Semi-Supervised Active Learning for COVID-19 Lung Ultrasound Multi-symptom Classification

1st Lei Liu, 1st Wentao Lei,
3th Xiang Wan, 4th Li Liu*
Shenzhen Research Institute of Big Data,
The Chinese University of Hong Kong, Shenzhen
Shenzhen, China
{leiliu,wentaolei}@link.cuhk.edu.cn,
wanxiang@sribd.cn, liuli@cuhk.edu.cn

1st Yongfang Luo, 2nd Cheng Feng
Department of Medical Ultrasonics,
National Clinical Research Center for Infectious Disease,
Shenzhen Third People's Hospital (Second Hospital Affiliated to Southern University of Science and Technology)
Shenzhen, China
luoyongfang2005@foxmail.com, chaosheng-01@szsy.sustech.edu.cn

***Abstract*—Ultrasound (US) is a non-invasive yet effective medical diagnostic imaging technique for the COVID-19 global pandemic. However, due to complex feature behaviors and expensive annotations of US images, it is difficult to apply Artificial Intelligence (AI) assisting approaches for the lung's multi-symptom (multi-label) classification. To overcome these difficulties, we propose a novel semi-supervised Two-Stream Active Learning (TSAL) method to model complicated features and reduce labeling costs in an iterative manner. The core component of TSAL is the multi-label learning mechanism, in which label correlation information is used to design a multi-label margin (MLM) strategy and a confidence validation for automatically selecting informative samples and confident labels. In this framework, a multi-symptom multi-label (MSML) classification network is proposed to learn discriminative features of lung symptoms, and a human-machine interaction (HMI) is exploited to confirm the final annotations that are used to fine-tune MSML. Moreover, a novel lung US dataset named COVID19-LUSMS is built, currently containing 71 clinical patients with 6,836 images sampled from 678 videos. Experimental evaluations show that TSAL can achieve superior performance to the baseline and the state-of-the-art using only 20% data. Qualitatively, visualization of the attention map confirms a good consistency between the model prediction and the clinical knowledge.**

***Index Terms*—COVID-19, Ultrasound Imaging, Multi-Label Classification, Active Learning, Semi-Supervised Learning**

I. Introduction

The novel coronavirus (COVID-19) has spread worldwide and is now officially a global pandemic. Typical diagnosing tools mainly include Computed tomography (CT) and X-ray, which are characterized by their relatively accurate performances [1]. However, due to the prevalence of COVID-19, in practice, deep learning-based CT or X-ray approaches remain several challenges. Firstly, CT and X-ray tools are generally inflexible and involve extra radiations. Secondly, images of CT and X-ray are not easy to collect from COVID-19 patients because the imaging procedures involve isolated patients, complex clinical equipment, and many other nontrivial processes.

In contrast, lung ultrasound (US) imaging is preferred as a mature tool for its fast, flexible, and reliable deployment, especially in emergencies [2]. More importantly, it is non-invasive and can work at the bedside. Recently, some works [3]–[5] focused on COVID-19 symptom detection based on lung US images. Indeed, based on lung US images, automatic AI assisting approaches for COVID-19 symptoms classification are significant for medical diagnoses. Therefore, we focus on lung US multi-symptom classification in this work.

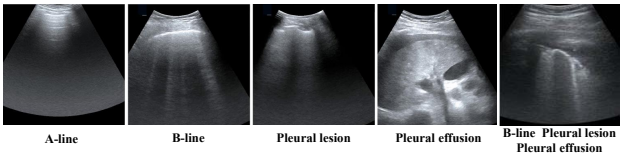


Fig. 1. Examples of the COVID19-LUSMS dataset.

In practice, the automatic classification of COVID-19 lung symptoms is difficult for twofold reasons. Firstly, the lung US images of COVID-19 patients may simultaneously present multiple symptoms, which exhibit complicated image features (see Fig. 1). One possible solution is the multi-label learning, which targets to judge whether an image possesses multiple characteristics denoted by labels [6], [7]. Secondly, it is expensive and tedious to collect and annotate numerous COVID-19 lung US images. To address this difficulty, a feasible solution is active learning [8], [9], which aims to achieve satisfactory performance given a limited labeling cost.

In this paper, we propose a novel semi-supervised Two-Stream Active Learning (TSAL) method, which works iteratively by sample selection, pseudo-label validation, human-machine interaction and CNN parameters updating. In TSAL, a multi-symptom multi-label (MSML) classification network is constructed as the basic model for feature learning. The sample stream works for informative sample selection by newly designed multi-label margin strategy (MLM), while the label stream is exploited to assign confident pseudo-label for selected images. Then HMI is used for confirming the final annotations to fine-tune the MSML. An overview of the proposed method is shown in Fig. 2.

Besides, a large-scale dataset of lung US images for COVID-19 is built for this work. Some examples are shown in Fig. 1. Experiments on this dataset show that our proposed

* Corresponding author

DOI 10.1109/ICTAI50040.2020.00191

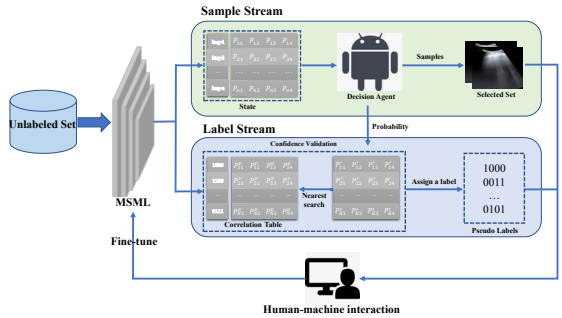


Fig. 2. Overview of the proposed MSML-TSAL. Sample stream is to select informative samples by state and decision agent, which are used to describe the prediction results of all unlabeled images and determine which image candidate will be selected. Label stream is to validate and assign pseudo labels for the selected images. Human-machine interaction is to confirm the final annotations for model updating. This iterative operation terminates when it reaches the annotation budget or required performance.

method achieves superior classification performance, outperforming the baseline models and the state-of-the-art (SOTA) using less than 20% of the labeled images. To the best of our knowledge, this is the first work on automatic multi-symptom multi-label classification for COVID-19 lung US images.

In summary, this work contains the following three contributions:

1) A novel large-scale dataset containing lung US images of COVID-19 is built specifically for this work. This dataset is annotated in the multi-label form by medical experts.
2) We propose a novel semi-supervised TSAL method, which effectively reduces the labeling cost. It exploits label correlation information to select informative samples and confident pseudo labels.
3) Experimental results show that our method achieves superior performance using less than 20% labeled data, compared with baselines and SOTA. Explainable analyses using attention map confirm that our results are well consistent with clinic expertise.

II. Related Work

A. Lung US techniques for COVID-19

Many studies [2], [10], [11] reported the superiority of US imaging in diagnosing pneumonia and related lung conditions. Sloun *et al.* [12] proposed a fully convolutional neural network (CNN) [13], [14] to identify and localize the B-lines in clinical lung ultrasonography. Born *et al.* [3] trained a POCOVID-Net on a 3-class dataset and achieved good accuracy in classification. [15] presented a novel deep network that simultaneously predicted the disease severity score associated with an input frame. However, they generally require a large-scale labeled US-image dataset, where annotations are expensive.

B. Multi-label image classification

As an important branch of classification, multi-label classification [16], [17] has been widely explored in recent years, supported by CNNs. It plays an important part in bridging the gap between low-level features and high-level semantic information [6]. Some multi-label classification methods applied CNNs to obtain competitive performances. [18] optimized top-k ranking objectives combined with convolutional architectures to learn a better feature representation. Hypotheses-CNN-Pooling [19] incorporated object segmentation hypotheses with max pooling to generate multi-label predictions. For lung US images, a COVID-19 patient may perform multiple symptoms simultaneously. Thus, COVID-19 symptoms classification can be formulated as a multi-symptom multi-label classification task.

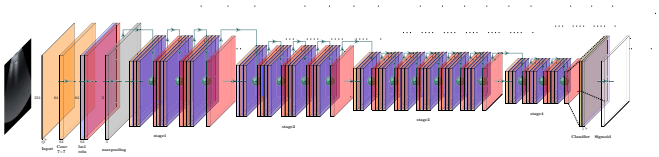


Fig. 3. Architecture of the MSML model.

C. Active learning

AL has been successfully deployed into semantic segmentation [20], image classification [21], human pose estimation [8], *etc*. These applications indicate that AL is a great choice for labeling efforts reduction. Besides, there have been many popular selection strategies in the literature, mainly including query by committee [22], expected error reduction [23], expected model change [24], and uncertainty sampling [25]. These strategies are usually exploited for single-label classification tasks, which are not suitable for COVID-19 lung US image classification with complex multi-label feature behaviors. In this work, we introduce AL into multi-symptom multi-label classification to actively reduce the labeling efforts.

III. Methodology

A. Multi-symptom multi-label model

The detailed architecture of the proposed MSML network is shown in Fig. 3. The model backbone is ResNet50 [13] pre-trained on ImageNet for its competitive performance and its relatively concise structure. Some modifications are made based on it, firstly, the hidden fully-connected layer is removed, and layers after the final convolutional layer are also removed to avoid too deep or too complicated architecture. Then it is followed by one specifically designed classifier consisting of average pooling fully-connected layer for multi-label tasks. To effectively learn the discriminative features of COVID-19 pulmonary symptoms, we adopt sigmoid cross-entropy loss, which can be calculated as:

$$\mathrm{L_{CE}} = -\frac{1}{N}\sum_{i=1}^{N}(y^{(i)}\log\hat{y}^{(i)} + (1-y^{(i)})\log(1-\hat{y}^{(i)})), \quad (1)$$

where $\hat{y}^{(i)} = 1/1 + e^{-x}$, and $y^{(i)}$ is the ground truth of the input, N is the batch size and x is the output of the last layer.

B. Two-stream deep active learning framework

The detailed framework of MSML-TSAL is presented in Fig. 2. Firstly, we regard all unlabeled images as the candidate pool. At each AL iteration t, in the sample stream, the MSML network provides a prediction state S_t for unlabeled images. Then decision agent makes an action for sample selection according to the state and selection strategy. Then pseudo labels are assigned by confidence validation in the label stream. The final annotations would be confirmed by HMI to fine-tune the MSML model. This iterative operation repeats until the expected performance of MSML or the empty candidate pool.

1) Sample stream: **State**: The state is utilized to describe the relationship between unlabeled images and the prediction capability of the model. Prediction probability has been widely exploited to measure the prediction capability of the model in AL tasks. In this work, we exploit output prediction probabilities to construct the state matrix. At each AL iteration t, the candidate pool is denoted by $D = \{d_1, d_2, ..., d_n\}$, where d_i is the ith unlabeled sample, and n is the pool size. The prediction probability vectors are extracted by MSML model. The ith prediction vector can be written as $p_i = (p_{i1}, p_{i2}, ..., p_{il})^{\mathrm{T}}$, where l is the number of the labels. The state matrix S_t can be denoted as $S_t = (p_1, p_2, ..., p_n)$.

Action The action is to select unlabeled images from the candidate pool. At each iteration t, the selected set $A_t = \{a_1, a_2, ..., a_{K_{max}}\}$ is decided by the decision agent according to state S_t, where a_i is the ith selected sample, K_{max} is the pre-defined annotation efforts (*i.e.*, account of the labeled samples) for each AL iteration. Once an action is executed, the selected samples are removed from the candidate pool.

Decision Agent: The decision agent is used to measure which image is worth annotating using selection strategies. Firstly, we note that there are few active strategies for the multi-label classification task. To adapt to this task, we redesign two classical strategies of multi-class classification, including Least Confidence [26] and Entropy [27].

Least confidence (LC): Lower confidence of an image illustrates that it is hard for the classifier to make a correct prediction. The calculation of LC is:

$$\mathrm{LC}(x) = \max_{1 \le i \le l} p(l_i \mid x), \tag{2}$$

where l is the number of labels (*i.e.*, symptoms). $p(l_i|x)$ denotes the prediction probability of symptom l_i appearing in image x.

Multi-label entropy (MLE): Higher entropy indicates that the image carries rich information. The MLE is denoted as:

$$\mathrm{MLE}(x) = \sum_{i=1}^{l} (p(l_i \mid x) \log p(l_i \mid x) + p(\overline{l_i} \mid x) \log p(\overline{l_i} \mid x)), \tag{3}$$

where $p(l_i|x)$ is the probability of symptom l_i appearing in image x and $p(\overline{l_i}|x) = 1 - p(l_i|x)$.

We find that LC prones to select the noisy samples and MLE doesn't consider the relation among different symptoms.

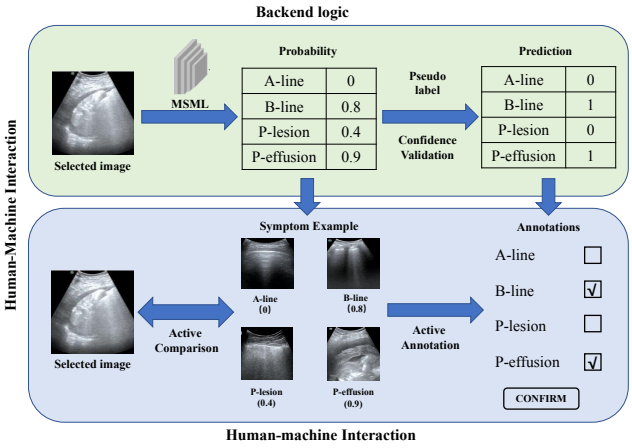


Fig. 4. Overview of the human-machine interaction.

Thus, we specially design a multi-label learning strategy called multi-label margin (MLM) to evaluate the informativeness of each unlabeled sample. The MLM is defined as:

$$\mathrm{MLM}(x) = \left| p(l_1 \mid x) - \max_{2 \le i \le l} p(l_i \mid x) \right|, \tag{4}$$

where $p(l_1|x)$ is the probability of A-line appearing in image x. For lung US images of COVID-19 patients, A-line denotes the health while others denote the disease. Thus, from the view of the medical knowledge, it is not reasonable that A-line appears with other symptoms simultaneously in an image. From the perspective of the model prediction, it is difficult to judge whether this image is healthy or unhealthy if the probability of A-line has a small margin with other symptoms, which may indicate that model has not learned effective information of this image. Thus, this margin intuitively can measure the informativeness of the unlabeled images.

2) Label stream: **Confidence validation**: Label correlations information has been widely employed for multi-label learning [28] by mining the potential relationships among different labels, which also can be exploited for confidence validation. For convenience, we call "0" or "1" of a single symptom as a single label, and call "0001" ("0110", "0111", *etc*) of four symptoms as label combination. In this work, we construct a correlation table to store the probability distribution information for each label combination.

Given the selected and labeled samples at each AL iteration t, several probability matrices $\{P_1, P_2, \cdots, P_N\}$ can be built according to different label combinations. N is the amount of the label combinations, which is $2^{l-1} + 1$. The matrix of nth label combination can be written as:

$$P_n = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1l} \\ p_{21} & p_{22} & \cdots & p_{2l} \\ \vdots & \vdots & \vdots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{ml} \end{pmatrix}, \ n \in \{1, 2, \ldots, N\}, \tag{5}$$

where l is the number of the labels, m is the number of the labeled samples with nth label combination.

The average probability vector p_n^{avg} for nth probability matrix P_n is calculated as:

$$p_n^{avg} = (\frac{1}{m}\sum_{j=1}^{m} p_{j1}, \frac{1}{m}\sum_{j=1}^{m} p_{j2}, \cdots, \frac{1}{m}\sum_{j=1}^{m} p_{jl}). \quad (6)$$

Then relationship vector for nth label combination $v_n = \{p_{n1}^v, p_{n2}^v, \cdots, p_{nl}^v\}$ can be calculated via normalization operation:

$$p_{ni}^v = \frac{p_n^{avg}(i)}{\sum_{j=1}^{l} p_n^{avg}(j)}, i = \{1, 2, \cdots, l\}, \quad (7)$$

where p_i^v is the ith value of v_n and $p_n^{avg}(i)$ is the ith value of p_n^{avg}. Vector v_n reflects the probability distribution for nth label combination. Vectors set $V = \{v_j | j = 1, 2, \cdots, N\}$ is defined as the correlation table.

Given a prediction probability vector $\hat{p}$, a normalization operation is also executed to transform $\hat{p}$ into the form of relationship vector as $\hat{p}^r$. Then we can search its the top nearest relationship vector in the correlation table, which is obtained through:

$$\mathrm{K}_{(\hat{\mathrm{p}}^{\mathrm{r}},\mathrm{V})} = \underset{v_{j\in(1,2,\cdots,N)}}{\arg\min} \; L(\hat{p}^r, v_j), \quad (8)$$

where L is the Manhattan distance, and K is the relationship vector v_n that has the minimum distance with $\hat{p}^r$. Then the label combination corresponding to $K_{(\hat{p}^r,V)}$ is the most confident label combination for $\hat{p}^r$.

To obtain a more confident correlation table, we update the table after each AL iteration with a constraint. For relationship vector v_n^{new} in the new table, it replaces the corresponding v_n in the previous table only if it satisfies the following condition:

$$v_n^{new} \cdot z_n \geq v_n \cdot z_n, \; n = \{1, 2, \ldots, N\}, \quad (9)$$

where z_n is the nth label combination (*e.g.*, (0, 0, 0, 1)).

3) Human-machine interaction: Given the selected samples and pseudo labels, a HMI is designed for annotators to judge the correctness of annotations and revise the annotations if they are not consistent with the symptom examples. To better understand the HMI, we illustrate the interface in Fig. 4. The first row is the pipeline of pseudo-label generation, which is the backend of the interface. The second row is the user interface, which exhibits examples for each label and the selected image. Besides, the annotations would be made as defaults according to the pseudo label. The human annotator only needs to judge whether the default annotations are consistent with the symptom examples.

4) CNN network updating: MSML is updated with a fine-tuning process. At each AL iteration t, the CNN is fine-tuned via selected samples. During fine-tuning, only weights of the last three layers in MSML are updated, while the remained weights are frozen to the values from the pre-training. When more images are selected and annotated, the model becomes more robust. The renewed network is exploited to update the state.

IV. Experiments

A. Implementation Details

We build the first version of the COVID-19 US dataset, called COVID19-LUSMS v1. US videos are collected in the Third People's Hospital of Shenzhen, China. A total of 71 COVID-19 patients are inspected, including 678 videos. Random rotation (up to 10 degrees) and horizontal flips are used as data augmentation transformations. The Stochastic Gradient Descent optimization is adopted. Learning rate is 2×10^{-3}, batch size is 32 and momentum is set as 0.9. K_{max} is 100 and the AL iteration limit is set as 20. For effective comparison, random strategy is implemented to randomly select samples for annotations.

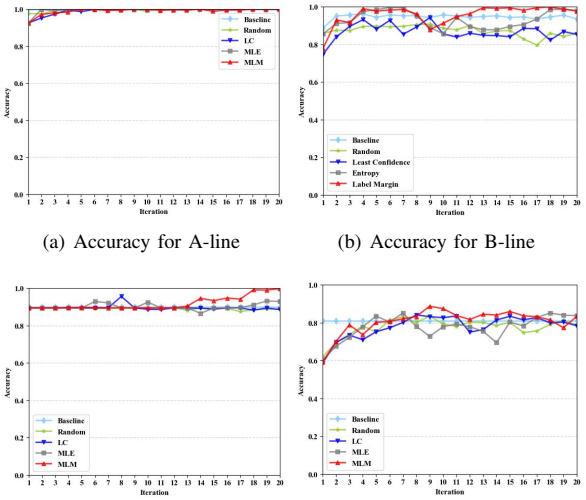
B. Quantitative Analysis

1) MSML: In Tab. I, we illustrate the performance comparisons on four baselines including POCOVID-Net [3], NNBD [12], VGG16 [14], and ResNet [13]. We have the following observations and discussion. (1) Accuracy: the proposed MSML model achieves 100%, 95.72%, 80.98%, and 90.09% accuracy for A-line, B-line, pleural lesion and pleural effusion, respectively. It shows that the MSML model almost outperforms all baseline models concerning accuracy. (2) Sensitivity: MSML model achieves 100%, 98.78%, 81.38%, and 6.08% sensitivity for A-line, B-line, pleural lesion and pleural effusion, respectively. It performs similar sensitivity with baseline models, which is mainly because of the distinct patterns of these symptoms. Besides, all these methods perform poor sensitivity for pleural effusion. We explain that multiple symptoms appearing simultaneously may cause complicated patterns, especially for pleural effusion. (3) Specificity: MSML model achieves 100%, 81.81%, 80.67%, and 100% specificity for A-line, B-line, pleural lesion and pleural effusion, respectively. It shows superior results on all symptoms compared with baseline models, which means it learns a better feature representation compared with baseline models.

2) MSML-TSAL: We report the performance of MSML-TSAL concerning four selection strategies in Tab. I. (1) Accuracy: MLE only uses 27.6% data to train a MSML model, whose accuracy outperforms the baseline models except for pleural effusion. MLM uses 14.7% labeled data to obtain a similar performance as the full training data. Only using 16.6% data, LC obtains comparable accuracy results as the full training set. (2) Sensitivity: all strategies achieve similar sensitivity using fewer images for A-line and B-line, but perform worse for pleural lesion, because the pleural lesion often appears together with other symptoms. It should be mentioned that, for pleural effusion, all strategies obtain near 0 sensitivity when using less than 30% data, because the image of pleural effusion is far less than others. (3) Specificity: these strategies merely exploit 16.6%, 27.6%, 16.6%, and 14.7% data to achieve similar or better specificity performance, among which the random strategy performs worst. Surprisingly, MSML-TSAL improves specificity for B-line and pleural effusion by a large margin, *e.g.*, the specificity for B-line is increased from

TABLE I
COMPARISONS WITH THE BASELINES AND SOTA. THE BOLDS ARE OF OUR METHOD.

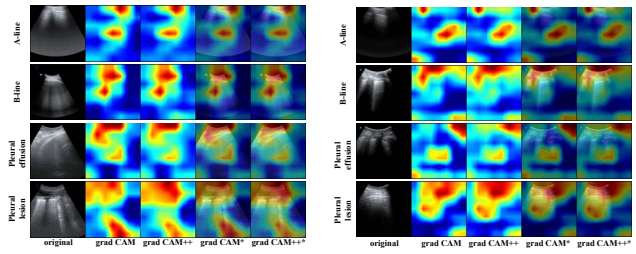
Method	A-line			B-line			P-lesion			P-effusion			data
	Acc	Sen	Spe	Acc	Sen	Spe	Acc	Sen	Spe	Acc	Sen	Spe	
VGG16	100	100	100	88.39	98.34	43.08	60.82	76.77	48.68	88.53	0	98.96	100%
ResNet34	100	100	100	90.88	96.87	63.63	73.36	81.71	67.00	89.31	0	99.84	100%
ResNet50	100	100	100	88.60	98.95	41.50	80.34	82.86	78.41	89.45	0	100	100%
ResNet101	100	100	100	90.45	98.17	55.33	79.91	82.37	78.04	89.45	0	100	100%
POCOVID-Net	100	100	100	84.97	90.35	60.47	80.84	79.90	81.55	91.02	14.86	100	100%
NNBD	100	100	100	90.31	99.91	46.64	71.86	68.20	74.65	89.45	0	100	100%
MSML+TSAL(Random)	99.85	100	100	90.74	98.52	65.61	83.47	77.92	98.87	89.38	0	99.92	16.6%
MSML+TSAL(MLE)	100	100	100	89.52	97.56	52.96	80.34	75.94	83.68	89.45	0	100	27.6%
MSML+TSAL(LC)	100	100	100	94.30	95.91	86.95	83.19	78.74	86.57	89.45	0	100	16.6%
MSML	**100**	**100**	**100**	**95.72**	**98.78**	**81.81**	**80.98**	**81.38**	**80.67**	**90.09**	**6.08**	**100**	**100%**
MSML+TSAL(MLM)	**100**	**92.38**	**100**	**98.50**	**98.79**	**92.49**	**83.26**	**76.77**	**96.36**	**89.45**	**0**	**100**	**14.7%**



(a) Accuracy for A-line (b) Accuracy for B-line

(c) Accuracy for P-effusion (d) Accuracy for P-lesion

Fig. 5. Comparisons of the proposed MSML-TSAL method with different selection strategies on accuracy. Baseline means using all labeled images in the dataset without using any selection strategy. Abscissa indicates the training iterations.



(a) Unselected samples for TSAL (b) Selected samples for TSAL

Fig. 6. Visualized results for MSML-TSAL. (a) The training weights of MSML-TSAL is used for the visualization of the unselected samples. (b) The training weights of MSML-TSAL is used for the visualization of the selected sample. (*) denotes that the attention map is overlapped on the original image.

81.81% to 92.49%. We deduce that the selection strategies can alleviate the unbalanced problem of data distribution.

In Fig. 5, we present the accuracy of MSML-TSAL combined with four selection strategies that achieve the best performance on the COVID19-LUSMS dataset. Almost all curves of different strategies achieve similar or better performance compared with the baseline, which uses the full training set (light blue curves). During the whole training process, these curves have several oscillations. We deduce that features in the COVID19-LUSMS dataset are complex and the amount of labeled samples in each iteration is limited. Besides, we note that MLM (red curves) in Fig. 5 performs the best among the four strategies, with its highest accuracy at the final AL iteration (*i.e.*, the 20th iteration). Concerning smoothness, we can see that the MLM performs the most stable changing tendency.

TABLE II
AMOUNT OF MANUALLY LABELED DATA.

Iteration	1st	2nd	3rd	4th	5th
HMI	100%	100%	100%	100%	100%
HMI + pseudo-label	100%	29%	11%	5%	1%
HMI + pseudo-label + CV	100%	2%	1%	0%	0%

C. Qualitative Analysis

We exploit Grad-CAM [29] to highlight the attention regions in the images. As shown in Fig. 6(a), attention regions from the MSML-TSAL model are consistent with the regions from the doctor. For example, based on the prior knowledge of doctors, A-line has an obvious horizontal-line region in the upper part of the images, while the lower part of the image is dark. Corresponding to the attention heatmaps, the visualized attention regions perform consistent results with doctors' diagnosis. From Fig. 6(b), we find that these images contain large dark regions, which are not pathological changing regions and may cause complicated characteristics for the model learning. These characteristics are not well learned via previous training, because the attention regions are likely to focus on the dark regions as shown in Fig. 6(b). Thus, these images should join in further training for their complicated information.

D. Component analysis for label stream

An ablation study is carried out to justify the efficiency of the label stream for pseudo-label assignment. As shown

in Tab. II, human annotators need to manually annotate all selected images in each iteration, using only HMI without pseudo-label and confidence validation. By pseudo-label, the manual annotations gradually decrease with the improving performance of MSML. Confidence validation can further reduce the manual annotations, *i.e.*, nearly zero after the first iteration. Through label stream, the selected samples can be automatically annotated, thus human annotators only need to confirm in HMI rather than manual annotations.

V. Conclusion

To achieve accurate classification of COVID-19 multiple symptoms of the lung US image with less annotated data, we innovatively propose a TSAL framework to effectively train the MSML model with less labeling efforts in a semi-supervised manner. Specifically, we design a MLM strategy and a confidence validation for TSAL by label correlations information. Moreover, a new large-scale lung US image dataset with multiple COVID-19 symptoms is built in this work. Quantitative and qualitative experimental results show that the TSAL model can achieve competitive performance, and we can train an effective MSML model merely using less than 20% data of the full training set. In future work, it is worthwhile to explore the reinforcement learning to learn a powerful and adaptive policy for image selection.

VI. Acknowledgements

This work is supported by grants from Special Research on Key Technology of "Emergency Prevention from COVID-19 Infection" (LGKCXGZX2020001) and Guangdong Medical Research Fund (Grant B2019163).

References

[1] D. Wang, B. Hu, C. Hu, F. Zhu, X. Liu, J. Zhang, B. Wang, H. Xiang, Z. Cheng, Y. Xiong *et al.*, "Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in wuhan, china," *JAMA*, vol. 323, no. 11, pp. 1061–1069, 2020.

[2] J. E. Bourcier, S. Braga, and D. Garnier, "Lung ultrasound will soon replace chest radiography in the diagnosis of acute community-acquired pneumonia," *Current Infectious Disease Reports*, vol. 18, no. 12, pp. 1534–3146, 2016.

[3] J. Born, G. Brändle, M. Cossio, M. Disdier, J. Goulet, J. Roulin, and N. Wiedemann, "Pocovid-net: Automatic detection of covid-19 from a new lung ultrasound imaging dataset (pocus)," 2020.

[4] Y. Huang, S. Wang, Y. Liu, Y. Zhang, C. Zheng, Y. Zheng, C. Zhang, W. Min, M. Yu, and M. Hu, "A preliminary study on the ultrasonic manifestations of peripulmonary lesions of non-critical novel coronavirus pneumonia (covid-19)," 2020.

[5] Q.-Y. Peng, X.-T. Wang, and L.-N. Zhang, "Findings of lung ultrasonography of novel corona virus pneumonia during the 2019–2020 epidemic," *Intensive Care Medicine*, pp. 1–2, 2020.

[6] L. Li, S. Wang, S. Jiang, and Q. Huang, "Attentive recurrent neural network for weak-supervised multi-label image classification," in *Proceedings of the ACM International Conference on Multimedia*, 2018.

[7] H. Guo, X. Fan, and S. Wang, "Human attribute recognition by refining attention heat map," *Pattern Recognition Letters*, vol. 94, pp. 38–45, 2017.

[8] B. Liu and V. Ferrari, "Active learning for human pose estimation," in *IEEE International Conference on Computer Vision*, 2017.

[9] A. Vezhnevets, J. M. Buhmann, and V. Ferrari, "Active learning for semantic segmentation with expected change," in *IEEE International Conference on computer vision and pattern recognition*, 2012, pp. 3162–3169.

[10] J. E. Bourcier, J. Paquet, M. Seinger, E. Gallard, J. P. Redonnet, F. Cheddadi, D. Garnier, J. M. Bourgeois, and T. Geeraerts, "Performance comparison of lung ultrasound and chest x-ray for the diagnosis of pneumonia in the ed," *American Journal of Emergency Medicine*, vol. 32, no. 2, pp. 115–118, 2014.

[11] A. S. Claes, P. Clapuyt, R. Menten, N. Michoux, and D. Dumitriu, "Performance of chest ultrasound in pediatric pneumonia," *European Journal of Radiology*, vol. 88, no. Complete, pp. 82–87, 2017.

[12] R. J. G. Van Sloun and L. Demi, "Localizing b-lines in lung ultrasonography by weakly supervised deep learning, in-vivo results," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 4, pp. 957–964, 2020.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE International Conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[15] S. Roy, W. Menapace, S. Oei, B. Luijten, E. Fini, C. Saltori, I. Huijben, N. Chennakeshava, F. Mento, A. Sentelli, E. Peschiera, R. Trevisan, G. Maschietto, E. Torri, R. Inchingolo, A. Smargiassi, G. Soldati, P. Rota, A. Passerini, R. J. G. Van Sloun, E. Ricci, and L. Demi, "Deep learning for classification and localization of covid-19 markers in point-of-care lung ultrasound," *IEEE Transactions on Medical Imaging*, 2020.

[16] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining*, vol. 3, no. 3, pp. 1–13, 2009.

[17] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, pp. 1819–1837, 08 2014.

[18] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe, "Deep convolutional ranking for multilabel image annotation," 2014.

[19] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan, "Hcp: A flexible cnn framework for multi-label image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1901–1907, 2016.

[20] Q. Sun, A. Laddha, and D. Batra, "Active learning for structured probabilistic models with histogram approximation," in *IEEE International Conference on computer vision and pattern recognition*, 2015.

[21] W. H. Beluch, T. Genewein, A. Nürnberger, and J. M. Köhler, "The power of ensembles for active learning in image classification," in *IEEE International Conference on computer vision and pattern recognition*, 2018.

[22] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, ser. COLT '92. New York, NY, USA: Association for Computing Machinery, 1992, p. 287–294. [Online]. Available: https://doi.org/10.1145/130385.130417

[23] N. Roy and A. Mccallum, "Toward optimal active learning through monte carlo estimation of error reduction," 01 2001.

[24] W. Cai, Y. Zhang, S. Zhou, W. Wang, C. Ding, and X. Gu, "Active learning for support vector machines with maximum model change," 09 2014, pp. 211–226.

[25] D. D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *Machine Learning Proceedings 1994*, W. W. Cohen and H. Hirsh, Eds. San Francisco (CA): Morgan Kaufmann, 1994, pp. 148 – 156. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B978155860335650026X

[26] B. Settles, "Active learning literature survey," University of Wisconsin–Madison, Computer Sciences Technical Report 1648, 2009.

[27] A. Holub, P. Perona, and M. C. Burl, "Entropy-based active learning for object recognition," in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2008, pp. 1–8.

[28] M. Zhang and Z. Zhou, "A review on multi-label learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, 2014.

[29] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *The IEEE International Conference on Computer Vision*, 2017.