Learning Based Clinical Diagnosis Framework for Inconsistently and Partially Labeled Dataset

Jae-Wook Jung, San-Eun Jeon, Jun-Pyo Hong
Dept. Intelligent Robot Engineering
Pukyong National University
Busan, South Korea

Ju-Hyeon Kim, Yeong-Kyun Bae, Jae-Hyun Lee
Dept. of Physical Medicine and Rehabilitation
College of Medicine, Kosin University Gospel Hospital
Busan, South Korea

***Abstract*— This paper presents a novel deep learning framework for autonomous clinical diagnosis by dealing with the training with poorly labeled clinical dataset. Partially labeled data and inconsistent labels from multiple annotators make the model hard to learn accurate diagnosis in frequently and drastically updated clinical dataset. Motivated by such difficulties, the proposed framework introduces the weighted combination of inconsistent labels by considering multiple annotators' expertise and adapt meta-learning approach for the quick adaptation to the updated dataset. Experimental results on the posterior pelvic tilt detection in a squat motion show the proposed approach outperforms the conventional learning approaches in terms of the convergence speed and the converged mean squared error.

Keywords—Meta-learning, clinical diagnosis, inconsistently and partially labeled dataset, squat exercise, posterior pelvic tilt

I. Introduction

Physiotherapy is an important treatment to assist in the recovery of many injuries, disabilities, and health conditions. Although proper physiotherapy treatment conducted under the supervision of medical specialists is beneficial for the speedy and successful recovery, it entails a large cost and the inconvenience of visiting medical facility. Furthermore, recent outbreak of COVID-19 makes it harder for patients to get proper treatment from medical facilities. For these reasons, the demand for rehabilitation monitoring systems is constantly increasing with the need of in-home physiotherapy [1], [2]

The autonomous diagnosis of medical disorder is considered as a key technology for the monitoring system. There has been great improvement in the autonomous diagnosis in virtue of recent advances in machine learning algorithms and hardware [3]. Although the previous work on the artificial intelligence (AI)-based diagnosis have successfully shown great potential for deep learning, it is not easy to achieve a high accuracy with deep learning approaches in practical systems with poor training dataset. In the field of medical diagnosis, it is hard to establish a large, high quality dataset due to expensive annotation [4], privacy [5], and scarcity of diseases [6]. Hence, dealing with the problems caused by poor dataset is essential to improve the deep learning-based diagnosis performance in practical scenarios.

Recently, crowdsourcing annotation via Amazon Mechanical Turk and Crowdflower has received attention as an effective solution to alleviate the annotation problem [7]. However, the crowdsourcing cannot clearly solve the annotation problem in the clinical diagnosis by introducing inconsistent labeling problem. Due to the nature of its symptom-based diagnosis, the results of clinical diagnosis are relatively more dependent on the annotator's expertise and personal experience compared to other fields. In addition, the limited number of data sources and annotators with expertise, the clinical dataset consists of a small number of data samples and can be drastically updated by the participation of new annotators and the additional data sources. Such frequent and drastic updates give rise to significant computational cost for re-training the deep neural network (DNN) model. For the computational cost reduction and the incremental performance improvement with the dataset update, the quick adaptation of DNN model to the updated dataset is required.

In order to cope with the partially and inconsistently labeled data, and dynamic dataset update, we propose a novel meta-learning based clinical diagnosis method that takes account of the annotator's expertise by introducing the weight on labels. Although there have been some previous works on learning from multiple annotators with varying expertise [8], [9], they have not provided the solution to deal with inconsistent and duplicated labels for the same data from multiple annotators. We apply the proposed method to the evaluation of squat exercise, which is one of the representative rehabilitation activities that help prevent injuries, strengthens core muscles, and improves balance and posture [10]. Specifically, based on inertial measurement unit (IMU) sensor data, the proposed method learns to detect the timing of posterior pelvic tilt, which is a critical factor for diagnosing low-back problem and evaluating athletic performance [11], in the descent phase of squatting. Experiment results show that the proposed clinical diagnosis method outperforms the conventional learning-based approaches, including transfer learning, in terms of not only the convergence speed of DNN model but also the timing gap with ground-truth. The contributions of our work can be summarized as follows:

- To the best of our knowledge, this is an initial work that tackles the practical challenges of deep learning for autonomous clinical diagnosis, such as a large variation in labeling pattern of annotators with varying expertise, partially labeled data samples, small dataset size, and dynamic dataset update.

978-1-6654-8550-0/22/$31.00 ©2022 IEEE

- We propose a novel deep learning framework for mitigating the problems caused by the partially and inconsistently labeled data, and dynamic dataset update in clinical diagnosis.
- We develop IMU-based wearable devices for sensing body movement. Based on the collected sensor data, we create a new squat dataset with the data annotation of physiatrists.
- The performance of the proposed methodology is evaluated with posterior pelvic tilt detection from the constructed squat dataset. Experiment results show that the proposed method can achieve better convergence speed and mean squared error (MSE) than the transfer learning-based approach developed for shortening the training time and mitigating the problems caused by small dataset size [12].
- Besides the clinical diagnosis, the proposed deep learning framework can also be utilized for solving other types of problems with partially and inconsistently labeled data, and dynamic dataset update.

II. PROBLEM DESCRIPTION

In a squat, the timing of posterior pelvic tilt provides useful information for diagnosing low-back problem and evaluating athletic performance. For autonomous disorder diagnosis and athletic performance evaluation, we consider the problem of detecting posterior pelvic tilt timing in the IMU sensor data of the squat movement.

Each squat data is constructed by V-dimensional sensing data of R IMU sensors for S sampling periods. The i-th squat data is denoted as $x_i \in \mathbb{R}^{R\times V\times S}$. Since the annotators are assumed to have different levels of experience and expertise in diagnosis, the annotation results for the same data can be different between annotators. To deal with such inconsistency, we combine the annotation results after assigning weights to annotators according to the level of annotator's experience and expertise. Specifically, the ground truth label for data x_i is defined as

$$y_i = \sum_{a \in \mathcal{A}} \frac{\lambda_a}{\sum_{a \in \mathcal{A}} \lambda_a} y_{i,a}, \tag{1}$$

where $\mathcal{A}$, λ_a, and $y_{i,a}$ denote the set of all annotators, the weight assigned to annotator a, and the annotation result of annotator a for data x_i, respectively. However, in practical scenarios, it is hard for each annotator to participate in the annotation of all data samples due to the limited processing capability of human. For this reason, only a subset $\mathcal{A}_i \subset \mathcal{A}$ of annotators annotate data x_i, and the corresponding combined label is represented by

$$\bar{y}_i = \sum_{a \in \mathcal{A}_i} \frac{\lambda_a}{\sum_{a \in \mathcal{A}_i} \lambda_a} y_{i,a}. \tag{2}$$

Hence, the combined label (2) is actually available for training the model instead of (1). Different annotator set of data leads to the reliability variation of the combined label $\bar{y}_i$ and it makes hard to learn the generalized rule for detecting posterior pelvic tilt with conventional deep learning technique.

Furthermore, due to the nature of clinical data, the size of clinical dataset is generally small. For this reason, the dataset can be drastically updated by the participation of new annotators and the additional data sources. Specifically, the dataset $\mathcal{D} = \{(\boldsymbol{x}_i, \{y_{i,a} : a \in \mathcal{A}_i\})\}_{i=1}^{N}$ can be updated by $\mathcal{D}'$ with more data samples $N' \geq N$ and $\mathcal{A}_i' \supset \mathcal{A}_i$. In addition, with the new annotators and re-arrangement of weights, the annotator weights $\Lambda = \{\lambda_1, \lambda_1, \ldots, \lambda_{|\mathcal{A}|}\}$ can be updated by $\Lambda' = \{\lambda'_1, \lambda'_2, \ldots, \lambda'_{|\mathcal{A}'|}\}$ for $\mathcal{A}' \supset \mathcal{A}$. Such drastic updates require additional training process, and it causes significant computational cost and time for re-training the DNN model in the conventional deep learning algorithm.

Eventually, the objective of our work is to learn initial model parameters θ that can be quickly adapted to the updates in $\mathcal{D}'$ and Λ' so as to well approximate the updated ground truth

$$y_i' \approx f_\theta(\boldsymbol{x}_i) \quad \text{for } (\boldsymbol{x}_i, \{y_{i,a} : a \in \mathcal{A}_i'\}) \in \mathcal{D}', \tag{3}$$

where y_i' denotes the ground truth label (1) computed with the updated annotator set $\mathcal{A}'$, and $f_\theta(\cdot)$ denotes DNN model with parameters θ.

III. QUICK MODEL ADAPTATION WITH META-LEARNING-BASED APPROACH

Meta-learning, also known as learning to learn, aims to learn a general purpose learning algorithm that can generalize across multiple tasks and enables new task to be learned quickly. For the quick adaptation to the updates $\mathcal{D}'$ and Λ', we adapt the model-agnostic meta-learning (MAML) approach, which learns an initialization of model parameters so that a new task can be learned with a few gradient update steps [13]. In other words, the initial model parameters θ that can quickly adapt to a new task $\mathcal{T}' = (\mathcal{D}', \Lambda')$ should be learned from the dataset $\mathcal{D}$.

First of all, we generate new tasks $\mathcal{T}^{(m)}$ by randomly trimming data samples and annotations and randomly re-arranging annotator weights of the original task $\mathcal{T} = (\mathcal{D}, \Lambda)$ as follows

$$\begin{aligned}\mathcal{T}^{(m)} &= \left(\mathcal{D}^{(m)}, \Lambda^{(m)}\right) \\ &= \left(\left\{\left(\boldsymbol{x}_i, \left\{y_{i,a} : a \in \mathcal{A}_i^{(m)}\right\}\right)\right\}_{i=1}^{N^{(m)}}, \left\{\lambda_1^{(m)}, \lambda_2^{(m)}, \ldots, \lambda_{|\mathcal{A}^{(m)}|}^{(m)}\right\}\right),\end{aligned} \tag{4}$$

where, $\mathcal{A}^{(m)} \subset \mathcal{A}$, $\mathcal{A}_i{}^{(m)} \subset \mathcal{A}_{i,}$ $N^{(m)} \leq N$, and $\lambda_k{}^{(m)}$ denote the total annotator set, annotator set of data x_i, number of data samples, and weight of annotator k, respectively, in the trimmed task m. Such tasks can be utilized to learn internal features applicable to various tasks. Specifically, the model parameters θ adapt to a task $\mathcal{T}^{(m)}$ via stochastic gradient descent (SGD)

$$\theta^{(m)} \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}^{(m)}}(f_\theta) \tag{5}$$

where α denotes a learning rate and $\mathcal{L}_{\mathcal{T}^{(m)}}(\cdot)$ denotes a loss function that evaluates the model parameters for a given task $\mathcal{T}^{(m)}$. To take account of the reliability of the combined label, weighted mean squared error (WMSE) is adopted as a loss function

$$\mathcal{L}_{\mathcal{T}^{(m)}}(f_\theta) = \frac{1}{|\mathcal{B}^{(m)}|} \sum_{i \in \mathcal{B}^{(m)}} \sum_{a \in \mathcal{A}_i^{(m)}} \lambda_a^{(m)} \left| f_\theta(\boldsymbol{x}_i) - \bar{y}_i^{(m)} \right|^2, \tag{6}$$

where $\mathcal{B}^{(m)}$ denotes a batch sampled from $\mathcal{D}^{(m)}$, and $\bar{y}_i{}^{(m)}$ denotes the combined label of x_i for participating annotator set $\mathcal{A}_i{}^{(m)}$ and annotator weight $\Lambda^{(m)}$. With the summation term of weights in (6), the data sample that is annotated by highly experienced annotators has more influence on the model parameter adaptation (5).

Based on the parameter adaptations for M trimmed tasks, we derive the generalized model parameters θ that enables model to adapt quickly to a new task by solving the following problem

$$\begin{aligned} &\arg\max_\theta \sum_{m=1}^{M} \mathcal{L}_{\mathcal{T}^{(m)}}\left(f_{\theta^{(m)}}\right) \\ &= \arg\max_\theta \sum_{m=1}^{M} \mathcal{L}_{\mathcal{T}^{(m)}}\left(f_{\theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}^{(m)}}(f_\theta)}\right). \end{aligned} \tag{7}$$

Based on (7), the optimized parameters θ across the trimmed tasks are derived via SGD as follows

$$\theta \leftarrow \theta - \beta \nabla_\theta \sum_{m=1}^{M} \mathcal{L}_{\mathcal{T}^{(m)}}\left(f_{\theta^{(m)}}\right). \tag{8}$$

Eventually, the proposed meta-learning-based model generalization for clinical diagnosis can be summarized as algorithm 1.

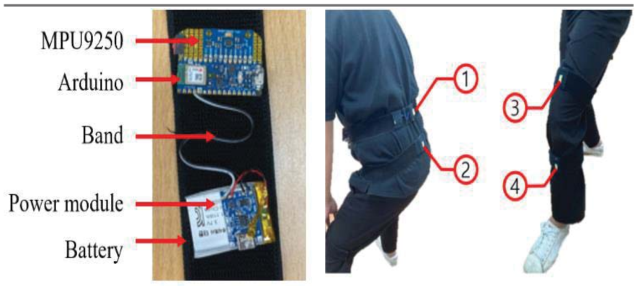
IV. Experimental results

A. Measurement Settings

To collect motion data, we develop wearable devices with IMU sensors. Fig. 1-(a) shows the structure of the developed wearable device. 9-axis IMU sensor MPU9250 can measure acceleration, angular velocity, and magnetic strength. Arduino Nano 33 BLE board is used to record the sensor data to database via Bluetooth. In order to get rid of any restrictions on movements, rechargeable lithium polymer ion battery is utilized for supplying power without wired connection. Based on the medical advice from physiatrist, the developed devices are placed on R = 4 parts of body as shown in fig. 1-(b). Two out of four devices are placed on top of lumbar spine L4 and S2. The other two devices are placed at the one-third point between the patella and pelvis bone on the left thigh and the midpoint of the patella and ankle bone on the left calf. Eventually, R = 4 IMU sensors measure the V = 3-dimensional sensing data of squat motion with a sampling frequency of 20 Hz for 5 seconds. Environmental parameters are summarized in table I.

Algorithm 1 Meta-learning-based model generalization for clinical diagnosis

```
Require: Original task $\mathcal{T} = (\mathcal{D}, \Lambda)$
Require: Learning rates $\alpha, \beta$
Initialize parameters $\theta$ randomly
while not done do
    Generate $M$ trimmed tasks $\mathcal{T}^{(m)}$ from $\mathcal{T}$
    for all $\mathcal{T}^{(m)}$ do
        Sample batch $\mathcal{B}^{(m)}$ of datapoints from $\mathcal{D}^{(m)}$
        Evaluate $\mathcal{L}_{\mathcal{T}^{(m)}}(f_\theta)$ with $\mathcal{B}^{(m)}$
        $\theta^{(m)} = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}^{(m)}}(f_\theta)$
        Sample batch $\tilde{\mathcal{B}}^{(m)}$ of datapoints from $\mathcal{D}^{(m)}$
    end for
    Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{m=1}^{M} \mathcal{L}_{\mathcal{T}^{(m)}}(f_{\theta^{(m)}})$ with $\tilde{\mathcal{B}}^{(m)}$ for each $m$
end while
```



(a) Components of device (b) Device placement

Fig. 1. Experiment setup

B. Dataset Construction

With the developed devices, the squat motion dataset is constructed through the collaboration with physiatrists in department of physical medicine & rehabilitation, Kosin University gospel hospital. The clinical dataset is composed of 1268 data samples that are measured from 4 subjects and annotated by 2 physiatrists.

In order to overcome the limited number of annotators, we create 8 virtual annotators with different annotation patterns by adding some noise to the labels annotated by physiatrist $a \in \{1, 6\}$. For instance, the virtual annotator $a' \in \mathcal{A}_i$ is assumed to annotate the data x_i as follows

$$y_{i,a'} = y_{i,a} + w_{i,a'}, \tag{9}$$

where $y_{i,a}$ denotes a label annotated by physiatrist $a \in \{1, 6\}$, and $w_{i,a'} \sim \mathcal{N}(\mu_{a'}, 1)$ denotes the variation from $y_{i,a'}$. Specifically, the labels of virtual annotator $a' \in \{2, 3, 4, 5\}$ or $a' \in \{7, 8, 9, 10\}$ are generated by adding noise to labels of physiatrist a = 1 or a = 6, respectively. The noise means are $\mu_2 = \mu_7 = 2$, $\mu_3 = \mu_8 = -2$, $\mu_4 = \mu_9 = 4$, $\mu_5 = \mu_{10} = -4$. Note that $\mathcal{A} = \{1, 2, 3, 4, 5\}$ and $\mathcal{A}' = \mathcal{A} \cup \{6, 7, 8, 9, 10\}$.

TABLE I. ENVIRONMENTAL PARAMETERS

Parameters	**Value** [unit]
No. of IMU sensors, R	4 [sensors]
Dimension of a sensing sample, V	3 [dimensions]
No. of samples in a data sample, S	100 [samples]
No. of annotators in original task, $\lvert\mathcal{A}\rvert$	5 [annotators]
No. of annotators in original task, $\lvert\mathcal{A}'\rvert$	10 [annotators]
No. of trimmed tasks, M	3 [tasks]
Original dataset size, N	400 [data samples]
Updated dataset size, N'	800 [data samples]

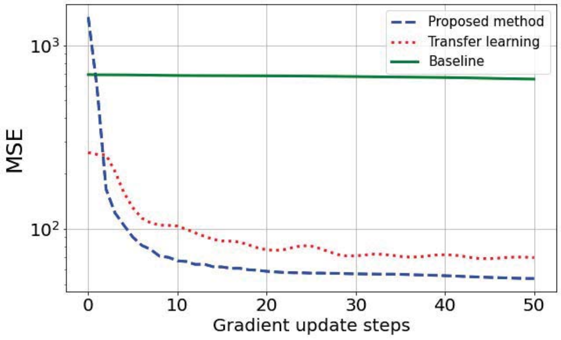


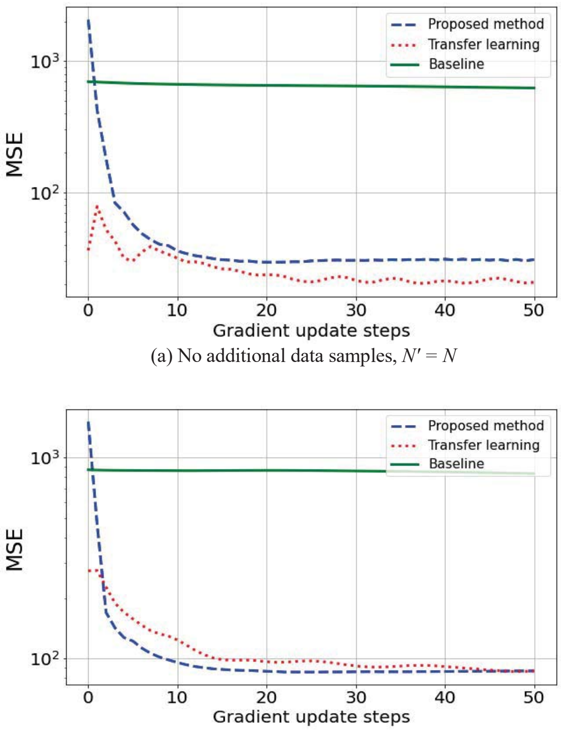
Fig. 2. Training progress comparison

In the original task $\mathcal{T}$, the annotation set $\mathcal{A}_i$ for data $i \in \{1, 2, \ldots, N\}$ is randomly generated while satisfying $|\mathcal{A}_i| = 3$. On the other hand, in a new task $\mathcal{T}'$, the annotation set $\mathcal{A}_i'$ for data $i \in \{1, 2, \ldots, N'\}$ is randomly generated by adding two additional annotators among $\{6, 7, 8, 9, 10\}$.

C. Performance Comparison

In this subsection, simulation results show the performances of the proposed framework and the conventional learning-based approaches in terms of the convergence speed and the converged MSE. We consider two conventional learning methods, denoted by *Baseline* and *Transfer learning* in figures. In the baseline method, a simple supervised learning is conducted for the updated task $\mathcal{T}'$ with randomly initialized model parameters. In the transfer learning method, the model is initialized with the model parameters trained for the original task $\mathcal{T}$ and is fine-tuned to the updated task $\mathcal{T}'$. For all learning methods, we adopt the same convolutional neural network (CNN) model consisting of two convolution layers and a single fully connected layer. In all simulation results, the inference accuracy is quantified by MSE between ground truth label and inference, $\mathbb{E}[|y_i' - f_\theta(\boldsymbol{x}_i)|^2]$.

Fig. 2 shows MSE of training model inference with respect to gradient update steps in the situation where the task is updated by new annotators, additional annotations, annotator weight re-arrangement, and new data samples according to section IV-B. All methods are shown to reduce the error as the training progresses; however, there are big performance gaps between them in the convergence speed and the converged MSE. Even though the transfer learning-based method is shown to achieve MSE comparable to the proposed method in the early stage of training process by exploiting the similarity between tasks, its initial parameters fully-fitted to the original task $\mathcal{T}$ makes it hard for the model to adapt to a new task $\mathcal{T}'$. On the other hand, from the observation that the proposed learning framework outperforms the other methods in terms of the convergence speed and converged MSE, we can see that the proposed framework is able to effectively derive the generalized initial model representation across various tasks.



(a) No additional data samples, $N' = N$

(b) No additional annotators, $\mathcal{A}' = \mathcal{A}$

Fig. 3. Training progress with restricted task updates

In order to see the effects of task update factors on the performance separately, fig. 3 shows training progresses with the updated task without (a) additional data samples, $N' = N = 400$, and (b) additional annotators, $\mathcal{A}' = \mathcal{A} = \{1, 2, 3, 4, 5\}$. All environmental parameters except the restriction factor are the same with fig. 2. From two panels of fig. 3, the proposed framework is shown to achieve significant performance gains in both types of task updates. Furthermore, we can see that the proposed framework is relatively more effective to the task update with additional annotators than the update with additional data samples.

V. CONCLUSIONS

In this paper, we have proposed a learning-based clinical diagnosis framework where the model training is conducted with the dataset in poor conditions with inconsistently and partially labeled data from multiple annotators. The proposed framework has dealt with such challenges by introducing the compromised label with the weighted combination of the inconsistent labels and adapting a meta-learning approach for the generalized initial model parameters. In addition, we have developed wearable devices with IMU sensors to construct a clinical dataset for posterior pelvic tilt detection in squat motion. Experimental results have shown that the proposed framework outperforms the conventional learning-based approaches in terms of convergence speed and MSE. It has also been shown that the proposed framework is relatively more effective for the task update with additional annotators than the update with additional data samples.

Acknowledgment

This research was supported by the National Research Foundation (NRF), Korea, under project BK21 FOUR (Smart Robot Convergence and Application Education Research Center)

References

[1] R. Komatireddy, A. Chokshi, J. Basnett, M. Casale, D. Goble, and T. Shubert, "Quality and quantity of rehabilitation exercises delivered by a 3-d motion controlled camera: A pilot study," *Int. J. Phys. Med. Rehabil.,* vol. 2, no. 4, p. 214, Aug. 2014.

[2] J. Lee, H. Joo, J. Lee and Y. Chee, "Automatic classification of squat posture using inertial sensors: Deep learning approach," *Sensors*, vol. 20, no. 2, p. 361, Jan. 2020.

[3] Y. Liao, A. Vakanski, and M. Xian, "A deep learning framework for assessing physical rehabilitation exercises," *IEEE Trans. Neural Syst. Rehabil, Eng*., vol. 28, no. 2, pp. 468-477, Jan. 2020.

[4] X. Li, L. Yu, Y. Jin, C. W. Fu, L. Xing, and P. A. Heng, "Difficulty-aware meta-learning for rare disease diagnosis," *in Proc. Int. Conf. Med. Imgage Comput. Comput.-Assist. Intervent. (MICCAI)*, Springer, pp. 357-366, Oct. 2020.

[5] G. Kaissis, A. Ziller, J. Passerat-Palmbach, T. Ryffel, D. Usynin, A. Trask, I. Lima, J. Mancuso, F. Jungmann, M.-M. Steinborn, A. Saleh, M. Makowski, D. Rueckert, and R. Braren, "End-to-end privacy preserving deep learning on multi-institutional medical imaging," *Nature Mach. Intell*., vol. 3, no. 6, pp. 473–484, Jun. 2021.

[6] R. Tanno, A. Saeedi, S. Sankaranarayanan, D. C. Alexander and N. Silberman, "Learning from noisy labels by regularized estimation of annotator confusion," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 11244–11253. Jun. 2019.

[7] F. Rodrigues and F. C. Pereira, "Deep learning from crowds," *in Proceedings of the 32nd AAAI Conf. Artif. Intell.*, Apr. 2018.

[8] Y. Yan, R. Rosales, G. Fung, R. Subramanian, and J. Dy, "Learning from multiple annotators with varying expertise," *Mach. Learn.*, vol. 95, no. 3, pp. 291–327, Jun. 2014.

[9] R. Xiaoqian, W. Gaoang, "Disjoint contrastive regression learning for multi-sourced annotations," *arXiv Prepr*. arXiv:2112.15411, 2021.

[10] L. Vecchio, H. Daewoud and S. Green "The health and performance benefits of the squat, deadlift, and bench press," *MOJ Yoga Phys. Ther.*, vol. 3, no. 2, pp. 40-47., Mar. 2018.

[11] T. Tani et al., "Posterior pelvic tilt from supine to standing in patients With symptomatic developmental dysplasia of the hip," *J. Orthop. Res.*, vol. 38, no. 3, pp. 578-587, Mar. 2020.

[12] G. Liang and L. Zheng, "A transfer learning method with deep residual network for pediatric pneumonia diagnosis," *Comput. Methods. Programs Biomed.*, vol. 187, no. 104964, Apr. 2020.

[13] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," *in Proc. 34th Int. Conf. Mach. Learn. (PMLR)*, pp. 1126-1135, Aug. 2017.