# PARALLEL TIME-SENSOR ATTENTION FOR ELECTRONIC HEALTH RECORD CLASSIFICATION

**Rachael DeVries**
Department of Biology
University of Copenhagen, Denmark
rachael.devries@bio.ku.dk

**Marie Lisandra Zepeda Mendoza**
Department of Machine Intelligence
Novo Nordisk Research Center Oxford Ltd.,
United Kingdom

**Ole Winther**
DTU Compute & Department of Biology
Technical University of Denmark
University of Copenhagen, Denmark

## ABSTRACT

When working with electronic health records (EHR), it is critical for deep learning (DL) models to achieve both high performance and explainability. Here we present the Parallel Attention Transformer (PAT), which performs temporal and sensor attention in parallel, is competitive to state-of-the-art models in EHR classification, and has a uniquely explainable structure. PAT is trained on two EHR datasets, compared to five DL models of different architectures, and its attention weights are used to visualize key sensors and time points. Our results show that PAT is particularly well-suited for healthcare and pharmaceutical applications, which have a strong interest in identifying key features to differentiate patient groups and conditions, and key times for intervention.

## 1 INTRODUCTION

Modelling Electronic Health Record data has gained significant attention in the medical, clinical and pharmaceutical fields (Sidey-Gibbons & Sidey-Gibbons, 2019; Moor et al., 2023; Rajpurkar et al., 2022). EHR data can be generated from multiple time points, allowing for longitudinal analyses (Cascarano et al., 2023). This can lead to the identification of: slow and fast progressors (Geifman et al., 2018), disease prevention and co-morbidity risk prediction (Li et al., 2020), and prediction of disease onset (Zhao et al., 2019), leading to more personalized, effective treatments.

Longitudinal data poses special challenges compared to single time point measurements. The data may contain missing values, irregularly sampled observations, and varying lengths of time intervals between observations (Miotto et al., 2018; Zhang et al., 2021; Yuan et al., 2022). Furthermore, datasets can vary in the number of patients, time points, and type of features (categorical, continuous, images, or text) (Johnson et al., 2016). These issues affect the choice and performance of algorithms (Javidi et al., 2022). Besides accuracy, different approaches have varying degrees of explainability (Subramanian et al., 2020), which is vital for developing drugs as well as for reliably informing the patients and regulatory agencies in healthcare settings (Liu et al., 2020).

Although Transformer models have been successful in various fields (Jumper et al., 2021; Yang et al., 2019), literature shows that they are often outperformed in the context of EHR data, both in accuracy and explainability (Lu & Uddin, 2021; Xiao et al., 2018). In this paper, we implement a Parallel Attention Transformer (PAT), inspired by Axial Attention Ho et al. (2019) and iTransformer Liu et al. (2023). Applied on EHR data, it has performance competitive to state-of-the-art (SOTA) models and allows for the analysis sensor impact on model predictions. PAT has the potential to improve patient outcomes by enabling more accurate and interpretable predictions, thus enhancing the quality of DL applied to healthcare.

# 2 BACKGROUND

## 2.1 PROBLEM FORMULATION

The dataset consists of tuples, $\mathcal{D} := \{(\boldsymbol{S}_1, \boldsymbol{p}_1, \boldsymbol{h}_1, y_1), ..., (\boldsymbol{S}_N, \boldsymbol{p}_N, \boldsymbol{h}_N, y_N)\}$. For each sample $i \in N$, there is a multivariate time series of dynamic features $\boldsymbol{S}_i \in \mathbb{R}^{T_i \times D}$, where $x^t_{i,d} \in \mathbb{R}$ is the reading at time $t$ for feature $d$, and the total number of time points $T_i$ is nonuniform and varies per sample. Time vectors $\boldsymbol{h}_i \in \mathbb{R}^{T_i}$ contain the observation times in hours, static feature vectors $\boldsymbol{p}_i$ are taken once at $t = 0$, and class labels are represented as $y \in \{1, ..., C\}$. For EHR data, dynamic features correspond to lab test results and medical instrument readings, and static features include demographic information, such as age, sex, and height. The objective is to predict $y_i$, given $(\boldsymbol{S}_i, \boldsymbol{h}_i, \boldsymbol{p}_i)$. This is complicated by $\boldsymbol{S}_i$ being sparsely sampled for each $i$, with only a subset of $D$ containing values for each $t$.

## 2.2 TRANSFORMERS FOR TIME SERIES

One approach to this task is using the Transformer by Vaswani et al. (2017) to perform self-attention between the time points for each individual. Briefly, self-attention quantifies relationships between the time point representations, generating weights that represent the importance of each to the model's final prediction. An attention layer transforms $\boldsymbol{S}$ into queries $Q$, keys $K$, and values $V$ with learned weight matrices implemented as linear projections. It then applies an attention mechanism, such as the scaled dot-product $\text{Attention}(Q, K, V) = \text{softmax}(QK^T/\sqrt{d_k})V$, where $d_k$ corresponds to the dimension of each key. The full model will first generate an embedding for each $t$ from all corresponding dynamic features $(x^t_{i,1}, ..., x^t_{i,D})$. The embeddings are then attended to, reduced to a single representation via a pooling layer, and fed to a classification head to predict $y_i$. The Transformer has been implemented in various baseline experiments for EHR classification, but is often outperformed by other model types (Zhang et al., 2021; Xu et al., 2023).

Transformers have been used to explore multidimensional data from novel perspectives, resulting in strong performance across several applications. In particular, Axial Attention by Ho et al. (2019) was designed for image/video prediction and generation, performing self-attention across multiple data dimensions using a shared pixel embedding. The SeFT model (Horn et al., 2020), like Axial Attention along the time dimension, applied self-attention to each sensor separately and used the time of the observation in the Transformer positional embedding function. The iTransformer (Liu et al., 2023), performed time series forecasting on dense datasets by attending to features rather than time points. This resulted in a SOTA model for densely observed time series data that improves with increasing lookback windows, which has previously been seen as a challenge with Transformers (Zeng et al., 2023; Das et al., 2023; Zhang & Yan, 2022; Ekambaram et al., 2023). For longitudinal datasets, this is a desirable quality.

# 3 METHODOLOGY

## 3.1 PARALLEL ATTENTION TRANSFORMER (PAT)

PAT draws from Axial Attention and iTransformer to learn from sparse irregular data (Figure 1). iTransformer's strategy of attending to sensor representations can retain information from early time points within each sensor embedding, and it is also possible to explore the impact of sensors on model decisions by analyzing attention weights. However, in comparison to the more dense and uniform datasets used in the development and testing of iTransformer (Wu et al., 2021; Lai et al., 2018; Liu et al., 2022), the model showed worse performance on sparse irregular EHR data, with the Transformer outperforming it (Experiments 4). By performing attention over both time points and sensors, PAT retains the ability to analyze sensor-model relationships while obtaining a classification performance that is better than Transformer and iTransformer.

Axial attention analyzes multiple input dimensions with a shared embedding, generated from the channel (color) dimension (Ho et al., 2019). In an image or video, color is agnostic to dimension, as the same channels are used to generate pixels in all rows and columns. For EHR data, each dynamic sensor can be represented by a unique distribution, and the relationship of an observation to other

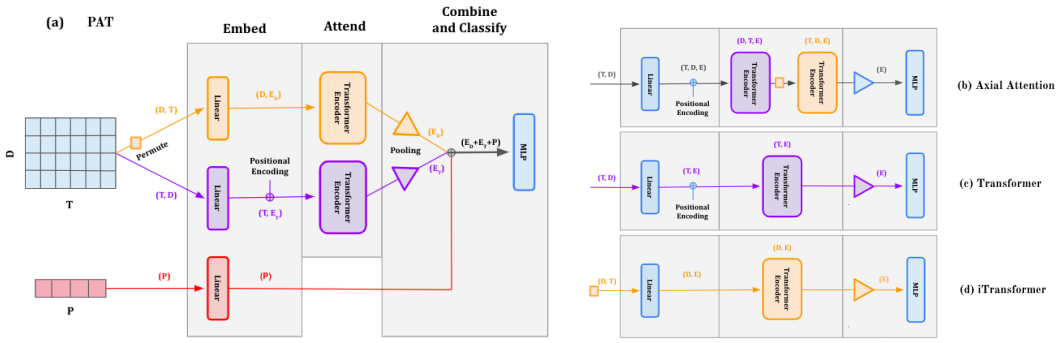



Figure 1: a) PAT architecture, which embeds, attends to, and classifies EHR data. The data's shape throughout the forward pass is shown by tuples containing time $T$ and dynamic feature $D$ dimensions, with $E$ and $\boldsymbol{p}$ representing embeddings and static vectors. Orange follows feature attention and purple follows temporal attention, and the attention operation always works on the second-to-last dimension. We also show architectures of b) Axial Attention as implemented by Lucidrains (2020), c) Transformer, and d) iTransformer. Input data and static features $\boldsymbol{p}$ handling for b), c), and d) are omitted for simplicity, as they are equivalent to PAT.

readings from the same sensor (across $t$) and to different sensors (across $d$) is more complicated. To retain the complexity of these relationships, PAT generates unique embeddings for each axis.

PAT has the following components: embed, attend, and combine/classify (Fig. 2a). The first two parts are applied separately to each axis, resulting in unique "tracks" for the time and the dynamic feature dimensions. We first permute $\boldsymbol{S}$ on the feature track to embed the correct dimension, and append binary mask vectors on both tracks to indicate the presence or absence of observations. The embeddings are then generated using a linear layer. A time encoding is added on the time track calculated from observation times in vectors $\boldsymbol{h}$ (Horn et al., 2020) using the standard positional encoding recipe (Vaswani et al., 2017). Sensor identity embeddings were not used for the dynamic feature track, as the relative position of sensors in the data is always the same (Liu et al., 2023).

Next, attention is performed across embeddings with a Transformer encoder block, which includes attention heads, normalization layers and a feed-forward network. At the final stage, the weighted embeddings are pooled into a single embedding for each dimension, concatenated from each track, and fed to a two-layer classification head. PAT also considers the static features, learning a separate embedding from them that is also concatenated prior to classification. PAT is trained to minimize the cross entropy between mortality predictions and the true outcomes, $L = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(1 - p(y_i))$, with $p(y_i)$ representing the predictions. Further details about PAT hyperparameters, which were selected by testing for optimal performance, can be found in Appendix Table 3.

## 4 Experiments

### 4.1 Mortality Prediction

PAT was trained and tested on two publicly-available EHR datasets for mortality prediction from Intensive Care Unit (ICU) stays, the PhysioNew/CinC Challenge 2012 (P12) Goldberger et al. (2000) and the Medical Information Mart for Intensive Care Database (MIMIC-III) Johnson et al. (2016). Further information about the datasets and preprocessing can be found in Appendix Section A.2. Both datasets were randomly split into 5 iterations of train/test/val groups with a ratio of 8/1/1, on each of which the evaluated models were separately trained and tested. The average and standard deviations of the AUPRC and AUROC can be seen in Table 1.

Table 1: Mortality prediction performance on P12 and MIMIC-III, averaged across 5 data splits.

| | P12 | | MIMIC-III | |
|---|---|---|---|---|
| | **AUPRC** ↑ | **AUROC** ↑ | **AUPRC** ↑ | **AUROC** ↑ |
| SeFT | 53.97 +/- 1.70 | 85.66 +/- 1.04 | **54.01 +/- 3.58** | 84.99 +/- 1.18 |
| GRU-D | **56.96 +/- 2.66** | **87.39 +/- 0.36** | 52.38 +/- 3.12 | 85.29 +/- 0.69 |
| IP-Nets | 55.18 +/- 1.99 | 86.39 +/- 0.35 | 53.29 +/- 2.00 | 85.12 +/- 0.61 |
| Transformer | 52.58 +/- 2.05 | 85.89 +/- 0.68 | 49.75 +/- 1.82 | 84.22 +/- 0.88 |
| iTransformer | 35.15 +/- 1.57 | 74.11 +/- 1.98 | 32.20 +/- 3.03 | 71.40 +/- 1.99 |
| PAT | 54.21 +/- 2.64 | 86.21 +/- 1.15 | 52.00 +/- 1.78 | **85.33 +/- 0.52** |

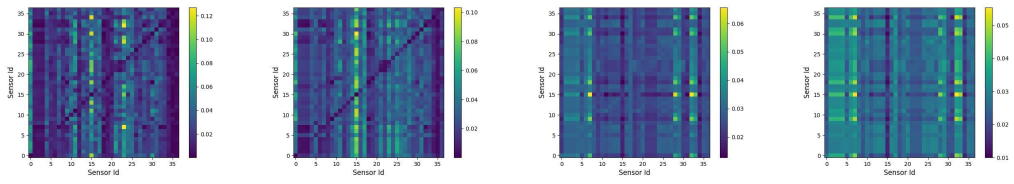
The evaluated models[1], chosen for their high performance on mortality or other time series prediction tasks, were: Transformer Vaswani et al. (2017), iTransformer Liu et al. (2023), SeFT Horn et al. (2020), GRU-D Che et al. (2018), and IP-Nets Shukla & Marlin (2019), each of which were trained and tested using their published hyperparameters (Appendix Table 3).

Transformer and iTransformer were also tested in order to compare performance when attending to the time or sensor axis alone. Results show that performing attention on both axes via PAT is an improvement over the single-axis attention in Transformer and iTransformer (Table 1). Furthermore, PAT results in AUPRC and AUROC values that are competitive to the tested models, although the top model for each data set and metric varied.

### 4.2 Attention Weights

Heatmaps of the average sensor attention weights for members of each label class $C = \{0, 1\}$ (where $C = 1$ is positive for mortality) were generated from PAT and iTransformer after training on P12 (Fig. 2). Average time attention weight heatmaps can be found in Appendix Figure 3.

Sensor attention in iTransformer is more uniformly distributed across dynamic features, while PAT shows high attention weights on a subset of features. Additionally, PAT identifies specific sensors contributing the most to the classification of each class, with $C = 1$ having a large weight on sensor 23 (non-invasive diastolic arterial blood pressure) (Appendix Table 2), while $C = 0$ does not. This could suggest better discrimination capabilities than iTransformer, which shows no such clear difference between classes. While it is not possible to draw conclusions on the clinical relevance of these findings without further study, these results demonstrate that PAT's architecture can provide clear model explanations.




(a) PAT $C = 1$ (b) PAT $C = 0$ (c) iTransformer $C = 1$ (d) iTransformer $C = 0$

Figure 2: P12 average sensor attention weights after Softmax normalization. Includes PAT averages for all a) positive and b) negative mortality class samples, and iTransformer averages for all c) positive and d) negative class samples. Attention weights sum to 1 across rows.

[1]RAINDROP by Zhang et al. (2021) was also tested, but not reported as parameter optimization is needed. Further details in Appendix A.3

## 5 DISCUSSION

Understanding model predictions on longitudinal clinical data is of high priority in the pharmaceutical and healthcare industries. However, many EHR classification models do not provide insights on both time points and sensors. This capability is crucial in differentiating one group of patients from another, and determining the best time for medical intervention. PAT achieves these goals by attending to both temporal and dynamic feature embeddings, and performs competitively with SOTA models on P12 and MIMIC-III.

It is important not only to be explainable, but to have explanations that are intuitive and discriminative. To this end, we show that PAT is able to identify specific features characterizing each class when visualizing the average attention weights from P12. While further work is needed to validate and contextualize the clinical relevance of these weights (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019), PAT's architecture shows great promise in further understanding the impact of times and sensors on patient health. Future studies may uncover new explanations through additional techniques, such as gradient based attributions.

## REFERENCES

Anna Cascarano, Jordi Mur-Petit, Jerónimo Hernández-González, Marina Camacho, Nina de Toro Eadie, Polyxeni Gkontra, Marc Chadeau-Hyam, Jordi Vitrià, and Karim Lekadir. Machine and deep learning for longitudinal biomedical data: a review of methods and applications. *Artificial Intelligence Review*, 56(Suppl 2):1711–1771, 2023.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1):6085, 2018.

Abhimanyu Das, Weihao Kong, Andrew Leach, Rajat Sen, and Rose Yu. Long-term forecasting with tide: Time-series dense encoder. *arXiv preprint arXiv:2304.08424*, 2023.

Vijay Ekambaram, Arindam Jati, Nam Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. Tsmixer: Lightweight mlp-mixer model for multivariate time series forecasting. *arXiv preprint arXiv:2306.09364*, 2023.

Nophar Geifman, Richard E Kennedy, Lon S Schneider, Iain Buchan, and Roberta Diaz Brinton. Data-driven identification of endophenotypes of alzheimer's disease progression: implications for clinical trials and therapeutic interventions. *Alzheimer's research & therapy*, 10(1):1–7, 2018.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *CoRR*, abs/1912.12180, 2019. URL http://arxiv.org/abs/1912.12180.

Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten M. Borgwardt. Set functions for time series. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4353–4363. PMLR, 13–18 Jul 2020.

Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357.

Hamed Javidi, Arshiya Mariam, Gholamreza Khademi, Emily C Zabor, Ran Zhao, Tomas Radivoyevitch, and Daniel M Rotroff. Identification of robust deep neural network models of longitudinal clinical measurements. *NPJ Digital Medicine*, 5(1):106, 2022.

Alistair EW 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):1–9, 2016.

John M. Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Zídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andy Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David A. Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583 – 589, 2021. URL https://api.semanticscholar.org/CorpusID:235959867.

Aleksander Krasowski, Joachim Krois, Adelheid Kuhlmey, Hendrik Meyer-Lueckel, and Falk Schwendicke. Predicting mortality in the very old: a machine learning analysis on claims data. *Scientific Reports*, 12, 10 2022. doi: 10.1038/s41598-022-21373-3.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pp. 95–104, 2018.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):7155, 2020.

Minhao Liu, Ailing Zeng, Muxi Chen, Zhijian Xu, Qiuxia Lai, Lingna Ma, and Qiang Xu. Scinet: Time series modeling and forecasting with sample convolution and interaction. *Advances in Neural Information Processing Systems*, 35:5816–5828, 2022.

Xiaoxuan Liu, Samantha Cruz Rivera, David Moher, Melanie J Calvert, Alastair K Denniston, Hutan Ashrafian, Andrew L Beam, An-Wen Chan, Gary S Collins, Ara DarziJonathan J Deeks, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the consort-ai extension. *The Lancet Digital Health*, 2(10):e537–e548, 2020.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*, 2023.

Haohui Lu and Shahadat Uddin. A weighted patient network-based framework for predicting chronic diseases using graph neural networks. *Scientific reports*, 11(1):22607, 2021.

Lucidrains. Implementation of axial attention - attending to multi-dimensional data efficiently, 2020. URL https://github.com/lucidrains/axial-attention.

Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. Ai in health and medicine. *Nature medicine*, 28(1):31–38, 2022.

Satya Narayan Shukla and Benjamin Marlin. Interpolation-prediction networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=r1efr3C9Ym.

Jenni AM Sidey-Gibbons and Chris J Sidey-Gibbons. Machine learning in medicine: a practical introduction. *BMC medical research methodology*, 19:1–18, 2019.

Indhupriya Subramanian, Srikant Verma, Shiva Kumar, Abhay Jere, and Krishanpal Anamika. Multi-omics data integration, interpretation, and its application. *Bioinformatics and biology insights*, 14:1177932219899051, 2020.

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.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1002. URL https://aclanthology.org/D19-1002.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems*, 34:22419–22430, 2021.

Cao Xiao, Edward Choi, and Jimeng Sun. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 25(10):1419–1428, 2018.

Yanbo Xu, Shangqing Xu, Manav Ramprassad, Alexey Tumanov, and Chao Zhang. Transehr: Self-supervised transformer for clinical time series data. In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pp. 623–635, 2023.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. *XLNet: generalized autoregressive pretraining for language understanding*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Xiaohan Yuan, Shuyu Chen, Chuan Sun, and Lu Yuwen. A novel early diagnostic framework for chronic diseases with class imbalance. *Scientific Reports*, 12(1):8614, 2022.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023.

Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, and Marinka Zitnik. Graph-guided network for irregularly sampled multivariate time series. *arXiv preprint arXiv:2110.05357*, 2021.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The Eleventh International Conference on Learning Representations*, 2022.

Juan Zhao, QiPing Feng, Patrick Wu, Roxana A Lupu, Russell A Wilke, Quinn S Wells, Joshua C Denny, and Wei-Qi Wei. Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction. *Scientific reports*, 9(1):717, 2019.

# A APPENDIX

## A.1 PAT ARCHITECTURE AND HYPERPARAMETER SELECTION

In addition to the model structure outlined in 3.1, there were several other components implementated in PAT. This includes masking out all missing time points for the temporal attention track, and utilizing the same mask in the temporal pooling layer. Because the number of time points $T$ varies between samples, $T_i$ was zero-padded to the largest $T$ in each dataset.

It should also be noted that while sensor identity embeddings were not required for this analysis, dynamic feature identity handling would be necessary when leaving out or have to generalize to new dynamic features.

Several iterations of hyperparameters were tested. Using early stopping, the model with the best AUROC on the validation set was selected. The tested parameters included learning rate, pooling layer, and number of heads and layers (Table 3). Due to time constraints, testing for batch and embedding sizes was not completed, but we aim to do so in the future.

## A.2 DATASETS AND PREPROCESSING

P12: The PhysioNet/Computing in Cardiology Challenge 2012 is a publicly-available EHR dataset containing sensor readings and lab results for de-identified ICU stays (Goldberger et al., 2000). We used the preprocessed dataset as described in Horn et al. (2020), which included one-hot encoding all passive sensor readings and normalizing active sensor readings by mean and standard deviation. After preprocessing, the dataset contained 11,988 samples, with 37 dynamic features and 8 static features over a maximum of 215 time points. The specific features and their corresponding sensor IDs as referenced in Figure 2 can be found in Table 2. Data split indices were originally generated by Zhang et al. (2021).

Table 2: P12 sensor IDs and corresponding sensors. Further detail on individual sensors can be found in Goldberger et al. (2000).

| **ID** | **Sensor** | **ID** | **Sensor** |
|---|---|---|---|
| 1 | Weight | 20 | MechVent |
| 2 | ALP | 21 | Mg |
| 3 | ALT | 22 | NIDiasABP |
| 4 | AST | 23 | NIMAP |
| 5 | Albumin | 24 | NISysABP |
| 6 | BUN | 25 | Na |
| 7 | Bilirubin | 26 | PaCO2 |
| 8 | Cholesterol | 27 | PaO2 |
| 9 | Creatinine | 28 | Platelets |
| 10 | DiasABP | 29 | RespRate |
| 11 | FiO2 | 30 | SaO2 |
| 12 | GCS | 31 | SysABP |
| 13 | Glucose | 32 | Temp |
| 14 | HCO3 | 33 | TroponinI |
| 15 | HCT | 34 | TroponinT |
| 16 | HR | 35 | Urine |
| 17 | K | 36 | WBC |
| 18 | Lactate | 37 | pH |
| 19 | MAP | | |

MIMIC-III: The Medical Information Mart for Intensive Care (MIMIC-III Clinical Database v.1.4) is another ICU stay database containing records from 53,423 patients, including demographics, vital sign measurements, and laboratory test results (Johnson et al., 2016). It is publicly available to researchers after completing a required training and data use agreement. Similarly to P12, we used

the dataset as processed by Horn et al. (2020), and split indices were obtained from Raindrop Zhang et al. (2021).

Both datasets are highly imbalanced for positive/negative class labels, with more negative labels present (for example, P12 was 14% positive). To address this imbalance, each epoch was constructed with all negative training samples and an equal number of positive samples, resampled from the training group up to 3 times.

### A.3 Baseline Comparison

All baseline models were chosen for their SOTA performance, and to represent several of the architectures types that exist for time series classification. Hyperparameters were chosen based on original model publications and implementations, and can be found in Table 3.

Created by Che et al. (2018), GRU-D is a variant of the Gated Recurrent Unit (GRU) model, which adds a decay mechanism to balance memory retention. It also feeds the model a mask to inform about present/missing feature values (a strategy we have also incorporated into PAT), and includes time intervals to the closest preceding measurement.

Set Functions for Time Series, or SeFT, was designed by Horn et al. (2020) to be time and memory-efficient by representing inputs as tuples containing the time point, dynamic feature, and value. This also makes it easy to query the relationship between single values and model predictions.

Interpolation Prediction Networks (IP-Nets) was designed by Shukla & Marlin (2019) and contains two components. The interpolation component generates interpolants for each dimension and time point in a time series, which is given as input to the prediction component to predict $y$. In the original publication, a GRU network is used as the prediction component.

As stated previously, the classic Transformer by Vaswani et al. (2017) and iTransformer by Liu et al. (2023) were also implemented and tested, containing the same masking components as PAT in order to be as comparable as possible.

There are other models that perform time-series classification on EHR datasets. This particularly includes graph model Raindrop by Zhang et al. (2021), which uses multi-dimensional attention for sensors and time points. For this study, Raindrop was also trained and tested for mortality prediction on P12 and MIMIC-III (resulting in, for example, an AUPRC and AUROC of 40.07 and 80.20 for MIMIC-III). However, there are no published hyperparameters for mortality prediction on either of these two datasets, and because of resource constraints, no parameter searches were performed to optimize Raindrop performance. We highlight this as a potential baseline comparison in future work.

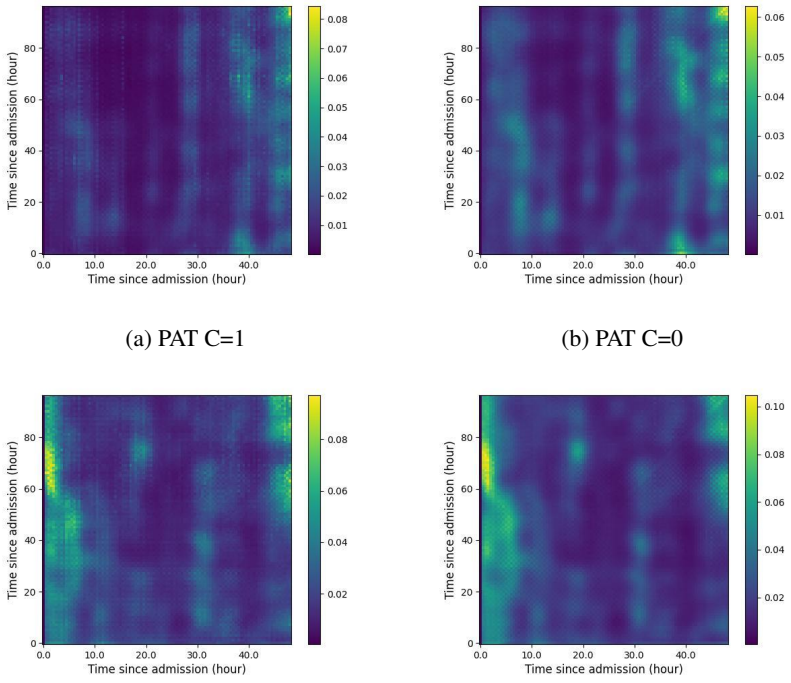
### A.4 Attention Weights

In addition to the average sensor attention weights, the average time attention weights were calculated and used to generate heatmaps from the best-performing PAT and Transformer p12 classification models (Fig. 3). The attention weights sum to 1 across each row, and due to irregularities between samples, time points were binned by hour.

When analyzing the time axis, it is interesting to note in the PAT results that the most strongly attended time points for classification are the latest ones (closest to the diagnosis) for the class 1 group, while class 0 patients are already distinguished from much earlier time points. In contrast, the Transformer time attention provides fewer insights, as weights are very similar between the two groups. Notably, Transformer pays the most attention to the earliest time points, which is contrary to evidence that a patient's phenotype is the clearest closest to the diagnosis (the outcome) (Krasowski et al., 2022). In other words, the measurements of the sensors close to the endpoint, are expected to be the most informative for the classification.

Table 3: Hyperparameters used for baseline model comparison.

| | **P12** | **MIMIC-III** |
|---|---|---|
| SeFT | attn_dropout: 0.5, phi_dropout: 0.2, rho_dropout: 0.0, heads: 4, lr: 0.00081, phi_layers: 4, phi_width: 128, psi_layers: 2, psi_width: 64, psi_latent_width: 128, dot_prod_dim: 128, latent_width: 32, rho_layers: 2, rho_width: 512, batch: 512, max_timescale: 100, positional_dims: 4 | attn_dropout: 0.1, phi_dropout: 0.1, rho_dropout: 0.1, heads: 4, lr: 0.00245, phi_layers: 3, phi_width: 64, psi_layers: 2, psi_width: 64, psi_latent_width: 128, dot_prod_dim: 128, latent_width: 256, rho_layers: 2, rho_width: 512, batch: 512, max_timescale: 1000, positional_dims: 8 |
| GRU-D | dropout:0.5, recurrent_dropout:0.5, n_units:100, lr:0.001, batch:32 | dropout:0.5, recurrent_dropout:0.5, n_units:64, lr:0.001, batch:32 |
| IP-Nets | dropout:0.2, recurrent_dropout:0.2, lr:0.001, batch:256, n_units:100, impute_stepsize:0.25, reconstruct_fraction:0.2 | dropout:0.2, recurrent_dropout:0.2, lr:0.001, batch:256, n_units:100, impute_stepsize:0.25, reconstruct_fraction:0.2 |
| Transformer | dropout:0.2, attn_dropout:0.2, heads:2, layers:1, pool:sum, dims:74, static_dims:12, lr:0.001, batch:128 | dropout:0.4, attn_dropout:0.0, heads:8, layers:2, pool:mean, dims:512, static_dims:5, lr:0.00204, batch:256 |
| iTransformer | dropout:0.2, heads:1, layers:1, pool:mean, dims:500, batch:64 | dropout:0.2, heads:1, layers:1, pool:mean, dims:500, batch:64 |
| PAT | dropout:0.3, time_heads:2, sensor_heads:1, layers:1, pool:max, dims1: 430 , dims2: 74, batch:64 | dropout:0.2, time_heads:2, sensor_heads:1, layers:1, pool:max, dims1: 500 , dims2: 32, batch:64 |




(a) PAT C=1

(b) PAT C=0

(c) Transformer C=1

(d) Transformer C=0

Figure 3: P12 average time attention weights after Softmax normalization. Includes PAT averages for all a) positive and b) negative mortality class samples, and Transformer averages for all c) positive and d) negative class samples.