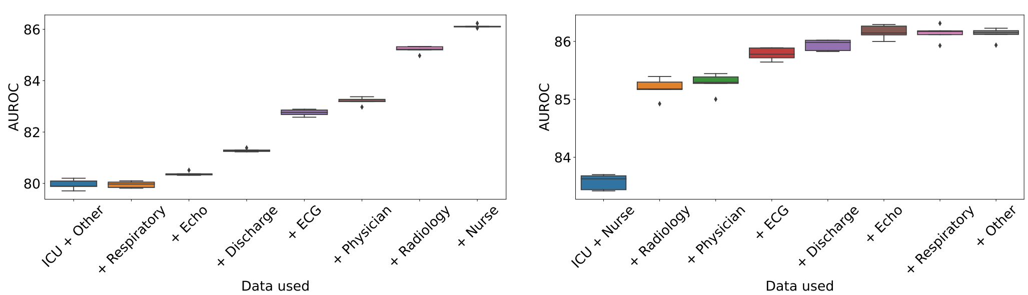

Figure 4: Micro-averaged AUROC note type ablation study results for the phenotyping task. *(left)* By increasing note type frequency. *(right)* by decreasing note type frequency.

B.2 Decreasing Frequency

We also provide an ablation study on note types by decreasing frequency, as a sanity check, in Figure 5 and Table 5. We draw the same conclusions on note types importance.

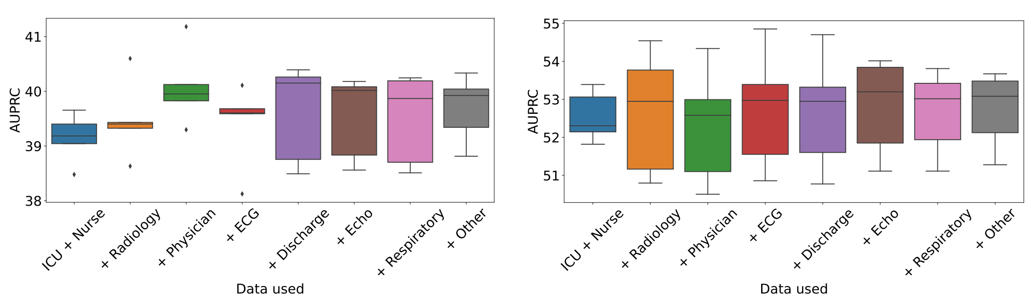

Figure 5: Note type ablation study results by adding types by decreasing frequency. *(left)* Decompensation. *(right)* In-hospital mortality.

Table 5: Note type ablation study results when adding clinical notes by decreasing frequency.

Task	Decompensation		IHM		Phenotyping	
	AUPRC	AUROC	AUPRC	AUROC	AUROC	
Model					Macro	Micro
ICU + Nurse	39.2 ± 0.4	92.3 ± 0.1	52.5 ± 0.7	86.6 ± 0.2	79.2 ± 0.2	83.6 ± 0.1
+ Radiology	39.5 ± 0.7	92.4 ± 0.1	52.6 ± 1.6	86.8 ± 0.9	81.4 ± 0.2	85.2 ± 0.2
+ Physician	40.1 ± 0.7	92.4 ± 0.2	52.3 ± 1.5	86.8 ± 0.8	81.5 ± 0.2	85.3 ± 0.2
+ ECG	39.4 ± 0.8	92.3 ± 0.2	52.7 ± 1.6	87.1 ± 0.8	82.2 ± 0.1	85.8 ± 0.1
+ Discharge	39.6 ± 0.9	92.3 ± 0.3	52.7 ± 1.5	87.1 ± 0.8	82.3 ± 0.1	85.9 ± 0.1
+ Echo	39.5 ± 0.8	92.2 ± 0.2	52.8 ± 1.3	87.1 ± 0.7	82.7 ± 0.1	86.2 ± 0.1
+ Respiratory	39.5 ± 0.8	92.2 ± 0.2	52.7 ± 1.1	87.1 ± 0.7	82.6 ± 0.2	86.1 ± 0.1
+ Other	39.7 ± 0.6	92.2 ± 0.2	52.7 ± 1.0	87.1 ± 0.6	82.6 ± 0.1	86.1 ± 0.1

C Interpretability results

In this section, we describe our attempt to analyze notes importance more in-depth through attention rollout and chunk level representations. We also provide a similar analysis for another illustrative example.

C.1 Deeper analysis on main manuscript example

In the example from the main manuscript, we found that the 7th note was strongly attended to in Figure 3. To further understand, what lead this note to significantly change the model prediction toward a imminent death, we ran attention rollout [31] and summed over the words [27] (see Figure 6). Additionally, we normalized by the number of words in each chunk when a note does not fit a single Clinical BERT pass (i.e., it is longer than 128 tokens). One interesting observation is how *dnr* is significant in the overall note representation. Indeed this stands for "do no resucitate", indicating a clear change of treatment in case of cardiac or respiratory arrest and thus a more probable death.

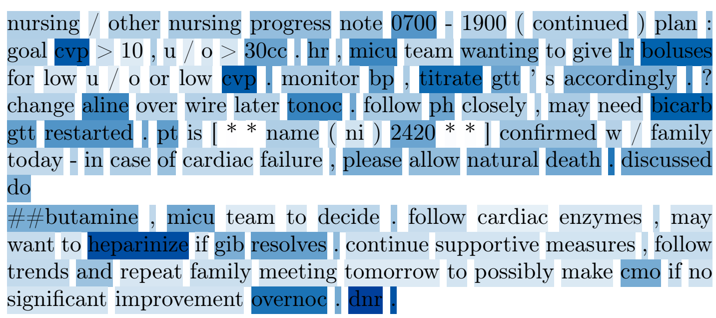

Figure 6: Attention rollout of Clinical BERT on 7th note from patient 11555

Chunk level representation of notes. In our model, if a note is longer than 128 tokens on which Clinical BERT was trained, we passed the chunks separately through Clinical BERT and averaged the global token embedding. For an additional level of interpretability, we use the same model architecture and train using all embedded chunks of the same note with the same time appended. For the identical patient, we see that chunk 12 is attended early on, and the 25th and 27th chunks are the strongest attended to when the prediction gap increases. Figure 8 shows the attention rollout results on the chunks separately. Once again in the most attended chunk (see Figure 8c), we see the nurses mentions to "please allow natural death".

C.2 Other Illustrative Example

To further illustrate our conclusions, we also present a second example corresponding to the patient with ID 12767. Figure 9 shows the predictions and the corresponding cross-attention from ICU

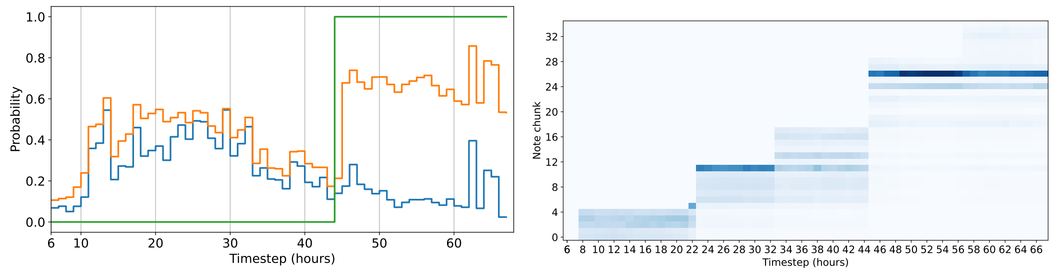

Figure 7: Illustrative example of chunk model's behaviour for decompensation. (*left*) Predictions for EHR-only (blue) and cross-modal (orange) models over time compared to ground truth labels (green). (*right*) Chunk cross-attention over time for the cross-modal model where EHR timesteps (x-axis) attend to clinical notes (y-axis). Both figures represent patient 11555.

was no insulin gtt x 1 / 2 hr this am . fsbs increased this pm , and ? restart of insulin gtt . soc) pt ' s family at bedside and has been updated throughout the day . pt was made dnr as of 18 : 00 , per pt ' s hcp [* * name (ni) * *] [* * name (ni) * *]) .

(a) Attention rollout for 12th chunk

##ative skin care done , frequent turns . access : r ij , l sc central lines in place , patent . l radial aline not functioning . micu team aware . social : son , daughter and in laws here for family meeting todday w / micu team . discussed very poor prognosis and likely long complicated recovery if he does survive this episode of severe sepsis and cardiogenci compromise . plan is to give pt 24hrs to follow labs and trends . will re - meet tomorrow at 3 or 4 to discuss goals of care . family clearly expressed that pt would not want trac

(b) Attention rollout for 25th chunk

nursing / other nursing progress note 0700 - 1900 (continued) plan : goal cvp > 10 , u / o > 30cc . hr , micu team wanting to give lr boluses for low u / o or low cvp . monitor bp , titrate gtt ' s accordingly . ? change aline over wire later tonoc . follow ph closely , may need bicarb gtt restarted . pt is [* * name (ni) 2420 * *] confirmed w / family today - in case of cardiac failure , please allow natural death . discussed do

(c) Attention rollout for 27th chunk

Figure 8: Attention rollout over Clinical BERT from nurse's note chunks for the patient with ID 11555

time series to text series over time. We observe that at the 63rd timestep when the multi-modal transformer's prediction deviates from the EHR-only transformer's prediction it attends to the newly available 8th note.

As in the previous example, we compute the attention rollout from this particular note in Figure 9. This short note states that the patient is most likely to be made DNR/DNI. We see that besides the DNR status, attention is seemingly strongly placed also on *unresponsive* and *tylenol*. Interestingly, we observe that the last note is highly attended to. Indeed, we see that the note already discusses the patient's death, highlighting the existing leakage of information. Hence, it is clear why the model is above 90% certain that the patient will die within the next 24 hours. We show this last note attention rollout in Figure 11. Interestingly, not necessarily *expired* is focused in contrast to *mortem*. We believe this is because the former is a clinical term, thus, more frequent in the corpus used to train clinical BERT.

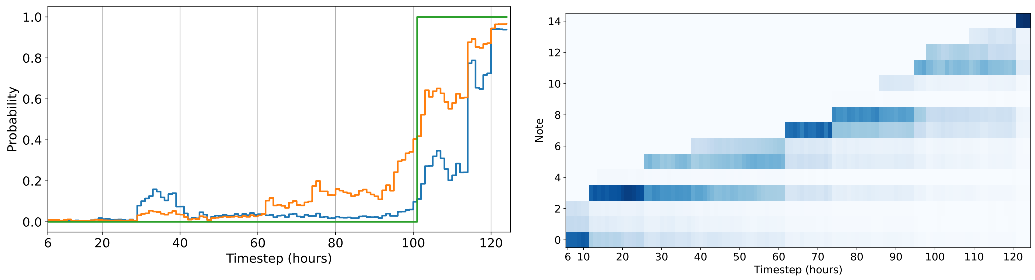

Figure 9: Illustrative example of model's behaviour for decompensation. (*left*) Predictions for EHR-only (blue) and cross-modal (orange) models over time compared to ground truth labels (green). (*right*) Cross-attention over time for the cross-modal model where EHR timesteps (x-axis) attend to clinical notes (y-axis). Both figures represent patient 12767.

nursing / other nursing progress notes : pt remains [* * name (ni) * *] / vented 40 % 5ps and 5peep . awaiting arrival of one more son to add in decision process of code status and extubation . lung sounds remain clear , sx for minimal amt ' s of thick tan secretions . 02 sat ' s cont in high 90 ' s . cv : hr 80 ' s to 90 ' s nsr no ectopy . bp stable with lopressor see careview for data . temp max 101 . tylenol given . gi : foley patent draining mod amt ' s of clear yellow urine . gu : ngt clamped . npo . no stool . abd soft distended , hypoactive bowel sounds . neuro : no chg in neuro status . cont to be unresponsive , does not open eyes . moving right arm and right leg , no movement on left side . pupils 2mm react briskly . iv fluid cont at 40 / hr . pt waiting till morning to most likely be made dnr / di .

Figure 10: Attention rollout of Clinical BERT of 8th note from the patient with ID 12767

As for the previous examples, we also explore the chunks of the notes directly. Figure 12 shows the ICU to nurse note chunks cross-attention over time. For this patient the 8th, 17th, 23rd, 25th, and 33rd chunks are strongly attended to over time. In Figure 13 we present the corresponding attention rollout results. In chunk 7, we find medically relevant observations like *obtundation* or *decerebrate like posturing* attended to. Obtundation describes the status when the patient has a reduced level of consciousness and alertness. Decerebrate posture is a body pose indicating severe brain damage. In the other chunks, the *DNR* status is again highlighted. Though, the cross-modal network actually does not have a higher likelihood of death for this patient at the time of this note. Hence, it might not have been able to learn about such medical knowledge yet. Interestingly the last note chunk, which strongly indicates that death is near, shows attention on the *namepatterns* and not the patient being placed on comfort measures and that they were extubated (cf. Figure 13e). Since we did not train end-to-end due to resource constraints, it is possible that the model was unable to capture the expected humanly relevant rationales correctly and represented this observation differently instead.

nursing / other pt expired mr [* * known lastname 5307 * *] had been made cmo yesterday . comfort maintained today using morphine . pt repositioned . vs monitored and family updated q2 - 3 hours on pt condition . their desire was to be updated if there was a change . rn shaving patient at 1615 and his hr dropped from 115 to the 60 ' s . and rr decr to 9 from rate of 30 . family called immediately . [* * first name11 (name pattern1) 2294 * *] [* * initials (
namepattern4) * *] [* * last name (namepattern4) 2295 * *] , rn spoke with [* * name (ni) 2489 * *] [* * name (ni) 5307 * *] , pt son and told him the family should come in . bp droppped to 72 / 20 , hr and rr continued to decr . rn called son again and told him that they would not make in in time , that pt was dying and rn would stay with him in the room . pt expired at 161
##5 with nurse [* * first name (titles) * *] [* * last name (titles) 64 * *] . family to arrive within an hour . neurosurg resident called . post - mortem care to be done and pt will be transported to morgue after family has arrived .

Figure 11: Attention rollout of Clinical BERT of last note from the patient with ID 12767.

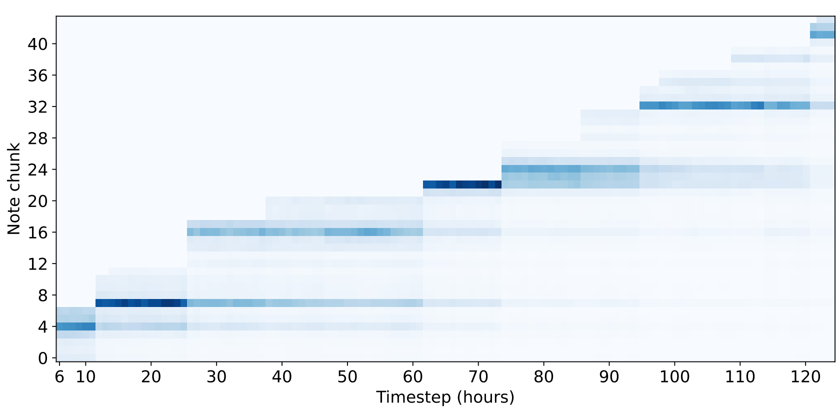

Figure 12: Chunk cross-attention over time for the cross-modal model where EHR timesteps (x-axis) attends to clinical notes (y-axis) for the patient with ID 12767

nursing / other night note : micub neuro : unchanged through night . responds to voice intermittently . able to hold one thumb up on command w / r hand , wiggle toes sl . on r . very weak . no movement on left side . neuro and neuro sicu in to eval pt . pupils [* * 1 - 21 * *] and sluggish . periods of deep obtundation to more wake , almost decerebrate like posturing on r , md ' s in while this was occuring . plan for f / u ct scan today

(a) Attention rollout for 8th chunk

doctors . son in particular seemed upset and eager for info . neuro [* * doctor first name * *] team member spoke with family . a - no real change in neuro exam , stable post parietal bleed . p - will wean nipride as we can once lopressor kicks in , goal bp < 150 / . for now will leave intubated until neuro exam brightens . check culture results . watch i ' s and o ' s , labs . ? can start tube feeds in am if still intubated . monitor neuro exam for changes . keep family advised as to all

(b) Attention rollout for 17th chunk

. gi : foley patent draining mod amt ' s of clear yellow urine . gu : ngt clamped . npo . no stool . abd soft distended , hypoactive bowel sounds . neuro : no chg in neuro status . cont to be unresponsive , does not open eyes . moving right arm and right leg , no movement on left side . pupils 2mm react briskly . iv fluid cont at 40 / hr . pt waiting till morning to most likely be made dnr / di .

(c) Attention rollout for 23rd chunk

be the furthest rehab he would get to , maybe recognizing family , chair , flacid on left . wife , son , and daughter want to uphold wishes of pt which is not to have non quality life , as his mother did for 6 yrs in a nsg home , s / p stroke . dnr order obtained . family waiting for last daughter to arrive from [* * name (ni) 794 * *] late tonight . family wishes to give pt . a few days to see if any improvements occur . dr . [* * last name (stitle) 530

(d) Attention rollout for 25th chunk

nursing / other nursing note 0700 - 1500 ; neuro ; exam slightly worse per neuro , family meeting held with [* * first name8 (namepattern2) * *] [* * last name (namepattern1) * *] md sicu team , [* * first name4 (namepattern1) * *] [* * last name (namepattern1) * *] social worker . and family in agreement with patients own wishes decided to place patient on comfort measures only . he was extubated to r / a at 12 15 and has

(e) Attention rollout for 33rd chunk

Figure 13: Attention rollout over Clinical BERT from nurse's note chunks for the patient with ID 12767