APPENDIX

A. PRELIMINARIES

Self-Supervised Learning. Learning effective data representations without any label supervision that are also easily extendable and robust across different downstream tasks has been a long-standing challenge. Most mainstream approaches can be categorized into one of two unsupervised learning classes: generative methods or discriminative methods [1]. Generative methods explicitly learn the underlying distribution of the data [2]. However, they are computationally expensive and sensitive to data volume [3], particularly in the small datasets commonly found in the medical field. Discriminative methods, especially the SSL, have recently attracted attention in data representation, and superior performance has been demonstrated in various research fields [4], [5]. They employ objective functions similar to those used in supervised learning.

This objective function aims to treat each sample within a dataset as a distinct class. The goal is to learn representations that bring similar samples closer together while pushing away dissimilar ones. This is known as **contrastive learning**. Its success relies on combining two elements: (i) *Pretext* and (ii) *Contrastive loss*.

- *Pretext* refers to setting up a label-like task to replace label supervision by exploiting the domain knowledge of data to learn a good representation. The objects can be local patches versus the global semantic meaning of text or image [6], various views of visual features [7], or past sequences versus the future in time series [8].

- *Contrastive loss* learns the similarity by maximizing the mutual information between objects, which is known as performing the *pretext* [9]. In our context, we are interested in the following form of mutual information:

$$I(\boldsymbol{x}, \boldsymbol{z}) := \mathcal{H}(\boldsymbol{z}) - \mathcal{H}(\boldsymbol{z}|\boldsymbol{x}),$$

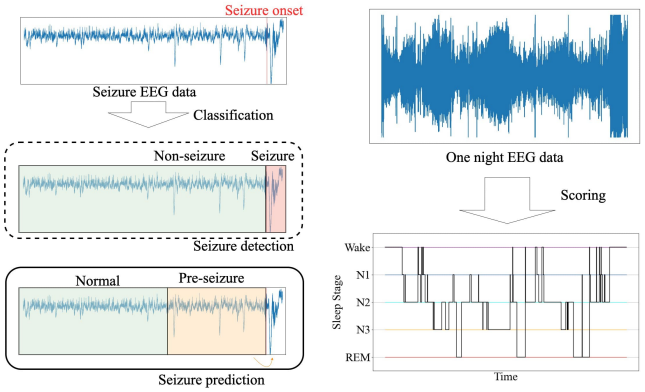
where $\mathcal{H}(\boldsymbol{z}), \mathcal{H}(\boldsymbol{z}|\boldsymbol{x})$ denote the Shannon entropy and conditional entropy, respectively. The effect of mutual information has been well-studied [10]–[12]. In an EEG study, the analysis given by [8] states that $I(\boldsymbol{x}, \boldsymbol{z})$ encourages the confident prediction of eventful brain waveforms while avoiding degenerate solutions.

InfoNCE Mutual Information Estimator. In practice, estimating mutual information is intractable especially when the data are high dimensional and continuous [13], [14]. Instead, researchers usually rely on approximations to evaluate the objective function [15]. The Information Noise Contrastive Estimation (InfoNCE) loss is often used as the objective [16]:

$$\mathcal{L}_{\text{InfoNCE}} := \mathbb{E}_{\mathbf{X}} \left[-\log \frac{\exp\left(\boldsymbol{z} \cdot \boldsymbol{z}^{(+)}/\tau\right)}{\sum_j \exp\left(\boldsymbol{z} \cdot \boldsymbol{z}^{(-)}/\tau\right)} \right], \qquad (1)$$

where $\tau > 0$ is the scaling parameter [17], $\boldsymbol{z}^{(+)}$ and $\boldsymbol{z}^{(-)}$ are positive and negative samples, respectively. The loss should have a lower value when $\boldsymbol{z}$ is similar to its positive variable $\boldsymbol{z}^{(+)}$ and dissimilar to all other *negative* variables $\boldsymbol{z}^{(-)}$. Eq.(1) is equivalent to identifying the target $\boldsymbol{z}$ among a set of $\boldsymbol{z}^{(-)}$ randomly sampled from the dataset.

Pre-Training to Fine-Tuning. Since self-supervised training techniques eliminate the need for manually labeled data, a prevailing paradigm is to utilize models pre-trained via self-supervision for fine-tuning downstream tasks. Conventionally, both the pre-trained and fine-tuned models adhere to an end-to-end mode: full model fine-tuning. Recent studies demonstrate incorporating intermediate layers of knowledge transfer can yield competitive gains. It has been shown that intermediate or partial network fine-tuning is most effective for low-resource target tasks [18]. *SplitSEE* follows this line and aims to find adequate intermediate layers that can well-serve downstream and decrease computational cost.



(a) Seizure prediction task (b) Sleep stage scoring task

Fig. 1: Description of (a) seizure prediction and (b) sleep stage scoring tasks. (a) Seizure prediction task involves predicting epileptic seizures using EEG data designed to alert patients before an event. (b) Sleep stage scoring task categorizes sleep patterns into distinct stages, facilitating an understanding of sleep quality and disorders.

REFERENCES

[1] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3733–3742.

[2] X. Chen, D. P. Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel, "Variational lossy autoencoder," in *International Conference on Learning Representations, ICLR*, 2017.

[3] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning*, 2020, pp. 1597–1607.

[4] M. Yang, Y. Yang, C. Xie, M. Ni, J. Liu, H. Yang, F. Mu, and J. Wang, "Contrastive learning enables rapid mapping to multimodal single-cell atlas of multimillion scale," *Nature Machine Intelligence*, pp. 1–14, 2022.

[5] Y.-H. H. Tsai, T. Li, M. Q. Ma, H. Zhao, K. Zhang, L.-P. Morency, and R. Salakhutdinov, "Conditional contrastive learning with kernel," in *International Conference on Learning Representations*, 2022.

[6] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," in *7th International Conference on Learning Representations, ICLR*, 2019, pp. 1–15.

[7] O. J. Hénaff, A. Srinivas, J. De Fauw, A. Razavi, C. Doersch, S. M. A. Eslami, and A. Van Den Oord, "Data-efficient image recognition with contrastive predictive coding," in *37th International Conference on Machine Learning*, 2020.

[8] D. Cai, J. Chen, Y. Yang, T. Liu, and Y. Li, "Mbrain: A multi-channel self-supervised learning framework for brain signals," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, p. 130–141.

[9] D. Babaev, N. Ovsov, I. Kireev, M. Ivanova, G. Gusev, I. Nazarov, and A. Tuzhilin, "Coles: Contrastive learning for event sequences with self-supervision," in *Proceedings of the 2022 International Conference on Management of Data*, 2022, p. 1190–1199.

[10] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Advances in Neural Information Processing Systems*, 2004, p. 529–536.

[11] F. Keller, E. Müller, and K. Böhm, "Estimating mutual information on data streams," in *Proceedings of the 27th International Conference on Scientific and Statistical Database Management*, 2015, p. 12.

[12] X. Chen and S. Wang, "Efficient approximate algorithms for empirical entropy and mutual information," in *Proceedings of the 2021 International Conference on Management of Data*, 2021, p. 274–286.

[13] J. Liu, Z. Li, Y. Yao, F. Xu, X. Ma, M. Xu, and H. Tong, "Fair representation learning: An alternative to mutual information," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, p. 1088–1097.

[14] D. McAllester and K. Stratos, "Formal limitations on the measurement of mutual information," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 2020, pp. 875–884.

[15] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 531–540.

[16] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2019.

[17] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9726–9735.

[18] C. Poth, J. Pfeiffer, A. Rücklé, and I. Gurevych, "What to pre-train on? Efficient intermediate task selection," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 10 585–10 605.