Asia-Pacific Journal of Information Technology and Multimedia 15(1) 2026: 140 – 150
https://doi.org/10.17576/apjitm-2026-1501-08

Weakly Supervised Semantic Segmentation for Tuberculosis Lung Cavity Diagnosis

Segmentasi Semantik Berpenyeliaan Lemah untuk Diagnosis Kaviti Paru-Paru akibat Tuberkulosis

Zhuoyi Tan[1] , Hizmawati Madzin[1] Wei Sun[1], Zeyu Ding[1], Fengzhou Cai[2], Tianyu Nie[3] , Mas Rina Mustaffa[1]*

[1]Universiti Putra Malaysia, Malaysia
[2]University of Southampton, United Kingdom
[3]Chongqing University, China

**Corresponding author: hizmawati@upm.edu.my*

Received 13 June 2025
Accepted 3 March 2026, Available online 30 June 2026

ABSTRACT

Tuberculosis is a worldwide disease that threatens human health, and its early diagnosis is critical for effective treatment. The lung cavity is an important indicator for TB diagnosis, and its detection can provide valuable diagnostic information about tuberculosis lesions. However, traditional supervised learning methods for lung cavity detection usually require large amounts of labeled data, and obtaining these data is a time-consuming and laborious task for tuberculosis images. To address this challenge, a weakly supervised method for lung cavity semantic segmentation is proposed. In this approach, EfficientNet is utilized for co-training with image-level multi-class classification labels to generate regions of interest related to lung cavities. These generated regions are subsequently refined to determine the locations of lung cavities. This research results show that CT images under weak supervision method effectively segment lung cavity lesions. which achieves good performance without pixel-wise full supervision (W), with IoU and DSC of 31.2 % and 44.7%, respectively. It shows that weak supervision methods are in performance and even beyond some fully supervised learning methods.

Keywords**:** Tuberculosis, weakly supervised segmentation, multi-task learning, classification

ABSTRAK

Tuberkulosis merupakan penyakit yang tersebar di seluruh dunia dan menjadi ancaman kepada kesihatan manusia, justeru diagnosis awal amat penting bagi memastikan rawatan yang berkesan. Kaviti paru-paru merupakan salah satu indikator penting dalam diagnosis tuberkulosis, dan pengesanannya dapat memberikan maklumat diagnostik yang bernilai berkaitan lesi tuberkulosis. Walau bagaimanapun, kaedah pembelajaran terselia tradisional bagi pengesanan kaviti paru-paru lazimnya memerlukan sejumlah besar data berlabel,

sedangkan pengumpulan data tersebut bagi imej tuberkulosis merupakan proses yang memakan masa dan memerlukan usaha yang tinggi.
Bagi mengatasi cabaran ini, kajian ini mencadangkan satu pendekatan berpenyeliaan lemah untuk segmentasi semantik kaviti paru-paru. Dalam kaedah ini, model EfficientNet terlebih dahulu digunakan untuk latihan bersama berdasarkan label pengelasan pelbagai kelas pada peringkat imej bagi menjana kawasan minat yang berkaitan dengan kaviti paru-paru. Kawasan yang dijana ini kemudiannya diperhalusi untuk mengenal pasti lokasi sebenar kaviti paru-paru. Keputusan eksperimen menunjukkan bahawa kaedah yang dicadangkan berupaya mengekstrak maklumat semantik utama dalam imej CT di bawah keadaan penyeliaan lemah serta berkesan dalam mensegmentasikan lesi kaviti paru-paru.

Kata kunci**:** Tuberkulosis, segmentasi berpenyelia lemah, pembelajaran berbilang tugas, klasifikasi.

INTRODUCTION

In the past 130 years, tuberculosis (TB) has remained a significant challenge to public health. In the diagnosis of TB, a critical radiological parameter is the lung cavity. Lung cavities typically result from the infection of lung tissue by My-cobacterium tuberculosis, leading to inflammation and necrosis, forming low-density regions in lung tissue on CT im- ages, containing necrotic lung tissue, pathogens, and other substances. However, in CT images, the density and grayscale values of these substances often resemble those of surrounding normal tissue, making accurate delineation of lung cavities a challenging task (Dartois & Rubin 2022).

In the broader context of machine learning applications in healthcare, recent literature demonstrates the growing effectiveness of automated classification systems in non-invasive disease diagnosis, a principle that strongly parallels the goals of automated medical image analysis. One of example is on detection of Dysphonia Disease by employing a Naive Bayes (NB) classifier alongside Mel-Frequency Cepstral Coefficient (MFCC) techniques for acoustic feature extraction (Al-Dhief et al. 2022). Evaluated on the Saarbrucken Voice Database, their method achieved an impressive detection accuracy of 81.48% notably outperforming complex deep learning models like Convolutional Neural Networks (CNNs), which only achieved 70% accuracy on the same dataset. Just as the objective assessment of voice pathologies seeks to reduce the cost, time, and reliance on invasive traditional medical examinations or subjective expert experience, the proposed weakly supervised semantic segmentation framework for tuberculosis detection similarly aims to optimize diagnostic efficiency by drastically reducing the time-consuming and laborious task of acquiring pixel-level manual annotations

Currently, researchers have proposed various methods to address the challenges associated with annotating medical im- ages. These methods include self-supervised learning (Chen et al. 2022, Wei et al. 2022), weakly supervised learning (Ru et al. 2022, Jiang et al. 2022, Chen et al. 2022), etc. In the realm of weakly supervised semantic segmentation (WSSS), Class Activation Maps (CAM) (Selvaraju et al. 2017) represent one of the most employed solutions. CAM involves training a classification network to leverage deep semantic information for segmentation tasks. For instance, a CAM-based approach, coupled with a series of post-processing steps, was applied for weakly supervised organ semantic segmentation (Chan et al., 2024).

However, CAM-based methods encounter a significant challenge due to the absence of pixel-level guidance during the classification network training process (Tan et al. 2022a, Tan et al

2022b). This often leads to a broad focus on features, which, in turn, hampers the effectiveness of CAM and subsequently impacts segmentation results.

In this paper, we propose a simple yet effective weakly supervised multi-task learning method for semantic segmentation of lung cavity lesions in CT images. Radiologists need only determine the presence and quantity of lung cavities in a CT scan without the need to meticulously delineate the boundaries of lung cavities on each slice, significantly saving annotation time. The basic idea of this method is first to generate pixel-level weakly supervised region information using CT image-level multi-class labels and then further process these generated region cues to identify the presence of lung cavities within them, as shown in Figure 1.

Our research contributions are two-fold: (1) Introducing an innovative weakly supervised semantic segmentation framework for lung cavity lesion recognition. (2) Conduct an extensive comprehensive evaluation of our proposed method and compare it with multiple mainstream deep learning models. Experimental results demonstrate that this weakly supervised semantic segmentation method performs admirably in lung cavity segmentation, with its performance even surpassing that of some fully supervised learning methods.

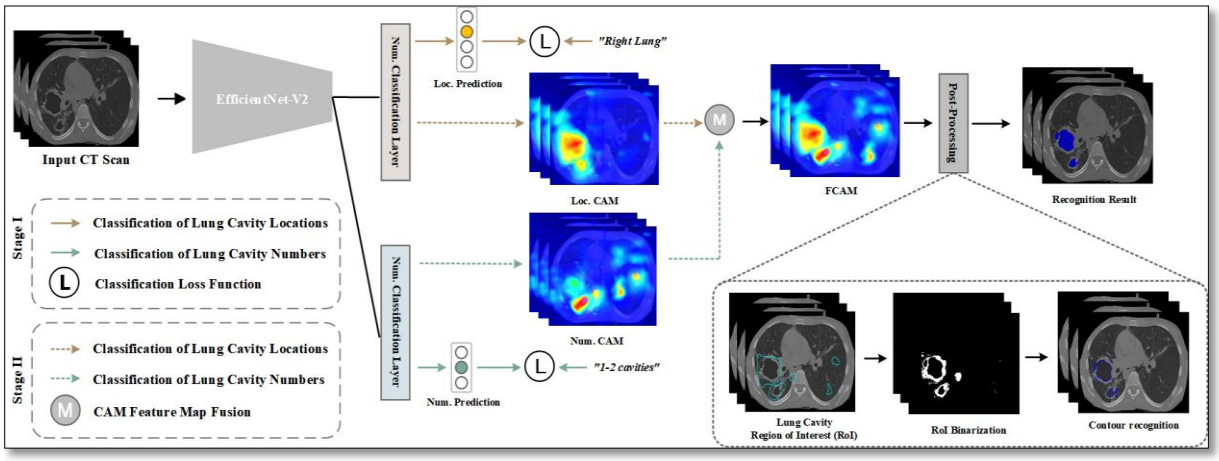

FIGURE 1. Overview of the proposed weakly supervised learning framework

METHOD

STAGE I: NETWORK TRAINING

In Stage I, this model training involves collaborative learning of two classification tasks: lung cavity location classification and quantity classification. This approach of collaborative learning offers distinct advantages as it allows the model to gain a comprehensive understanding of lung cavity features by simultaneously learning both spatial locations and quantities.

For the classification loss, similar to previous work (Ru et al., 2022), the top-layer hidden features are passed through two separate classification layers. Then, the multi-label soft margin loss is adapted as the classification function to calculate the loss between the class probability vector p_c and the true image-level label y.

$$\mathcal{L}_c = \frac{1}{N}\sum_{n=1}^{N}(y^n \log(p_c^n) + (1 - y^n)\log(1 - p_c^n)) \quad (1)$$

where *n* is the total number of classes and *c* represents the type of task. When *c* takes on the values num and loc, it corresponds to the classification tasks for the number and position of lung cavities, respectively.

Finally, the optimization objective function for model training is as follows:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{num} + \lambda_2 \mathcal{L}_{loc} \quad (2)$$

where λ_1 and λ_2 are weighting factors that balance the contributions of the quantity and position classification tasks during training.

In addition, during the collaborative training process of multi-objective lung cavity classification tasks, the gradient of the objective function is often dominated by the gradient of a specific classification task. However, this dominance tends to compromise the performance of other classification tasks. Furthermore, improvements in the performance of dominant tasks may be overestimated due to high curvature issues, while performance degradation in non-dominant tasks may be underestimated. These challenges make it difficult for the optimizer to optimize the overall objective effectively. A common approach to address these issues is the use of gradient clipping techniques, such as PCGrad (Yu et al. 2020).

However, PCGrad (Yu et al. 2020) employs a linear projection method, which often falls short in capturing the intricate relationships between different tasks and is susceptible to the influence of locally optimal solutions. To empower the model to share information more effectively among complex tasks and enhance the performance of collaborative learning, PCGrad variant is refined based on nonlinear projection, as outlined in Figure 2.

Algorithm Improved PCGrad Update Rule with Nonlinear Projection

Require: Convolutional neural network parameters θ, image classification tasks $\mathcal{B} = \{\text{Task}_k\}$

Initialize the model with parameters θ
for $\text{Task}_i \in \mathcal{B}$ **do**
  Load and preprocess images and labels for Task_i
  Compute the loss $\mathcal{L}_i(\theta)$ for Task_i
  Compute the gradient of the loss with respect to θ: $\mathbf{g}_i = \nabla_\theta \mathcal{L}_i(\theta)$
  Set $\mathbf{g}_i^{\text{PC}} = \mathbf{g}_i$ *// Initialize $\mathbf{g}_i^{PC}$ with the original gradient*
  for $\text{Task}_j \overset{\text{uniformly}}{\sim} \mathcal{B} \setminus \text{Task}_i$ in random order **do**
    if $\mathbf{g}_i^{\text{PC}} \cdot \mathbf{g}_j < 0$ **then**
      // Apply nonlinear projection using EfficientNet-v2
      Initialize nonlinear projection function parameters θ_{proj}
      Extract intermediate feature representations x from the top layer of EfficientNet-v2
      Pass x through EfficientNet-v2 with parameters θ_{proj} to obtain the projected feature representation y
      Compute the scaling factor α as a function of y and $\mathbf{g}_j$: $\alpha = \text{sigmoid}(W \cdot y + b)$
      Compute the projected gradient $\mathbf{g}_i^{\text{PC}}$ as a weighted combination of y and $\mathbf{g}_j$:
      $\mathbf{g}_i^{\text{PC}} = \alpha \cdot y + (1 - \alpha) \cdot \frac{\mathbf{g}_i^{\text{PC}} \cdot \mathbf{g}_j}{\|\mathbf{g}_j\|^2} \mathbf{g}_j$
return updated parameters θ after PCGrad update

FIGURE 2. Refined PCGrad gradient conflict optimization algorithm.

STAGE II: WEAKLY SUPERVISED LUNG CAVITY RECOGNITION

In Stage II, class activation maps (CAM) are generated based on the pre-trained weights obtained in Stage I. Subsequently, post-processing is performed on the generated CAM to identify lung cavities.

GENERATION OF FUSION CLASS ACTIVATION MAPS

Accurate recognition of lung cavities relies heavily on the precise extraction of regions of interest. To precisely pin- point this specific area, a fusion technique known as Fusion Class Activation Maps (FCAM) is introduced. To elaborate, FCAM is generated through the amalgamation of CAMs from two separate branches: one tasked with quantifying lung cavity quantity and the other dedicated to classifying their positions. By integrating CAMs from these two distinct tasks, a more comprehensive characterization of lung cavity features can be achieved. The operational principle of our approach is outlined as follows:

Firstly, to capture a given class c in the activation map A_k of the classification layer, the weights w_k^c is calculated. These weights signify the influence of the given category c on each spatial position in the activation map A_k. The weights are obtained through a process known as Global Average Pooling. The formula for calculating w_k^c is given by:

$$w_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial s_c}{\partial A_k(i,j)} \tag{3}$$

where Z is the size of the activation map A_k, and $\partial A_k(i,j)$ represents the gradient of the given category c with respect to the activation map A_k for the model's output.

Secondly, the weight w_k^c corresponding to each layer is multiplied by the activation value A_k of that layer. By summing up these weighted activation values across all layers, the resultant value M_c represents the contribution of the given category c to the heat map. Moreover, to accentuate the regions of interest and suppress irrelevant areas in the generated heat map Mc, a Rectified Linear Unit (ReLU) operation is applied. The overall calculation is expressed as:

$$M^c = ReLU(\sum_k w_k^c \cdot A_k) \tag{4}$$

Finally, given the CAM $M \in R^{h \times w \times C}$, the fusion CAM ($M^F(i,j)$) is defined as follows:

$$M^F(i,j,:) = max\{M^{num}(i,j,:), M^{loc}(i,j,:)\}$$

where, $M^F(i,j,:)$ denotes the amalgamated CAM that fuses both quantity-related ($M^{num}(i,j,:)$) and location-based ($M^{loc}(i,j,:)$) CAM.

FCAM POST-PROCESSING

The post-processing of FCAM can be divided into the following three steps:

In the first step, dependable foreground and background information are extracted from the target CAM $M^F(i,j)$, utilizing a background score β_t (where $0 < \beta_t < 1$) (Ru et al. 2022). Specifically, the following formula to post-process $M^{total}(i,j)$ is applied:

$$Y(i,j) = \begin{cases} argmax\left(M^{i,j,:}\right), & if\ max\left(M^{i,j,:} \geq \beta_t\right) \\ 0, & if\ max\left(M^{i,j,:} < \beta_t\right) \end{cases} \quad (5)$$

where, $M^{i,j,:}$ represents the channel values in the CAM at position (i,j), and $max\left(M^{i,j,:}\right)$ represents the maximum channel value at that position. If the maximum channel value is greater than or equal to the threshold β_t, $Y(i,j)$ is set to the index of the maximum value in the corresponding channel (i.e., the predicted class label). If the maximum channel value is less than or equal to the threshold β_l, $Y(i,j)$ is set to 0, representing the background. This step helps filter out unreliable weakly supervised information while retaining information strongly correlated with both background and foreground.

In the second step, a binary thresholding operation on $Y(i,j)$ is performed to facilitate the subsequent lung cavity contour recognition. The result after binary thresholding, denoted as $I(i,j)$, is as follows:

$$I(i,j) = \begin{cases} 255, & if\ Y(i,j) > T \\ 0, & \text{otherwise} \end{cases},$$

where T represents the threshold value. After binary thresholding, $I(i,j)$ will have pixel values of either 0 or 255, making it suitable for further lung cavity contour recognition.

In the third step, contour detection algorithms provided by OpenCV (Bradski 2022) is employed to recognize the contours of lung cavities from the input image $I(i,j)$. To be specific, systematically traverse through all the closed contours c_i (where $i = 1, \ldots, n$) in $I(i,j)$ and assess whether they qualify as parent contours. Typically, parent contours correspond to the external boundaries of the lung cavity wall. If a contour c_i is identified as a parent contour, it is subsequently removed from the contour array. Following this, within the remaining contour array, the hollow contours are identified, designating it as the region representing the lung cavity area.

EVALUATION METRIC

There are several evaluation metrics used in this research such as Accuracy, F-measure (Albadr, 2023), Intersection-Over-Union (IoU) and Dice Similarity Coefficient (DSC) (Latiff, 2025).

Accuracy: The ability of the algorithm to correctly differentiate three categories “No cavities,” “1 - 2 cavities,” and “More than 2 cavities” Mathematically, this can be stated as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP}$$

Which
TP: true positive, TN: true negative, FP: false positive, FN: false negative

$$\text{F-Measure (F1-score)} = \frac{(2\ x\ precision\ x\ recall)}{(precision+recall)}$$

Intersection over Union (IoU) is a fundamental metric used to quantify the accuracy of an object detector by measuring the overlap between two boundaries.

$$\text{IoU} = \frac{Area\ of\ Intersection}{Area\ of\ Union}$$

It determines if a prediction is a True Positive (high overlap) or False Positive (no overlap). It is considered a correct detection, though higher percentage require better precision.

The Dice Similarity Coefficient (DSC) is a spatial overlap index used to measure the similarity between two sets of data ranging from 0 (no overlap) to 1 (perfect overlap). It is widely used in medical image segmentation to validate algorithms by calculating the ratio of twice the intersection as follows:

$$\text{DSC} = 2\ X\ \frac{TP}{2\ X\ TP+FP+FN}$$

DATA

The experimental data for this study is sourced from the ImageCLEF2022 Tuberculosis dataset (Kozlovski et al. 2022), comprising a total of 5463 CT slice images containing lung cavity lesion tissues. All images have a resolution of 512 × 512. Due to the absence of number, location, and semantic segmentation annotations for the lung cavity lesion tissues in the ImageCLEF2022 dataset, we enlisted the expertise of medical professionals from the Hospital Pengajar Universiti Putra Malaysia (HPUPM) to manually annotate these images, as shown in Figure 3. For the quantity classification task, it encompasses three categories: “No cavities,” “1 - 2 cavities,” and “More than 2 cavities.” As for the location classification task, it consists of four categories: “None,” “Left lung,” “Right lung,” and “Both lungs.” Furthermore, this dataset is divided into training and testing sets in an 8:2 ratio.

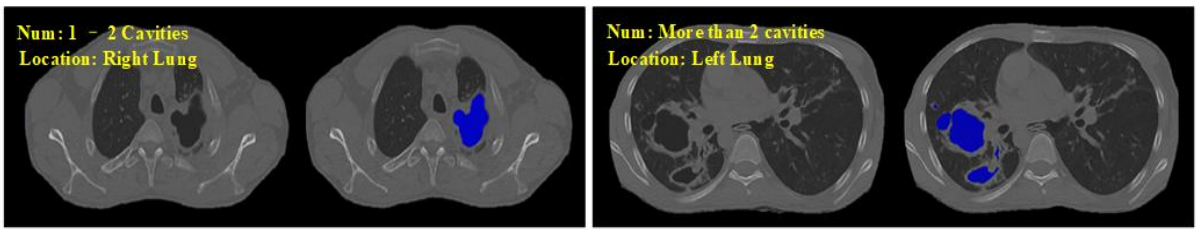

FIGURE 3. Examples of experiment data. Blue area represents lung cavity lesion.

EXPERIMENTS AND RESULTS

All experiments were conducted in the PyTorch development environment utilizing an NVIDIA A100 GPU.

Stage I. In this stage, joint training for lung cavity position and quantity classification is conducted using the EfficientNet-v2 network. In the objective functions of the joint training, the values of balancing factors λ1 and λ2 are set to 1. Adam optimizer is used with a learning rate set to 0.0005. The training comprised 125 epochs, each with a mini-batch size of 16. During data preprocessing, techniques such as mixup (Zhang et al. 2017), color transformations,

and random rotations to augment the training data (Tan et al. 2020) are applied. Finally, the best-performing weights from the model trained over 125 epochs is selected for use in the generation of class activation maps in stage II.

Stage II. In this stage, post-processing on the generated CAMs for lung cavity generation is applied. The default hyperparameter settings are as follows: for extracting regions of interest corresponding to the lung cavity in the class activation maps, the background score βl is set to 0.55. For generating binary images, the segmentation threshold T that separates the foreground from the background is set to 125.

Table 1 presents the classification results for both the quantity and location of lung cavity lesions. To assess the effectiveness of this method, two single-task learning models are established as controls: Swin Transformer (Liu et al. 2021) and EfficientNet-v2 (Tan & Le 2021). The evaluation metrics used are ACC(Accuracy) and F-measure (F1-score). From the table, the method proposed in this paper, which involves joint training of lung cavity quantity and location prediction classification tasks (EfficientNet-v2†), achieved the best ACC and F1-score performance in both quantity and location classification tasks. Specifically, EfficientNet-v2† achieved an ACC of 0.733 and an F1 of 0.816 in the quantity classification task, representing improvements of 1.9% and 1.7%, respectively, compared to EfficientNet-v2. Furthermore, in the location classification task, EfficientNet-v2† attained an ACC of 0.746 and an F1 of 0.837, corresponding to improvements of 2.5% and 2.4% compared to EfficientNet-v2. The experimental results demonstrate that the joint training approach presented in this paper effectively enhances the accuracy and predictive performance of lung cavity lesion classification.

TABLE 1. The classification results of the number and location of lung cavity lesions. The dagger symbol (†) represents the joint training of the lung cavity's quantity and position prediction classification tasks.

Methods	Metrics	Tasks	
		Number	Location
Swin Transformer (Liu et al. 2021)	ACC	0.721	0.729
	F1	0.807	0.826
EfficientNet-v2 (Tan et al. 2021)	ACC	0.714	0.721
	F1	0.799	0.813
EfficientNet-v2† (Ours)	ACC	0.733	0.746
	F1	0.816	0.837

Table 2 presents the performance of deep learning architecture for medical image segmentation in the task of lung cavity lesion segmentation. Sup. denotes supervision type. F: full supervision. W: weakly supervision. The evaluation metrics used are Intersection-Over-Union (IoU), Dice Similarity Coefficient (DSC) as explained in section 2.2.3. The methods encompass in this research article are Unet (Ronneberger et al. 2022), Mobilenet-v2 (Sandler et al. 2018), DeepLabv3+ (Chen et al. 2017), Swin Transformer, EfficientNet-v2, and our proposed weakly supervised approach.

From the table, it shows that EfficientNet-v2 exhibits superior performance under the fully supervised (F) setting, achieving an IoU of 38.9% and a DSC of 53.9 %. However, the proposed weakly supervised method also achieves good performance without pixel-wise full supervision (W), with IoU and DSC of 31.2 % and 44.7%, respectively. In addition, this weak supervision method is in performance and even beyond some fully supervised learning methods. This result shows that the weakly supervised method constructed in this article can still effectively

segment the lung cavity diseased tissue although pixel-level (segmentation) supervised annotation is limited, effectively reducing the cost of manual annotation.

TABLE 2. Performance comparison for deep learning architecture for medical image segmentation.

Methods	*Sup*	IoU (%)	DSC (%)
UNet (Ronneberger et al. 2022)		29.6	41.7
Mobilenet-v2 (Sandler et al. 2018)		30.5	43.5
DeepLabv3+ (Chen et al. 2017)	*F*	30.9	44.2
Swin Transformer (Liu et al. 2021)		37.2	52.1
EfficientNet-v2 (Tan et al. 2021)		38.9	53.9
Ours Weakly Supervised Method	W	31.2	44.7

CONCLUSION

In this paper, an innovative weakly supervised approach is proposed for segmenting lung cavity lesion tissue. Our experiments demonstrate that, when compared to single-task learning, this joint training strategy significantly enhances the network's performance in classifying the location and quantity of lung cavities. Furthermore, the category activation maps produced by the classification network trained in this collaborative manner effectively recognize regions of interest within the lung cavity. With the information from these areas, it can be more accurately identify the lung cavity structure in CT slice images. This weakly supervised method is expected to become an effective way to obtain efficient pixel-level lung cavity annotation in the future.

Conflicts of Interest. The authors have no relevant financial or non-financial interests to disclose.

Acknowledgments. This work was supported by the Malaysia Ministry of Higher Education, Fundamental Research Grant Scheme. [grant number FRGS/1/2022/ICT01/UPM/02/1]

REFERENCES

Albadr, M. A. A., Tiun, S., Ayob, M., Nazri, M. Z. A., & AL-Dhief, F. T. (2023). Grey wolf optimization-extreme learning machine for automatic spoken language identification. *Multimedia Tools and Applications*, *82*(18), 27165-27191.

Al-Dhief, F. T., Latiff, N. M. A., Malik, N. N. N. A., Baki, M. M., Sabri, N., & Albadr, M. A. A. (2022). Dysphonia detection based on voice signals using naive bayes classifier. In *2022 IEEE 6th international symposium on telecommunication technologies (ISTT)* (pp. 56-61). IEEE.

Bradski, G. (2000). The openCV library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, *25*(11), 120–123. https://www.researchgate.net/publication/322276713_The_openCV_library.

Chan, L., Hosseini, M. S., Rowsell, C., Plataniotis, K. N., & Damaskinos, S. (2019). Histosegnet: Semantic segmentation of histological tissue type in whole slide images. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10662–10671. https://doi.org/10.1109/ICCV.2019.01076.

Chen, L.-C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*. https://doi.org/10.48550/arXiv.1706.05587.

Chen, Q., Yang, L., Lai, J. H., & Xie, X. (2022). Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4288–4298. https://doi.org/10.48550/arXiv.2203.02909.

Dartois, V. A., & Rubin, E. J. (2022). Anti-tuberculosis treatment strategies and drug development: Challenges and priorities. *Nature Reviews Microbiology, 20*(11), 685–701. https://doi.org/10.1038/s41579-022-00731-y.

Jiang, P. T., Yang, Y., Hou, Q., & Wei, Y. (2022). L2g: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16886–16896. https://doi.org/10.48550/arXiv.2204.03206.

Kozlovski, S., Dicid, Y. D., Kovalev, V., & Müller, H. (2022). Overview of imagecleftuberculosis 2022: Ct-based cavern detection and report. *Proceedings of CLEF 2022: Conference and Labs of the Evaluation Forum*, 3180(96). https://api.semanticscholar.org/CorpusID:251471986.

Latiff, N. M. A. A., Al-Dhief, F. T., Sazihan, N. F. S. M., Baki, M. M., Malik, N. N. N. A., Albadr, M. A. A., & Abbas, A. H. (2025). Voice pathology detection using machine learning algorithms based on different voice databases. *Results in Engineering*, *25*, 103937.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022. https://doi.org/10.48550/arXiv.2103.14030.

Ronneberger, O., Fischer, P., & Brox, T. (2022). Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*. https://doi.org/10.1007/978-3-319-24574-4_28.

Ru, L., Zhan, Y., Yu, B., & Du, B. (2022). Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16846–16855. https://doi.org/10.48550/arXiv.2203.02664.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4510–4520. https://doi.org/10.48550/arXiv.1801.04381.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision*, 618–626. https://doi.org/10.48550/arXiv.1610.02391.

Tan, M., & Le, Q. (2021). Efficientnetv2: Smaller models and faster training. *International Conference on Machine Learning*, 10096–10106. https://proceedings.mlr.press/v139/tan21a.html.

Tan, Z., Hu, Y., Luo, D., Hu, M., & Liu, K. (2020). The clothing image classification algorithm based on the improved Xception model. *International Journal of Computational Science and Engineering, 23*(3), 214–223. https://doi.org/10.1504/IJCSE.2020.111426

Tan, Z., Madzin, H., & Ding, Z. (2022a). Image quality assessment based on multi-model ensemble class-imbalance repair algorithm for diabetic retinopathy uw-octa images.

MICCAI Challenge on Mitosis Domain Generalization, 118–126. https://doi.org/10.1007/978-3-031-33658-4_11.

Tan, Z., Madzin, H., & Ding, Z. (2022b). Semi-supervised semantic segmentation methods for uw-octa diabetic retinopathy grade assessment. *MICCAI Challenge on Mitosis Domain Generalization*, 97–117. https://doi.org/10.1007/978-3-031-33658-4_10.

Wei, C., Fan, H., Xie, S., Wu, C. Y., Yuille, A., & Feichtenhofer, C. (2022). Masked feature prediction for self-supervised visual pre-training. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14668–14678. https://doi.org/10.48550/arXiv.2112.09133.

Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., & Finn, C. (2020). Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems, 33*, 5824–5836. https://dl.acm.org/doi/10.5555/3495724.3496213.

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). Mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*. https://doi.org/10.48550/arXiv.1710.09412.