A Survey of Few-Shot Learning for Biomedical Time Series

Chenqi Li, Timothy Denison, Tingting Zhu

***Abstract*—Advancements in wearable sensor technologies and the digitization of medical records have contributed to the unprecedented ubiquity of biomedical time series data. Data-driven models have tremendous potential to assist clinical diagnosis and improve patient care by improving long-term monitoring capabilities, facilitating early disease detection and intervention, as well as promoting personalized healthcare delivery. However, accessing extensively labeled datasets to train data-hungry deep learning models encounters many barriers, such as long-tail distribution of rare diseases, cost of annotation, privacy and security concerns, data-sharing regulations, and ethical considerations. An emerging approach to overcome the scarcity of labeled data is to augment AI methods with human-like capabilities to leverage past experiences to learn new tasks with limited examples, called few-shot learning. This survey provides a comprehensive review and comparison of few-shot learning methods for biomedical time series applications. The clinical benefits and limitations of such methods are discussed in relation to traditional data-driven approaches. This paper aims to provide insights into the current landscape of few-shot learning for biomedical time series and its implications for future research and applications.**

***Index Terms*—Few-Shot Learning, Biomedical Time Series**

I. Introduction

Data plays a fundamental role in the success of deep learning algorithms. For biomedical applications, access to extensively labeled datasets is often a luxury due to ethical and privacy concerns associated with data acquisition and sharing. Moreover, the long-tail distribution of rare diseases leads to natural data imbalance and bias toward majority classes. Inter-subject variability further poses generalization challenges to new patients. To address these challenges, more data must be collected for the long-tail classes and individual patients to improve generalization. Alternatively, deep learning algorithms need to learn and evolve with limited labeled samples, mirroring how human adults and children can transfer past experiences to guide the learning of new tasks using a few examples. Few-shot learning emerges as a promising paradigm to augment models with capabilities to generalize effectively even when confronted with a scarcity of labeled data. The paradigm strongly resonates with meta-learning, learning how to learn, which aims to transfer knowledge from previous tasks to accelerate the learning of new tasks.

Several surveys have provided an overview of the taxonomy of few-shot learning techniques from different perspectives [53, 127, 103, 152, 142, 140], focusing primarily on theoretical background and/or computer vision applications. For biomedical applications, existing surveys have covered biomedical imaging [99, 65] and text [34], and have not yet considered time series. Thus, the goal of this survey is to fill the gap and provide a literature review of few-shot learning techniques for biomedical time series applications, which play an increasingly important role in medical diagnostics, brain-computer interfaces, and wearable computing. This survey aims to answer the following research questions: (1) How are few-shot learning problems defined and how do they differ from traditional deep learning pipelines? (2) What is the taxonomy of few-shot learning methods for biomedical time series data? (3) What are the applications of few-shot learning methods to biomedical time series problems and what benefits do they provide in a clinical setting? (4) What are the key challenges and future directions of few-shot learning for biomedical time series?

Department of Engineering Science, University of Oxford, United Kingdom
Email: {chenqi.li, timothy.denison, tingting.zhu}@eng.ox.ac.uk

II. Background and Concepts

A. Problem Definition

Unlike traditional machine learning and deep learning setups that divide the dataset into training, validation, and test subsets for different stages of the model development pipeline, few-shot learning setups are limited by the number of labeled data available and cannot afford such a division. Models trained on small datasets may memorize the specific samples instead of learning the patterns and generalizing them beyond the training set. Under the few-shot learning setting for classification, datasets are available in the form of the support set $\mathcal{S} = \{(\mathbf{x}_{s,i}, y_{s,i})\}_{i=1}^{N\times K}$ and the query set $\mathcal{Q} = \{(\mathbf{x}_{q,j}, y_{q,j})\}_{j=1}^{M}$ where $\mathbf{x}_{s,i}$ represents the i-th input of the support set of $N\times K$ samples and $\mathbf{x}_{q,j}$ represents the j-th input sample of the query set of M samples. $y_{s,i}, y_{q,j} \in \{1, 2, ..., N\}$ represents the corresponding ground-truth label that belongs to one of the N classes. The goal of few-shot learning is to leverage the information available from the support set $\mathcal{S}$ to estimate the label of each query sample from the query set $\mathcal{Q}$. Such setups are commonly described as N-way-K-shot problems, where N indicates the number of classes and K indicates the number of samples available for each class, resulting in a total of $N \times K$ samples in the support set. Figure 1 provides an overview of the setup and pipeline for few-shot classification with biomedical time series. The aforementioned setup can be extended beyond classification tasks. For example, the ground truth $y_{s,i}, y_{q,j} \in \mathbb{R}$ are continuous targets for regression tasks as opposed to categorical labels for classiciation tasks. In the case of forecasting, the ground-truth consists of sequences $\mathbf{y}_{s,i}, \mathbf{y}_{q,j} \in \mathbb{R}^h$, where h indicates the forecast horizon. Without categorical classes, such tasks are referred to as K-shot learning problems.

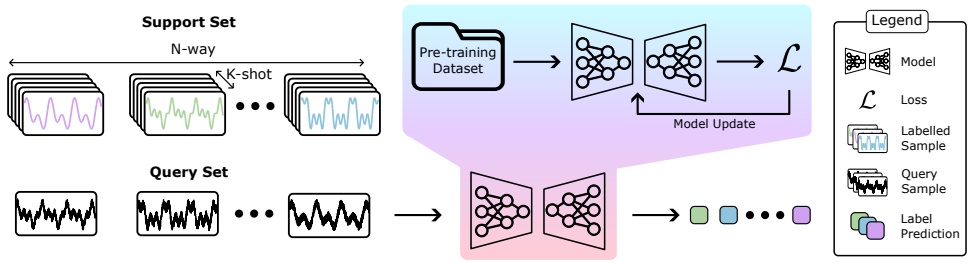

Fig. 1. **Overview of few-shot learning setup.** During the pre-training stage (optional), the model is trained to learn meaningful representations and initialize its parameters to facilitate future adaption to target tasks. Subsequently, the model learns a new task using the N-way-K-shot support set and performs inference on the query set to provide label prediction.

B. Episodic Pre-Training

While the N-way-K-shot problem defines the setup of the support set, no constraint is placed on the use of different but potentially related datasets, $\mathcal{D}_{pre} = \{(\mathbf{x}_l, y_l)\}_{l=1}^{L_{pre}}$ consisting of L_{pre} input-label pairs, to pre-train the few-shot learning model. The pre-trained model can then be adapted using $\mathcal{S}$ before making inferences on the query set $\mathcal{Q}$. Algorithm 1 provides a pseudo-code for the supervised episodic pre-training paradigm. Note that during each episode of pre-training, a support set $\mathcal{S}'$ and a query set $\mathcal{Q}'$ are sampled from the pre-training dataset $\mathcal{D}_{pre}$ to mimic the few-shot learning setting. The class, query, and support size n, m, k of the sampled datasets are typically set to be greater than or equal to the few-shot learning class, query, and support size N, M, K. For a traditional learning pipeline, the loss is computed between the ground truth label and the prediction of the model in the query sample, $\mathcal{L}(y_{q,j}, f_\theta(\mathbf{x}_{q,j}))$. For episodic pre-training, the model makes predictions using an additional support set $\mathcal{S}'$, as shown in line 5 of Algorithm 1. $\mathcal{S}'$ can be used to fine-tune the model or simply provide anchors for comparison against query samples. Episodic pre-training can also be performed in a semi-supervised manner, where the label is available for some samples and missing for others. Similarly, self-supervised episodic pre-training follows a similar procedure, but the label for each sample is generated using the dataset itself as opposed to human labeling.

Algorithm 1 Algorithm for Supervised Episodic Pre-Training

Require: $\mathcal{D}_{pre}$ pre-training dataset
Require: n, m, k class, query, support size of the sampled dataset
Require: $episodes$ number of episodes
1: Randomly initialize θ
2: **for** $episode \leftarrow 1$ to $episodes$ **do**
3: Sample support set $\mathcal{S}' = \{(\mathbf{x'}_{s,i}, y'_{s,i})\}_{i=1}^{n \times k}$ from $\mathcal{D}_{pre}$
4: Sample query set $\mathcal{Q}' = \{(\mathbf{x'}_{q,j}, y'_{q,j})\}_{j=1}^{m}$ from $\mathcal{D}_{pre}$
5: Evaluate $\mathcal{L}(y'_{q,j}, f_\theta(\mathbf{x'}_{q,j}, \mathcal{S}'))$ $\quad \forall j = 1, ..., m$
6: Update model parameters θ based on $\nabla_\theta \mathcal{L}$
7: **end for**

Few-shot learning problems can be characterized based on the type of knowledge transfer a model is expected to bridge between the pre-training problem and the N-way-K-shot problem. For biomedical applications, common types of knowledge transfer include cross-session, cross-subject, cross-dataset, cross-class, cross-task, and no pre-training. In the case of cross-session few-shot learning, the model is pre-trained using data from multiple sessions of a particular subject. The model is expected to generalize to a query set from a previously unseen session of the same subject, using a N-way-K-shot support set from the same unseen session to assist prediction. Similarly, a model for cross-subject (dataset, class, task) few-shot learning is pre-trained using data from multiple subjects (datasets, classes, tasks) and is expected to generalize to unseen subjects (datasets, classes, tasks) with few samples as support. Sometimes, no pre-training is performed, the model is expected to learn with few-shot support samples from scratch and make predictions on the query set. Evidently, the type of knowledge transfer defines various challenges in few-shot learning, and it is important to consider this difference when designing and comparing few-shot methods.

To address the data shortage in the support set and facilitate the transfer of knowledge from the pre-training dataset to the N-way-K-shot problem, few-shot learning approaches propose modifications to different parts of the training pipeline to overcome the limited dataset and can be categorized into five types: data-based, model-based, metric-based, optimization-based, and hybrid methods. A taxonomy of few-shot learning is provided in Figure 2. The specific methods and relevant applications will be elaborated in further detail in Section IV.

III. Search Strategy

In this work, biomedical time series are defined as time-sequential biological data of varying modalities to improve healthcare or assist clinical practices, such as diagnosis or risk prediction. We design the search key to capture three important characteristics of the relevant literature: “few shot learning”, “biomedical”, and “time series”. After identifying and screening relevant literature using a variety of digital libraries, a total of 55 records are included in this review. Detailed search strategy and procedure are described in Appendix A.

IV. Few-Shot Learning Taxonomy & Applications

A. Data-Based Few-Shot Learning

Data-based methods (DBMs) aim to directly address the data shortage problem by increasing the quantity and diversity of the support set by generating synthetic samples using techniques such as generative adversarial networks (GANs) or data augmentation.

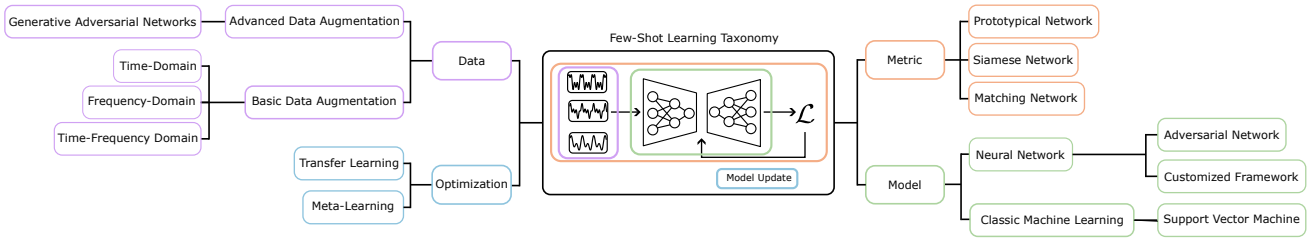

Fig. 2. **Taxonomy of few-shot learning for biomedical time series.** Different few-shot learning methods propose modifications to varying parts of the training pipeline, each highlighted using a distinct color. Data-based methods directly address the data shortage problem by generating synthetic samples to increase the size and diversity of the support set. Metric-based methods focus on learning the similarity between samples in a representation space. Model-based methods design model architectures to improve generalization across tasks with few samples. Optimization-based methods guide model convergence to parameter spaces that can be quickly adjusted by fine-tuning with a few samples.

TABLE I
DATA-BASED FEW-SHOT LEARNING METHODS FOR BIOMEDICAL TIME SERIES

	Problem			Pre-Training	Knowledge Transfer	Few-Shot Learning Setup				Open Access
Reference	Task	Modality	Method	Type of Pre-Training	Type of Transfer	FSL Dataset	Number of Classes	Support Set Size	Evaluation Metric	Code Availability
[160]	Sleep Staging	EEG	GAN	-	N/A	SleepEDF [61]	5	1 patient 1 night	AC, F1, CK	✗
[51]	Visual Stimuli	EEG	WGAN	Supervised	Classes	[129]	10	0-shot	AC	✗
[109]	Mortality Heart Failure Survival	EHR	GAN	Self-Supervised	Tasks	MIMIC-IV [57] Synthea [148] All of Us [143]	2 2 2	1-50% 1-50% 1-50%	AUROC, AUPRC, F1	✓
[134]	Indoor Localization	RSSI ACC	DA	-	N/A	[18]	Varying	1-shot	F1	✗
[30]	Sleep Posture	IMU	DA	-	N/A	Collected (Private)	12	1-shot	AC, F1	✗
[69]	Visual Stimuli	EEG	GAN	Supervised	Subjects	EEG-ImageNet40 [129]	40	{1,2,3,4,5}-shot	AC	✓

All tasks are classification tasks. Abbreviations: Electroencephalogram (EEG), Electronic Health Record (EHR), Radio Signal Strength Indication (RSSI), Accelerometry (ACC), Inertial Measurement Unit (IMU), Accuracy (AC), Cohen's Kappa (CK), Area Under Receiver Operating Characteristic Curve (AUROC), Area Under Precision-Recall Curve (AUPRC)

1) GAN: The core concept of GAN [37] involves adversarial training of separate discriminator and generator networks, to produce synthetic samples that match the distribution of real data. The goal of the generator G is to mimic the distribution of the training data and create synthetic data $G(\mathbf{z})$ as a function of random noise $\mathbf{z}$ drawn from a latent space distribution $p_{\mathbf{z}}(\mathbf{z})$. Meanwhile, the discriminator D aims to distinguish between training data $\mathbf{x}$ drawn from the training data distribution $p_d(\mathbf{x})$ and synthetically generated data $G(\mathbf{z})$. The discriminator learns to estimate the probability that the given input is real or synthetic. Networks are backpropagated together using an adversarial loss function $V(D,G)$, in which the discriminator aims to maximize while the generator aims to minimize:

$$\min_G \max_D V(D,G) = \mathop{\mathbb{E}}_{\mathbf{x}\sim p_d(\mathbf{x})}[\log D(\mathbf{x})] + \mathop{\mathbb{E}}_{\mathbf{z}\sim p_z(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))]$$

Several variants of GAN have since been proposed to address the challenges associated with the training and generation process of GAN. Notably, mode collapse and vanishing gradients are widely recognized training challenges for GAN [6]. To address these challenges and improve training stability, Wasserstein generative adversarial networks (WGANs) proposed to use Wasserstein distance as a training objective to measure the similarity between the real and synthetic data distributions [7]. On another note, conditional generative adversarial network (CGAN) pass a variable or label y as an additional input to the discriminator and the generator, to control the generation process and synthesize data that conform to specific conditions or constraints [87]. Many works have focused on using GANs to diversify and expand the size of the few-shot support set, in hopes of using the expanded support set to train a better-performing model. For example, You et al. [160] proposed SleepGAN which consists of a WGAN and a relational memory generator to generate synthetic electroencephalogram (EEG) sleep epochs and transition sequences from the sleep stage. To accelerate training and generation performance, the authors adopted the generator and discriminator architectures of ConSinGAN [44]. To demonstrate the effectiveness of SleepGAN in generating high-quality samples from few-shot real data, a downstream sleep staging model was trained using a combination of the patient's few-shot real data and synthetically generated epochs. An accuracy of 81.1% was observed on SleepEDF [61], compared to 77.5% when training the same sleep staging model using the patient's few-shot real data alone. Similarly, Lee et al. [69] generated synthetic EEG samples for the problem of source-free subject adaptation. Source-free subject adaptation refers to the transfer of knowledge from source subjects to target subjects, using few-shot target subject data and a pre-trained model on the source subject data, without access to any source subject data itself. The authors employed a modified version of GAN. The discriminator is a frozen network pre-trained on the source subject data, and it supervises the generator to synthesize source subject samples that align with the source distribution. The generated source samples can then be combined with few-shot target samples for subject-independent

feature learning. Without relying on real source subject data, few-shot subject transfer is achieved while protecting patient privacy and bypassing barriers to data sharing. Learning to synthesize model representations, as opposed to raw time series, can also be considered a form of data augmentation to supplement the few-shot support set. Hwang et al. [51] proposed a three-stage framework for zero-shot learning of visual stimuli classification, which aims to predict the image class presented to the subject based on the given EEG response. First, a gated recurrent unit (GRU)-based encoder is trained to extract features from real EEG signals. Second, a WGAN conditioned on word2vec [86] vectors of a given semantic class is trained to generate synthetic features that mimic the real features extracted by the GRU from real EEG signals. Thirdly, the word2vec semantic vector of each unseen class is provided as input to the WGAN, and the generated features are used to train a classifier for zero-shot classification, achieving 39.65% zero-shot ten-class accuracy on [129]. The pipeline relies heavily on the assumption that WGAN is capable of generalizing and generating GRU feature vectors for unseen classes. Similarly, Poulain et al. [109] proposed to simulate representations of bidirectional encoder representations from transformers (BERT) outputs using a multilayer perceptron (MLP)-based GAN. The generator learns to simulate synthetic representation, while the discriminator learns to perform two tasks: differentiating between real and synthetic representations through an unsupervised loss and producing classification predictions through a supervised loss. Therefore, the discriminator not only provides supervision to the generator but also acts as the classifier for downstream tasks. In addition to the few-shot labeled dataset, the authors further leverage an unlabelled dataset to provide a more comprehensive representation of the real samples for comparison against synthetic samples. Experimental results on multiple electronic health record (EHR) predictive tasks show that the proposed network surpasses state-of-the-art EHR predictive models, especially under the few-shot setting. However, the assumption that an unlabeled dataset is available is often invalid in the few-shot learning setting, limiting the generalization of this approach to other tasks and datasets.

2) Data Augmentation: Data augmentation (DA) for time series can be divided into time, frequency, and time-frequency domain-based methods. Time domain methods manipulate time series directly, such as noise injection, flipping, cropping, scaling, imputation, amplitude scaling, and up/downsampling. Frequency domain methods apply the Fourier transform to convert signals from the time domain to the frequency domain, perturbing amplitude and frequency representations to introduce variations. Time-frequency domain methods introduce variations to time-frequency representations obtained through a variety of transformations, such as wavelet transform, short-time Fourier transform (STFT), and spectrogram extraction. Suwannaphong et al. [134] studied two augmentation techniques for the localization of indoor patients using radio signal strength indication (RSSI) sensors, namely, adding Gaussian noise and dropping out channels. The results show that the augmentation techniques are simple but effective in boosting the one-shot classification with random forest, reducing the data collection time from hours to minutes. In a different application, Elnaggar et al. [30] investigated DA for inertial measurement unit (IMU) sensor. Due to the mathematical constraints of the quaternion representation, simply adding noise to the quaternion creates non-sensible synthetic data. To overcome this challenge, the quaternion representations are converted to axis-angle representations, and Gaussian noise is injected into axis and angle components to generate near-realistic postural data. The classification of sleep posture is subsequently performed through an ensemble of soft margin support vector machine (SVM). Experimental results show that noise augmentation helps to improve the robustness of the classifier against intra-posture similarity, given the appropriate degree of noise level. It should be noted that both DA-based approaches from Section IV-A2 are applied to traditional machine learning models that typically require less data than deep learning models. DA is likely insufficient as a standalone method to help overcome the data shortage barrier in few-shot learning settings for deep learning models.

3) Summary: DBMs focus on addressing data scarcity by generating realistic synthetic samples or feature representations to complement the few-shot support set, contributing to performance gains in a variety of downstream tasks. However, a major concern for DBMs is that the support set does not capture a representative distribution of the data. Thus, the distribution of the augmented set will be similar to the distributions of the un-augmented set, potentially leading to overfitting on the support set. Overall, DBMs are simple, interpretable, computationally inexpensive, and do not require any pre-training. Good domain knowledge of the problem is often necessary to ensure the augmented data is meaningful and beneficial for model learning. GANs can synthesize more diverse samples that are not limited to transformations of existing samples. GANs are also more flexible, and capable of synthesizing both raw time series and model representations.

B. Metric-Based Few-Shot Learning

Metric-based methods (MBMs) focus on learning the similarity or distance between data points through an embedding network (EN). EN projects data points to a representation space where instances of similar labels are clustered closely, while those of different labels are positioned farther apart.

1) Siamese Network (SN): SN consists of parallel weight-sharing ENs and similarity functions that project and measure similarity between paired input data points in the representation space [15]. Given a N-way-K-shot support set $\mathcal{S} = \{(\mathbf{x}_{s,i}, y_{s,i})\}_{i=1}^{N\times K}$ and a query set $\mathcal{Q} = \{(\mathbf{x}_{q,j}, y_{q,j})\}_{j=1}^{M}$, SNs compute the similarity between the query and support samples using a similarity metric (SM) consisting of a distance function and a sigmoid activation. The similarity between the j-th query sample $\mathbf{x}_{q,j}$ and the i-th support sample $\mathbf{x}_{s,i}$ can be computed as:

$$s(\mathbf{x}_{q,j}, \mathbf{x}_{s,i}) = \frac{1}{1 + \exp(-||f_\theta(\mathbf{x}_{q,j}), f_\theta(\mathbf{x}_{s,i})||)}$$

where $||f_\theta(\mathbf{x}_{q,j}), f_\theta(\mathbf{x}_{s,i})||$ is the distance function (e.g. Euclidean distance, cosine similarity) between embeddings of

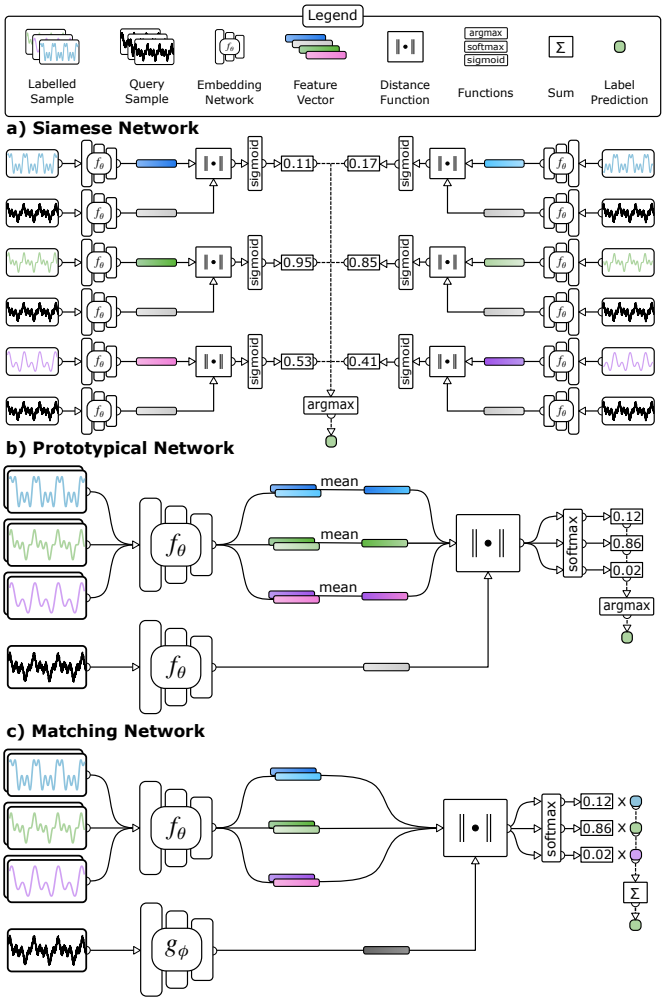

Fig. 3. **Comparison of metric-based methods in few-shot learning with a 3-way-2-shot setup.** All methods employ an embedding network to project support and query samples into the representation space, to measure the similarity between support and query samples. a) Siamese network computes the similarity between each pair of support-query feature vectors to determine the most likely class. b) Prototypical network computes class-wise prototypes by taking the average of all support sample feature vectors from each class. Similarity is then computed between the class-wise prototypes and the query feature vector. c) Matching network directly computes the similarity between every pair of support and query feature vectors, but the support and query samples are projected using two different embedding networks. The network output is a linear combination of similarity between each support-query pair and class of the support sample.

paired inputs to the SN and f_θ represents the EN parameterized by parameters θ. The classification prediction $\hat{y}_{q,j}$ for query sample $\mathbf{x}_{q,j}$ is defined as:

$$\hat{y}_{q,j} = y_{s,\hat{i}} \quad \text{where} \quad \hat{i} = \underset{i}{\operatorname{argmax}}\, s(\mathbf{x}_{q,j}, \mathbf{x}_{s,i})$$

The choice of parallel weight-sharing EN and SM is highly flexible and dependent on the application and modality. Common choices of ENs include convolutional neural network (CNN) [10] & transformer [9] for images & time series, long short-term memory (LSTM) [90] for text, and graph neural network (GNN) [55] for graphs. SNs can be trained using the cross-entropy loss, but contrastive and triplet loss that are designed to learn effective and discriminative embeddings are also widely used to train SNs. Given an EN f_θ parameterized by θ, the contrastive loss between the i-th support pair $(\mathbf{x}_{s,i}, y_{s,i})$ and the j-th query pair $(\mathbf{x}_{q,j}, y_{q,j})$ can be defined as:

$$\mathcal{L}_c = (1-y)\cdot||f_\theta(\mathbf{x}_{s,i}), f_\theta(\mathbf{x}_{q,j})||^2 + y\cdot\max(0, m - ||f_\theta(\mathbf{x}_{s,i}), f_\theta(\mathbf{x}_{q,j})||^2)$$

where y is a binary label for whether input pairs are similar $y=0$ if $y_{s,i}=y_{q,j}$ or dissimilar $y=1$ if $y_{s,i}\neq y_{q,j}$, and $m>0$ is a threshold that will not push dissimilar pairs further apart if the distance between the dissimilar pairs is sufficiently large. The triplet loss further extends the contrastive loss to include comparison against both similar and dissimilar samples simultaneously. Given the j-th query pair $(\mathbf{x}_{q,j}, y_{q,j})$ as an anchor, the i^+-th support pair $(\mathbf{x}_{s,i^+}, y_{s,i^+})$ as a similar example to the anchor where $y_{q,j} = y_{s,i^+}$, and the i^--th support pair $(\mathbf{x}_{s,i^-}, y_{s,i^-})$ as a dissimilar example to the anchor where $y_{q,j}\neq y_{s,i^-}$, the triplet loss is defined as:

$$\mathcal{L}_t = \max(0, ||f_\theta(\mathbf{x}_{q,j}), f_\theta(\mathbf{x}_{s,i^+})|| - ||f_\theta(\mathbf{x}_{q,j}), f_\theta(\mathbf{x}_{s,i^-})|| + \alpha)$$

where α is a margin that enforces a minimum difference between the embedding distance from the anchor to the positive sample and from the anchor to the negative sample, to avoid convergence to the trivial solution of projecting all samples to the same point in the embedding space. The effectiveness of Siamese CNN for few-shot learning has been verified across many biomedical time series applications of varying modalities. Major differences across Siamese CNN lie in the choice of ENs and SMs. 1D CNN-based SNs are by far the most common choice among applications to biomedical time series. Gupta et al. [40] calculated similarity based on sigmoid activation of L1 distance between feature pairs from 1D CNN EN and show superior few-shot learning electrocardiogram (ECG) classification performance than dynamic time warping and LSTM-based methods. Soroushmojdehi et al. [128] adopted a similar setup but added a fully connected layer, fine-tuned with the support set, between the EN and the L1 distance function. On the other hand, Li et al. [72] chose to add a fully connected layer between the L1 distance function and the sigmoid activation. Ng et al. [96] also adopted a similar architecture as [40] and replaced the sigmoid activation function with a two-layer fully connected network with two output neurons for direct binary classification of input pairs as similar or dissimilar. Classification with SNs relies heavily on the assumption that the labels for each data point are accurate and often do not hold in emotion recognition datasets that involve fine-grained, self-reported labels. To overcome this problem, Zhang et al. [164] proposed to replace the averaging of similarity score for each class with a Bayesian fusion-based distance module to minimize overfitting to mislabelled data points. All the aforementioned Siamese 1D CNN methods demonstrate improvements in few-shot learning performance compared to existing backbones trained without the Siamese framework. Many works have also sought to convert raw time series to alternative representations. For example, Munia et al. [91] proposed to symbolize EEG signal as 1D linear binary pattern (LBP) histogram representation, in an attempt to capture wave morphology of local neighborhoods and summarize structural patterns of the signal. The authors further trained a SN with 1D CNN EN and sigmoid-activated fully connected layer as SM. The results show improvements over

baseline CNN under the few-shot learning setting but do not show consistent improvements when trained using all available data. In another work, Tam et al. [136] mapped individual time series segments of each electromyography (EMG) channel into a 2D muscle activity heatmap based on the physical layout of the channels. A personalized SN with 2D CNN EN and cosine SM demonstrate strong few-shot learning transfer across user sessions, with near-perfect performance for at least half of the participants. Similarly, Wang et al. [150] mapped time series segments into four 2D muscle activity heatmaps, arranged by the physical layout of the channels. Each heatmap corresponds to one of the four time-domain features originally proposed by [49]: mean absolute value, number of zero-crossings, number of slope sign changes, and waveform length. SN based on 2D CNN EN and sigmoid-activated Euclidean distance function was found to demonstrate strong generalization to new classes and subjects for recognizing hand gesture.

2) Matching Network: Matching networks (MNs) [144] follow a similar procedure as SNs, but instead of comparing pairs of query and support samples, the query sample is compared against the entire support set. Furthermore, the EN for support and query samples may not necessarily share the same weights or architecture. Given a N-way-K-shot support set $\mathcal{S} = \{(\mathbf{x}_{s,i}, y_{s,i})\}_{i=1}^{N\times K}$ and query set $\mathcal{Q} = \{(\mathbf{x}_{q,j}, y_{q,j})\}_{j=1}^{M}$, MNs compute the similarity between the support and query sample using a SM consisting of a distance function and a softmax activation. The similarity between the j-th query sample $\mathbf{x}_{q,j}$ and the i-th support sample $\mathbf{x}_{s,i}$ is:

$$s(\mathbf{x}_{q,j}, \mathbf{x}_{s,i}) = \frac{\exp(-||f_\theta(\mathbf{x}_{q,j}), g_\phi(\mathbf{x}_{s,i})||)}{\sum_{i'=1}^{N\times K} \exp(-||f_\theta(\mathbf{x}_{q,j}), g_\phi(\mathbf{x}_{s,i'})||)}$$

where f_θ and g_ϕ are ENs for query and support samples parameterized by θ and ϕ respectively. The final prediction $\hat{y}_{q,j}$ for query sample $x_{q,j}$ then becomes a linear combination of labels weighted by similarity score:

$$\hat{y}_{q,j} = \sum_{i=1}^{N\times K} s(\mathbf{x}_{q,j}, \mathbf{x}_{s,i}) \cdot y_{s,i}$$

The similarity score can be interpreted as an attention mechanism that weighs the label of each support sample. One of the first applications of the MN to biomedical time series was performed by Compagnon et al. [24] using a GRU-based EN and a softmax activated Euclidean or cosine distance function. The EN consists of a GRU encoder to produce embeddings for the support and query samples, which are further strengthened by a bidirectional GRU and an attention GRU to produce context embeddings for measuring similarity. The best performance was observed when the encoder network was sequence-to-sequence pre-trained on a similar dataset, which achieved a comparable result to baseline networks trained on the entirety of the target dataset. Similarly, Bhosale et al. [12] also employed Euclidean distance to match the query sample to the support set, but the EN consists of 3D convolution encoder and bi-directional LSTM to learn temporal relations. An episodic pre-training paradigm was adopted and different sampling techniques for forming the support and query sets were explored. Random sampling draws samples randomly without any constraint and is the most widely used sampling technique for episodic pre-training. Subject-dependent sampling draws support samples that belong to the same subject as the query samples, while subject-independent sampling draws support samples that belong to different subjects as the query samples. Quantitative results show that the choice of sampling strategy does indeed affect few-shot learning performance. To achieve the best test performance, the episodic pre-training strategy should closely mimic the test scenario. In other words, if support and query samples are to be drawn from different subjects during testing, the subject-independent sampling strategy should be employed during episodic pre-training. Meanwhile, if support and query samples are to be drawn from the same subject during testing, subject independent sampling strategy should be adopted over the other sampling strategies. Instead of matching time series, McCartney et al. [82] designed a multi-modal MN for brain computer interface (BCI) image retrieval. The task involves identifying the visual stimuli of a given EEG signal. A variety of manually selected filters and correlation metrics were used to extract features from EEG and corresponding visual stimuli images. Linear regression models are then trained to map EEG features to visual stimulus features and determine the most likely visual stimulus through the nearest neighbor. Wang et al. [151] proposed to further extend the MN to steady state visually evoked potential (SSVEP) classification under a generalized zero-shot learning setting, where query samples can belong to both seen and unseen classes during training. SSVEP describes the brain's natural response to visual stimulation of specific frequencies. For a given EEG signal, the SSVEP classification problem is to predict the modulation frequency of a visual stimulus, typically in the form of a flickering screen at set frequencies. The authors propose to replace the support set EEG samples with the sine wave template of the luminance modulation function of the visual stimuli for all classes. Consequently, the MN consists of two separate ENs, one for projecting query EEG signals to feature vectors and the other for projecting a sine wave template of visual stimuli to feature vectors of the same representation space. Capitalizing on the use of a visual stimuli template and prior knowledge of the problem setup, the proposed MN is capable of achieving multi-fold improvements to existing zero-shot or training-free methods, as well as comparable results to SOTA methods trained on the full dataset.

3) Prototypical Network: Prototypical networks (PNs)[125] are similar to the MNs, but instead of comparing the query embedding with every support embedding, the query embeddings are compared against class-wise prototype embeddings. Given a N-way-K-shot support set $\mathcal{S} = \{(\mathbf{x}_{s,i}, y_{s,i})\}_{i=1}^{N\times K}$, PNs project support samples of a given class into embeddings of the feature space and construct class-wise prototypes by computing the average over the feature vectors. For class n, the prototype can be constructed as follows:

$$\mathbf{c}_n = \frac{1}{|\mathcal{S}_n|} \sum_{(\mathbf{x}_{s,i}, y_{s,i})\in\mathcal{S}_n} f_\theta(\mathbf{x}_{s,i})$$

where $\mathcal{S}_n$ denotes the subset of $\mathcal{S}$ that belongs to class n, $|\mathcal{S}_n|$ denotes the number of samples of the subset, and $|\mathcal{S}_n| = N$ for N-way-K-shot support set $\mathcal{S}$. A SM computes the similarity between the j-th query sample $\mathbf{x}_{q,j}$ against each of the N prototypes $\mathbf{c}_n$, using a combination of the distance function and softmax activation:

$$s(\mathbf{x}_{q,j}, \mathbf{c}_n) = \frac{\exp(-||f_\theta(\mathbf{x}_{q,j}), \mathbf{c}_n||)}{\sum_{n'=1}^{N} \exp(-||f_\theta(\mathbf{x}_{q,j}), \mathbf{c}_{n'}||)}$$

The final prediction $\hat{y}_{q,j}$ for the j-th query sample $\mathbf{x}_{q,j}$ is:

$$\hat{y}_{q,j} = \underset{n}{\operatorname{argmax}}\, s(\mathbf{x}_{q,j}, \mathbf{c}_n)$$

PNs are less computationally expensive than SNs, as the similarity between the query feature vector is computed against N prototypes in PNs, as opposed to $N \times K$ supporting feature vectors in SNs. In the case of one-shot learning, MNs and PNs are equivalent. The major difference between prototypical, MNs, and SNs lies in how the similarity between the support and query embeddings is computed. SNs compute the distance between pairs of query and support embeddings. MNs compute the distance between query embedding and every support sample embedding. PNs compute the distance between query embedding and prototype embedding of each class. Figure 3 provides a visual illustration of the differences between the MBMs using a 3-way-2-shot setup. The choice of EN and SM for PNs is similarly diverse. Salekin et al. [115] implemented a prototypical network based on 1D CNN EN and Euclidean distance function. The authors conclude from the experimental results that, under a class-imbalanced few-shot learning setting, PNs do not suffer from overfitting like conventional supervised classifiers. Using a similar setup, Hernandez-Galvan et al. [43] adopted an EN that consists of a mixture of 1D convolution and GRU layers, as well as softmax activated Euclidean distance function. Meanwhile, Tang et al. [138] chose to extract features using handcrafted symbolic Fourier transformation word histograms and a fully connected layer. A softmax-activated Euclidean distance function between the query embeddings and prototypes is chosen as the SM. The authors further take advantage of the interpretability of the handcrafted features and prototypes to facilitate model explanation, data analysis, and problem understanding. Despite the popularity of the Euclidean distance function as part of the SM, Pan et al. [100] also adopted the cosine distance function and a multi-scale 1D CNN as the EN to extract features of different scales and contextual characteristics. An ablation study found that cosine distance as a SM led to better model performance than Euclidean distance, potentially since cosine distance increases the distance between samples of different classes [131]. This is contrary to the findings when PNs are first proposed, where the authors found that squared Euclidean distance greatly improves results and conjecture that cosine distance performs relatively poorly due to it not being a Bregman divergence [125]. Therefore, the choice of SM plays an important role in the model's performance and could vary depending on many factors, such as the architecture of the EN. Hyperdimensional computing presents an alternative feature extractor to 1D CNN-based ENs. Moin et al. [88] chose to extract features using a hyperdimensional computing framework that consists of three components: mapping module, encoder, and associative memory. The mapping module converts raw biosignals into hyperdimensional vectors, and the encoder module extracts spatial and temporal features using hyperdimensional vector space operations, such as point-wise multiplication, point-wise addition, scalar multiplication, and permutation. The associative memory module stores prototypes of encoded features for each class. During testing, the most similar hyperdimensional vector to the prototype hyperdimensional vectors is identified using cosine distance. Adopting a similar ideology, Burrello et al. [17] combined hyperdimensional computing with LBP to achieve near-perfect seizure detection performance with one and few-shot samples. Local LBP representation of the brain state is constructed over time for all electrodes. Then, LBP representations are manipulated by hyperdimensional operations to produce complex representation vectors, which form the basis for constructing prototypes of each class. Hamming distance is selected as the distance function for the SM during inference to select the best matching prototype to the query sample. The use of LBP brings transparency to the learning procedure. It translates the learned codes into spatial localization of seizure-generating regions, which provides useful insights to support clinical decision-making. Several studies explored alternative methods of constructing prototypes from feature vectors of the support set. Sun et al. [132] argued that some relationship between prototypes exists, thus, inter-class relationships amongst the prototypes can improve model learning. Consequently, the author proposes a self-attention-based prototype enhancement module to generate enhanced prototypes that take into account the relationship among prototypes. Furthermore, to adapt to a class incremental few-shot learning setting, a prototype non-overlapping loss has been proposed to encourage the separation of prototypes of new classes from prototypes of old classes. Similarly, Suo et al. [133] explored learning the relationship amongst prototypes with supervision from a hierarchy graph. Simply taking the average of feature vectors from the EN as prototypes can be sensitive to outliers given limited support samples of each class. The authors observed that hierarchy graphs capture relationships amongst classes, where leaf node classes that share the same ancestral nodes are likely to share similar characteristics. Therefore, prototypes of leaf node classes that belong to the same ancestral node should be closer than prototypes of leaf node classes that belong to different ancestral nodes [133]. Thus, an attention-based procedure has been proposed to enhance prototype representations using information from the hierarchy graph. However, a hierarchy graph is not readily available, limiting the applicability of this method to different applications. Instead of computing prototypes as the average across feature vectors of the support set, Ahn et al. [1] drew inspiration from [154] and constructed prototypes by pooling the feature vectors of each class from the support set as a feature matrix. The SM between the feature matrix prototype and the feature vector of the query sample is calculated as a solution of ridge regression on the feature map.

TABLE II
METRIC-BASED FEW-SHOT LEARNING METHODS FOR BIOMEDICAL TIME SERIES

	Problem			Pre-Training	Knowledge Transfer	Few-Shot Learning Setup				Open Access
Reference	Task	Modality	Method	Type of Pre-Training	Type of Transfer	FSL Dataset	Number of Classes	Support Set Size	Evaluation Metric	Code Availability
[151]	SSVEP	EEG	Matching	Supervised	Classes	BETA [73] Benchmark [153]	8 8	0-shot 0-shot	AC	✗
[24]	Human Activity	IMU	Matching	Self-Supervised	Datasets	MobiActV2 [20]	12	1-shot	AC	✗
[132]	Arrhythmia Face Outline Hand Gesture	ECG FO ACC	Prototypical	-	N/A	MIT-BIH [36] UCR FaceAll [25] UCR UWave [25]	9 14 24	5-shot 5-shot 5-shot	AC	✗
[43]	Speech Imagery	EEG	Prototypical	Supervised, Epiodic	Subjects	KaraOne [167] ASU [97]	2, 11 2, 8	{3,5,7,10}-shot {3,5,7,10}-shot	AC	✗
[1]	Stridor	Audio	Prototypical	Supervised	Tasks	Collected (Private)	2	{4,6,8}-shot	AC, SP, SE, PR, F1	✓
[138]	Myocardial Infarction Arrhythmia	ECG	Prototypical	-	N/A	UCR ECG200 [25] UCR TwoLeadECG [25])	2 2	{6,8}-shot {6,8}-shot	AC	✓
[100]	Consciousness	EEG	Prototypical	Supervised, Episodic	Sessions	Collected (Private)	2	{5,10,15}-shot	AC, RE, PR, F1	✗
[115]	ASD	EEG	Prototypical	Supervised, Episodic	Subjects	Collected (Private)	4, 8	3-shot	AC	✗
[133]	Disease	EHR	Prototypical	Supervised, Episodic	Classes	MIMIC-III [58]	3, 5	{1,5}-shot	AC	✗
[72]	Arrhythmia	ECG	Siamese	-	N/A	MIT-BIH [36]	5	{1,5,10,30,50}-shot	AC	✗
[164]	Emotion	EDA BVP TEMP	Siamese	-	N/A	CASE [121] RCEA [165] CEAP-360VR [159]	2, 5 2, 5 2, 5	{1,5,10}-shot {1,5,10}-shot {1,5,10}-shot	AC, Macro-F1	✓
[91]	Seizure	EEG	Siamese	Supervised	Sessions	CHB-MIT [123]	2	1-shot	SP, SE, AC, F1, PR	✓
[150]	Hand Gesture	EMG	Siamese	Supervised	Classes	Collected (Private)	5	{1,5}-shot	AC	✗
[40]	Arrhythmia	ECG	Siamese	Supervised	Datasets	MIT-BIH [36]	5	5-shot	AC	✗
[128]	Hand Gesture	EMG	Siamese	Supervised	Classes	Collected [128]	6	5-shot	AC	✗
[136]	Hand Gesture	EMG	Siamese	Supervised	Sessions	Collected (Private)	6	{0-20}-shots	AC	✗
[96]	Atrial Fibrillation	ECG	Siamese	Supervised	Subjects	MIT-BIH [36] LT-AF [105]	2 2	{1,3,5,7,9,11}-shot {1,3,5,7,9,11}-shot	Macro-F1, AC, SE, SP	✗
[88]	Hand Gesture	EMG	Prototypical	-	N/A	Collected [88]	5	{1-10}-shot	AC	✓
[12]	Emotion	EEG	Matching	Supervised	Subjects	DEAP [64]	2	{5,15,20,25}-shot	AC	✗
[82]	Visual Stimuli	EEG	Matching	Supervised	Classes	Trento [92] [60]	60 6	0-shot 0-shot	CMC, AUROC	✗
[17]	Seizure	EEG	Prototypical	-	N/A	Collected [17]	2	{2-14}-shot	SP, SE, AC	✗

All tasks are classification tasks. Abbreviations: Electroencephalogram (EEG), Inertial Measurement Unit (IMU), Electrocardiogram (ECG), Face Outline (FO), Accelerometry (ACC), Electronic Health Record (EHR), Electrodermal Activity (EDA), Blood Volume Pulse (BVP), Temperature (TEMP), Electromyography (EMG), Accuracy (AC), Specificity (SP), Sensitivity (SE), Precision (PR), Recall (RE), Cumulative Match Curve (CMC), Area Under Receiver Operating Characteristic Curve (AUROC)

4) Summary: MBMs learn to compare and create flexible models that are not anchored to specific classes. By making contrastive pairs and sets, MBMs are more robust to inter-subject variability and class imbalance than traditional supervised learning. Although diverse combinations of ENs and SMs have been proposed, there is little consensus on the choice of ENs and SMs. As a general guideline, the ENs are usually selected from existing backbones that are designed for the task or modality in question to extract relevant features. Subsequently, the choice of SM can be treated as a hyper-parameter optimization. In general, PNs are computationally inexpensive and have the lowest memory requirement to store class prototypes. However, it assumes that the EN can project each class to a tight cluster. When class samples are diverse and do not form unimodal clusters, matching and SNs would be more suitable. Furthermore, when the number of support samples for each class is extremely low, such as one-shot, MN outperforms SN [144].

C. Model-Based Few-Shot Learning

Model-based few-shot learning methods rely on designing a model architecture that can generalize well to different few-shot learning tasks.

1) Adversarial Networks: Adversarial learning-based models design modules that compete against each other and improve together during the learning process. While GAN presented in Section IV-A1 is a subset of adversarial learning-based models, research works presented in this section focus on designing adversarial models that generalize across domains, as opposed to supplementing the few-shot support set with synthetic data or features. Phunruangsakao et al. [106] proposed an adversarial domain adaptation framework that guides the model to extract subject-invariant features. The proposed framework first pre-trains a classification model using samples from the source subjects. The pre-trained model is then frozen and queried using samples from both the source and target subjects. The features extracted by the CNN backbone of the classification model are provided as input to a discriminator, which attempts to distinguish the origin of the feature as the source or target subject. To train the discriminator and the backbone network of the classification model in an adversarial manner, the weights of one network are frozen while the weights of the other are updated. This process encourages the CNN backbone to extract features that are invariant between the source and the target subjects, effectively enabling positive transfer that adapts knowledge from the source domain to boost performance in the few-shot target domain. Zhu et al. [168] adopted a similar adversarial domain adaptation framework in an unsupervised manner. Instead of pre-training on source data and asynchronously updating the discriminator and backbone network, Zhu et al. proposed train a discriminator and an autoencoder simultaneously. The discriminator learns to distinguish between source and target subject samples using the latent space representation of the encoder. At the same time, a reconstruction loss guides the training of the autoencoder. Thus, the encoder is encouraged to extract subject-invariant features and allows the classifier to be trained on source-subject data to the target subject.

TABLE III
MODEL-BASED FEW-SHOT LEARNING METHODS FOR BIOMEDICAL TIME SERIES

	Problem			Pre-Training	Knowledge Transfer	Few-Shot Learning Setup				Open Access
Reference	Task	Modality	Method	Type of Pre-Training	Type of Transfer	FSL Dataset	Number of Classes	Support Set Size	Evaluation Metric	Code Availability
[106]	Motor Imagery	EEG	Adversarial	Supervised	Subject	BCI-IV 2a [139] BCI-IV 2b [139]	4 2	? ?	AC, CK	✗
[168]	Seizure	EEG	Adversarial	Unsupervised	Subject	[147]	2	{0,1,2,3}-shot	AUROC	✗
[162]	Seizure	EEG	SVM	-	N/A	CHB-MIT [123] Collected (Private)	2 2	1-shot 1-shot	SE, SP	✗
[80]	Emotion	EEG	Custom	Supervised	Classes	SEED-V [76]	5	5-shot	AC	✗
[101]	HAR	ACC, Video	Custom	Supervised, Episodic	Classes	[126] Stanford ECM [93]	5 5	{1,3}-shot {1,3}-shot	AC	✗
[41]	HAR	EMG, ACC	Custom	-	N/A	CSL-SHARE [74] UniMiB SHAR [85]	22 17	{1-500}-shot {1-500}-shot	AC	✗
[149]	Blood Pressure	PPG	SVM	-	N/A	Private (Collected)	Regression	{10,15,20,30,40,45}-shot	MAE, MPE, PC, SD	✗
[28]	Motor Imagery	EEG	Custom	Supervised	Classes	Private (collected)	3 (1 unseen)	0-shot	AC	✗
[111]	Hand Gesture	EMG	Custom	Supervised, Episodic	Repetitions Subjects Classes	Ninapro DB2 [107]	50 (5-way) 50 (5-way) 34 (5-way)	{1,5}-shot {1,5}-shot {1,5}-shot	AC	✗

All tasks, except for blood pressure prediction, are classification tasks. Abbreviations: Electroencephalogram (EEG), Accelerometry (ACC), Electromyography (EMG), Photoplethysmography (PPG), Accuracy (AC), Cohen's Kappa (CK), Area Under Receiver Operating Characteristic Curve (AUROC), Sensitivity (SE), Specificity (SP), Mean Absolute Error (MAE), Mean Percentage Error (MPE), Pearson Correlation (PC), Standard Deviation of Estimation Error (SD)

2) SVM: SVM is a class of supervised machine learning algorithms that attempts to find the optimal decision boundary or hyperplane that separates different classes within a given feature space. The optimal hyperplane maximizes the distance between the hyperplane and the closest data points to the hyperplane of each class, commonly referred to as support vectors. While SVM hyperplane can only handle linear decision boundaries, the use of "kernel trick" to project input features to high dimensional space enables the separation of non-linear patterns. The concept of SVM can be further expanded to regression problems, also known as support vector regression (SVR). Instead of treating the hyperplane as a decision boundary that maximizes the distance to data points of each class, the hyperplane serves to fit the data points. The "kernel trick" can be applied similarly to fit non-linear regression problems. In [162], Zhang et al. proposed an epilepsy system on a chip (SoC) consisting of a time-channel averaging feature extractor and a SVM classifier with a second-order polynomial guided kernel. One-shot learning in a patient recruited from a local hospital achieved a vector-based sensitivity of 39.5% and a specificity of 98.9% on seven seizures in four hours. Online tuning using three additional false negatives and one false positive as support vectors further boosted the vector-based sensitivity to 71.9%. In another application, a modified SVR algorithm was proposed for blood pressure regression using photoplethysmography (PPG) signals [149]. Namely, an error feedback model fine-tuning mechanism is incorporated into the SVR algorithm, which achieves good few-shot learning performance and can improve the long-term monitoring capability by adapting to new samples without model retraining. Using the proposed framework, a personalized blood pressure prediction model can be constructed with few samples and continue to model blood pressure variations after three months of constructing the model. SVM remains a popular choice for few-shot learning in wearable and edge devices due to its low memory & computation burden as well as its ease of online tuning with additional samples. While deep neural network-based few-shot learning methods achieve SOTA performance for cross-dataset, cross-task performance, SVM excels at building patient-specific models, especially when training from scratch is necessary due to a lack of access to external data sources to facilitate knowledge transfer. However, SVM relies on manual feature selection, which is laborious and also heavily influences the performance of the model.

3) Customized Models and Frameworks: In addition to existing models, many researchers have also explored novel, customized frameworks to achieve few-shot learning. Some focused on designing feature extraction and classifier modules that can generalize with few-shot samples and others re-imagined the classification framework in its entirety. For example, Pan et al. [101] explored graph networks for multi-modal egocentric activity recognition. To extract effective features and recognize new classes from few-shot multi-modal samples, a two-stream graph network was proposed. The heterogeneous graph-based multi-modal association module captures dynamic and complementary information from multi-modal data. The knowledge-aware classifier module captures semantic relations between activity classes and objects to assist generalization to unseen classes. An ablation study of different modules shows that the heterogeneous graph network is useful for fusing multi-modal information and the knowledge-aware classifier shows improvements over the linear classifier. Comparison against existing baselines with the likes of LSTM, MNs, and PNs further demonstrate the effectiveness of the approach. Incremental few-shot learning was explored by Ma et al. [80] for emotion recognition. The key idea behind the framework is to add new weights to the final linear classifier for each new class introduced by the incremental learning stages. During the base stage, the graph convolutional network (GCN) feature extractor and a linear classifier are pre-trained using abundant data from the base classes. During incremental stages, the GCN feature extractor and linear classifier weights from the previous stages are frozen. New linear classifier weights for incremental classes are added and fine-tuned with the support set. Entropy and subspace regularization were introduced to help the model learn from new samples and avoid catastrophic forgetting of old classes during the few-shot fine-tuning process. On a different note,

Duan et al. [28] explored zero-shot outlier detection based on local density. The proposed framework consists of two stages: projection to target space and novelty detection. In the first stage, a common spatial pattern feature extractor and a two-layer fully connected network are used to project EEG time series to semantic feature space. In the second stage, outlier classification is done in a zero-shot manner by determining whether the features belong to an unseen class, based on the feature manifold of seen classes. Without any information on the unseen class, the proposed method can achieve accuracy similar to that of a classifier trained with full access to all classes. Delving into an alternative perspective, Hartmann et al. [41] proposed a few-shot learning framework based on high-level features. Individual classes are defined through a combination of different high-level features, and separate classifiers are trained to provide binary classification for each of the high-level features. With the pre-trained classifiers for high-level features, zero-shot learning for new classes becomes a natural extension of the framework. An inherent advantage of the framework lies within its model interpretability and ease of performing error attribution analysis. However, this comes at the cost of time and domain knowledge required to define such high-level features. The extension of zero-shot learning beyond the finite combination of high-level features presents another shortcoming of the framework. In another investigation, Rahimian et al. [111] proposed an architecture for processing support and query samples similar to language models, where the input of the model receives concatenated support samples with labels followed by query samples with null labels. The network consists of attention and dilated causal 1D convolution modules to process the relationship between inputs. Extensive experiments show that the network can generalize to new classes and new repetitions, although there is a lack of comparison against existing baselines and few-shot methods.

4) Summary: Model-based methods approach the design of model architecture from creative perspectives. However, many model-based frameworks make strong assumptions about problem setup, which does not hold when applied to other applications. Adversarial networks excel at learning domain- and subject-invariant features, which often demand pre-training. SVMs, on the other hand, are capable of learning from few-shot samples directly. Their training and inference are computationally efficient, rendering them ideal choices for edge and wearable devices. Customized frameworks provide additional functionalities such as incremental learning, multi-modal learning, and outlier detection, which more closely mimics how humans learn.

D. Optimization-Based Few-Shot Learning

Optimization-based few-shot learning methods focus on developing innovative ways to train models. The non-convex nature of deep neural networks means that there are many local optimums that the model weights can converge to. Optimization-based methods focus on designing the optimization objective and training procedure to guide the model to converge to weight spaces that generalize well to new tasks when fine-tuned with few-shot samples. These methods are closely associated with meta-learning, learning how to learn.

1) Transfer Learning: Transfer learning (TL) aims to reuse a model trained on one task and adapt it to a related task. Typically, a model is pre-trained on the source task and subsequently fine-tuned on the target task with a relatively smaller dataset. Model-level transfer refers to fine-tuning the entirety of the model, and it assumes that the input and output structures of the source and target tasks are identical. Feature-level transfer refers to freezing feature extractors and fine-tuning task-specific heads. Pre-training on large source datasets helps the model to learn meaningful and generalizable features across tasks, providing an initialization to boost performance on the target task. Feature-level TL is a popular method amongst few-shot learning applications to biomedical time series. Meyer et al. [84] applied TL to a 1D UNet for model personalization. To adapt to new subjects, the pre-trained encoder and decoder are frozen while the classifier with fully connected layers is re-initialized and trained using a few samples from the new subject. Training using only datasets from the source subject, the proposed 1D UNet outperforms the baseline SVM on the targeted subject test set. Its performance is further boosted when fine-tuned with few-shot samples from the target subject in question. Nazari et al. [95] studied the choice of classifier in the performance of few-shot TL for seizure detection. After pre-training the 2D CNN backbone and classifier using data from 15 source subjects, fine-tuning the SVM classifier led to a superior performance on the target subject than using a fully connected layer as a classifier. Bhaskarpandit et al. [11] provided a benchmark on the number of samples needed for TL to be effective when applied to myocardial infarction classification. A variety of pre-trained backbone networks were studied, including VGG16 [124], DenseNet [48], InceptionV3 [135], ResNet [42], and EfficientNet [137]. Unsurprisingly, as the number of support samples for fine-tuning increased, the few-shot learning performance improved. With around 5 or fewer support samples, TL fails, but with around 20 support samples, performance surpasses 99% accuracy on [62] and the marginal benefit of additional samples diminishes. Other works have attempted to modify the pre-training process to facilitate fine-tuning. Lv et al. [79] studied two types of pre-training strategies for TL: cross-subject mixture transfer and model averaging transfer. Cross-subject mixture transfer mixes all source subjects' data for pre-training, while model averaging transfer averages the weights of models individually trained on each source subject's data. Experiments were carried out with a 2D CNN consisting of two convolution layers and a fully connected layer. The results show that cross-subject mixture transfer is useful and provides the best performance boost when the fully connected layers are fine-tuned, which agrees with the established norm of fine-tuning only the classifier and not the feature extractors. On the other hand, model averaging transfer provides worse performance than cross-subject mixture transfer, despite its similarity to the FedAvg [83] algorithm, widely adopted as a baseline for federated learning. Hur et al. [50] proposed to extend TL to a multi-task, multi-source setting and pre-train the source model

in a self-supervised manner. To overcome the heterogeneity of medical codes and schemes of EHRs used by different hospitals and datasets, the authors propose to convert EHRs to text representation, thus eliminating the pre-processing and feature engineering procedure for each hospital. Then, multi-source pre-training with unlabeled data is performed using different self-supervised learning strategies: SimCLR [21], Wav2Vec 2.0 [8], MLM [26] and SpanMLM [59]. Few-shot samples from the target hospital or dataset enable fine-tuning of the multi-source model, adapting to any hospital without restraints on the data format. The authors found that TL with self-supervised pre-training methods outperforms training from scratch, except for SpanMLM. Furthermore, SimCLR is the best self-supervised pre-training method, which focuses on learning patient-level representations, as opposed to event-level or token-level representations with other self-supervised pre-training strategies. In addition, multi-source pre-training was found to yield better TL performance than single-source pre-training. Similarly, Saeed et al. [114] also applied self-supervised pre-training to facilitate TL. The authors introduce a variety of self-supervised pre-training tasks to encode underlying signal semantics of multi-modal time series from different sensor modalities. Examples of self-supervised pre-training tasks include differentiating between clean and blended signals, feature prediction from masked windows, and signal transformation recognition. Experimental results show that self-supervised networks serve as great initialization to boost performance with limited labeled data. For some datasets, self-supervised pre-training can outperform a fully supervised model trained with the entire dataset. Moreover, learned representations from self-supervised training are highly transferable across related datasets.

2) MAML: Model-Agnostic Meta-Learning (MAML) [32] was first proposed by Finn et al. in 2017 as a framework to initialize model parameters that are easy to fine-tune. The problem consists of a set of T tasks $\mathcal{T} = \{\mathcal{T}_t\}_{t=1}^{T}$ and corresponding set of datasets $\mathcal{D} = \{\mathcal{D}_t\}_{t=1}^{T}$ with each $\mathcal{D}_t = \{(\mathbf{x}_l, y_l)\}_{l=1}^{L_t}$ consisting of L_t input-label pairs. MAML aims to find an initialization for model parameters θ that can be adapted quickly to perform well on any arbitrarily chosen task from $\mathcal{T}$, after fine-tuning with few-shot samples from its corresponding dataset from $\mathcal{D}$. The pseudo-code for the MAML algorithm is outlined in Algorithm 2.

In general, MAML consists of two weight update computations. MAML starts by sampling a batch of B tasks from the distribution on all tasks $p(\mathcal{T})$. For the t-th sampled task $\mathcal{T}_t$, the inner loop computes the adapted parameters using a small number of support samples $\mathcal{D}_t^{inner} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{I}$ drawn from $\mathcal{D}_t$ where $(\mathbf{x}_i, y_i)$ are the i-th input label pair of $\mathcal{D}_t^{inner}$. The adapted parameter for each task $\mathcal{T}_t$ can be computed as:

$$\theta_t' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_t}(f_\theta)$$

Although the formulation in the algorithm indicates that the adapted parameters involve a single gradient update step, the adapted model parameters can be obtained after taking several gradient steps in the inner loop. The outer loop then updates the model parameters using another set of small samples $\mathcal{D}_t^{outer} = \{(\mathbf{x}_j, y_j)\}_{j=1}^{J}$ drawn from $\mathcal{D}_t$ where $(\mathbf{x}_j, y_j)$ are the j-th input label pair of $\mathcal{D}_t^{outer}$, to enhance the model's generalization ability across all tasks. This is also known as the meta-update step, which updates the model using the gradient over the loss of the adapted parameters across all sampled tasks. The meta-update can be written as:

$$\theta \leftarrow \theta - \beta \nabla_\theta \sum_{t=1}^{B} \mathcal{L}_{\mathcal{T}_t}(f_{\theta_t'})$$

Note that the loss of the inner loop weight update is computed using the meta-model θ on each of the sampled datasets $\mathcal{D}_t^{inner}$ for each task $\mathcal{T}_t$, while the outer loop weight update is computed using the adapted model parameters θ_t' on another sampled dataset $\mathcal{D}_t^{outer}$ over all sampled tasks. Since the inner loop computation of θ_t' involves gradients over θ, the outer loop computation requires second-order gradients. To reduce the computational complexity, first-order model-agnostic meta-learning (FOMAML) that omits the second-order gradient computations was found to achieve nearly the same performance as MAML. Drawing inspiration from FOMAML, Reptile [98] proposes to use the difference in parameter weights between the adapted parameters and meta-parameters as the gradient for meta-update. Theoretical analysis and experimental results show that first-order meta-learning algorithms can achieve similar performance as MAML with full second-order gradients. MAML and Reptile have been directly applied to several biomedical time series applications. Early efforts by Lan et al. [67] proposed a spatial-temporal gaze graph-based model for video-oculography (VOG) and trained it using the MAML paradigm for quick adaptation to few-shot samples from new subjects. In MAML terminology, the authors treat each task as a randomly sampled subset of the source dataset, consisting of N-way-K-shot support and query sets that contain disjoint samples. This trains the model in a way that is easy to fine-tune and adapt to new datasets from new subjects, such that when it is fine-tuned on an unseen support set, it can achieve good performance on the unseen query set. Experimental results show that training the

Algorithm 2 Algorithm for MAML [32]

Require: $\mathcal{T} = \{\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_T\}$ set of tasks
Require: $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_T\}$ set of corresponding datasets
Require: $p(\mathcal{T})$ distribution over tasks
Require: α, β step size hyperparameters
Require: I, J, B batch size hyperparameters
randomly initialize θ
while not done **do**
  Sample B tasks from $\mathcal{T}$ following $p(\mathcal{T})$
  for $t \leftarrow 1$ to B **do**
    Sample $\mathcal{D}_t^{inner} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{I}$ from $\mathcal{D}_t$
    Evaluate $\mathcal{L}_{\mathcal{T}_t}$ and $\nabla_\theta \mathcal{L}_{\mathcal{T}_t}(f_\theta)$ using $\mathcal{D}_t^{inner}$
    Compute adapted parameters with gradient descent: $\theta_t' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_t}(f_\theta)$
    Sample $\mathcal{D}_t^{outer} = \{(\mathbf{x}_j, y_j)\}_{j=1}^{J}$ from $\mathcal{D}_t$
  end for
  Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{t=1}^{B} \mathcal{L}_{\mathcal{T}_t}(f_{\theta_t'})$ using $\mathcal{D}_t^{outer}$
end while

TABLE IV
OPTIMIZATION-BASED FEW-SHOT LEARNING METHODS FOR BIOMEDICAL TIME SERIES

	Problem			Pre-Training	Knowledge Transfer	Few-Shot Learning Setup				Open Access
Reference	Task	Modality	Method	Type of Pre-Training	Type of Transfer	FSL Dataset	Number of Classes	Support Set Size	Evaluation Metric	Code Availability
[156]	Motor Imagery	EEG	MAML	Supervised, Episodic	Subjects	BCI-IV 2a [139] [22] PhysionetMI [117]	4 2 4	{1,5}-shot {1,5}-shot {1,5}-shot	AC	✗
[67]	Gait Impairment	VOG	MAML	Supervised, Episodic	Subjects	SedentaryActivity [130] JapaneseDocument [66] Collected [67]	6 8 5	{5,10}-shot {5,10}-shot {5,10}-shot	F1, AC	✗
[71]	Motor Imagery Motor Imagery ERP ERP	EEG	MAML	Supervised, Episodic	Subjects	BCI-IV 2a [139] BNCI-Horizon 002-2014 [139] BNCI-Horizon 009-2014 [5] BCI Challenge [81]	4 2 2 2	{0,2,4,6}-shot {0,2,4,6}-shot {0,2,4,6}-shot {0,2,4,6}-shot	AUROC, AC	✓
[166]	Clinical Risk	EHR	MAML	Supervised, Episodic	Tasks	OHSU Hospital (Private)	2	20%	AUROC, F1	✓
[161]	Blood Pressure	EHR	MAML	Supervised, Episodic	Subjects	MIMIC-III [58]	Regression	10-shot	RMSE, MAE	✓
[84]	HAR	IMU, VOG	TL	Supervised, Episodic	Subjects	Collected (Private)	7	3-shot	AC, Macro-F1	✗
[79]	Gait	EEG, EMG	TL	Supervised	Subject	Collected (Private)	4	20%	AC	✗
[95]	Seizure	EEG	TL	Supervised	Subjects	CHB-MIT [122]	2	3-shot	SE, FPR, AUROC	✗
[50]	ICU Outcome	EHR	TL	Semi-Supervised	Dataset	MIMIC-III [58] eICU [108] MIMIC-IV [57]	Varying Varying Varying	{0,10,30,50,70}% {0,10,30,50,70}% {0,10,30,50,70}%	AUROC	✓
[11]	Myocardial Infarction	ECG	TL	Supervised	Tasks	[62]	3	{0,10,30,50,70}% {0,10,30,50,70}% {0,10,30,50,70}%	AC	✗
[114]	HAR HAR HAR HAR HAR Sleep Staging Stress	ACC, GYR ACC, GYR ACC, GYR ACC, GYR ACC, GYR EEG, EOG HR, SC	TL	Self-Supervised	Datasets	HHAR MobiAct MotionSense UCI HAR HAPT Sleep-EDF MIT DriverDb	6 11 6 6 12 5 2	{5,10}-shot {5,10}-shot {5,10}-shot {5,10}-shot {5,10}-shot {5,10}-shot {5,10}-shot	F1	✗

All tasks, except for blood pressure prediction, are classification tasks. Abbreviations: Electroencephalogram (EEG), Video-Oculograph (VOG), Electronic Health Record (EHR), Inertial Measurement Unit (IMU), Accelerometry (ACC), Gyroscope (GYR), Electrooculography (EOG), Heart Rate (HR), Skin Conductance (SK), Accuracy (AC), Area Under Receiver Operating Characteristic Curve (AUROC), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Sensitivity (SE), False Positive Rate (FPR).

proposed gaze graph model using MAML enables superior adaptation performance to the few-shot support and query set than a model trained using TL. Similarly, Youssef et al. [161] benchmarked MAML against TL for blood pressure prediction. Unlike [67], each task consists of data from a specific subject. In particular, the authors found that given a fixed model architecture and pre-training dataset, pre-training using MAML and pre-training using a traditional approach achieves the same performance on an unseen target subject without fine-tuning. However, after fine-tuning both models with few samples from the target subject, MAML pre-trained model significantly outperforms the traditional pre-trained model. This reiterates the fact that MAML does not necessarily pre-train a model that achieves good performance on all tasks it was pre-trained on, but rather a model that is easy to fine-tune and adapt quickly to new tasks. This is particularly important for applications like model personalization to address the challenges associated with inter-subject variability. In contrast, Wu et al. [156] applied Reptile to pre-train EEGNet [68] and found no improvement over pre-training EEGNet using traditional one-step model parameter update. The authors hypothesize that it is possible since biomedical time series signals such as EEG have fewer samples and higher variance than computer vision and natural language processing datasets, which makes it more challenging to learn representations that are transferable between subjects. A mere application of meta-learning algorithms would not suffice for all biomedical time series applications. Several works have also proposed to modify the MAML framework for better adaptation to biomedical applications. Li et al. [71] proposed a modified version of MAML, where the inner loop takes a few gradient steps based on data from one subject and calculates the meta-loss based on another. The inner loop is repeated for randomly paired subjects from the entire dataset before a meta-step is taken in the outer loop using the average of meta-loss from the inner loop. The modification of MAML to take into account differences across subjects enables better few-shot adaptation to new subjects compared to existing baselines like PN [125] and the original MAML [32]. Furthermore, models trained using the proposed MAML achieve better model calibration than models trained using the original MAML, where the model's predicted probabilities better reflect the true likelihood of the model's prediction. Zhang et al. [166] proposed another variation of MAML. In the computation of gradients for meta-update, the authors propose to incorporate the inner loop loss for the source tasks, in addition to the loss on target task observed in traditional MAML. The experimental results show that the proposed modification of MAML outperforms the original MAML, and the authors conclude that the incorporation of supervised knowledge from the source domain enhances meta-learning knowledge transfer.

3) Summary: Optimization-based methods are highly flexible in terms of the model architecture, which allows them to be wrapped around any existing model. On the other hand, optimization-based methods rely heavily on access to extensive external pre-training datasets. The pre-training process can be computationally heavy and hyper-parameter sensitive, especially for meta-learning methods that involve inner- and outer-loop optimizations. When choosing between TL and MAML, it is important to consider the availability of computational resources. TL is less computationally heavy during pre-training while MAML requires less fine-tuning computation. Moreover,

TL works well when the pre-training and target few-shot tasks are similar while meta-learning would be more beneficial when adapting to a multitude of few-shot tasks.

E. Hybrid Methods

Hybrid methods to be presented in this section combine the aforementioned ideas to further improve few-shot learning performance. A popular choice is to train MBMs through meta-learning optimization algorithms. Li et al. [70] combined the idea of MAML with PNs. Treating each task of MAML as a dataset from a specific subject, the authors use MAML to pre-train a backbone network and calculate prototypes for each class using the support set. Then, the prototypes guide the selection and pseudo-labeling process of unlabeled target-subject data, to provide additional samples for fine-tuning and adapting to the target subject. Experiments were conducted in three BCI applications and demonstrated improvements over the baseline models and MAML itself. The authors conclude that the meta-learned class centers are useful for guiding the artificial labeling and selection of unlabeled data for improved semi-supervised adaption to target subjects. Narwariya et al. [94] proposed to train a MN with the reptile meta-learning framework. The proposed framework combines the Reptile meta-learning approach with the triplet loss function to pre-train a residual neural network. Each task of the reptile is treated as a dataset from various domains, such as healthcare and activity recognition. The main advantage of MNs and MBMs is that it is not restrained to the classes they are pre-trained on, and Reptile helps to pre-train a MN that is easy to fine-tune to new tasks with new classes. Therefore, the reptile-based cross-domain meta-learning enables the pre-trained model to quickly adapt to unseen domains, as well as adapt to tasks with varying numbers of classes without introducing additional class parameters. Along the same lines of combining MBMs and meta-learning, Liu et al. [75] took inspiration from [141] to combine PNs with TL. Triantafillou et al. [141] observed that a prototypical network with a Euclidean distance measure can be rewritten as a linear classifier for the embedding of the query.

$$\begin{aligned} -\|f_\theta(\mathbf{x}_{q,j}) - \mathbf{c}_n\|^2 &= -f_\theta(\mathbf{x}_{q,j})^T f_\theta(\mathbf{x}_{q,j}) + 2\mathbf{c}_n^T f_\theta(\mathbf{x}_{q,j}) - \mathbf{c}_n^T \mathbf{c}_n \\ &= 2\mathbf{c}_n^T f_\theta(\mathbf{x}_{q,j}) - \|\mathbf{c}_n\|_2^2 + constant \end{aligned}$$

where $constant$ is a scalar that does not affect the prediction, $f_\theta(\mathbf{x}_{q,j})$ is the embedding of the query sample $\mathbf{x}_{q,j}$, and $\|\mathbf{c}_n\|_2$ is the Euclidean norm of the prototype of class n. Therefore, the prototypical network classifier can be rewritten as a linear layer as a function of $f_\theta(\mathbf{x}_{q,j})$, with the n-th neuron having weights $\mathbf{W}_n = 2\mathbf{c}_n$ and biases $\mathbf{b}_n = -\|\mathbf{c}_n\|^2$. Liu et al. pre-trained the EN of the prototypical network using the source dataset, followed by initializing the linear layer using the pre-trained class prototypes for fine-tuning. The proposed paradigm was found to outperform existing few-shot learning baselines, including PNs, MNs, relation networks, and MAML. In a different study, Liu et al. [77] chose to apply TL to SNs. The proposed Siamese autoencoder is first trained in a semi-supervised manner, through a combination of mean squared error (MSE) for signal reconstruction and cross-entropy for two individual classifiers: side predictor (left or right ear) and class predictor (normal or tinnitus). The pre-trained encoder is then fine-tuned in an unsupervised manner with the side predictor classifier. The advantage of this framework is the automatic calibration process that does not require annotation of few-shot samples, as fine-tuning involves the use of information from the left/right ear, which can be obtained directly from data collection. With a different stream of thought, Akbari et al. [2] combined the idea of model-based methods and meta-learning. The authors propose an encoder-decoder network, where the encoder is shared between subjects and the decoder is personalized. While the encoder weights are trained through backpropagation, the decoder weights are generated by a weight calibration network. The weight calibration network is a meta-learning module that receives the support set as input and generates the weights and biases for a linear layer. The authors claim that with limited samples, the weight calibration network can update the weights of the decoder better than backpropagation-based fine-tuning, which tends to overfit without sufficient data. Furthermore, personalization with a weight-calibration network only requires a forward pass to update the decoder weights, which is particularly appealing for low-power edge devices. However, the weight calibration network is unstable, in the sense that the weight generation process is sensitive to any small perturbations of the support set. Thus, it is paramount to pre-train a weight calibration network that can generalize well across different distributions of support sets.

As discussed in Section IV-A3, augmenting the dataset with synthetic samples using (DBMs) alone is insufficient to overcome the unique challenges associated with biomedical time series, such as inter-patient variability. Thus, a few works take a step further and design more complex networks to complement DA. For example, Aldahr et al. [3] proposed to augment SNs with GAN DA. The framework consists of three components: diversity-enhanced DA with GAN, temporal-aware feature extraction with graph theory, and SN. The GAN model is enhanced with a diversity factor to avoid generating redundant samples by performing a diversity analysis between synthetic & real samples and between synthetic samples [104, 31]. The graph theory method transforms EEG samples into complex weighted graphs, which capture intrinsic hidden patterns and information with minimal information loss. Finally, the SN learns inter-patient seizure patterns through pairs of similar and dissimilar samples from different subjects. Experimental results show that the proposed framework addresses data scarcity and inter-patient variability problems of EEG for seizure detection, outperforming existing methods with less data. Similarly, Kim et al. [63] combined the DA technique SMOTE with a custom-designed model consisting of ResNet and bi-directional LSTM. SMOTE was applied to address the problem of class imbalance during pre-training by generating synthetic samples via interpolation between samples of minority classes. An ablation study demonstrates the usefulness of each module when trained on the full dataset. Compared to ShallowConvNet [118] and DeepConvNet [146], the proposed model achieves better few-shot adaptation for cardiac arrhyth-

TABLE V
HYBRID FEW-SHOT LEARNING METHODS FOR BIOMEDICAL TIME SERIES

	Problem			Pre-Training	Knowledge Transfer	Few-Shot Learning Setup				Open Access
Reference	Task	Modality	Method	Type of Pre-Training	Type of Transfer	FSL Dataset	Number of Classes	Support Set Size	Evaluation Metric	Code Availability
[2]	Modality Translation	ECG, Bio-Z	Model + Optimization	Supervised, Episodic	Subjects	Private	N/A	{2,5}-shot	PRD, WDD, PC	✗
[70]	ERP Emotion Sleep Staging	EEG	Metric + Optimization	Supervised, Episodic	Subjects	ERP [45] SEED [29] SleepEDF [61]	N/A	5-shot 10-shot 10-shot	AC	✗
[3]	Seizure	EEG	Data + Metric	Supervised	Subjets	Bonn [4] CHBMIT [122]	3 3	60%	RE, SP, AC, AUROC	✗
[63]	Arrhythmia	ECG	Data + Model	Supervised	Dataset	MIT-BIH [36] CinC DB [23]	4 2	{0,5,10,20}-shot {0,5,10,20}-shot	PR, RE, SP, F1, GM	✗
[89]	Gait	Pressure	Metric + Model	Semi-Supervised	Subjects	Collected [89]	2	10-shot	TNR, TPR, AC	✗
[75]	Arrhythmia	ECG	Metric + Optimization	Supervised	Dataset	MIT-BIH [36] PTB [13]	2, 4 2, 4	{1,5,10}-shot {1,5,10}-shot	AC	✗
[94]	Arrhythmia	ECG	Metric + Optimization	Supervised, Episodic	Tasks	UCR ECG200 [25] UCR ECG5000 [25] UCR ECGFiveDays [25]	2 5 2	{2,5,10,20}-shot {2,5,10,20}-shot {2,5,10,20}-shot	AC	✗
[77]	Tinnitus	EEG	Metric + Optimization	Semi-Supervised, Episodic	Dataset	[116] [39]	2	2 per subject	F1, AC	✗

All tasks, except modality translation, are classification tasks. Abbreviations: Electrocardiogram (ECG), Bio-Impedance (Bio-Z), Electroencephalogram (EEG), Percentage Root Mean Square Difference (PRD), Weighted Diagnostic Distortion (WDD), Pearson Correlation (PC), Accuracy (AC), Recall (RE), Specificity (SP), G-Mean (GM), True Negative Rate (TNR), True Positive Rate (TPR)

mia classification datasets, but the results were not desirable for clinical application. While the aforementioned works have been restricted to the task of classification, Moon et al. [89] explored few-shot open set recognition through a prototypical network with SVM classifier. Open set recognition differs from traditional closed set recognition problems, where test samples belong to the same training classes for closed set recognition problems and test samples can contain different classes than the training dataset. To do so, the authors propose a two-stage network. First, an encoder-decoder is trained in a self-supervised manner using triplet and reconstructed prototype loss to minimize the feature embedding distance of inputs from the same subject and maximize the distance of inputs from different subjects. Second, the pre-trained encoder is frozen and few-shot labeled samples are forward passed to construct support vectors for each subject and a decision boundary for a one-class SVM is extracted for test predictions. If the embedding of the query sample falls outside the decision boundary of all training classes, then it is considered unrecognized.

1) Summary: Hybrid methods aim to synergize strengths or complement weaknesses of individual methods. They provide performance gains compared to existing few-shot learning baselines. This also increases the complexity of the system, which reduces interpretability and the ease of error attribution.

V. DISCUSSION

A. Challenges and Potential Solutions

1) Lack of Benchmarks and Standardized Evaluation Metrics: While many few-shot learning methods have been proposed for biomedical time series, it is rather difficult to compare them. A major obstacle to fair comparison is that proposed methods address different clinical applications using different datasets that are often privately collected. Furthermore, most works compare the proposed method against baseline networks that are not designed for few-shot learning. Although it is clear that few-shot learning methods outperform baseline methods not designed for learning from a few samples, it is unclear which few-shot learning architecture performs better. The differences in experimental setups further prevent fair comparison amongst methods. Some methods involve pre-training on external data sources, while others do not. Furthermore, while most work report performance using the N-way-K-shot paradigm, N and K differ significantly across applications and datasets and can even be different when using the same dataset. Furthermore, the N-way-K-shot support and query sets are usually artificially sampled from a larger dataset. Due to the small sample size, the evaluation results can differ significantly depending on which samples are drawn. When support and query sets are artificially generated, it is important to repeat the sampling across multiple trials and report the mean and standard deviation of the results, in order to showcase how sensitive the methods are to different support set samples. Moreover, many studies have found that, in general, as the number of support set samples available increases, few-shot learning performance increases and plateaus. Conducting such ablation is helpful to the community to better understand how the newly proposed and existing methodologies would fare at different levels of data scarcity. This calls for a need to develop standardized few-shot learning datasets to provide a platform to compare and benchmark different few-shot learning methods in a fair way.

2) How to Choose Pre-Training Dataset: For some types of few-shot learning methods, such as optimization-based methods, there is a heavy reliance on pre-training the model with a pre-training dataset. However, there is limited research on how the choice of pre-training datasets impacts knowledge transfer and generalization to the target few-shot task. Often, it is preferred to pre-train using datasets similar to the problem or application at hand, but access to such datasets can be difficult, especially if it is an emerging application or niche field. Given the availability of diverse physiological datasets from databases like PhysioNet [35], perhaps it would be preferable to pre-train using a range of datasets across different applications to help the few-shot learning method learn generalizable knowledge. Guidelines on choosing datasets for pre-training could offer insights into the impact of task similarity and the number of pre-training tasks on few-shot knowledge transfer, providing direction for practitioners navigating the

complexities of implementing few-shot learning algorithms.

B. Emerging and Promising Fields of Few-Shot Learning

1) Foundation Models: Foundation models are general-purpose models that have been trained on data from a variety of domains. LLM, as an example of a foundation model for text, has transformed AI research. As the amount of training data and the size of LLM continue to increase, LLM demonstrates incredible generalization abilities, along with other emerging capabilities. A surprising example is in-context learning, where the model is asked to perform a previously unseen task based on demonstrations as part of the input prompt. LLM is capable of identifying patterns from the prompt and using the next token prediction to generalize to unseen NLP tasks [16]. More recently, similar ideas were adopted for time series data. By converting time series into a text representation, LLM can achieve a forecasting performance similar to or exceeding that of purpose-built forecasting methods [56, 38]. With recent meta-learning attempts to design better prompts for few-shot in-context learning [46], the idea can also be embraced to optimize conversion from time series to text representation and prompts. With the success of LLM, recent efforts have also sparked much interest in developing foundation models for time series. For example, LaBraM [54] drew inspiration from LLM and developed a Large EEG Model. Although LaBraM is relatively small compared to large vision and language models with only 5.8 to 369 million parameters, LaBraM achieves SOTA performance in a variety of downstream tasks after fine-tuning. More recently, MOIRAI [155] was proposed as a universal time series forecasting foundation model that has been trained on 27 billion observations. Similarly to LLMs, MOIRAI also demonstrates strong out-of-distribution generalization capabilities, achieving competitive or superior zero-shot forecasting performance than SOTA full-shot models. Leveraging the powerful emergent capabilities of foundation models presents many exciting research directions for few-shot learning in the time series domain.

2) Multi-Modality: The vast majority of current methods focus on a singular modality, with a limited number of methods that target multimodal data for recognition of human activity. Rapid advancements in wearable technologies [119, 158] provide means to improve the accessibility of multimodal data sets in a variety of applications. Research on how to fuse multi-modal information effectively and how to learn with unaligned data pave the way for more robust few-shot learning systems that take advantage of complementary information sources. In addition to combining different time series sensors (e.g. EEG and VOG), multi-modality can also combine different data modalities (e.g. time series and text) to complement each other. A notable example is CaFo [163], a Cascade of Foundation models that leverages diverse prior knowledge of varying pre-trained foundation models to improve few-shot learning. The authors unify the domain knowledge of CLIP [110], DINO [19], DALL-E [112] & GPT-3 [16] through a "Prompt, Generate, then Cache" pipeline and demonstrated state-of-the-art few-shot learning performance on 11 datasets. Intuitively, triangulating across data modalities encourages the model to learn more discriminative and informative representations, encouraging collaboration of pre-trained knowledge from different modalities to improve few-shot learning performance.

3) Neural Adaptive Processes: While existing few-shot learning methods for biomedical time series focus on adapting model parameters through gradient computations, adapting network parameters using a weight adaptation network has received increasing interest [113]. CNAPs introduced the idea of using an adaptation network that modulates classifier parameters using the support set as input. The idea of hypernetworks, networks that generate weights for classifiers, has also shown promising results for few-shot learning [120]. Since backpropagation and gradient computations are not required for adaptation to new tasks, such approaches are computationally efficient, paving the way for personalized continual learning models on wearable devices.

4) Beyond Classification: Existing exploration of few-shot learning for biomedical time series has focused on classification and regression tasks. However, few-shot learning for other tasks such as forecasting [52, 14, 157], anomaly detection [78, 102, 27], and clustering [145, 47] remain largely unexplored for biomedical time series. This phenomenon can be attributed to a few factors. First, classification tasks are well-studied problems, and most benchmarks in computer vision, natural language processing (NLP), and biomedical time series focus on classification tasks. Few-shot learning naturally builds on these benchmarks and focuses on classification. Furthermore, many tasks are fundamentally more challenging than classification. For example, regression and forecasting tasks require the prediction of continuous values as opposed to discrete targets. Few-shot data might not provide sufficient samples to help the model capture the input-output relationship to generalize effectively. Furthermore, some few-shot learning methods cannot be easily extended to other tasks. For example, metric-learning methods that rely on comparison between classes fail when data labels are no longer discrete. Despite these challenges, all of these tasks form important foundations for the development of robust AI systems and warrant further exploration.

VI. Conclusion

This paper provides an introduction to the concepts of few-shot learning followed by a comprehensive review of different learning methods for biomedical time series applications. The methodologies were categorized based on how the few-shot learning algorithms modify traditional deep learning training pipelines to overcome the shortage of labeled data, namely data-based, model-based, metric-based, optimization-based, and hybrid methods. The clinical advantages and disadvantages of each method were discussed with respect to traditional deep learning methods. Visual illustrations and tables summarize the key characteristics of concepts and relevant works, presenting concise information for practitioners to choose the most appropriate few-shot learning methods for their custom applications. Trends, challenges, and potential directions for future work were discussed, highlighting exciting areas and opportunities for the research community.

ACKNOLWEDGEMENTS

Tim Denison was funded by Royal Academy of Engineering and supported by the NIHR Oxford Health Biomedical Research Centre (NIHR203316). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. Tingting Zhu was supported by the Royal Academy of Engineering under the Research Fellowship scheme. Chenqi Li is supported by the Cyril and Phillis Long Scholarship at The Queen's College in partnership with the Clarendon Fund.

APPENDIX A
DETAILED SEARCH STRATEGY

We identify and screen relevant literature using a variety of digital libraries, including ACM, PubMed, and Google Scholar. Since few-shot learning is an emerging field, especially with application to biomedical time series, no date-of-publication constraint was enforced. The search key was designed to capture three important characteristics of the relevant literature: "few shot learning", "biomedical", and "time series". In this work, biomedical time series are defined as time-sequential biological data of varying modalities to improve healthcare or assist clinical practices, such as diagnosis or risk prediction. The final search key uses a combination of synonyms for each characteristic:

1) "shot learning"
2) "clinical" OR "medical" OR "biomedical" OR "physiological"
3) "EHR" OR "Electronic Health Records" OR "EEG" OR "Electroencephalogram" OR "ECG" OR "Electrocardiogram" OR "PPG" OR "Photoplethysmography" OR "EMG" OR "Electromyogram" OR "Electrooculography" OR "EOG" OR "Electrodermal Activity" OR "Actigraph" OR "Accelerometer" OR "Gyroscope" OR "IMU" OR "Inertial Measurement Unit" OR "Gait" OR "Watch" OR "Polysomnography" OR "Voice" OR "Wearables"

ACM and PubMed returned a total of 125 and 10 articles, respectively and were all included for subsequent screening. Google Scholar returned 5,500 search results ranked in descending order of relevance. 580 articles were included for subsequent screening, after which no relevant search result has appeared over two consecutive pages. Articles from all searched databases are aggregated to remove duplicates and pre-prints. The title and abstract screening excluded records not related to few-shot learning or application to biomedical time series, but have qualified for the search criteria due to the appearance of keywords in references, future work, or background sections. Full-text screening excluded records involving biomedical images or text modalities, as existing reviews [65, 34, 33] have extensively covered them. A total of 55 records are included in this review.

REFERENCES

[1] J. H. Ahn et al. "Automatic Stridor Detection Using Small Training Set via Patch-Wise Few-Shot Learning for Diagnosis of Multiple System Atrophy". In: *Scientific Reports* 13.1 (2023), p. 10899. DOI: 10.1038/s41598-023-37620-0.

[2] A. Akbari, J. Martinez, and R. Jafari. "A Meta-Learning Approach for Fast Personalization of Modality Translation Models in Wearable Physiological Sensing". In: *IEEE journal of biomedical and health informatics* 26.4 (2021), pp. 1516–1527.

[3] R. S. Aldahr, M. Alanazi, and M. Ilyas. "Addressing Inter-Patient Variability in EEG: Diversity-Enhanced Data Augmentation and Few-Shot Learning-based Epilepsy Detection". In: *2022 International Conference on Healthcare Engineering (ICHE)*. 2022, pp. 1–7.

[4] R. G. Andrzejak et al. "Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state". In: *Physical Review E* 64.6 (2001), p. 061907.

[5] P. Aricò et al. "Influence of P300 latency jitter on event related potential-based brain–computer interface performance". In: *Journal of neural engineering* 11.3 (2014), p. 035008.

[6] M. Arjovsky and L. Bottou. "Towards principled methods for training generative adversarial networks". In: *arXiv preprint arXiv:1701.04862* (2017).

[7] M. Arjovsky, S. Chintala, and L. Bottou. "Wasserstein generative adversarial networks". In: *International conference on machine learning*. PMLR. 2017, pp. 214–223.

[8] A. Baevski et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations". In: *Advances in neural information processing systems* 33 (2020), pp. 12449–12460.

[9] W. G. C. Bandara and V. M. Patel. "A transformer-based siamese network for change detection". In: *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. 2022, pp. 207–210.

[10] L. Bertinetto et al. "Fully-convolutional siamese networks for object tracking". In: *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14*. Springer. 2016, pp. 850–865.

[11] S. Bhaskarpandit. "How Much Data Is Enough? Benchmarking Transfer Learning for Few Shot ECG Image Classification". In: (2023).

[12] S. Bhosale, R. Chakraborty, and S. K. Kopparapu. "Calibration Free Meta Learning Based Approach for Subject Independent EEG Emotion Recognition". In: *Biomedical Signal Processing and Control* 72 (2022), p. 103289.

[13] R. Bousseljot, D. Kreiseler, and A. Schnabel. "Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet". In: (1995).

[14] L. Brinkmeyer et al. "Few-shot forecasting of time-series with heterogeneous channels". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2022, pp. 3–18.

[15] J. Bromley et al. "Signature verification using a" siamese" time delay neural network". In: *Advances in neural information processing systems* 6 (1993).

[16] T. Brown et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[17] A. Burrello et al. "Hyperdimensional Computing with Local Binary Patterns: One-shot Learning of Seizure Onset and Identification of Ictogenic Brain Regions Using Short-Time iEEG Recordings". In: *IEEE Transactions on Biomedical Engineering* 67.2 (2019), pp. 601–613.

[18] D. Byrne et al. "Residential wearable RSSI and accelerometer measurements with detailed location annotations". In: *Scientific data* 5.1 (2018), pp. 1–14.

[19] M. Caron et al. "Emerging properties in self-supervised vision transformers". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 9650–9660.

[20] C. Chatzaki et al. "Human daily activity and fall recognition using a smartphone's acceleration sensor". In: *Information and Communication Technologies for Ageing Well and e-Health: Second International Conference, ICT4AWE 2016, Rome, Italy, April 21-22, 2016, Revised Selected Papers 2*. Springer. 2017, pp. 100–118.

[21] T. Chen et al. "A simple framework for contrastive learning of visual representations". In: *International conference on machine learning*. PMLR. 2020, pp. 1597–1607.

[22] H. Cho et al. "EEG datasets for motor imagery brain–computer interface". In: *GigaScience* 6.7 (2017), gix034.

[23] G. D. Clifford et al. "AF classification from a short single lead ECG recording: The PhysioNet/computing in cardiology challenge 2017". In: *2017 Computing in Cardiology (CinC)*. IEEE. 2017, pp. 1–4.

[24] P. Compagnon et al. "Learning Personalized ADL Recognition Models from Few Raw Data". In: *Artificial Intelligence in Medicine* 107 (2020), p. 101916. DOI: 10.1016/j.artmed.2020.101916.

[25] H. A. Dau et al. "The UCR time series archive". In: *IEEE/CAA Journal of Automatica Sinica* 6.6 (2019), pp. 1293–1305.

[26] J. Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[27] K. Ding et al. "Few-shot network anomaly detection via cross-network meta-learning". In: *Proceedings of the Web Conference 2021*. 2021, pp. 2448–2456.

[28] L. Duan et al. "Zero-Shot Learning for EEG Classification in Motor Imagery-Based BCI System". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 28.11 (2020), pp. 2411–2419.

[29] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu. "Differential entropy feature for EEG-based emotion classification". In: *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE. 2013, pp. 81–84.

[30] O. Elnaggar et al. "Sleep Posture One-Shot Learning Framework Based on Extremity Joint Kinematics: In-silico and in-Vivo Case Studies". In: *Information Fusion* 95 (2023), pp. 215–236.

[31] F. V. Farahani, W. Karwowski, and N. R. Lighthall. "Application of graph theory for identifying connectivity patterns in human brain networks: a systematic review". In: *frontiers in Neuroscience* 13 (2019), p. 585.

[32] C. Finn, P. Abbeel, and S. Levine. "Model-agnostic meta-learning for fast adaptation of deep networks". In: *International conference on machine learning*. PMLR. 2017, pp. 1126–1135.

[33] Y. Ge et al. "Few-Shot Learning for Medical Text: A Review of Advances, Trends, and Opportunities". In: *Journal of Biomedical Informatics* 144 (2023), p. 104458. DOI: 10.1016/j.jbi.2023.104458.

[34] Y. Ge et al. *Few-Shot Learning for Medical Text: A Systematic Review*. 2022. DOI: 10.48550/arXiv.2204.14081.

[35] A. L. Goldberger et al. "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals". In: *Circulation* 101.23 (2000 (June 13)). Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215, e215–e220.

[36] A. L. Goldberger et al. "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals". In: *circulation* 101.23 (2000), e215–e220.

[37] I. Goodfellow et al. "Generative adversarial nets". In: *Advances in neural information processing systems* 27 (2014).

[38] N. Gruver et al. "Large language models are zero-shot time series forecasters". In: *arXiv preprint arXiv:2310.07820* (2023).

[39] H. Guest et al. "Tinnitus with a normal audiogram: Relation to noise exposure but no evidence for cochlear synaptopathy". In: *Hearing research* 344 (2017), pp. 265–274.

[40] P. Gupta, S. Bhaskarpandit, and M. Gupta. "Similarity Learning Based Few Shot Learning for ECG Time Series Classification". In: *2021 Digital Image Computing: Techniques and Applications (DICTA)*. 2021, pp. 1–8.

[41] Y. Hartmann, H. Liu, and T. Schultz. "High-Level Features for Human Activity Recognition and Modeling". In: *Biomedical Engineering Systems and Technologies*. Ed. by A. C. A. Roque et al. Vol. 1814. 2023, pp. 141–163. DOI: 10.1007/978-3-031-38854-5_8.

[42] K. He et al. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1026–1034.

[43] A. Hernandez-Galvan, G. Ramirez-Alonso, and J. Ramirez-Quintana. "A Prototypical Network for Few-Shot Recognition of Speech Imagery Data". In: *Biomedical Signal Processing and Control* 86 (2023), p. 105154.

[44] T. Hinz et al. "Improved techniques for training single-image gans". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, pp. 1300–1309.

[45] U. Hoffmann et al. "An efficient P300-based brain–computer interface for disabled subjects". In: *Journal of Neuroscience methods* 167.1 (2008), pp. 115–125.

[46] Y. Hou et al. "MetaPrompting: Learning to learn better prompts". In: *arXiv preprint arXiv:2209.11486* (2022).

[47] G. Huang, H. Larochelle, and S. Lacoste-Julien. "Centroid networks for few-shot clustering and unsupervised few-shot classification". In: *arXiv preprint arXiv:1902.08605* 3.7 (2019).

[48] G. Huang et al. "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[49] B. Hudgins, P. Parker, and R. N. Scott. "A new strategy for multifunction myoelectric control". In: *IEEE transactions on biomedical engineering* 40.1 (1993), pp. 82–94.

[50] K. Hur et al. "GenHPF: General Healthcare Predictive Framework for Multi-task Multi-source Learning". In: *IEEE Journal of Biomedical and Health Informatics* (2023).

[51] S. Hwang et al. "EZSL-GAN: EEG-based Zero-Shot Learning Approach Using a Generative Adversarial Network". In: *2019 7th International Winter Conference on Brain-Computer Interface (BCI)*. 2019, pp. 1–4.

[52] T. Iwata and A. Kumagai. "Few-shot learning for time-series forecasting". In: *arXiv preprint arXiv:2009.14379* (2020).

[53] S. Jadon and A. Jadon. *An Overview of Deep Learning Architectures in Few-Shot Learning Domain*. 2023.

[54] W.-B. Jiang, L.-M. Zhao, and B.-L. Lu. "Large brain model for learning generic representations with tremendous EEG data in BCI". In: *arXiv preprint arXiv:2405.18765* (2024).

[55] M. Jin et al. "Multi-scale contrastive siamese networks for self-supervised graph representation learning". In: *arXiv preprint arXiv:2105.05682* (2021).

[56] M. Jin et al. "Time-llm: Time series forecasting by reprogramming large language models". In: *arXiv preprint arXiv:2310.01728* (2023).

[57] A. Johnson et al. "Mimic-iv". In: *PhysioNet. Available online at: https://physionet. org/content/mimiciv/1.0/(accessed August 23, 2021)* (2020).

[58] A. E. Johnson et al. "MIMIC-III, a freely accessible critical care database". In: *Scientific data* 3.1 (2016), pp. 1–9.

[59] M. Joshi et al. "Spanbert: Improving pre-training by representing and predicting spans". In: *Transactions of the association for computational linguistics* 8 (2020), pp. 64–77.

[60] B. Kaneshiro et al. "A representational similarity analysis of the dynamics of object processing using single-trial EEG classification". In: *Plos one* 10.8 (2015), e0135697.

[61] B. Kemp et al. "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG". In: *IEEE Transactions on Biomedical Engineering* 47.9 (2000), pp. 1185–1194.

[62] A. H. Khan, M. Hussain, and M. K. Malik. "ECG Images dataset of Cardiac and COVID-19 Patients". In: *Data in Brief* 34 (2021), p. 106762.

[63] Y. K. Kim et al. "Automatic Cardiac Arrhythmia Classification Using Residual Network Combined with Long Short-Term Memory". In: *IEEE Transactions on Instrumentation and Measurement* 71 (2022), pp. 1–17.

[64] S. Koelstra et al. "Deap: A database for emotion analysis; using physiological signals". In: *IEEE transactions on affective computing* 3.1 (2011), pp. 18–31.

[65] J. Kotia et al. "Few Shot Learning for Medical Imaging". In: *Machine Learning Algorithms for Industrial Applications*. Ed. by S. K. Das et al. 2021, pp. 107–132. DOI: 10.1007/978-3-030-50641-4_7.

[66] K. Kunze et al. "I know what you are reading: recognition of document types using mobile eye tracking". In: *Proceedings of the 2013 international symposium on wearable computers*. 2013, pp. 113–116.

[67] G. Lan et al. "GazeGraph: Graph-Based Few-Shot Cognitive Context Sensing from Human Visual Behavior". In: *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 2020, pp. 422–435. DOI: 10.1145/3384419.3430774.

[68] V. J. Lawhern et al. "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces". In: *Journal of neural engineering* 15.5 (2018), p. 056013.

[69] P. Lee et al. "Source-Free Subject Adaptation for EEG-based Visual Recognition". In: *2023 11th International Winter Conference on Brain-Computer Interface (BCI)*. 2023, pp. 1–6.

[70] J. Li et al. "A Novel Semi-Supervised Meta Learning Method for Subject-Transfer Brain–Computer Interface". In: *Neural Networks* 163 (2023), pp. 195–204.

[71] S. Li et al. "Meta-Learning for Fast and Privacy-Preserving Source Knowledge Transfer of EEG-Based BCIs". In: *IEEE Computational Intelligence Magazine* 17.4 (2022), pp. 16–26.

[72] Z. Li, H. Wang, and X. Liu. "A One-Dimensional Siamese Few-Shot Learning Approach for ECG Classification under Limited Data". In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference* 2021 (2021), pp. 455–458. DOI: 10.1109/EMBC46164.2021.9630622.

[73] B. Liu et al. "BETA: A large benchmark database toward SSVEP-BCI application". In: *Frontiers in neuroscience* 14 (2020), p. 627.

[74] H. Liu, Y. Hartmann, and T. Schultz. *CSL-SHARE: A multimodal wearable sensor-based human activity dataset*. 2021.

[75] T. Liu et al. "Few-Shot Learning for Cardiac Arrhythmia Detection Based on Electrocardiogram Data from Wearable Devices". In: *Digital Signal Processing* 116 (2021), p. 103094.

[76] W. Liu et al. "Comparing Recognition Performance and Robustness of Multimodal Deep Learning Models for Multimodal Emotion Recognition". In: *IEEE Transactions on Cognitive and Developmental Systems* (2021).

[77] Z. Liu et al. "Side-Aware Meta-Learning for Cross-Dataset Listener Diagnosis with Subjective Tinnitus". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 30 (2022), pp. 2352–2361.

[78] Y. Lu et al. "Few-shot scene-adaptive anomaly detection". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer. 2020, pp. 125–141.

[79] Y. Lv et al. "An Exploratory Study of Transfer Learning Frameworks in the Context of Few Available Shots of Neurophysiological Signals". In: *Computers and Electrical Engineering* 101 (2022), p. 108091.

[80] T.-F. Ma, W.-L. Zheng, and B.-L. Lu. "Few-Shot Class-Incremental Learning for EEG-Based Emotion Recognition". In: *Neural Information Processing*. Ed. by M. Tanveer et al. Vol. 1792. 2023, pp. 445–455. DOI: 10.1007/978-981-99-1642-9_38.

[81] J. Mattout et al. "BCI Challenge". In: *Kaggle* (2015).

[82] B. McCartney et al. "A Zero-Shot Learning Approach to the Development of Brain-Computer Interfaces for Image Retrieval". In: *Plos one* 14.9 (2019), e0214342.

[83] B. McMahan et al. "Communication-efficient learning of deep networks from decentralized data". In: *Artificial intelligence and statistics*. PMLR. 2017, pp. 1273–1282.

[84] J. Meyer et al. "U-HAR: A Convolutional Approach to Human Activity Recognition Combining Head and Eye Movements for Context-Aware Smart Glasses". In: *Proceedings of the ACM on Human-Computer Interaction* 6.ETRA (2022), pp. 1–19. DOI: 10.1145/3530884.

[85] D. Micucci, M. Mobilio, and P. Napoletano. "Unimib shar: A dataset for human activity recognition using acceleration data from smartphones". In: *Applied Sciences* 7.10 (2017), p. 1101.

[86] T. Mikolov et al. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[87] M. Mirza and S. Osindero. "Conditional generative adversarial nets". In: *arXiv preprint arXiv:1411.1784* (2014).

[88] A. Moin et al. "An EMG gesture recognition system with flexible high-density sensors and brain-inspired high-dimensional classifier". In: *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE. 2018, pp. 1–5.

[89] J. Moon et al. "Explainable Gait Recognition with Prototyping Encoder–Decoder". In: *Plos one* 17.3 (2022), e0264783.

[90] J. Mueller and A. Thyagarajan. "Siamese recurrent architectures for learning sentence similarity". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 30. 1. 2016.

[91] M. S. Munia et al. "Imbalanced Eeg Analysis Using One-Shot Learning with Siamese Neural Network". In: *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*. 2021, pp. 4–12.

[92] B. Murphy, M. Baroni, and M. Poesio. "EEG responds to conceptual stimuli and corpus semantics". In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. 2009, pp. 619–627.

[93] K. Nakamura et al. "Jointly learning energy expenditures and activities using egocentric multimodal signals". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 1868–1877.

[94] J. Narwariya et al. "Meta-Learning for Few-Shot Time Series Classification". In: *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*. 2020, pp. 28–36. DOI: 10.1145/3371158.3371162.

[95] J. Nazari et al. "Epilepsy Seizure Prediction with Few-Shot Learning Method". In: *Brain Informatics* 9.1 (2022), p. 21. DOI: 10.1186/s40708-022-00170-8.

[96] Y. Ng et al. "Few-Shot Transfer Learning for Personalized Atrial Fibrillation Detection Using Patient-Based Siamese Network with Single-Lead ECG Records". In: *Artificial Intelligence in Medicine* 144 (2023), p. 102644.

[97] C. H. Nguyen, G. K. Karavas, and P. Artemiadis. "Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features". In: *Journal of neural engineering* 15.1 (2017), p. 016002.

[98] A. Nichol, J. Achiam, and J. Schulman. "On first-order meta-learning algorithms". In: *arXiv preprint arXiv:1803.02999* (2018).

[99] E. Pachetti and S. Colantonio. *A Systematic Review of Few-Shot Learning in Medical Imaging*. 2023.

[100] J. Pan et al. "Multiple Scale Convolutional Few Shot Learning Networks for Online P300-based Brain-Computer Interface and Its Application to Patients with Disorder of Consciousness". In: *IEEE Transactions on Instrumentation and Measurement* (2023).

[101] J. Pan et al. "Few-Shot Egocentric Multimodal Activity Recognition". In: *ACM Multimedia Asia*. 2022, pp. 1–7. DOI: 10.1145/3469877.3490603.

[102] G. Pang et al. "Explainable deep few-shot anomaly detection with deviation networks". In: *arXiv preprint arXiv:2108.00462* (2021).

[103] A. Parnami and M. Lee. *Learning from Few Examples: A Summary of Approaches to Few-Shot Learning*. 2022.

[104] D. Pascual et al. "Synthetic epileptic brain activities using GANs". In: *Machine Learning for Health (ML4H) at NeurIPS* (2019).

[105] S. Petrutiu, A. V. Sahakian, and S. Swiryn. "Abrupt changes in fibrillatory wave characteristics at the termination of paroxysmal atrial fibrillation in humans". In: *Europace* 9.7 (2007), pp. 466–470.

[106] C. Phunruangsakao, D. Achanccaray, and M. Hayashibe. "Deep Adversarial Domain Adaptation with Few-Shot Learning for Motor-Imagery Brain-Computer Interface". In: *IEEE Access* 10 (2022), pp. 57255–57265.

[107] S. Pizzolato et al. "Comparison of six electromyography acquisition setups on hand movement classification tasks". In: *PloS one* 12.10 (2017), e0186132.

[108] T. J. Pollard et al. "The eICU Collaborative Research Database, a freely available multi-center database for critical care research". In: *Scientific data* 5.1 (2018), pp. 1–13.

[109] R. Poulain, M. Gupta, and R. Beheshti. "Few-Shot Learning with Semi-Supervised Transformers for Electronic Health Records". In: *Proceedings of Machine Learning Research* 182 (2022), pp. 853–873.

[110] A. Radford et al. "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

[111] E. Rahimian et al. "Few-Shot Learning for Decoding Surface Electromyography for Hand Gesture Recognition". In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021, pp. 1300–1304.

[112] A. Ramesh et al. "Zero-shot text-to-image generation". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 8821–8831.

[113] J. Requeima et al. "Fast and flexible multi-task classification using conditional neural adaptive processes". In: *Advances in Neural Information Processing Systems* 32 (2019).

[114] A. Saeed, V. Ungureanu, and B. Gfeller. "Sense and Learn: Self-supervision for Omnipresent Sensors". In: *Machine Learning with Applications* 6 (2021), p. 100152.

[115] A. Salekin and N. Russo. "Understanding Autism: The Power of EEG Harnessed by Prototypical Learning". In: *Proceedings of the Workshop on Medical Cyber Physical Systems and Internet of Medical Things*. 2021, pp. 12–16. DOI: 10.1145/3446913.3460317.

[116] R. Schaette and D. McAlpine. "Tinnitus with a normal audiogram: physiological evidence for hidden hearing loss and computational model". In: *Journal of Neuroscience* 31.38 (2011), pp. 13452–13457.

[117] G. Schalk et al. "BCI2000: a general-purpose brain-computer interface (BCI) system". In: *IEEE Transactions on biomedical engineering* 51.6 (2004), pp. 1034–1043.

[118] R. T. Schirrmeister et al. "Deep learning with convolutional neural networks for EEG decoding and visualization". In: *Human brain mapping* 38.11 (2017), pp. 5391–5420.

[119] J. R. Sempionatto et al. "An epidermal patch for the simultaneous monitoring of haemodynamic and metabolic biomarkers". In: *Nature Biomedical Engineering* 5.7 (2021), pp. 737–748.

[120] M. Sendera et al. "Hypershot: Few-shot learning by kernel hypernetworks". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023, pp. 2469–2478.

[121] K. Sharma et al. "A dataset of continuous affect annotations and physiological signals for emotion analysis". In: *Scientific data* 6.1 (2019), p. 196.

[122] A. Shoeb. *Chb-mit scalp eeg database*. 2000.

[123] A. H. Shoeb. "Application of machine learning to epileptic seizure onset detection and treatment". PhD thesis. Massachusetts Institute of Technology, 2009.

[124] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[125] J. Snell, K. Swersky, and R. Zemel. "Prototypical networks for few-shot learning". In: *Advances in neural information processing systems* 30 (2017).

[126] S. Song et al. "Multimodal multi-stream deep learning for egocentric activity recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2016, pp. 24–31.

[127] Y. Song et al. "A Comprehensive Survey of Few-shot Learning: Evolution, Applications, Challenges, and Opportunities". In: *ACM Computing Surveys* 55.13s (2023), 271:1–271:40. DOI: 10.1145/3582688.

[128] R. Soroushmojdehi et al. "Transfer Learning in Hand Movement Intention Detection Based on Surface Electromyography Signals". In: *Frontiers in Neuroscience* 16 (2022), p. 977328.

[129] C. Spampinato et al. "Deep learning human mind for automated visual classification". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 6809–6817.

[130] N. Srivastava, J. Newn, and E. Velloso. "Combining low and mid-level gaze features for desktop activity recognition". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2.4 (2018), pp. 1–27.

[131] J. Sun et al. "Euler common spatial patterns for EEG classification". In: *Medical & Biological Engineering & Computing* 60.3 (2022), pp. 753–767.

[132] L. Sun et al. "Few-Shot Class-Incremental Learning for Medical Time Series Classification". In: *IEEE Journal of Biomedical and Health Informatics* (2023).

[133] Q. Suo et al. "TAdaNet: Task-Adaptive Network for Graph-Enriched Meta-Learning". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020, pp. 1789–1799. DOI: 10.1145/3394486.3403230.

[134] T. Suwannaphong, R. McConville, and I. Craddock. "Radio Signal Strength Indication Augmentation for One-Shot Learning in Indoor Localisation". In: *Proceedings of the 1st ACM Workshop on Smart Wearable Systems and Applications*. 2022, pp. 7–12. DOI: 10.1145/3556560.3560714.

[135] C. Szegedy et al. "Rethinking the inception architecture for computer vision". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826.

[136] S. Tam et al. "Siamese Convolutional Neural Network and Few-Shot Learning for Embedded Gesture Recognition". In: *2022 20th IEEE Interregional NEWCAS Conference (NEWCAS)*. 2022, pp. 114–118.

[137] M. Tan and Q. Le. "Efficientnet: Rethinking model scaling for convolutional neural networks". In: *International conference on machine learning*. PMLR. 2019, pp. 6105–6114.

[138] W. Tang, L. Liu, and G. Long. "Interpretable Time-Series Classification on Few-Shot Samples". In: *2020 International Joint Conference on Neural Networks (IJCNN)*. 2020, pp. 1–8.

[139] M. Tangermann et al. "Review of the BCI competition IV". In: *Frontiers in neuroscience* (2012), p. 55.

[140] S. Tian et al. "A survey on few-shot class-incremental learning". In: *Neural Networks* 169 (2024), pp. 307–324.

[141] E. Triantafillou et al. "Meta-dataset: A dataset of datasets for learning to learn from few examples". In: *arXiv preprint arXiv:1903.03096* (2019).

[142] G. Tsoumplekas et al. "Toward green and human-like artificial intelligence: A complete survey on contemporary few-shot learning approaches". In: *arXiv preprint arXiv:2402.03017* (2024).

[143] A. of Us Research Program Investigators. "The "All of Us" research program". In: *New England Journal of Medicine* 381.7 (2019), pp. 668–676.

[144] O. Vinyals et al. "Matching networks for one shot learning". In: *Advances in neural information processing systems* 29 (2016).

[145] V. Viswanathan et al. "Large language models enable few-shot clustering". In: *arXiv preprint arXiv:2307.00524* (2023).

[146] M. Völker et al. "Deep transfer learning for error decoding from non-invasive EEG". In: *2018 6th International Conference on Brain-Computer Interface (BCI)*. IEEE. 2018, pp. 1–6.

[147] J. B. Wagenaar et al. "Collaborating and sharing data in epilepsy research". In: *Journal of Clinical Neurophysiology* 32.3 (2015), pp. 235–239.

[148] J. Walonoski et al. "Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record". In: *Journal of the American Medical Informatics Association* 25.3 (2018), pp. 230–238.

[149] D. Wang et al. "Personalized Modeling of Blood Pressure with Photoplethysmography: An Error-Feedback Incremental Support Vector Regression Model". In: *IEEE Internet of Things Journal* (2023).

[150] X. Wang et al. "Similarity Function for One-Shot Learning to Enhance the Flexibility of Myoelectric Interfaces". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 31 (2023), pp. 1697–1706.

[151] X. Wang et al. "A Generalized Zero-Shot Learning Scheme for SSVEP-Based BCI System". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 31 (2023), pp. 863–874.

[152] Y. Wang et al. "Generalizing from a Few Examples: A Survey on Few-shot Learning". In: *ACM Computing Surveys* 53.3 (2021), pp. 1–34. DOI: 10.1145/3386252.

[153] Y. Wang et al. "A benchmark dataset for SSVEP-based brain–computer interfaces". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25.10 (2016), pp. 1746–1752.

[154] D. Wertheimer, L. Tang, and B. Hariharan. "Few-shot classification with feature map reconstruction networks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 8012–8021.

[155] G. Woo et al. "Unified training of universal time series forecasting transformers". In: *arXiv preprint arXiv:2402.02592* (2024).

[156] X. Wu and R. H. Chan. "Does Meta-Learning Improve EEG Motor Imagery Classification?" In: *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. 2022, pp. 4048–4051.

[157] F. Xiao et al. "Meta-learning for few-shot time series forecasting". In: *Journal of Intelligent & Fuzzy Systems* 43.1 (2022), pp. 325–341.

[158] C. Xu et al. "A physicochemical-sensing electronic skin for stress response monitoring". In: *Nature Electronics* (2024), pp. 1–12.

[159] T. Xue et al. "CEAP-360VR: A Continuous Physiological and Behavioral Emotion Annotation Dataset for 360 VR Videos". In: *IEEE Transactions on Multimedia* (2021).

[160] Y. You et al. "A Few-Shot Learning-Based EEG and Stage Transition Sequence Generator for Improving Sleep Staging Performance". In: *Biomedicines* 10.12 (2022), p. 3006. DOI: 10.3390/biomedicines10123006.

[161] P. Youssef et al. "Model Personalization with Static and Dynamic Patients' Data". In: *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*. 2022, pp. 324–333.

[162] M. Zhang et al. "A Patient-Specific Closed-Loop Epilepsy Management SoC with One-Shot Learning and Online Tuning". In: *IEEE Journal of Solid-State Circuits* 57.4 (2022), pp. 1049–1060.

[163] R. Zhang et al. "Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 15211–15222.

[164] T. Zhang et al. "Few-Shot Learning for Fine-Grained Emotion Recognition Using Physiological Signals". In: *IEEE Transactions on Multimedia* (2022).

[165] T. Zhang et al. "Rcea: Real-time, continuous emotion annotation for collecting precise mobile video ground truth labels". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1–15.

[166] X. S. Zhang et al. "MetaPred: Meta-Learning for Clinical Risk Prediction with Limited Patient Electronic Health Records". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019, pp. 2487–2495. DOI: 10.1145/3292500.3330779.

[167] S. Zhao and F. Rudzicz. "Classifying phonological categories in imagined and articulated speech". In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2015, pp. 992–996.

[168] B. Zhu and M. Shoaran. "Unsupervised Domain Adaptation for Cross-Subject Few-Shot Neurological Symptom Detection". In: *2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)*. 2021, pp. 181–184.