MULTI-SOURCE DOMAIN GENERALIZATION FOR ECG-BASED COGNITIVE LOAD ESTIMATION: A PLUG-IN METHOD AND BENCHMARK

Jiyao Wang, Ange Wang, Haolong Hu, Kaishun Wu, Dengbo He

Hong Kong University of Science and Technology (Guangzhou), China

ABSTRACT

Electrocardiography (ECG) for objective cognitive load estimation gained increasing attention, and offers a more feasible and non-invasive alternative to traditional methods such as electroencephalography (EEG). Despite the promise of ECG signal, application in real-world scenarios is hampered by the domain shift present in data collected in controlled environments versus real-world settings. We propose a novel plug-in generalizable framework, CogDG-ECG, assessed on a first-introduced multi-source domain generalization (MSDG) protocol for generalized cognitive load estimation. CogDG-ECG bridges the domain gap by extracting domain-invariant features through adversarial learning, and estimating instance-specific unseen features by synthesizing plausible feature statistical variations. A new benchmark based on three public datasets and MSDG protocol was established, which demonstrates the superiority of our proposed method.

***Index Terms*—** Cognitive load estimation, ECG, multi-source domain generalization, deep learning

1. INTRODUCTION

Researchers have shown increased interest in objectively assessing cognitive workload based on physiological measurements. But, previous studies that used physiological indicators for high cognitive load (e.g., electroencephalography (EEG) [1], galvanic skin response (GSR) [2]) predominantly rely on invasive devices. Presently, the acquisition of electrocardiography (ECG) data through non-invasive wearable devices, such as steering wheel-integrated sensors and smartwatches, has gradually become more feasible [3]. Currently, several hand-crafted cardiac indicators (e.g., heart rate, heart rate variability) were affirmed to be significant to cognitive load [4]. Thus, we try to make estimation based on ECG signals in this work. Although some research leveraged manual-processed indicators from ECG [5, 6], few tried to utilize ECG signals for cognitive load estimation directly, which saves time cost for feature preprocess and is closer to real-world demand [7].

Meanwhile, advanced methods [8, 9] have been proposed to detect cognitive load, but their effectiveness is limited in real-world scenarios. The reason is that most training data is collected in controlled laboratory environments, which lack the variations present in real-world settings. Seeing Table 1, in three public datasets for academic purposes, there are still notable differences such as individuals, acquisition devices, and state-induced tasks. As a result, models might struggle to generalize well to unseen testing domains due to domain shift [10]. Recent attempts to address this issue with transfer learning, aiming to achieve generalizability across different subjects [11] or state-induce tasks [12, 13], which were also believed as two sources of domain shifts [12]. However, there are still obstacles to overcome. Existing methods primarily focus on resolving a single generalization task, while real-world scenarios often involve multiple types of domain shifts. Additionally, current generalizable measures rely on domain adaptation protocol (i.e., data from the target domain can be assessed), which is not feasible in practical deployment.

To address these gaps, we propose a novel plug-in framework called CogDG-ECG for generalized cognitive load estimation evaluated on multi-source domain generalization (MSDG) protocol. The principle of this framework involves extracting domain-invariant features and estimating instance-specific unseen features through two regularization techniques. In cognitive load detection, domain-invariant information refers to the correlations between physiological attributes and load levels. We regularize the domain-invariant feature extraction from domain variations using adversarial learning. Additionally, by introducing uncertainty to domain attribute-related features, we aim to synthesize plausible feature statistical variations to enhance the network's robustness to out-of-distribution (OOD) attributes across domains. Through these procedures, we effectively reduce the distance between feature distributions from different datasets, thereby improving accuracy in cross-domain cognitive load estimation. In terms of training, CogDG-ECG is jointly optimized using the two aforementioned regularizations and a classification loss, which is more suitable for multi-dataset training.

The main contributions are summarized as follows: (1) We proposed an end-to-end plug-in method CogDG-ECG, for resolving the domain shift of large-scale training in cognitive load estimation with ECG. To the best of our knowledge, it is

This work was supported by the startup funding of the Hong Kong University of Science and Technology (Guangzhou) (Corresponding author: Dengbo He).

Table 1. Dataset statistics.

Datasets	Age range	Cognitive load task	Scenario	ECG Device	Sampling frequency
DMTD [14]	24.5±5.95	Counting back	Driving	Biopac	1000Hz
CLAS [15]	20 to 27	Mathematical calculation	Interactive activity	Shimmer3	256Hz
MNBD [16]	27.6±4.45	Modified 2-back	Driving	Becker Meditec	240Hz

the first time MSDG in cognitive load estimation was introduced as a new challenge; (2) To align domain-agnostic features and stimulate unseen instance-level variations, we incorporate adversarial loss and contrastive loss to regularize the network and enhance the generalizability when facing OOD samples; (3) We establish a benchmark for cognitive load estimation with ECG under MSDG protocol, which shows the competitive performance of our proposal.

2. METHOD AND MATERIALS

2.1. Overall Framework

We formally introduce a plug-in domain generalization (DG) framework for ECG-based cognitive load estimation, CogDG-ECG. In brief, given in total N ECG signals $X_S = \{x_i\}_{i=1}^N, x_i \in \mathbb{R}^W$ from M source domains. Each domain is labeled with one distinct domain signal $\{d_i\}_{i=1}^M$. To stimulate the application scenario, we directly input signals into CogDG-ECG, which is formulated as $f(X_S;\theta)$. The W is the time window size of the signal, and the θ is the parameter to optimize. The MSDG protocol assumes that $f(X;\theta)$ can only be trained in source domains, but tested without assessing data X_T in the target domain. Our goal is to optimize θ to narrow the shifted feature space from multiple source domains to the target domain, and learn the mapping $f(X_T;\theta)$ to target cognitive workload level $Y = \{y_i\}_{i=1}^N$.

The overall architecture of CogDG-ECG is illustrated in Figure 1. Firstly, a substitutable backbone network $Enc(*)$ is initialized to obtain under-optimized representations $O = \{o_i\}_{i=1}^N \in \mathbb{R}^{N\times D}$ from source ECG signals X_S, where D is the dimension of representation. Secondly, to enhance the capacity of precise estimation and robustness to unseen domain shifts, we parallelly fed feature O into and trained the model with several key components: (1) **Adversarial Domain Alignment** block is introduced to align multi-domain representation spaces to united domain-agnostic distribution; (2) **Uncertainty Variation Estimation** block to obtain the insight to OOD space while ensuring that the augmented features maintain plausibility; (3) an estimation head with fully connected (FC) layers to output cognitive load label Y.

2.2. Adversarial Domain Alignment

To improve the generalization of the network, we force the network to align representations from different domains with an adversarial process. The key idea is a game between two components - a feature extractor and a domain discriminator. The feature extractor aims to learn invariant representations across M source domains. Meanwhile, the discriminator tries to identify which particular source domain the sample comes from. This adversarial process makes the feature extractor learn representations with minimal differences between domains, such that the discriminator cannot reliably tell them apart. Specifically, after inputting representation O into the domain classifier, we optimize the parameters θ for distinguishing domain by maximizing $\mathcal{L}_{ADA}$ and optimizing the parameters of the domain discriminator to the opposite.

$$\mathcal{L}_{ADA} = -\sum_{i=1}^{M} \mathbb{I}_{[i=d]} \log p_i. \tag{1}$$

In equation (1), $\mathbb{I}$ is the binary indicator, and p_i indicates the predicted probability of domain i outputted by the domain classifier. Besides, considering the training instability of adversarial paradigm [17], we apply the gradient reversal layer (GRL) [10]. During backpropagation, the GRL reverses the gradients flowing from the discriminator to the feature extractor. As a result, the feature extractor is optimized to fool the discriminator and extract domain-agnostic features.

2.3. Uncertainty Variation Estimation

In our approach, we address the challenge of target domain shift by estimating the variation of features. However, determining an appropriate range of variations is difficult when the target domain is unknown. To overcome this issue, we want the generated auxiliary features not seen in previous iterations to fulfill two important criteria: diversity and plausibility [18]. Therefore, we present an adaptive and straightforward non-parametric method to constrain the model to avoid straying too far from the characteristics of existing source domains, striking a balance between novelty and plausibility.

$$O^{'} = \mu(O) + \epsilon\sigma(O), where\ \epsilon \sim \mathcal{N}(0,1). \tag{2}$$

The variances of feature space ($\sigma(*)$) were leveraged to determine the plausible variation range. To control the direction of attribute shift, we assume that the statistical discrepancies follow a standard normal distribution [19]. Therefore, we sample K noise (ϵ) from the Gaussian distribution. This process stimulates uncertain domain shifts in the target domain and produces novel and plausible estimates of uncertain features ($O^{'} = \{o_i^{'}\}_{i=1}^K$) from extended O distribution, as described in Equation (2). This approach facilitates the

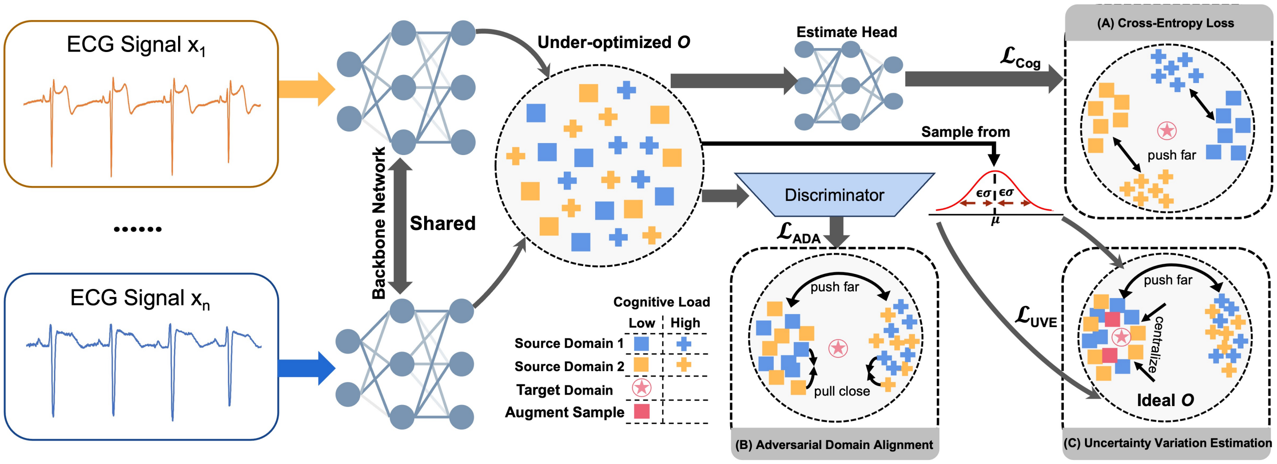

Fig. 1. The overall architecture of proposed CogDG-ECG framework. (A) Feature space O is optimized by cross-entropy loss, while domain shift still exists; (B) $\mathcal{L}_{ADA}$ narrows the distance between samples with same label but from different domains; (C) $\mathcal{L}_{UVE}$ augments feature space and further pushes close to the potential OOD sample.

exploration of various combinations of directions and intensities that adhere to the two criteria, capitalizing on the inherent characteristics of the Gaussian distribution. To suppress unpredictable features, we propose a loss function based on contrastive learning [20], statistic-guided Uncertainty Variation Estimation (**UVE**) loss as follows:

$$\mathcal{L}_{UVE}(O,O^{'}) = \\ -\sum_{k\in K}\log(\frac{\exp(\text{sim}(o_i,o^{'}_k)/\tau)}{\sum_{j\in N}\mathbb{I}_{[y_j\neq y_i]}\exp(\text{sim}(o_i,o_j)/\tau)})]. \tag{3}$$

Where $\text{sim}(*)$ is the similarity measurement, which is instantiated as cosine similarity. And τ is the temperature control factor. and In Eq.(3), given one sample's representation o_i as the anchor, we generate K auxiliary features $O^{'}$ based on anchor o_i as positive samples to pull close, and the rest in the batch with different cognitive workload level to push far.

Finally, we incorporate $\mathcal{L}_{Cog}$ to classify the cognitive workload level based on O. Specifically, we use the cross-entropy loss $\mathcal{L}_{Cog}$ to maximize the probability of the correct cognitive load level. To prevent irrelevant regularizations in the early iterations, we introduce an adaptation factor λ [21]. The network is jointly trained using the overall loss, which includes two trade-off parameters (w_1 and w_2).

$$\mathcal{L}(X_S,Y_S) = \mathcal{L}_{Cog} + \lambda(w_1\mathcal{L}_{ADA} + w_2\mathcal{L}_{UVE}). \tag{4}$$

2.4. Datasets and Evaluation Protocol

Dataset Two public datasets, namely DMTD [14] and CLAS [15], along with our proprietary dataset MNBD [16], were employed for our benchmark. The basic statistical information is in Table 1. The DMTD dataset encompasses 40 males, 49 females, and 1 other. During a 20-minute conditional automated driving, cognitive load was induced in half of the participants, while the other half was not. The CLAS dataset comprises 62 volunteers with no gender balance. Cognitive load was induced through a series of mathematical tasks (4 seconds for each). The MNBD dataset consists of 33 drivers comprising 18 male and 15 female drivers. Cognitive load states were induced through three levels of modified n-back tasks in manual driving. In order to maintain task consistency, we used the state induced by no-task and 2-back tasks as the low and high load levels, respectively. We performed a simple pre-processing on the ECG signals. First, we downsampled the DMTD and CLAS to 240Hz. And we de-noised all ECG signals using a bandpass filter with cutoff frequencies between 3Hz and 45Hz. The time window is 5 seconds with a step size of 1/60 seconds [16].

Evaluation Protocol Our MSDG protocol is implemented by dividing the above three datasets into two piles: source domains and the target domain. Two datasets were combined and shuffled for source domains, and the left dataset constructed the target domain. Besides, we selected three wide-used metrics for performance evaluation: Accuracy (ACC), F1-score (F1), and Sensitivity (SEN). We presented the average of 5-time tests with different random seeds for each model. Meanwhile, the paired t-test is applied to check the significance of the performance difference.

3. EXPERIMENTS AND RESULTS

Several typical methods were chosen for comparison, including machine learning (ML) methods (i.e., SVM, KNN, LDA [12], and LightGBM [22]) and deep learning (DL) methods (i.e., ANN, LSTM, and TCN [23]). Furthermore, there were various DG methods (i.e., AD [10], DSU [19], IFL [24]) from

Table 2. Cognitive load estimation results on MSDG protocol. In this and following tables, $^+$ means it is based on the best DL baseline on each target domain, and * indicates the significant difference (p<0.05) between our method and the best baseline.

		Machine Learning				Deep Learning			Domain Generalization			
Target	**Metric**(%)	**LDA**	**KNN**	**SVM**	**LightGBM**	**LSTM**	**ANN**	**TCN**	**AD**$^+$	**DSU**$^+$	**IFL**$^+$	**Ours**$^+$
CLAS	ACC	51.75	62.93	67.88	57.23	67.98	69.31	68.23	68.30	72.28	70.31	**74.06***
	F1	61.69	75.95	80.53	70.81	79.92	81.57	79.02	80.22	83.72	82.23	**84.99***
	SEN	51.16	76.47	87.38	66.92	89.97	90.12	88.95	90.86	96.32	92.90	**99.30***
DMTD	ACC	50.75	52.75	51.88	54.18	56.11	56.83	58.12	59.22	60.03	58.16	**61.21***
	F1	49.90	59.82	53.72	62.15	62.99	68.33	69.77	69.90	69.24	67.52	**70.07**
	SEN	42.77	65.55	46.15	68.47	69.01	70.75	74.01	74.78	75.52	71.38	**76.14**
MNBD	ACC	59.25	60.80	55.61	64.74	62.41	65.08	60.37	66.40	68.63	67.11	**69.33***
	F1	63.21	69.75	66.18	73.55	73.26	72.80	70.12	72.29	74.59	73.45	**75.81***
	SEN	71.79	77.55	73.80	85.92	85.71	87.35	80.04	87.79	90.02	89.83	**91.24***

Table 3. Accuracy (%) in ablation test. Base means the backbone network without $\mathcal{L}_{ADA}$ and $\mathcal{L}_{UVE}$. * indicates if there is a significant difference with the complete method.

Target	**Base**	**w/o** $\mathcal{L}_{ADA}$	**w/o** $\mathcal{L}_{UVE}$	**Ours** $^+$
CLAS	69.31*	73.37*	70.22*	**74.06**
DMTD	58.12*	60.34*	58.84*	**61.21**
MNBD	65.08*	67.96*	67.10*	**69.33**

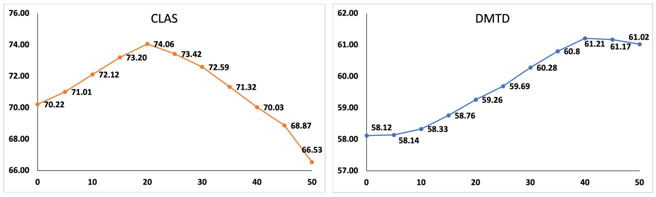

Fig. 2. Impacts of the hyperparameter K tested on two target domains CLAS and DMTD. Accuracy is used for evaluation and in the y-axis.

other fields taken into comparison. Note that, as there was no standard format for ANN, we independently implemented ANN, which consists of one up-sample linear layer with Relu activation, a layer-normalization layer, and one down-sample linear layer. Furthermore, we set $D, \tau, w1, w2$ to 256, 0.1, 0.1, 0.0001 according to the experimental results.

Cognitive load estimation The comparison results presented in Table 2 show that DG methods usually achieved more competitive performance than classic DL and ML models. Nevertheless, CogDG-ECG can significantly outperform the best baseline at most target domains. Moreover, we obtain some other interesting findings. For DL methods, although TCN is more advanced and complex than ANN, its performance is worse on CLAS and MNBD. We assume that is because the DMTD with far more samples contains redundant information than the target domain. Excessive model complexity causes the model to overfit on the source domains. Besides, insignificant improvement of our method on DMTD reflects the issue of the data volume gap between the source domain and target domain. Parameter space optimized from limited data seems to be underfitted, which is worth further study.

Ablation test In this part, we provide the results of ablation tests in Table 3. Particularly, the variants without $\mathcal{L}_{ADA}$ or $\mathcal{L}_{UVE}$ show significant performance degradation compared to the complete CogDG-ECG. It proves that only one regularization cannot transfer the under-optimized feature space to the ideal space. Extracting domain-invariant or augmenting instance-specific features solely might make the model unprepared for OOD samples in the target domain, or the feature space of the same label is still separated into different domains, respectively. Therefore, jointly applying regularizations based on aligning domain-agnostic and generating auxiliary features from uncertainty is necessary for this task.

Hyper-parameter sensitive study As mentioned above, we notice the influence of data-volume differences in domains. Therefore, we tested our proposed model with different K on CLAS and DMTD. The hyperparameter K determines how many auxiliary features are estimated from the extended feature distribution with the same cognitive level and are chosen as positive samples in $\mathcal{L}_{UVE}$. Seeing Figure 2, we can find the best performance on CLAS belongs to the model with 20 novel features, while DMTD requires 40. It indicates that the underfit issue over DMTD can be alleviated. On the other hand, it also elaborates that the generated auxiliary features belong to invisible target domain space to some extent. Besides, the convex curves in both two figures show that too many or few positive samples will cause improper activations or insufficient training of the network, respectively.

4. CONCLUSION

This paper introduces CogDG-ECG, a novel plug-in framework for cognitive load estimation using ECG data, which demonstrates superior performance compared to baseline models and previous DG methods. The properties of our method such as end-to-end training and plug-ins, are beneficial for large-scale training in industrial applications. Furthermore, the new proposed protocol and benchmark bring a new challenge, and facilitate future studies as well.

5. REFERENCES

[1] H. Yang, J. Wu, Z. Hu, and C. Lv, "Real-time driver cognitive workload recognition: Attention-enabled learning with multimodal information fusion," *IEEE Transactions on Industrial Electronics*, 2023.

[2] N. Nourbakhsh, Y. Wang, F. Chen, and R.A. Calvo, "Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks," in *Proceedings of the 24th Australian Computer-Human Interaction Conference*, 2012, pp. 420–423.

[3] S. Liu, *The Empathetic Car: Detecting Emotion and Well-being of Drivers under Naturalistic Condition*, Ph.D. thesis, ETH Zurich, 2022.

[4] P. Ayres, J.Y. Lee, F. Paas, and J.J. van Merriënboer, "The validity of physiological measures to identify differences in intrinsic cognitive load," *Frontiers in psychology*, vol. 12, pp. 702538, 2021.

[5] S. Solhjoo, M.C. Haigney, E. McBee, J.J. van Merrienboer, L. Schuwirth, A.R. Artino Jr, A. Battista, T.A. Ratcliffe, H.D. Lee, and S.J. Durning, "Heart rate and heart rate variability correlate with clinical reasoning performance and self-reported measures of cognitive load," *Scientific reports*, vol. 9, no. 1, pp. 14668, 2019.

[6] D. He, Z. Wang, E.B. Khalil, B. Donmez, G. Qiao, and S. Kumar, "Classification of driver cognitive load: exploring the benefits of fusing eye-tracking and physiological measures," *Transportation research record*, vol. 2676, no. 10, pp. 670–681, 2022.

[7] P.C. Chanel, M.D. Wilson, and S. Scannella, "Online ecg-based features for cognitive load assessment," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 2019, pp. 3710–3717.

[8] A. Tjolleng, K. Jung, W. Hong, W. Lee, B. Lee, H. You, J. Son, and S. Park, "Classification of a driver's cognitive workload levels using artificial neural network on ecg signals," *Applied ergonomics*, vol. 59, pp. 326–332, 2017.

[9] S. Yang, J. Kuo, M.G. Lenné, M. Fitzharris, T. Horberry, K. Blay, D. Wood, C. Mulvihill, and C. Truche, "The impacts of temporal variation and individual differences in driver cognitive workload on ecg-based detection," *Human factors*, vol. 63, no. 5, pp. 772–787, 2021.

[10] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

[11] Y. Zhou, P. Wang, P. Gong, F. Wei, X. Wen, X. Wu, and D. Zhang, "Cross-subject cognitive workload recognition based on eeg and deep domain adaptation," *IEEE Transactions on Instrumentation and Measurement*, 2023.

[12] Y. Zhou, Z. Xu, Y. Niu, P. Wang, X. Wen, X. Wu, and D. Zhang, "Cross-task cognitive workload recognition based on eeg and domain adaptation," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 50–60, 2022.

[13] K. Guan, Z. Zhang, T. Liu, and H. Niu, "Cross-task mental workload recognition based on eeg tensor representation and transfer learning," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023.

[14] Q. Meteier, M. Capallera, E. De Salis, L. Angelini, S. Carrino, M. Widmer, O. Abou Khaled, E. Mugellini, and A. Sonderegger, "A dataset on the physiological state and behavior of drivers in conditionally automated driving," *Data in brief*, vol. 47, pp. 109027, 2023.

[15] V. Markova, T. Ganchev, and K. Kalinkov, "Clas: A database for cognitive load, affect and stress recognition," in *2019 International Conference on Biomedical Innovations and Applications (BIA)*. IEEE, 2019, pp. 1–4.

[16] D. He, B. Donmez, C.C. Liu, and K.N. Plataniotis, "High cognitive load assessment in drivers through wireless electroencephalography and the validation of a modified n-back task," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 4, pp. 362–371, 2019.

[17] M.Z. Zaheer, J.h. Lee, M. Astrid, and S.I. Lee, "Old is gold: Redefining the adversarially learned one-class classifier training paradigm," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14183–14193.

[18] J. Kang, S. Lee, N. Kim, and S. Kwak, "Style neophile: Constantly seeking novel styles for domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7130–7140.

[19] X. Li, Y. Dai, Y. Ge, J. Liu, Y. Shan, and L.Y. Duan, "Uncertainty modeling for out-of-distribution generalization," *arXiv preprint arXiv:2202.03958*, 2022.

[20] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," *arXiv preprint arXiv:1911.05722*, 2019.

[21] H. Lu, Z. Yu, X. Niu, and Y. Chen, "Neuron structure modeling for generalizable remote physiological measurement," *arXiv preprint arXiv:2303.05955*, 2023.

[22] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.

[23] S. Bai, J.Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[24] K. Tang, M. Tao, J. Qi, Z. Liu, and H. Zhang, "Invariant feature learning for generalized long-tailed classification," in *European Conference on Computer Vision*. Springer, 2022, pp. 709–726.