

Exploring Hidden Patterns: A Priori Class Labels in Contrastive Learning for Phenotype Discovery

Annina Helmy^{1,2,*}Rafael Morand^{1,2,3,*}Markus Schmidt¹,
Claudio L. A. Bassetti¹, Stavroula Mougiakakou^{3,+}, and Athina Tzovara^{1,4,+}

¹ Sleep-Wake Epilepsy Center, Center of Experimental Neurology, Department of Neurology, Inselspital, Bern University Hospital, University of Bern, Switzerland

² Graduate School for Cellular and Biomedical Sciences, University of Bern, Switzerland

³ ARTORG Center for Biomedical Engineering Research, University of Bern, Switzerland

⁴ Institute of Computer Science, University of Bern, Switzerland

*These authors contributed equally to this work. + Co-senior authors.

Abstract. The diagnosis of complex conditions remains challenging when biomarkers are lacking and diagnostic criteria rely on subjective clinical judgment. We propose a novel contrastive clustering framework for phenotype discovery, combining instance- and class-level learning with soft-priors to guide representation learning. Paired with consensus clustering, our method guides the identification of subgroups in heterogeneous populations. We apply this approach to a dataset of electroencephalography and physical activity data from patients with Central Disorders of Hypersomnolence, a clinically ambiguous spectrum that lacks biomarkers and exhibits overlapping symptoms. To validate generalizability, we also test the framework on an open-source dermatological image dataset characterized by distinctly defined diagnostic categories. Our results highlight the potential of our methodology for data-driven discoveries across a range of clinical contexts, whilst incorporating expert clinical knowledge.

Keywords: exploratory data analysis · cluster analysis · health care

1 Introduction

A major challenge in medicine is the number of disorders that share similar clinical features, making accurate diagnosis difficult. Consequently, diagnoses often rely on subjective criteria, which can lead to inadequate treatment. Sleep disorders illustrate this diagnostic challenge particularly well. Central Disorders of Hypersomnolence (CDH), including Narcolepsy Type 1 (NT1), are characterized by excessive daytime sleepiness despite adequate sleep. Although NT1 has biomarkers, the larger spectrum of CDH, often referred to as the narcoleptic borderland (NBL), remains heterogeneous and difficult to classify [1]. This clinical overlap results in imprecise diagnoses, making it difficult to identify reliable biomarkers using supervised learning methods. Recent advances in unsupervised

learning methods allow the discovery of hidden disease patterns and potential biomarkers without requiring prior labeling [2, 3]. However, these approaches neglect valuable clinical knowledge provided by experts. Hence, we propose a contrastive learning framework designed to reveal hidden patterns by incorporating clinical insights *a priori*.

We apply our framework to time series data from polysomnography recordings and physical activity from a CDH cohort and healthy controls. To address the challenges of modeling temporal dynamics, we transform the time series data into images, allowing us to leverage pre-trained computer vision models. To validate generalization, we also tested our framework on curated skin lesions with well-defined diagnostic labels.

2 Methods

We propose a new framework for contrastive learning that utilizes prior knowledge. The pipeline (Fig.1) includes a sampler for contrastive triplets, a pretrained ResNet-18 model as backbone for feature extraction with a non-linear projection head for the contrastive loss, and finally a consensus clustering module.

2.1 Contrastive Learning with Class Guidance

Our framework employs the contrastive triplet loss (Eq. 1) with the goal of extracting meaningful features by grouping similar data and scattering dissimilar data in the representation space as

$$\mathcal{L}_{\text{triplet}} = \max(\text{dist}(a, p) - \text{dist}(a, n) + m, 0) \quad (1)$$

where a is the anchor sample, p and n are the positive and negative samples, respectively, and m is the margin (Fig. 1a). To include prior knowledge in the learning process, we introduce a combined loss function that employs instance and class labels to capture both patient- and disease-level structures weighted by the amplitude of each respective margin (Eq. 2).

$$\mathcal{L}_{\text{combined}} = \left(1 - \frac{m_{\text{I}}}{m_{\text{C}} + m_{\text{I}}}\right) \cdot \mathcal{L}_{\text{instance-triplet}} + \frac{m_{\text{I}}}{m_{\text{C}} + m_{\text{I}}} \cdot \mathcal{L}_{\text{class-triplet}} \quad (2)$$

We fine-tuned the last layers (depending on the dataset) of the pretrained backbone, along with training the added projection head for $k = 60$ folds, each consisting of a training split (60%), a validation split (30%), and a holdout set (10%). We used the ADAM optimizer and a Cosine Annealing scheduler with warm restarts (max. 50 epochs), applied early stopping, and performed hyperparameter tuning for the learning rate, weight decay, and the triplet margins.

2.2 Consensus Clustering

After training, we extracted the representations at the output of the backbone, reduced their dimensionality using PCA, and applied K-Means clustering to obtain cluster labels per fold for $k \in [2, 6]$. These labels were then aggregated across folds to perform consensus clustering.

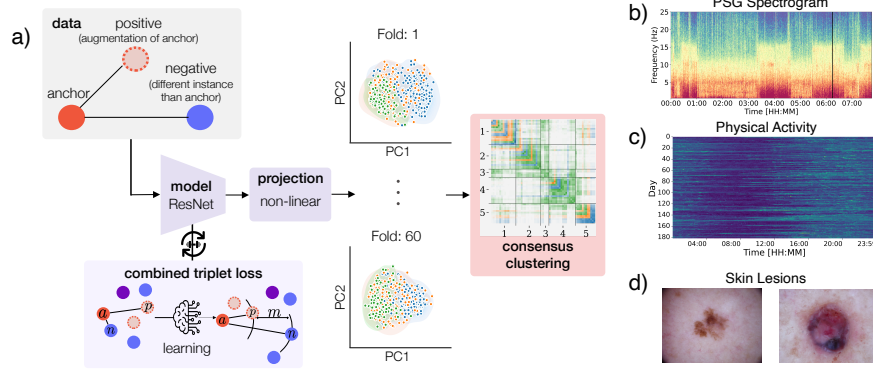


Fig. 1. Overview of model training and data modalities. a) Framework using triplet loss, followed by PCA and consensus clustering. b) Polysomnography spectrogram showing frequency power over time. c) Physical activity data restructured by time of day and day of year. d) Two examples of skin lesions.

2.3 Datasets

We evaluated our framework on three datasets. Two of them —polysomnography and physical activity —were collected as part of the iSPHYNCS study, which investigates the clinical and phenotypical characteristics of CDH [4]. To assess the applicability of our method to other domains, we included the HAM10000 dataset with well-characterized lesion images of skin lesions [5].

The Polysomnography dataset includes spectrograms of two central EEG channels from $n = 180$ individuals (59.4% NBL, 24.4% NT1, 16.2% HC) from overnight recordings. To ensure consistency across the cohort, we standardized the spectrograms to a duration of 6.25 hours, starting at sleep onset (Fig. 1b).

The Physical Activity dataset contains 182 days of continuous heart rate, step count, and calorie measurements with an activity tracker (Fitbit Inspire HR/2/3) for $n = 90$ individuals (61.1% NBL, 18.9% NT1, 20.0% HC). We rearranged the time series to represent images by stacking measurements from the same time of day in a second dimension (Fig. 1c).

The Skin Lesions dataset is a subset ($n = 442$ images) of the HAM10000 dataset [5] (Fig. 1d). We included three classes (melanoma: mel; melanocytic nevi: nv; actinic keratoses: akiec) to reflect the class distribution in the other two datasets.

3 Results

We present results for the most representative value of k for each dataset, selected based on the qualitative separation of diagnostic group (Fig. 2). Smaller values for k resulted in bulks, and larger values resulted in excessively fine-grained partitions. The training converged for all datasets (Fig. 2, loss curves).

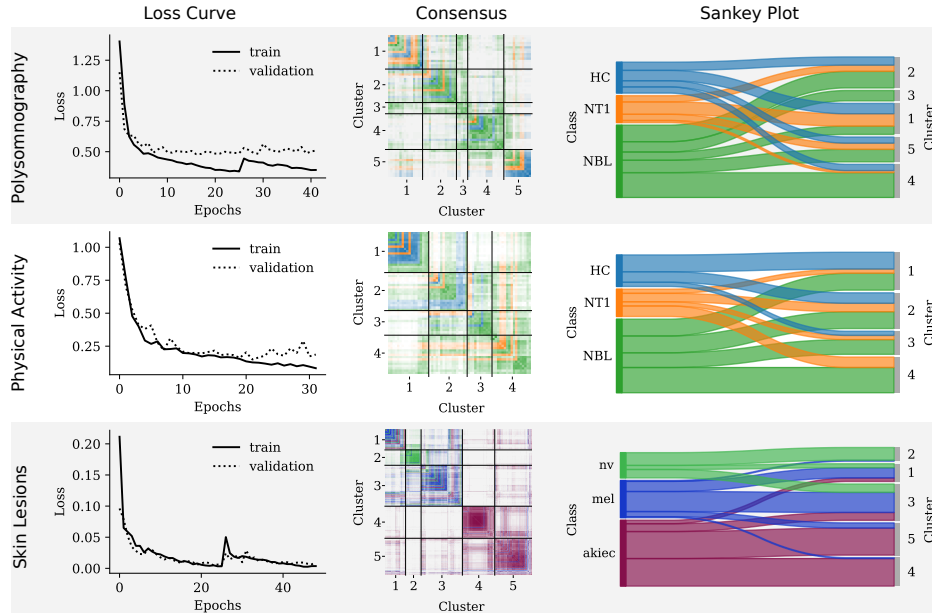


Fig. 2. Learning curves, consensus plots, and Sankey plots for each dataset.

For polysomnography and physical activity, the Sankey plots indicate partial separation between cohorts. In particular, clusters 3 and 4 in the polysomnography dataset predominantly contain patients from the NBL. In physical activity, group 1 contains mainly healthy controls and NBL patients, while group 4 contains a balanced mix of both patient groups. However, the consensus plots reveal local clustering instability in both datasets, suggesting heterogeneity in the underlying phenotypes. The skin lesions dataset shows clearer class separations, confirming the ability of the method to identify subgroups in a well-defined context (Fig. 2). The *akiec* category splits into two distinct clusters with high consensus in clusters 4 and 5, suggesting potential subtypes. In contrast, clusters 1 and 3 show more intermixed groups, indicating possible phenotype overlap.

4 Discussion and conclusion

We presented a new framework for contrastive learning that utilizes prior expert knowledge for phenotype discovery. The skin lesion dataset demonstrated the anticipated behavior, with diagnostic groups forming defined clusters. The *akiec* group divides into separate clusters for $k \geq 4$, possibly indicating subtypes consistent with known clinical variants [6]. The polysomnography and physical activity datasets revealed clusters beyond clinical labels, potentially indicating undetected phenotypes in CDH. In both datasets, we identified clusters consisting predominantly of patients from the NBL, grouped with either HC or patients

with NT1. The next step must involve a rigorous statistical assessment to identify patterns in the data that drive cluster formation for hypothesis generation. Addressing limitations, we note that the sample sizes were relatively small compared to state-of-the-art machine learning methods. Larger, more diverse datasets could improve generalization, stability, and feature discovery. In addition, ResNet-18 might be too complex for these sparse datasets; smaller pretrained models with full tuning of the entire model may be more effective.

In conclusion, contrastive learning combined with the integration of clinical expertise can uncover latent structures in clinical data. This method enables a new approach to hypothesis generation by leveraging domain knowledge while retaining exploratory capacity. We intend to extend the framework towards explainability through network visualizations and automated biomarker identification.

Acknowledgments. This project is financially supported by two Swiss National Science Foundation project grants (320030_185362 and 32003B_215721). We express our gratitude to the participants and consortium members of the iSPHYNCS study.

Disclosure of Interests. The authors have no competing interests to declare.

Ethical Approval The study protocol of the iSPHYNCS study received ethical approval from the relevant authorities in each participating country (Ethics Committees ID: 2019-00788 in Switzerland, 202/2022 in Germany, and NL84710.058.23 in the Netherlands). The trial is registered at ClinicalTrials.gov (ID: NCT04330963). All participants provided informed consent in accordance with the Declaration of Helsinki.

References

- [1] Zeeshan Khan et al. “Central Disorders of Hypersomnolence: Focus on the Narcolepsies and Idiopathic Hypersomnia”. In: *Chest* (2015). DOI: 10.1378/chest.14-1304.
- [2] Taedong Yun et al. “Unsupervised representation learning on high-dimensional clinical data improves genomic discovery and prediction”. In: *Nature Genetics* (2024). DOI: 10.1038/s41588-024-01831-6.
- [3] Haohui Lu et al. “Unsupervised machine learning for disease prediction: a comparative performance analysis using multiple datasets”. In: *Health and Technology* (2024). DOI: 10.1007/s12553-023-00805-8.
- [4] Anelia Dietmann et al. “The Swiss Primary Hypersomnolence and Narcolepsy Cohort study (SPHYNCS): Study protocol for a prospective, multicentre cohort observational study”. In: *Journal of Sleep Research* (2021). DOI: 10.1111/jsr.13296.
- [5] Philipp Tschandl et al. “The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions”. In: *Nature Scientific Data* (2018). DOI: 10.1038/sdata.2018.161.
- [6] Clarissa Prieto Herman Reinehr et al. “Actinic keratoses: review of clinical, dermoscopic, and therapeutic aspects”. In: *Anais Brasileiros de Dermatologia* (2019). DOI: 10.1016/j.abd.2019.10.004.