

PAPER • OPEN ACCESS

A multi-source domain fine-tuning framework for deep generalization performance in physiological time series analysis

To cite this article: Eran Zvuloni *et al* 2026 *Mach. Learn.: Health* **2** 015002

View the [article online](#) for updates and enhancements.

You may also like

- [DUDE: deep unsupervised domain adaptation using variable nEighbors for physiological time series analysis](#)
Jeremy Levy, Noam Ben-Moshe, Uri Shalit et al.
- [Improving generalization performance of electrocardiogram classification models](#)
Hyeongrok Han, Seongjae Park, Seonwoo Min et al.
- [Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020](#)
Erick A Perez Alday, Annie Gu, Amit J Shah et al.

MACHINE LEARNING

Health



PAPER

OPEN ACCESS

RECEIVED
23 August 2025

REVISED
3 November 2025

ACCEPTED FOR PUBLICATION
27 November 2025

PUBLISHED
19 January 2026

Original content from this work may be used under the terms of the [Creative Commons Attribution 4.0 licence](#).

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.



A multi-source domain fine-tuning framework for deep generalization performance in physiological time series analysis

Eran Zvuloni^{1,7,*} , Guido Gagliardi^{2,3,7} , Antônio H Ribeiro⁴ , Antonio Luiz P Ribeiro⁵ , Maarten De Vos^{2,6}  and Joachim A Behar^{1,*} 

¹ Faculty of Biomedical Engineering, Technion-IIT, Haifa, Israel

² Department of Electrical Engineering, KU Leuven, Leuven, Belgium

³ Department of Information Engineering, University of Pisa, Pisa, Italy

⁴ Department of Information Technology, Uppsala University, Uppsala, Sweden

⁵ Department of Internal Medicine, Faculdade de Medicina, Telehealth Center, Hospital das Clínicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

⁶ Department of Development and Regeneration, KU Leuven, Leuven, Belgium

⁷ These authors contributed equally to this work.

* Authors to whom any correspondence should be addressed.

E-mail: eranzvuloni@gmail.com and jbehar@technion.ac.il

Keywords: deep learning, generalization, physiological time series, multi-source domain training

Supplementary material for this article is available [online](#)

Abstract

A major challenge in translating medical AI systems into clinical practice is their limited generalization. In the field of physiological time series analysis, we propose a fine-tuning framework that leverages multiple small annotated datasets from diverse domains to improve out-of-distribution generalization performance (OOD-GP). Through an ablation study, we demonstrate the performance of our framework by evaluating the role of incorporating a greater number of independent datasets for fine-tuning to improve OOD-GP. Our experiments involve thirteen publicly available electrocardiogram and electroencephalogram datasets across four distinct tasks. In addition, we develop a method to measure the alignment of the latent space of target domains. We use this method to interpret our results, suggesting that multi-source domain training facilitates the learning of robust cross-domain features while minimizing learning of shortcut features. To support further research, we provide reproducible source code, establishing a framework and benchmark for studies on OOD-GP (<https://github.com/EranZvuloni/MSD-generalization>).

1. Introduction

Continuous physiological time series are essential in the medical field to monitor vital signs of patients, such as heart rate and blood pressure, as well as to evaluate the function of various physiological systems, including cardiovascular activity and brain function [1–4]. These measurements provide health-care professionals with critical insights, enabling the identification of significant events or changes in a patient's health status and facilitating timely interventions when necessary. Continuous physiological time series are defined as uninterrupted measurements recorded at extremely short intervals, ranging from milliseconds to seconds, between data points [5].

Deep learning (DL) for continuous physiological time series analysis has demonstrated performance on par with or exceeding that of classical machine learning models, which depend on manually crafted features [6–8]. Furthermore, DL has achieved results comparable to or better than expert interpretations of continuous physiological time series in medical tasks [9–14] and has even allowed the execution of medical tasks beyond human capabilities [15–19]. These impressive results are typically evaluated on examples from the same datasets used to develop the DL models, which are assumed to share the same distribution as the development set, i.e. the model's source domain. Unfortunately, a particularly challenging and common scenario arises when models are evaluated on examples from external datasets,

i.e. target domains. When the supports of the source and target domains do not fully overlap, DL models encounter out-of-distribution (OOD) examples that differ due to unseen conditions, populations, or environments. As a result, these models often demonstrate moderate to poor OOD generalization performance (OOD-GP) [17, 20–24], and in practice, even minor deviations from the training domain can lead to incorrect predictions [25].

Current strategies to address domain shifts, such as transfer learning and domain adaptation, often assume the availability of labeled or unlabeled data from the test domain during the training phase [25]. However, access to test domain data, especially labeled data, is often limited in real-world applications. This limitation is common, for example, when developing a DL model to analyze data from portable biosensors. During the development stage, gold-standard annotations are typically used to train the model. However, upon deployment, when fine-tuning is necessary for model maintenance, only the data recorded from the portable biosensors may be available, without the corresponding gold-standard annotations. Collecting additional examinations, which often require medical expertise, is not only impractical, but it can also be economically infeasible, particularly when considering multiple target domains.

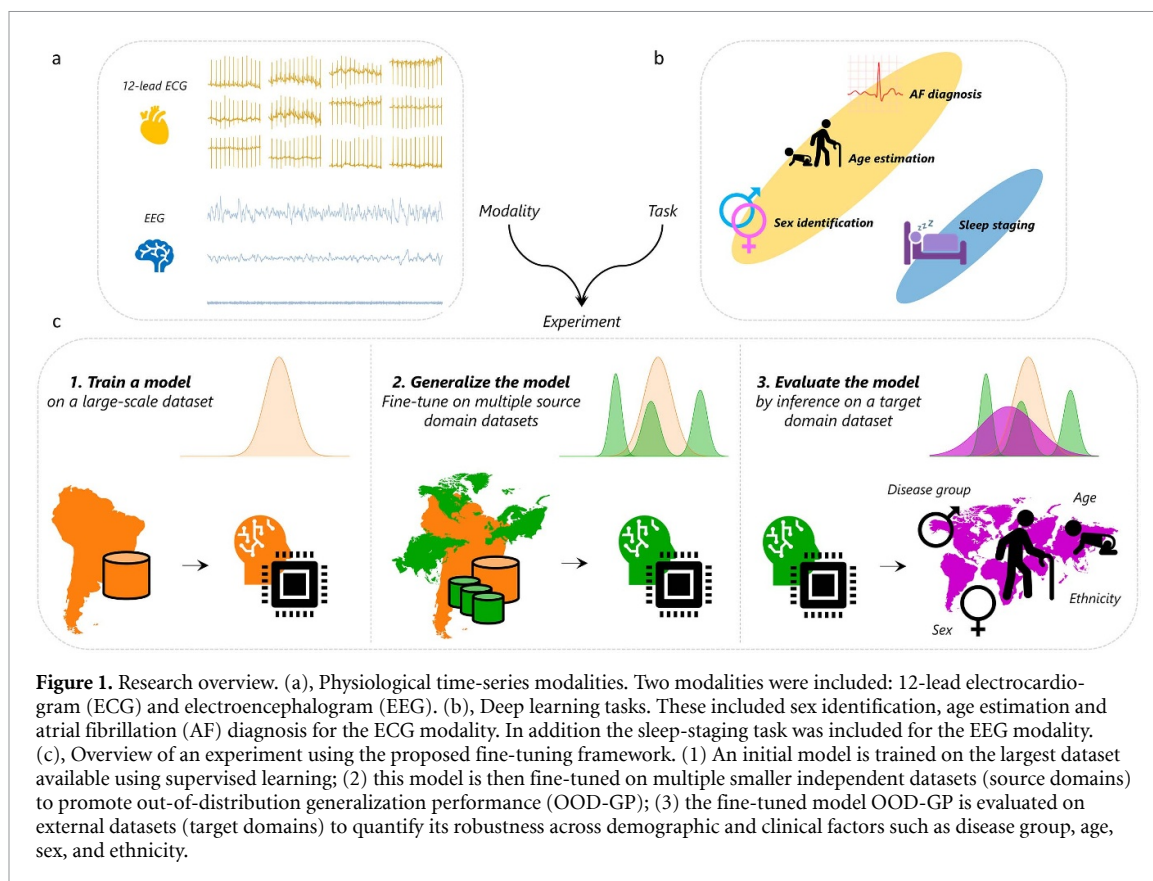
Recently, large-scale foundation models have emerged as a promising direction in physiological signal analysis, leveraging self-supervised pretraining on diverse datasets to improve generalization across tasks and populations. In electrocardiogram (ECG) research, models such as HuBERT-ECG [26] and the work by Song *et al* [27] demonstrated the benefits of large, heterogeneous pretraining, while Gedon *et al* [28] showed that single-domain pretraining alone does not guarantee robust OOD-GP. These studies highlight that data diversity, rather than scale alone, is key for robustness—a challenge this work addresses from a complementary perspective through multi-source supervised fine-tuning.

The increasing availability of open medical repositories of physiological time series datasets, such as PhysioNet [29] and the National Sleep Research Resource (NSRR) [30], has made it easier to access multiple independent datasets for model development. However, a significant challenge remains: the lack of a streamlined approach to integrate these smaller datasets into a unified framework capable of maximizing OOD-GP. This research presents a novel fine-tuning framework that leverages multiple small annotated datasets from diverse domains to improve OOD-GP. To demonstrate the value of this approach, we investigate and quantify how the number of independent datasets affects the OOD-GP of a lightweight DL model in physiological time series analysis. In addition, we interpret these findings by analyzing the latent space of the model. In practice, we conducted experiments using 14 ECG and electroencephalogram (EEG) datasets (13 of which are publicly available) for 4 different tasks (figure 1(a)). Our research focuses on reasonable domain shifts, predominantly arising from factors such as large population variability, limited sampling (e.g. 1000 patients) or inconsistencies in standards and data quality, even when standardized electrode positions are used.

In this research, we make the following contributions: (1) we propose a new model fine-tuning framework with a focus on OOD generalization. This framework is evaluated on 320 different models. (2) We release an open-source family of domain-adapted models for analyzing physiological time series signals, derived from a large pretrained backbone and fine-tuned across multiple cohorts and tasks. These models are optimized for OOD-GP and provide strong, ready-to-use starting points for both research and model development in ECG analysis. (3) We perform ablation studies illustrating that our use of multiple cohorts is essential for obtaining robust performance. (4) We develop a method for measuring the latent space alignment of target domains and use it to interpret our results. (5) Our source code and models are made open access and designed generically to enable the incorporation of additional building blocks, as well as compatibility with other datasets and models.

2. Results

The experiments in this work used a multi-source domain (MSD) fine-tuning approach. To demonstrate its capabilities, an experimental framework was designed, which involved an overall of 14 independent datasets, equivalently denoted domains. Each domain may vary in its population sample, medical facility, and measurement procedure. The experiments were carried out on two types of continuous physiological time series modalities: the 12-lead ECG and the EEG. The ECG and EEG modalities measure the electrical activity of the heart and brain, respectively. The two modalities were used in four DL tasks (figure 1): the tasks of sex identification (binary classification), age estimation (regression) and atrial fibrillation (AF) diagnosis (binary classification) were evaluated on the ECG modality; the sleep staging (multi-class classification) was evaluated on the EEG modality.

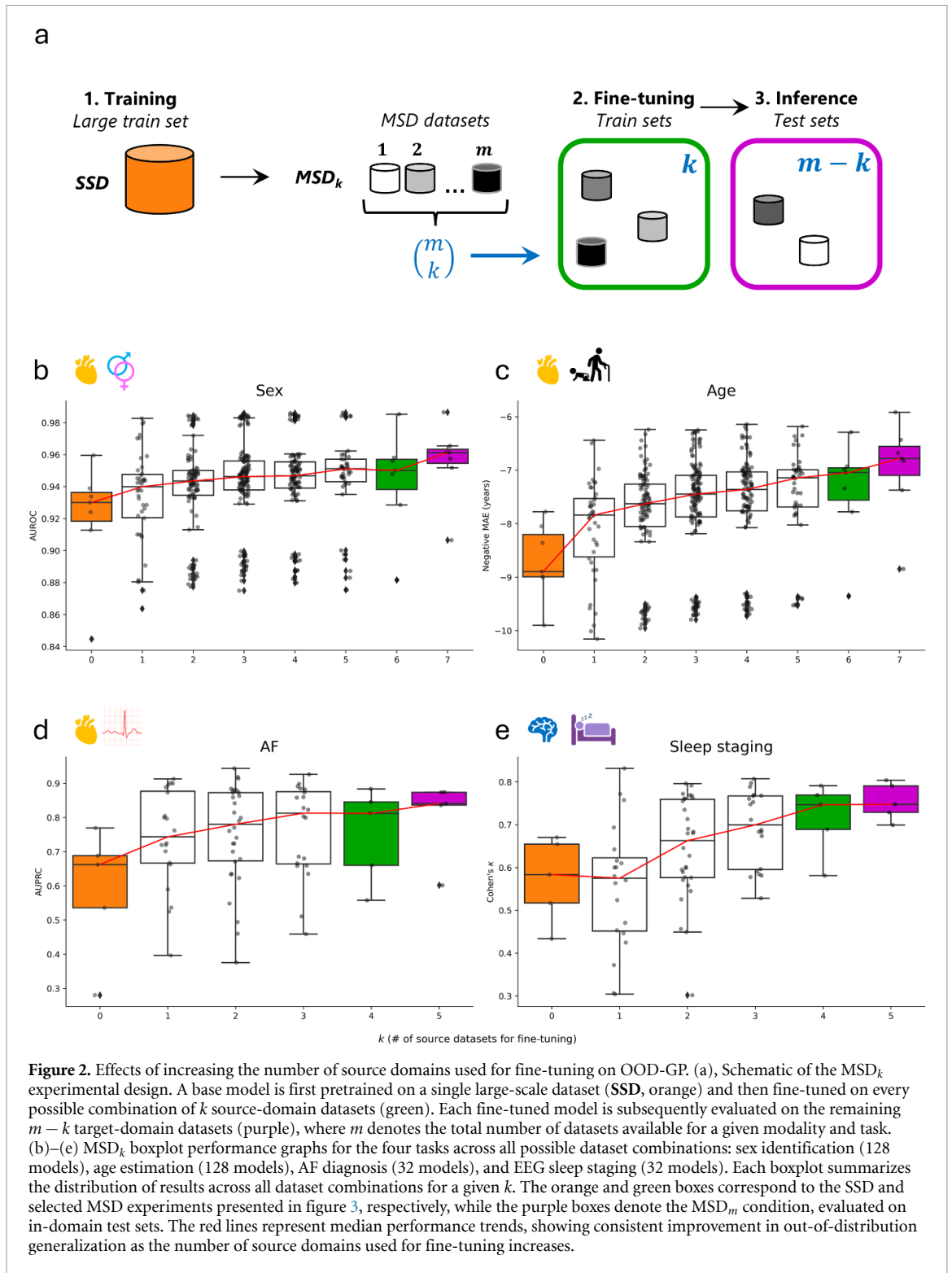


For each modality (ECG or EEG), experiments were performed for specific tasks (figure 1). In these experiments, models pre-trained using a single-source domain (SSD) approach were compared with those fine-tuned using a MSD approach. A main dataset, defined as the largest available dataset, was used alongside m smaller datasets (figure 2(a)). In the SSD approach, the model was trained on the main dataset, and inference was performed on the remaining m datasets. In the MSD approach, models (MSD_k) were obtained by fine-tuning the SSD model on k smaller datasets, and performance was evaluated on the remaining $m - k$ datasets (figure 2(a)). Of particular interest, MSD_{m-1} and MSD_m (figure 3(a)) were treated as models fine-tuned on all-but-one and on all available datasets, respectively. This setup simulated a common scenario where a pretrained model is fine-tuned on smaller datasets to improve OOD-GP for specific tasks. To ensure fair comparisons, datasets were split into train and test sets, and test set examples were excluded from training in all experiments.

For the ECG experiments, the Telehealth Network of Minas Gerais (TNMG) dataset (2.3 M recordings) was used for SSD training, while seven smaller datasets (3000–32 000 recordings) were used for MSD fine-tuning and OOD-GP evaluation for the tasks of sex identification ($m = 7$), age estimation ($m = 7$), and AF diagnosis ($m = 5$). For EEG experiments, the Sleep Heart Health Study (SHHS) dataset (5793 recordings) served as the SSD training dataset, with five smaller datasets (200–2900 recordings) were used for MSD fine-tuning. Datasets details are provided in tables 1 and 2.

2.1. Effects of increasing the number of source domains on OOD-GP

For the two modalities and overall four tasks, it was observed that OOD-GP (reported on target domains) increased as the number of source domains involved in MSD fine-tuning increased (figures 2(b)–(e), $k \geq 1$). In all ECG tasks (figures 2(b)–(d)), the SSD model ($k = 0$) performed worse than any of the MSD models, while in the EEG sleep staging task (figure 2(e)), a decrease in performance was noted at $k = 1$, with this performance gap closing from $k \geq 2$ onward. The performance drop is due to subject variability from training on MSD domains with limited data and differing conditions. For instance, the MASS dataset has fewer recordings than the SSD dataset (200 vs 5793) and varying subject ages (mean 38 vs 63) and environments (lab vs home). Fine-tuning only on the MASS dataset might lead to overfitting, reducing performance on other MSD datasets. The MSD_m models demonstrated the highest performance across all tasks, representing the potential upper bound achievable when all datasets are utilized as source domains.



2.2. Value of MSD fine-tuning on OOD-GP

In figure 3, a detailed comparison is presented between the SSD and the MSD_{*m*-1} models, with results reported on the test set of all target domains. Across the 24 OOD generalization experiments conducted for the two modalities and four tasks, MSD_{*m*-1} outperformed the SSD approach in 23 out of 24 experiments (96%). In one experiment, sex classification on PTB-XL, the SSD model outperformed MSD_{*m*-1}. With respect to the source domain dataset served to train the SSD models, the MSD_{*m*-1} fine-tuned models showed reduced performance compared to the SSD model itself on the source domain test sets (TNMG and SHHS columns). In addition, the four MSD_{*m*} models, serving as upperbound experiments by utilizing all datasets available for training in each task, performed better or equivalently to all other experiments. Since all *m* datasets are included in the fine-tuning phase, the MSD_{*m*} configuration has

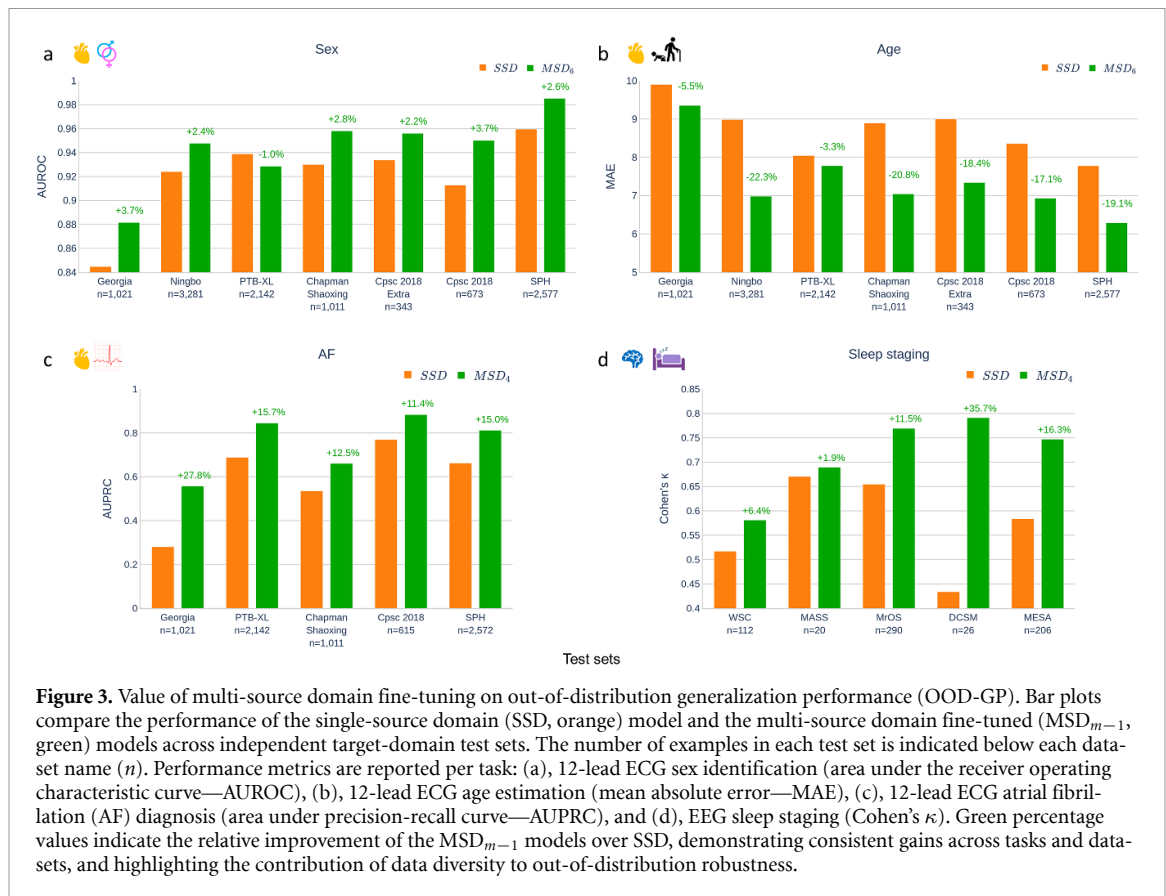


Figure 3. Value of multi-source domain fine-tuning on out-of-distribution generalization performance (OOD-GP). Bar plots compare the performance of the single-source domain (SSD, orange) model and the multi-source domain fine-tuned (MSD_{m-1} , green) models across independent target-domain test sets. The number of examples in each test set is indicated below each dataset name (n). Performance metrics are reported per task: (a), 12-lead ECG sex identification (area under the receiver operating characteristic curve—AUROC), (b), 12-lead ECG age estimation (mean absolute error—MAE), (c), 12-lead ECG atrial fibrillation (AF) diagnosis (area under precision-recall curve—AUPRC), and (d), EEG sleep staging (Cohen's κ). Green percentage values indicate the relative improvement of the MSD_{m-1} models over SSD, demonstrating consistent gains across tasks and datasets, and highlighting the contribution of data diversity to out-of-distribution robustness.

no remaining unseen target-domain test sets ($m - k = 0$). Therefore, the results of MSD_m represent an upper performance bound, reflecting the best achievable outcome when training and evaluation domains fully overlap.

2.3. Latent space analysis of target domains

To interpret the results, the latent space representations in the fine-tuned MSD models were analyzed for each 12-lead ECG task and the EEG sleep staging task. For each model, the Wasserstein distances (W) were measured between the latent space representations of each target domain and the SSD domain (i.e. TNMG or SHHS). The Wasserstein distance serves as a measure of divergence between two probability distributions; a reduction in this measure indicates an increased likelihood that the two distributions are similar and converge towards identity. Accordingly, a smaller distance between the target domain and the SSD domain suggests that the model has successfully acquired a unified joint representation for both the SSD and the target domain. These distances were computed for each k , i.e. number of source domain datasets, used to fine-tune the models. In figure 4(a), a $W(k)$ curve is shown for each task, illustrating latent space alignment as a function of k . Overall, in all of the tasks, there is noticeable reduction in W with the growth of k . Specifically, for the sex identification, age estimation, and AF diagnosis tasks, the alignment was affected in a lower, moderate, and higher manner, respectively. The alignment of the EEG sleep staging task was affected similarly to the sex identification task. Figure 4(b) provides an example of a t-distributed stochastic neighbor embedding (t-SNE) visualization of the latent space distributions for the TNMG and Georgia datasets, focusing on models where the Georgia dataset served as the target domain. This visualization demonstrates how as k increases, the representation of the Georgia distribution becomes more aligned with that of TNMG.

3. Discussion

The primary experimental findings of this research show that MSD fine-tuning is a valuable approach to improve OOD-GP. Experiments were performed for two modalities (ECG and EEG), four tasks and on a total of fourteen datasets (tables 1 and 2). Overall, the experiments involved more than 2400 000 12-lead ECG recordings and more than 110 000 hours of continuous EEG recordings, and the training of a

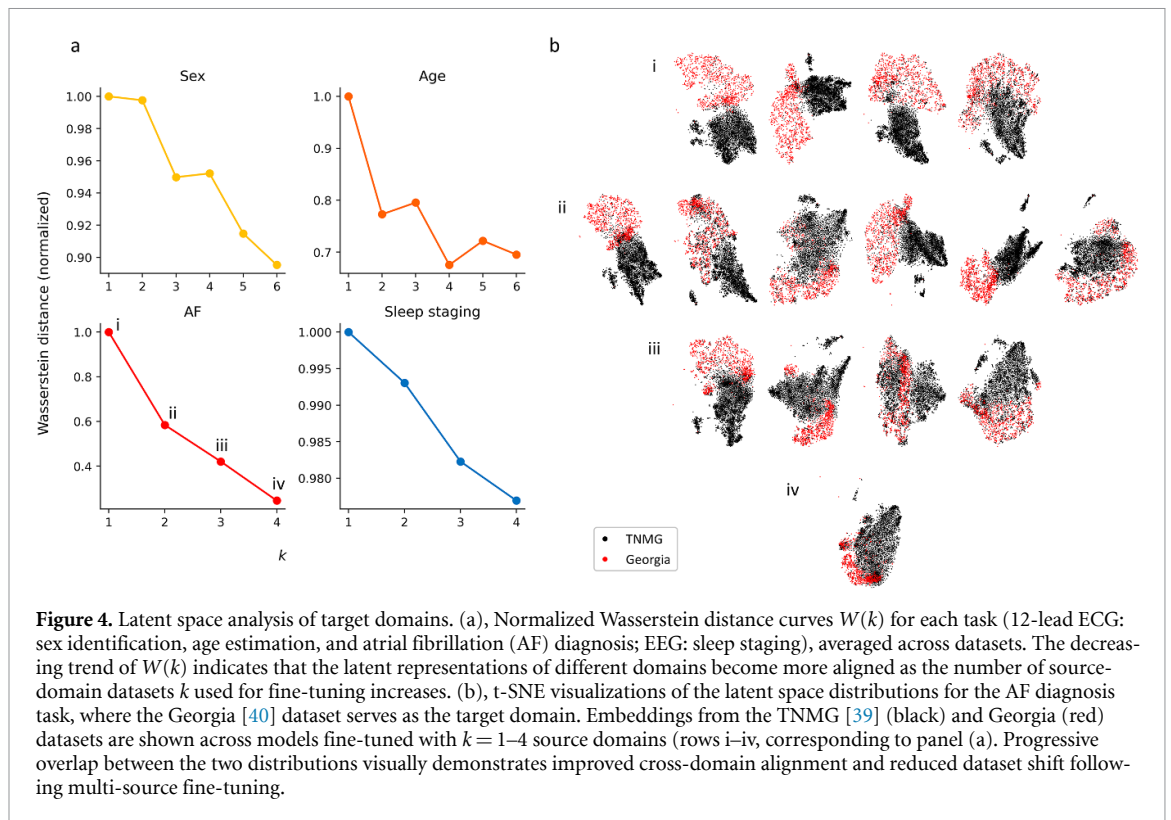


Figure 4. Latent space analysis of target domains. (a), Normalized Wasserstein distance curves $W(k)$ for each task (12-lead ECG: sex identification, age estimation, and atrial fibrillation (AF) diagnosis; EEG: sleep staging), averaged across datasets. The decreasing trend of $W(k)$ indicates that the latent representations of different domains become more aligned as the number of source-domain datasets k used for fine-tuning increases. (b), t-SNE visualizations of the latent space distributions for the AF diagnosis task, where the Georgia [40] dataset serves as the target domain. Embeddings from the TNMG [39] (black) and Georgia (red) datasets are shown across models fine-tuned with $k=1-4$ source domains (rows i-iv, corresponding to panel (a)). Progressive overlap between the two distributions visually demonstrates improved cross-domain alignment and reduced dataset shift following multi-source fine-tuning.

total of 320 models. MSD fine-tuning improved OOD-GP by up to 35% (figure 3). Furthermore, empirical results demonstrated that OOD-GP increases as a function of a growing number of source domains used for fine-tuning (figure 2). These results highlight the value of the MSD fine-tuning approach and provide further encouragement to the scientific community to open datasets from various geography, ethnicity, and medical centers in order to develop fair and generalizable AI models that benefit humanity at the largest.

Notably, in one instance of sex identification on the PTB-XL dataset (figure 3), the SSD model slightly outperformed the MSD_{m-1} configuration by approximately 1%. This deviation represents a single case out of 24 experiments ($\approx 96\%$) and therefore does not affect the overall trend favouring MSD learning. It is hypothesized that this exception arises from the intrinsic nature of the sex classification task, which is comparatively simple and less domain-sensitive. As a result, the features relevant for sex identification were likely already well captured during large-scale pretraining on TNMG, leaving little room for additional benefit from multi-source fine-tuning. Consistent with this, the observed MSD improvement in sex classification was modest (1%–3.7%), whereas more complex and domain-dependent tasks such as age estimation, AF diagnosis, and sleep staging showed substantially larger gains (up to 35.7%).

The latent space analysis offers valuable insights into the nature of the features extracted from the target domain, showing that they become increasingly common across all datasets as the number of source domain datasets used for fine-tuning increases. This suggests that models trained with MSD data are learning features that are more representative of the underlying modality and task, rather than overfitting to the characteristics of a specific dataset. Further explainability analyses, particularly those aimed at identifying specific, well-documented features, could provide additional value and represent a promising direction for future research.

It is also evident that the trends in figure 4(a) differ between tasks. While the AF detection and sleep staging tasks show a monotonic behavior, the sex and age prediction tasks exhibit non-monotonic trends. This may be due to the fact that the labels in the sex and age tasks are ‘clean,’ meaning they are exact and free from human inter-rater variability, whereas tasks like diagnosis or sleep staging are more prone to such variability in annotations. As a result, the sex and age tasks may lead to more similar feature distributions across domains.

One may question whether the performance improvement achieved with MSD_{m-1} over SSD is attributable to the diversity of datasets from different sources or simply the result of a larger number of training examples. To address this, our experimental design utilized the largest dataset available as the main

source domain for training an initial supervised model. Specifically, in the ECG experiments, the TNMG dataset contained two orders of magnitude more examples than the other datasets, and in the EEG experiments the SHHS dataset contained more than double the recordings of the other datasets. This design, combined with our quantitative analysis and t-SNE visualizations of the latent space, strongly supports the conclusion that the performance gains are driven by the inclusion of multiple fine-tuning domains rather than solely by the volume of training data.

Another important consideration is the potential variability in the contribution of each dataset, as certain datasets may negatively affect fine-tuning by shifting the feature distribution due to significant distribution shifts or low-quality labels. This research relied on widely used and well-established datasets (PhysioNet, NSRR), which minimizes this risk. As shown in figure 2, this approach resulted in a generally monotonic improvement in OOD-GP across tasks. Additionally, using a relatively large number of source-domain datasets is expected to mitigate potential adverse effects from any single dataset by promoting a more balanced and robust feature representation. However, exploring methods to select datasets, or even specific recordings within a dataset, to maximize the value of MSD training remains an open question and represents a natural extension of our research.

MSD fine-tuning is one approach to enhancing OOD-GP, and additional techniques could be integrated into the proposed framework to further improve generalization performance. Specifically, in the field of physiological time series, methods such as domain alignment [31] and data augmentation strategies [32, 33] can help reduce domain discrepancies and enhance model robustness. Furthermore, recent foundation models, pre-trained using self-supervised learning (SSL) on large physiological time series datasets [34–36], offer a promising direction. These foundation models, when fine-tuned for specific downstream tasks, have demonstrated improved generalization. The experiments carried out relied on a model architecture that was pretrained using supervised learning on the largest dataset available to us for the given modality and task. In future work, there will be value in evaluating the use of supervised learning pre-trained models against foundation models trained with SSL. SSL involves training models on pretext tasks where labels are either unnecessary or can be autogenerated. This approach leverages vast amounts of unlabeled data to learn versatile feature representations, which can then be fine-tuned for a particular task, post the pre-training phase. Foundation models for physiological time series, such as single and 12-lead ECG, EEG and photoplethysmogram signals, have recently shown promising results [34–36], making this an encouraging direction for further enhancement of the framework presented in this work. Although these prominent models have shown high performance in ECG diagnosis and EEG sleep staging tasks, they did not delve into explaining if an exhaustive usage of data contribute to the formation of model’s latent space representations, and systematically with respect to the diversity of datasets.

Moreover, the use of foundation models in ECG analysis has recently been shown to enhance OOD-GP across various tasks for both modalities. In the studies of HuBERT-ECG [26] and Song *et al* [27] leveraged multiple datasets to pretrain their models using SSL and performed successfully when testing the foundation models on the down-stream tasks. On the other hand, Gedon *et al* [28], who pre-trained only with the full TNMG dataset, demonstrated how the SSD pre-training could not improve performance on the external PTB-XL or the Cpsc datasets. Our research sheds light on why these foundation models pre-trained on diverse datasets may or may not achieve superior OOD-GP. Future directions also include developing strategies for better dataset and recordings selection. Indeed, within the current experimental settings, all recordings from all source domains are included in model training. This may be detrimental since model training possibly includes recordings with low quality signals and datasets with a low level of label curation [37]. Future work could also explore integrating the proposed multi-source fine-tuning framework with large-scale self-supervised or foundation model pretraining to further enhance cross-domain transferability. Extending the approach to additional physiological modalities such as photoplethysmography or electromyography (EMG) would allow evaluating its generality beyond ECG and EEG. Finally, developing standardized multi-domain benchmarks for physiological time-series data could support systematic comparison and reproducibility across studies investigating OOD generalization. Another future direction involves extending the proposed framework to incremental fine-tuning settings, where new datasets become available over time. As observed in figure 2, the performance variance among experiments with identical k values suggests that the order and selection of source domains may influence generalization. Investigating strategies for optimal dataset sequencing or adaptive fine-tuning could therefore further improve model robustness in practical, evolving data environments.

The development and accessibility of open datasets and source code for benchmarking have been pivotal in advancing the field of machine learning. A notable example in computer vision is the ImageNet competition [38], which provided an open-access dataset and pre-trained models that significantly accelerated progress. Our research highlights the importance of utilizing multiple independent open

datasets to train generalizable AI models, particularly in the domain of physiological time series analysis. Additionally, it offers an interpretation of these findings through latent space analysis. To facilitate further research, we provide source code to reproduce all our experiments. This open-source resource serves as a valuable framework for developing and benchmarking algorithmic approaches to address challenges in OOD-GP.

4. Methods

4.1. Problem definition

A supervised classification or regression task is considered, for which a continuous physiological time series is the input. Distribution shifts are assumed to be present in the different datasets, which are treated as different domains, given their origin may vary in population, geographic area or medical practice. A domain is considered as source domain, denoted \mathcal{D}_s , when it is used for developing a model and its examples take part in the DL model training and fine-tuning. Examples and labels are assumed to be i.i.d distributed given by $(x_i^s, y_i^s)_{i=1}^{N_s} \sim \mathcal{D}_s$, where N_s is the number of examples in the source domain. A domain is considered as target domain, denoted \mathcal{D}_t , if it is exclusive for model evaluation, and thus not accessible to the training process. Equivalently, $(x_i^t, y_i^t)_{i=1}^{N_t} \sim \mathcal{D}_t$. The objective of the MSD fine-tuning approach was to build a model that generalizes well on a given target domain \mathcal{D}_t . Each dataset was split into a fixed train, validation and test sets; thus, the same examples were used in the same training and evaluation sets throughout all experiments.

4.2. 12-lead ECG datasets

A total of eight datasets were included and are summarized in table 1. These datasets were taken from three different sources: (1) the TNMG dataset [39], (2) six out of eight datasets (table 1 #2,4–8) were taken from the PhysioNet 2021 Challenge [40, 41], and (3) the Shandong Provincial Hospital (SPH) dataset [42]. The potential PTB (516 recordings) and INCART (72 recordings) were excluded from the Challenge datasets, because of having a low number of recordings. Three common ECG tasks [7, 10, 43, 44] were considered namely sex identification, age estimation and AF diagnosis. AF diagnosis was chosen as a diagnostic task, over other arrhythmias, because labeled data is readily available across most datasets and AF is a very common arrhythmia which is subject to a high research interest [7, 10, 15, 16]. AF labels were available in six out of eight of the ECG datasets (table 1). Several exclusion criteria were applied: patient age was below 18 and above 100, the signal length was less than 7 s [10]), the file was corrupted, or the signal included NaN values. In total, 2.6%, 2.6% and 2.0% of the recordings were excluded for the sex, age and AF tasks, respectively. In order to standardize the input size to the DL model, all the 12-lead ECG datasets were either given or resampled to 500 Hz, and additionally were either padded or trimmed to 10 s length. This resulted in each ECG example being 5000 samples long. The TNMG set split was 90%, 5%, 5% for train, validation and test respectively, whereas for all the other datasets the split was 80%, 10%, 10%.

4.3. EEG datasets

A total of six polysomnography (PSG) datasets were included in the experiments (table 2). Four out of six are available on the NSRR [30] (SHHS [45], MESA [46], MrOS [47] and WSC [48]); the remaining two datasets (MASS [49] and DCSM [50]) are publicly available. All these datasets share overnight PSG recordings from a variety of individuals with various comorbidities (e.g. obstructive sleep apnea). All datasets have reference manually annotated sleep stages. Technicians annotated sleep stages following the rules outlined by the American Academy of Sleep Medicine. Manual sleep staging was expensive and time-consuming but closely linked to sleep disorders and other neurodegenerative diseases, like dementia and Parkinson's [51]. 1.6% of the recordings were discarded due to corrupted or partially unavailable data. Sleep stages annotations were available for each 30 s window of PSG. PSG channels used for manual sleep scoring included the EEG signal and other modalities such as electrooculography (EOG) and EMG. These stages were divided into Wake, N1, N2, N3 (deep sleep) and REM.

4.4. DL for 12-lead ECG analysis

The 12-lead ECG DL model used was based on residual blocks comprising of 1d-convolution layers as originally described in Ribeiro *et al* [10]. The model used was built from a deeper and wider than the original architecture (see supplementary figure 1 for the full elaboration), comprising of larger number of latent space features to allow a more comprehensive late representations. Similar to a previous work [7], the final activation layer was modified according to the task at hand. A sigmoid activation function was used for the binary sex and AF classification tasks and the identify activation function

Table 1. Summary of the 12-lead ECG datasets used in this study, detailing the number of available recordings, demographic distributions, atrial fibrillation (AF) label prevalence, and dataset origin.

#	Dataset	# of recordings after exclusions ^a	Age (median, Q1–Q3)	Sex (%F)	AF prevalence (%)	Origin
1	TNMG [39]	2322 513	54 (41, 67)	60.27	1.8	Brazil
2	Ningbo ^b [40]	32 804	63 (50, 73)	43.45	0	China
3	SPH [42]	25 770	50 (37, 62)	44.64	2.61	China
4	PTB-XL [40]	21 411	62 (50, 72)	47.48	6.77	Europe
5	Georgia [40]	10 206	62 (51, 72)	46.29	5.55	USA
6	Chapman Shaoxing [40]	10 103	62 (51, 72)	43.98	17.62	China
7	Cpsc 2018 [40]	6723	64 (49, 75)	46.14	18.49	China
8	Cpsc 2018 Extra ^b [40]	3412	65 (55, 75)	46.63	0	China

^a The number is given for the age task. Other tasks' numbers vary in less than 1%.

^b The Ningbo and Cpsc 2018 Extra datasets were not used in the diagnosis task due to the low prevalence of AF label.

Table 2. Summary of EEG datasets used in this study, including the number of recordings, demographic characteristics, total number of annotated sleep epochs, and dataset origin.

#	Dataset	# of recordings after exclusions	Age (median, Q1–Q3)	Sex (%F)	Sleep epochs	Origin
1	SHHS [45]	5793	63.0 (55.0, 72.0)	52.36%	5782 715	USA
2	MrOS [47]	2898	76.0 (72.0, 80.0)	0.00%	3624 129	USA
3	MESA [46]	2055	68.0 (62.0, 76.0)	53.58%	2448 749	USA
4	WSC [48]	1113	50.5 (56.4, 62.0)	46%	1015 564	USA
5	DCSM [50]	255	NA	NA	578 939	Denmark
6	MASS [49]	200	38.3 (std: 18.9)	51.50%	228 870	Canada

was used for the age regression task. For the sex and AF binary classification tasks a class-weighted binary cross entropy loss was used. For the age estimation regression task, the mean absolute error (MAE) loss was used [7]. A model was trained with a learning rate scheduler with respect to the validation set loss function score. The learning rate was initialized with $lr = 10^{-3}$ and was reduced by a factor of 10 if there was not improvement for 5 consecutive epochs. If there was no improvement for 7 consecutive epochs, the model was considered having converged and the training stopped. Then, the best model was selected according to the loss score on the validation set. The batch size was 256 in all experiments. There was no additional hyper-parameter tuning.

4.5. DL for EEG analysis

The three-channel version of the SeqSleepNet model [13] was used for sleep staging in this study. SeqSleepNet is a sequence-to-sequence model that takes a sequence of L consecutive sleep epochs as input and classifies one sleep stage for each epoch. It features a hierarchical neural network architecture, with an attention-based LSTM encoder operating at the epoch level and another LSTM-based encoder functioning at the sequence level. The version of SeqSleepNet used in this study is the PyTorch 2 implementation available in the *PhysioEx* Python library [52], taking as input 3 channels, i.e. EEG EOG and EMG, with sequence-length $L = 21$ (see supplementary figure 2 for the full elaboration).

We apply a standard state-of-the-art preprocessing pipeline to all sleep datasets [13, 53, 54], designed to isolate the frequency bands relevant for sleep analysis. Accordingly, all channels are resampled to 100 Hz; EEG and EOG signals are band-pass filtered between [0.3, 40] Hz, and EMG is high-pass filtered at 10 Hz. If wake epochs were overrepresented compared to sleep epochs, they were reduced by cutting the initial and final part of the recording [54]. Recordings were segmented into 30 s windows of three channels. These segments were then transformed into time-frequency images by dividing each epoch into two-second windows with 50% overlap, multiplied with a Hamming window, and transformed to the frequency domain using a 256-point fast Fourier transform. Finally, the amplitude spectrum was log-transformed [13]. The recordings were finally randomly divided into train, validation and test sets with 70%–15%–15% ratio. In all of the EEG datasets there was only one recording for each of the patients.

The model was trained for a sleep-staging task using cross-entropy loss over 20 epochs. In the SSD training experiment, the learning rate was initialized at $lr = 10^{-4}$, while in the fine-tuning experiment, it was set at $lr = 10^{-5}$. A learning rate scheduler, specifically the PyTorch reduce-learning-rate-on-plateau

method, was employed to adjust the learning rate based on validation loss during training. The scheduler was configured with patience of 5 epochs and a factor of 10. A batch size of 32 was used consistently across all experiments. The best model was selected throughout the training process according to the validation loss. The best model was used for the evaluation.

4.6. Latent space analysis

The curves shown in figure 4 illustrate the latent space alignment of the target domains. These curves, denoted as $W(k)$, represent the Wasserstein distance as a function of the number of datasets (k) used during the fine-tuning step, with one curve corresponding to each task. The Wasserstein distance was calculated for the latent space of each fine-tuned model, comparing the target domain datasets with the SSD dataset used for pretraining. Here, the SSD dataset serves as a reference point (origin) for model representations, with the datasets' latent representations measured with respect to it, and evolving as k changes. The derivation of these curves is based on the following mathematical formulation:

For each task, the total number of fine-tuned models is given by $J = \sum_{k=1}^m \binom{m}{k}$, where m is the number of datasets available for the fine-tuning step for all MSD_k experiments. Each dataset is denoted with $d \in \{0, 1, \dots, m\}$, where $d=0$ is the large dataset used to train the model in SSD approach, and $d > 0$ are the fine-tuning datasets in the MSD approach. The latent space representation Z_d^i for each dataset and model $j \in \{1, \dots, J\}$ is given by: $Z_d^j = E^j(X_d)$, where E^j is the encoder of model j , Z_d^j is the latent space representation of the test set examples when each example is represented by a row-stack of the latent space features, and X_d is the stack of test set examples from dataset d . Accordingly, the Wasserstein distance between each dataset representation Z_d^j and its SSD dataset representation Z_0^j (computed with the same E^j) is defined as: $w_d^j = \mathcal{W}(Z_0^j, Z_d^j)$. Thus, for each k , the distances corresponding to all models fine-tuned with k datasets are averaged per target domain dataset and then normalized by the distance corresponding to $k=1$:

$$W_{d \in \mathcal{D}_t}^k = \frac{1}{\sum_{i=j}^J 1_k(j)} \sum_{i=j}^J 1_k(j) \cdot w_{d \in \mathcal{D}_t}^j \quad (1)$$

followed by:

$$\bar{W}_d^k = \frac{W_d^k}{W_d^1} \quad (2)$$

where the indicator function $1_k(j)$ is applied to include only the models fine-tuned with k datasets. Finally, an overall median is computed across datasets, giving the distance as function of k :

$$W(k) = \text{median}_{d \in \{1, 2, \dots, m\}} \bar{W}_d^k. \quad (3)$$

The Wasserstein metric distance was computed via the Python Optimal Transport library [55]. To visualize the latent space, a t-SNE representation was used, computed with the scikit-learn Python library [56].

4.7. Performance measures

For each task, an adequate performance measure was selected. For the 12-lead ECG sex identification binary classification task, the area under receiver operating characteristic curve (AUROC) was used, similar to a previous work [43]. For the 12-lead ECG age estimation task, the MAE metric was used similarly to previous works [7, 43]. For the 12-lead ECG arrhythmia diagnosis task, the area under precision recall curve (AUPRC) was computed, similarly to previous works [7, 10]. For the EEG sleep staging task, Cohen's kappa was selected, similar to a previous work [53, 57].

Data availability statement

The 12-lead ECG datasets: TNMG dataset is partially publicly available via [Zenodo](#). The Physionet 2021 Challenge datasets, including Georgia, PTB-XL, Ningbo, Chapman Shaoxing, Cpsc 2018 and Cpsc 2018 Extra, are available via [PhysioNet.org](#). The SPH dataset is available via [Springer Nature](#). **The EEG datasets:** SHHS, WSC, MrOS and MESA datasets are available via [sleepdata.org](#). MASS is available via [CARSM](#). DCSM is available via [ERDA, University of Copenhagen](#).

Supplementary data available at <https://doi.org/10.1088/3049-477X/ae2ac5/data1>.

Code availability

The source code to reproduce all our experiments is made open source at URL: <https://github.com/EranZvuloni/MSD-generalization>.

Acknowledgments

The Sleep Heart Health Study (SHHS) was supported by National Heart, Lung, and Blood Institute cooperative Agreements U01HL53916 (University of California, Davis), U01HL53931 (New York University), U01HL53934 (University of Minnesota), U01HL53937 and U01HL64360 (Johns Hopkins University), U01HL53938 (University of Arizona), U01HL53940 (University of Washington), U01HL53941 (Boston University), and U01HL63463 (Case Western Reserve University).

The Multi-Ethnic Study of Atherosclerosis (MESA) Sleep Ancillary study was funded by NIH-NHLBI Association of Sleep Disorders with Cardiovascular Health Across Ethnic Groups (R01 HL098433). MESA is supported by NHLBI funded Contracts HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168 and N01-HC-95169 from the National Heart, Lung, and Blood Institute, and by cooperative Agreements UL1-TR-000040, UL1-TR-001079, and UL1-TR-001420 funded by NCATS.

The National Heart, Lung, and Blood Institute provided funding for the ancillary MrOS Sleep Study, ‘Outcomes of Sleep Disorders in Older Men,’ under the following Grant Numbers: R01 HL071194, R01 HL070848, R01 HL070847, R01 HL070842, R01 HL070841, R01 HL070837, R01 HL070838, and R01 HL070839.

The Wisconsin Sleep Cohort (WSC) Study was supported by the U.S. National Institutes of Health, National Heart, Lung, and Blood Institute (R01HL62252), National Institute on Aging (R01AG036838, R01AG058680), and the National Center for Research Resources (1UL1RR025011). The National Sleep Research Resource was supported by the U.S. National Institutes of Health, National Heart Lung and Blood Institute (R24 HL114473, 75N92019R002).

EZ acknowledges The Miriam and Aaron Gutwirth Memorial Fellowship, Technion EVPR Fund: Hittman Family Fund and the Israel Data Science Initiative. This research received funding from the Flemish Government (AI Research Program), FWO PhD SB Grants (1SH4Z24N), HORIZON-HLTH-2022-STAYHLTH: ‘Artificial intelligence-based Parkinson’s disease risk assessment and prognosis (AI-PROGNOSIS),’ under Grant Agreement 101080581 and FWO Project G046925N: ‘Task- and device-agnostic Artificial Intelligence (AI) for EEG analysis.’

Dr Ribeiro is supported in part by CNPq (National Council for Scientific and Technological Development, grants 310790/2021-2, 409604/2022-4, 445011/2023-8, and 408659/2024-6) and FAPEMIG (Minas Gerais State Foundation for Research Support, grant RED 00192-23) and is a member of the CIAA-S (Innovation Center on Artificial Intelligence for Health), and the IATS-CARE (Institute for Health Assessment and Translation for Chronic and Neglected Diseases of High RElevance).

Conflict of interest

The authors declare no competing interests.

Author contributions

Eran Zvuloni  0000-0002-5633-7927

Conceptualization (equal), Data curation (equal), Formal analysis (equal), Investigation (equal), Methodology (equal), Project administration (equal), Visualization (equal), Writing – original draft (equal), Writing – review & editing (equal)

Guido Gagliardi  0000-0003-2020-6439

Formal analysis (equal), Investigation (equal), Methodology (equal), Visualization (equal), Writing – original draft (equal), Writing – review & editing (equal)

Antônio H Ribeiro  0000-0003-3632-8529

Data curation (equal), Methodology (equal), Resources (equal), Writing – original draft (supporting), Writing – review & editing (supporting)

Antonio Luiz P Ribeiro  0000-0002-2740-0042

Data curation (equal), Investigation (equal), Resources (equal), Writing – original draft (supporting), Writing – review & editing (supporting)

Maarten De Vos  0000-0002-3482-5145

Funding acquisition (equal), Investigation (equal), Methodology (equal), Supervision (equal), Writing – original draft (equal), Writing – review & editing (equal)

Joachim A Behar  0000-0001-5956-7034

Conceptualization (equal), Formal analysis (equal), Funding acquisition (equal), Investigation (equal), Methodology (equal), Project administration (equal), Supervision (equal), Validation (equal), Writing – original draft (equal), Writing – review & editing (equal)

References

- [1] Ishaque S, Khan N and Krishnan S 2021 Trends in heart-rate variability signal analysis *Front. Dig. Health* **3** 639444
- [2] Buxi D, Redouté J-M and Yuce M R 2015 A survey on signals and systems in ambulatory blood pressure monitoring using pulse transit time *Physiol. Meas.* **36** R1–R26
- [3] Subha D P, Joseph P K, Acharya R U and Lim C M 2010 EEG signal analysis: a survey *J. Med. Syst.* **34** 195–212
- [4] Berkaya S K, Uysal A K, Gunal E S, Ergin S, Gunal S and Gulmezoglu M B 2018 A survey on ECG analysis *Biomed. Signal Process. Control* **43** 216–35
- [5] Lehman L W H, Adams R P, Mayaud L, Moody G B, Malhotra A, Mark R G and Nemati S 2015 A physiological time series dynamics-based approach to patient monitoring and outcome prediction *IEEE J. Biomed. Health Inform.* **19** 1068–76
- [6] Levy J, Álvarez D, del Campo F and Behar J A 2021 Machine learning for nocturnal diagnosis of chronic obstructive pulmonary disease using digital oximetry biomarkers *Physiol. Meas.* **42** 054001
- [7] Zvuloni E, Read J, Ribeiro A H, Ribeiro A L P and Behar J A 2023 On merging feature engineering and deep learning for diagnosis, risk prediction and age estimation based on the 12-lead ECG *IEEE Trans. Biomed. Eng.* **70** 2227–36
- [8] Biton S et al 2023 Generalizable and robust deep learning algorithm for atrial fibrillation diagnosis across geography, ages and sexes *npj Dig. Med.* **6** 44
- [9] Hannun A Y, Rajpurkar P, Haghpanahi M, Tison G H, Bourn C, Turakhia M P and Ng A Y 2019 Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network *Nat. Med.* **25** 65–69
- [10] Ribeiro A H et al 2020 Automatic diagnosis of the 12-lead ECG using a deep neural network *Nat. Commun.* **11** 1760
- [11] Sel K, Mohammadi A, Pettigrew R I and Jafari R 2023 Physics-informed neural networks for modeling physiological time series for cuffless blood pressure estimation *npj Digi. Med.* **6** 110
- [12] Danker-Hopfe H et al 2009 Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard *J. Sleep Res.* **18** 74–84
- [13] Phan H, Andreotti F, Cooray N, Chén O Y and De Vos M 2019 SeqSleepNet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging *IEEE Trans. Neural Syst. Rehabil. Eng.* **27** 400–10
- [14] Mikkelsen K B, Phan H, Rank M L, Hemmsen M C, De Vos M and Kidmose P 2022 Sleep monitoring using ear-centered setups: investigating the influence from electrode configurations *IEEE Trans. Biomed. Eng.* **69** 1564–72
- [15] Attia Z I et al 2019 An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction *Lancet* **394** 861–7
- [16] Biton S, Gendelman S, Ribeiro A H, Miana G, Moreira C, Ribeiro A L P and Behar J A 2021 Atrial fibrillation risk prediction from the 12-lead electrocardiogram using digital biomarkers and deep representation learning *Eur. Heart J. - Dig. Health* **2** 576–85
- [17] Kotzen K, Charlton P H, Salabi S, Amar L, Landesberg A and Behar J A 2023 SleepPPG-Net: a deep learning algorithm for robust sleep staging from continuous photoplethysmography *IEEE J. Biomed. Health Inform.* **27** 924–32
- [18] Holmstrom L et al 2023 Deep learning-based electrocardiographic screening for chronic kidney disease *Commun. Med.* **3** 1–8
- [19] Lin C et al 2024 Artificial intelligence-enabled electrocardiography contributes to hyperthyroidism detection and outcome prediction *Commun. Med.* **4** 1–11
- [20] Perez Alday E A et al 2020 Classification of 12-lead ECGs: the PhysioNet/computing in cardiology challenge 2020 *Physiol. Meas.* **41** 124003
- [21] Levy J, Álvarez D, Campo F D and Behar J A 2023 Deep learning for obstructive sleep apnea diagnosis based on single channel oximetry *Nat. Commun.* **14** 1–12
- [22] Ballas A and Diou C 2023 Towards domain generalization for ECG and EEG classification: algorithms and benchmarks *IEEE Trans. Emerg. Top. Comput. Intell.* **8** 44–54
- [23] Aar J F V D, Ende D A V D, Fonseca P, Van Meulen F B, Overeem S, Van Gilst M M and Peri E 2023 Deep transfer learning for automated single-lead EEG sleep staging with channel and population mismatches *Front. Physiol.* **14** 1287342
- [24] Heremans E R, Phan H, Borzé P, Buyse B, Testelmans D and De Vos M 2022 From unsupervised to semi-supervised adversarial domain adaptation in electroencephalography-based sleep staging *J. Neural Eng.* **19** 6
- [25] Zhou K, Liu Z, Qiao Y, Xiang T and Loy C C 2023 Domain generalization: a survey *IEEE Trans. Pattern Anal. Mach. Intell.* **45** 4396–415
- [26] Coppola E, Savardi M, Massucci M, Adamo M, Metra M and Signoroni A 2024 HuBERT-ECG: a self-supervised foundation model for broad and scalable cardiac applications *medRxiv* (<https://doi.org/10.1101/2024.11.14.24317328>)
- [27] Song J, Jang J-H, Lee B T, Hong D, Kwon J-M, Jo Y-Y and Lee T 2024 Foundation models for electrocardiograms *Proc. ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (ACM KDD 2025)* vol 1 p 6
- [28] Gedon D, Ribeiro A H, Wahlstrom N and Schon T B 2021 First steps towards self-supervised pretraining of the 12-lead ECG *Comput. Cardiol.* **48** 2021
- [29] Goldberger A L et al 2000 PhysioBank, PhysioToolkit and PhysioNet components of a new research resource for complex physiologic signals *Circulation* **101** e215–20
- [30] Zhang G-Q et al 2018 The national sleep research resource: towards a sleep data commons *Journal of the American Med. Assoc.* **25** 1351–8
- [31] Guo Z, Ding C, Do D H, Shah A, Lee R J, Hu X and Rudin C 2023 SiamAF: learning shared information from ecg and ppg signals for robust atrial fibrillation detection
- [32] Do E, Boynton J, Lee B S and Lustgarten D 2022 Data augmentation for 12-lead ECG beat classification *SN Comput. Sci.* **3** 1–17

- [33] Pan Q, Li X and Fang L 2020 Data augmentation for deep learning-based ECG analysis *Feature Engineering and Computational Intelligence in ECG Monitoring* pp 91–111 (available at: https://link.springer.com/chapter/10.1007/978-981-15-3824-7_6)
- [34] Li J, Aguirre A, Moura J, Liu C, Zhong L, Sun C, Clifford G, Westover B and Hong S 2024 An electrocardiogram foundation model built on over 10 million recordings with external evaluation across multiple domains
- [35] Fox B, Jiang J, Wickramaratne S, Kovatch P, Suarez-Farinas M, Shah N A, Parekh A and Nadkarni G N 2025 A foundational transformer leveraging full night, multichannel sleep study data accurately classifies sleep stages *Sleep* **48** zsaf061
- [36] Abbaspourazad S, Elachqar O, Miller A C, Emrani S, Nallasamy U and Shapiro I 2023 Large-scale training of foundation models for wearable biosignals *12th Int. Conf. on Learning Representations, ICLR 2024*
- [37] Huang Z, MacLachlan S, Yu L, Contreras L F H, Truong N D, Ribeiro A H and Kavehei O 2024 Generalization challenges in electrocardiogram deep learning: insights from dataset characteristics and attention mechanism *Future Cardiol.* **20** 209–20
- [38] Krizhevsky A, Sutskever I and Hinton G E 2012 ImageNet classification with deep convolutional neural networks *Advances in Neural Information Processing Systems* vol 25
- [39] Ribeiro A L P et al 2019 Tele-electrocardiography and bigdata: the CODE (Clinical Outcomes in digital electrocardiography) study *J. Electrocardiol.* **57** S75–S78
- [40] Reyna M A et al 2021 Will two do? Varying dimensions in electrocardiography: the physionet/computing in cardiology challenge 2021 *2021 Computing in Cardiology (CinC) (Brno)* vol 2021 (IEEE) , pp 1–4
- [41] Reyna M A et al 2022 Issues in the automated classification of multilead ecgs using heterogeneous labels and populations *Physiol. Meas.* **43** 084001
- [42] Liu H, Chen D, Chen D, Zhang X, Li H, Bian L, Shu M and Wang Y 2022 A large-scale multi-label 12-lead electrocardiogram database with standardized diagnostic statements *Sci. Data* **9** 272
- [43] Attia Z I et al 2019 Age and sex estimation using artificial intelligence from standard 12-lead ECGs *Circ. Arrhythm. Electrophysiol.* **12** 1–11
- [44] Lima E M et al 2021 Deep neural network-estimated electrocardiographic age as a mortality predictor *Nat. Commun.* **12** 5117
- [45] Quan S F et al 1997 The sleep heart health study: design, rationale and methods *Sleep* **20** 1077–85
- [46] Chen X et al 2015 Racial/ethnic differences in sleep disturbances: the multi-ethnic study of atherosclerosis (MESA) *Sleep* **38** 877–88
- [47] Blackwell T et al Men Study Group 2011 Associations between sleep architecture and sleep-disordered breathing and cognition in older community-dwelling men: the osteoporotic fractures in men sleep study *J. Amer. Geriatr. Soc.* **59** 2217–25
- [48] Young T, Palta M, Dempsey J, Peppard P E, Nieto F J and Hla K M 2009 Burden of sleep apnea: rationale, design and major findings of the wisconsin sleep cohort study *WMJ: Official Publication of the State Medical Society of Wisconsin* vol 108 p 246
- [49] O'reilly C, Gosselin N, Carrier J and Nielsen T 2014 Montreal archive of sleep studies: an open-access resource for instrument benchmarking and exploratory research *J. Sleep Res.* **23** 628–35
- [50] Perslev M, Darkner S, Kempfner L, Nikolic M, Jennum P J and Igel C 2021 U-sleep: resilient high-frequency sleep staging *npj Dig. Med.* **4** 72
- [51] Lee Y J, Lee J Y, Cho J H and Choi J H 2022 Interrater reliability of sleep stage scoring: a meta-analysis *J. Clin. Sleep Me.* **18** 193–202
- [52] Gagliardi G, Alfeo A L, Cimino M G, Valenza G and De Vos M 2025 Physioex: a new python library for explainable sleep staging through deep learning *Physiol. Meas.* **46** 025006
- [53] Phan H et al 2023 L-seqsleepnet: whole-cycle long sequence modelling for automatic sleep staging *IEEE J. Biomed. Health Inform.* **27** 4748–57
- [54] Phan H, Mikkelsen K, Chén O Y, Koch P, Mertins A and De Vos M 2022 Sleeptransformer: automatic sleep staging with interpretability and uncertainty quantification *IEEE Trans. Biomed. Eng.* **69** 2456–67
- [55] Flamary R et al 2021 POT: Python Optimal Transport *J. Mach. Learn. Res.* **22** 1–8
- [56] Pedregosa F et al 2011 Scikit-learn: machine learning in python *J. Mach. Learn. Res.* **12** 2825–30
- [57] Kontras K, Chatzichristos C, Phan H, Suykens J and De Vos M 2024 Core-sleep: a multimodal fusion framework for time series robust to imperfect modalities *IEEE Trans. Neural Syst. Rehabil. Eng.* **32** 840–9