Interactive Learning for Assisting Whole Slide Image Annotation

Ashish Menon[(✉)], Piyush Singh, P. K. Vinod, and C. V. Jawahar

International Institute of Information Technology, Hyderabad, India
{ashish.menon,piyush.singh}@research.iiit.ac.in,
{vinod.pk,jawahar}@iiit.ac.in

Abstract. Owing to the large dimensions of the histopathology whole slide images (WSI), visually searching for clinically significant regions (patches) is a tedious task for a medical expert. Sequential analysis of several such images further increases the workload resulting in poor diagnosis. A major impediment towards automating this task using deep learning models is that it requires a huge chunk of laboriously annotated data in the form of WSI patches. Our work suggests a novel CNN-based, expert feedback-driven interactive learning technique to mitigate this issue. The proposed method seeks to acquire labels of the most informative patches in small increments with multiple feedback rounds to maximize the throughput. It requires the expert to query a patch of interest from one slide and provide feedback to a set of unlabelled patches chosen using the proposed sampling strategy from a ranked list. The experiments on a large patient cohort of colorectal cancer histological patches (100K images with nine classes of tissues) show a significant reduction ($\approx 95\%$) in the amount of labelled data required to achieve state-of-the-art results when compared to other existing interactive learning methods (35%–50%). We also demonstrate the utility of the proposed technique to assist a WSI tumor segmentation annotation task using the ICIAR breast cancer challenge dataset (≈ 12.5K patches per slide). The proposed technique reduces the scanning and searching area to about 2% of the total area of WSI (by seeking labels of ≈ 250 informative patches only) and achieves segmentation outputs with 85% IOU. Thus our work helps avoid the routine procedure of exhaustive scanning and searching during annotation and diagnosis in general.

Keywords: Deep learning · Interactive learning · Whole slide image annotation · Histopathology · Sampling strategies · Expert-in-the-Loop

1 Introduction

Histopathology is considered the gold standard for cancer diagnosis [6,21]. A histopathologic Whole Slide Image (WSI) represents a digitized image of a tissue sample characterized by a large size of up to 10^9 pixels at maximum resolution. A significant bottleneck in WSI diagnosis is locating certain classes of tissues

[1] or regions of prognostic importance within the WSI. Hence it becomes essential to automate the detection of such regions using deep learning models. The requirement of large amounts of annotated data for deep learning models and their robustness to adapt to different datasets is an issue of concern. Thus, a model that interactively learns with minimal expert involvement without having labelled data upfront would be an appropriate solution.

1.1   Why Patch Level Analysis of WSI?

Interactive learning methods proposed in the past had experts providing feedback as pen strokes on specialized devices or drawing regions of interest by carefully delineating tissue regions [14], marking nuclei and cellular level details [16] on the WSI. These methods needed continuous involvement of the pathologist in correcting the model's prediction to obtain nuclei and cellular level features used as input for an ML model. With the advancements in deep learning and its capability to obtain powerful representations, recent focus has shifted to analysing WSI tiles(patches). Recent approaches have shown the effectiveness of patch level analysis of a WSI to detect regions of prognostic values such as survival prediction [8,24], mutation prediction [5,13], tumor grading and staging [12,22]. Several medical imaging challenges have provided patch-level annotated datasets highlighting regions of biological relevance [2,3,18] to facilitate patch analysis using deep learning methods.

1.2   Related Work

Interactive learning methods for patch labelling include active learning-based techniques to acquire the most informative samples for training, thereby reducing the need for large amounts of annotated data. One such technique selects the most informative samples from a relatively smaller cohort of patients with 5000 patches [10] using a variational drop-out-based uncertainty sampling [20]. At each step 160 most informative images were selected to be reviewed by the expert. This method achieved a scale reduction of 45% in the training set. A novel method of identifying the most informative patches was proposed [23] using conditional random fields in a spatially adaptive manner. This method showed a scale reduction in the training set by 38.0%. These approaches often required large batches of patches to be reviewed, which can be an overhead to the expert and the overall workflow. A deep active learning work for biomedical image segmentation [25] proposed a framework combining fully convolutional networks and active learning to determine the most representative and uncertain areas for annotation. They obtained state-of-the-art segmentation performance using only 50% of training data for gland segmentation in Colon histology images, with gland segmented image patches of a WSI as ground truth. An attention gated FCN (ag-FCN) and distribution-discrepancy based active learning algorithm [11] was demonstrated for gland segmentation. This work achieved state-of-the-art results using 50% training data and beat the state-of-the-art using full training data. DeepScribble [4] proposed an interactive segmentation method that

corrects the segmented boundaries from deep neural networks with user interactions. This method used two networks, one trained with annotated WSI patches generated an initial segmentation output. While the other refines the segmentation iteratively based on user inputs.

1.3   Our Contribution

Most of the previously mentioned interactive learning have used uncertainty based criteria to select the most informative samples. These methods have reported using patches or have used carefully delineated tumorous subregions of a WSI (gland segmentation). We select the most informative samples using distance metric learning combined with a classifier approach in our proposed technique. We also demonstrate the utility of our method in helping annotation (tumor segmentation) of an unannotated WSI. The key contributions of the proposed technique include

(i) Put forward a novel method of sampling from a ranked list of patches to pick the most informative samples to be labelled, resulting in a significant scale reduction in the training set (95%–97%) to achieve state-of-the-art results.

(ii) Demonstrate the utility of an ImageNet pre-trained model (last few layers trainable) without any architectural modification for this task, thereby resulting in quick and memory-efficient training

(iii) Assisting annotation of large WSI for a segmentation task by obtaining segmentation mask predictions with minimal expert efforts

(iv) Our method can also give segmentation predictions on multiple WSI using a single slide understudy without going through all the slides.

2   Proposed Interactive Learning Technique

Our method is developed by assuming that there is a database of patches of undiagnosed or unseen slides. An expert is ready to search for clinically important patches of one such slide and provide feedback to a set of patches chosen by a deep learning model. The proposed strategy of sampling from a ranked list is used to select these patches. Based on the expert requirement, the patches chosen for feedback could be the patches from the same slide or patches from multiple slides. The feedback input can either be relevant/irrelevant or explicit class labels of the patches and thus avoids complex expert interactions at the cellular and nuclei level details as seen in the previous interactive learning strategies [16].

Inspired by the idea of [19], our work demonstrates the utility of pre-trained CNNs for this task. When finetuned with the proposed sampling strategy, the last few layers of an ImageNet pre-trained model representation generalize well on unseen data within a minimal number of expert feedback rounds (less labelled data). The major steps of the proposed work is explained in the following subsections, before which we will introduce the following frequently used terms.
Retrieval: nearest neighbour retrieval using the deep learning representations with images ranked in the increasing order of their distance from the query image

in the higher dimensional space. **Session**: for every query, there is a session that consists of r rounds of review. **Review**: the step where we obtain relevance feedback $(0/1)$ or explicit class labels by presenting to the expert a set of K samples. **Sampling from ranked list**: to obtain the most informative samples to be reviewed at each review step. **Finetuning**: the trainable layers of the ImageNet pre-trained ResNet-18 model (layer4.1) is finetuned with the samples reviewed. Two models are trained, one for retrieval and the other for classification using a triplet loss with hard negative mining and cross-entropy loss respectively (Fig. 1).
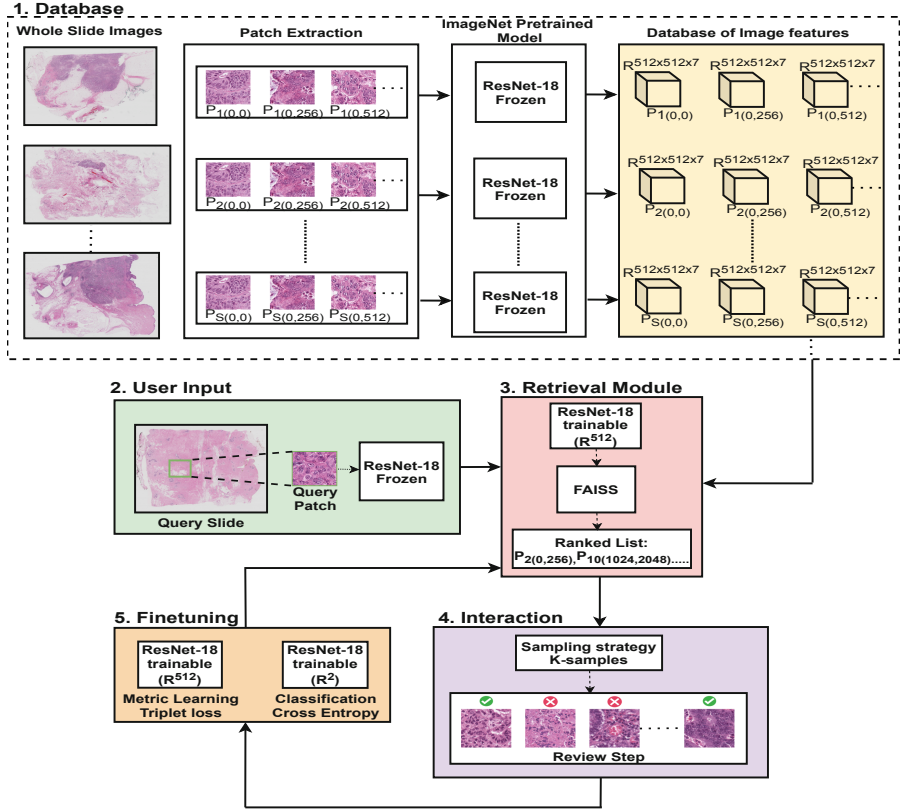

Fig. 1. Block diagram of the proposed technique. **1. Database:** The unannotated WSI are stored in the form of features ($R^{(512 \times 512 \times 7)}$) extracted from an ImageNet pretrained ResNet-18. **2. User Input:** Highlights a rectangular patch of interest on the WSI as the input. **3. Retrieval Module:** Nearest neighbour retrieval powered by FAISS [7]. **4. Interaction:** The expert provides feedback either as relevant/irrelevant or as explicit class label. Sampling from the ranked list ensures that the feedback is provided for the most informative samples. **5. Finetuning:** ResNet18 used as a feature extractor and as classifier are trained using the samples reviewed.

2.1 Assisting WSI Annotation

In this paper, we perform the task of annotating tumorous regions of multiple WSIs as a use case of the proposed interactive learning technique. The underlying assumption is that the group of patches deemed relevant to the query patch or provided with a specific class label by the expert are closely clustered in a higher-dimensional space. We pose the annotation task to be equivalent to a WSI patch classification using a deep learning model. The goal is to achieve a satisfactory annotation with as few feedback sessions as possible. Towards this, we propose a novel sampling strategy by sampling from a ranked list of patches from the search database, ranked based on the Euclidean distance to the query in a higher dimensional space learnt by another deep learning model using a distance metric learning approach. Both models used for classification and distance metric learning are trained using the same set of reviewed samples across feedback sessions.

The annotation starts by querying a patch of interest from the WSI. Nearest neighbour retrieval is performed using an ImageNet pre-trained ResNet-18 representation (R^{512}) to obtain a ranked list of patches from the search database arranged in the increasing order of their distance to the query patch. This is followed by sampling K patches from the ranked list, review and finetuning. The same steps are repeated for the subsequent rounds using the finetuned model representation and a refined query for retrieval. Query refinement is performed by assuming that the original query lies far from the cluster of relevant images in the feature space [19]. It is important to note that the retrieval is performed by leaving out the already reviewed samples from the database at each step. A new query would mark the beginning of the next session, and the same procedure is carried forward. Finetuning is done at every round/session using the data accumulated from the start till that particular round. Each session is restricted to 'r' rounds of review; the model presents K samples to be reviewed by the expert for every round. Thus the expert ends up having to review $r \times K$ images per query. At the end of N sessions, we would have $N \times r \times K$ reviewed images with true labels.

2.2 Sampling Strategy

It is important to provide the most informative samples (samples that are not obvious for the model) for review during the interactive feedback. A strategy of sampling from a ranked list is proposed to pick such samples at every review step. We experiment on the following sampling strategies: **random**, **top-K** and **front-mid-end** sampling to pick K samples from the front middle and the end of the ranked list. The **Closest Negative Farthest Positive (CNFP)** sampling uses predictions of a classifier trained on the samples reviewed till $(k-1)^{th}$ step, to pick the $K/2$ closest negatives and $K/2$ farthest positives from the ranked list. Finally, we also investigate a **hybrid** scheme, which uses the combination of CNFP and front-mid-end. **Entropy-based** sampling, which is a standard uncertainty measure for sampling by using the entropy of classifier prediction and choosing the samples with the highest entropy for review.

Algorithm 1: Pseudocode for retrieval and annotation using the proposed interactive learning technique. **S**:Images of Search database, **Q**:Images of Query database, **M**:Number of images in the search database, **N**:Number of images in the query database, $\mathbf{f_{ret}}$:Pre-trained deep learning model used for retrieval, $\mathbf{f_{cl}}$:Pre-trained deep learning model used for classification **r**:Number of feedback rounds/steps, **K**:Number of images to be reviewed per step, $\mathbf{\hat{f}}$:model that has learnt the new feature space

---

Input: Query image $\mathbf{Q_{sample}}$, deep learning models $\mathbf{f_{ret}}, \mathbf{f_{cl}}$, number of feedback rounds **r**, number of images to be reviewed per step **K**

Output: $Y = \{y_i\}_{i=1}^{M}$; `// labels for image in the database`

for Q_{sample} **in** Q **do**

    $q = f_{ret}(Q_{sample}), s = f_{ret}(S)$; `// embedding computation ` (R^{512})

    $\hat{S} = retrieval(q, s)$; `// ranked list of images from search database`

    for $i \leftarrow 1$ **to** r **do**

        $\hat{S}_K = SAMPLE(\hat{S}, K)$

        S_{0i} , $S_{1i} = feedback(\hat{S}_K)$

        $S_0.append(S_{0i})$, $S_1.append(S_{1i})$

        $\hat{f}_{ret} \leftarrow train(f_{ret}, S_0, S_1, Q)$

        $\hat{f}_{cl} \leftarrow train(f_{cl}, S_0, S_1, Q)$

        $S \leftarrow \{S\} - \{\hat{S}_K\}; q \leftarrow \hat{f}_{ret}(Q); q \leftarrow mean(q, \hat{f}_{ret}(S_{1i}))$

        $\hat{S} = retrieval(q, \hat{f}_{ret}(S))$

    end

end

if *annotation* **then**

    $\hat{Y} = \hat{f}_{cl}(S)$; `// Prediction step using classifier`

Algorithm 2: Sampling strategy. S represents unannotated set of Images, s represents the sampling strategy, K represents number of images to sample, M indicates number of images in the list

---

Input: Set of Images S , Sampling strategy s, number of samples K, number of images in the list M

Output: S_K ; `// K number of samples`

if $s == top - k$ **then**

    $S_K \leftarrow \{S_i\}_{i=1}^{K}$

if $s == front - mid - end$ **then**

    $N_{front} = \lceil M/3 \rceil; N_{end} = \lfloor M/3 \rfloor; N_{mid} = M - N_{front} - N_{end}$

    $S_K \leftarrow \{S_i\}_{i=1}^{N_{front}} \cup \{S_i\}_{i=M/2-N_{mid}/2}^{M/2+N_{mid}/2} \cup \{S_i\}_{i=M-N_{end}}^{M}$

if $s == CNFP$ **then**

    $S^{'} \leftarrow S[:: -1]$; `// Reversed ranked list`

    $S_1 \leftarrow \{S_i^{'} \ni f_{cl}(S_i^{'}) == 1\}_{i=1}^{M}, S_0 \leftarrow \{S_i \ni f_{cl}(S_i) == 0\}_{i=1}^{M}$

    $S_k \leftarrow \{S_{1i}\}_{i=1}^{i=K/2} \cup \{S_{0i}\}_{i=1}^{i=K/2}$

3   Implementation

3.1   Dataset and Database Formation

To validate and demonstrate our work, we use two publicly available datasets.

CRC Dataset [9]: Consists of 100,000 non overlapping image patches from H&E stained slides of human colorectal cancer (CRC). All images are 224×224 pixels. It consists of the following tissue classes cancer-associated stroma **STR**, colorectal adenocarcinoma epithelium **TUM**, adipose **ADI**, mucus **MUC**, smooth muscle **MUS**, debris **DEB**, lymphocytes **LYM**, background **BACK**, normal colon mucosa **NORM**. These were extracted from 86 cancer tissue slides. The prognostic importance of some of these tissues was demonstrated in predicting the overall survival prediction of colorectal cancer patients [8]. Due to the availability of ground truth patch labels, a pseudo feedback [19] technique was used to automate the manual feedback. We created a query database of 10 random patches from each class to simulate querying a patch of interest. The remaining patches formed the search database (Table 1).

Table 1. The dataset distribution of the CRC dataset.

Class labels	MUC	MUS	NORM	STR	TUM	ADI	BACK	DEB	LYM
Search DB	8886	13526	8753	10436	14307	10397	10556	11502	11547
Held out test set	1035	592	741	421	1233	1338	847	339	634
Query DB	10	10	10	10	10	10	10	10	10

ICIAR BACH Challenge Dataset [18]: is used to demonstrate the application of the proposed technique towards slide annotation. ICIAR dataset is composed of H&E stained Breast cancer histology microscopy and WSI. It provides annotations of pixel coordinates belonging to 4 different classes **normal, benign, invasive, insitu** for 10 WSIs. We group the insitu, invasive and benign classes as tumor class. Patches of 256×256 were extracted from these WSI using a sliding window approach with no overlap at the maximum magnification resolution. A patch was labelled as tumor if at least 50% of it consisted of the annotated pixels. Noisy and background patches are filtered out [24] during the patch extraction process. Patches with the fractal structure were rejected by considering only those patches with at least ten connected components present in their binarized format. During annotation of a given slide, 10 random tumor patches from it formed the query database, and the remaining patches of the slide formed the search database. The search database can contain patches of one slide or could be expanded to contain patches of multiple slides based on the use case. This step expands the annotation or finds tumorous patches across multiple slides using a single slide under study.

3.2   Classification and Metric Learning

An ImageNet pre-trained ResNet-18 was used as the base model. We first save the image features obtained from the frozen layers (upto layer4.0, $R^{512\times7\times7}$), which are fed as input to the respective trainable modules during training. So the database consists of images features instead of images. Benefits of this setup include quick and memory efficient training, avoid overfitting, and non-linearity introduced during training by layer 4.1. The images were normalized using the mean and standard deviation calculated across all the RGB channels on the entire dataset before obtaining the features.

The trainable module for the metric learning consisted of layer4.1 and the global average pooling (GAP) layer of ResNet-18. Thus the metric learning output would be an R^{512} embedding. The trainable module for the classification step consisted of layer4.1, Global Average pooling layer (GAP), followed by the fully connected layer with N neurons as output. Depending on the type of feedback, N could vary from 2 to the number of classes under analysis. Thus the classifier output would be an R^N embedding.

Every session consists of 5 rounds of review, with 5 images reviewed per round for the ICIAR dataset annotation and 10 images reviewed per round for the CRC dataset. If the type of feedback is relevant/irrelevant, two sets of images are maintained, a relevant set and an irrelevant set. Relevant set consists of those reviewed patches that share the same label as the queried patch, and the irrelevant set consists of patches otherwise. We used 50 training epochs and an Adam optimizer with a learning rate of 0.0001 for the metric learning and the classification steps. The metric learning was performed using triplet loss with hard triplet mining following the implementation suggested in [15] with the margin for triplet loss set to 0.2. The classifier was trained using the cross-entropy loss function. FAISS [7] was used for the nearest neighbour retrieval using Euclidean distance.

4   Results and Discussion

4.1   Evaluation on CRC Dataset

We validate the efficacy of the proposed approach on a held-out test set of 7180 patches from slides belonging to 50 patients that had no overlap with patients of the training set. Figure 2 indicates that the proposed interactive learning technique can achieve results for classification and retrieval in par with [17] with fewer labelled data. This was obtained using a standard ImageNet pre-trained ResNet-18 (with weights frozen up to layer4.1) for finetuning without additional architectural modification. Among all the proposed sampling strategies, the CNFP sampling strategy (Ref. 2.2) gives the best result. The state-of-the-art performance was obtained within 80 sessions which evaluates to an average of requiring 10 query images from each class and providing feedback to 4000 patches ($\approx 4.34\%$ of the labelled samples per class). The proposed CNFP sampling technique also performs better than the standard entropy-based sampling
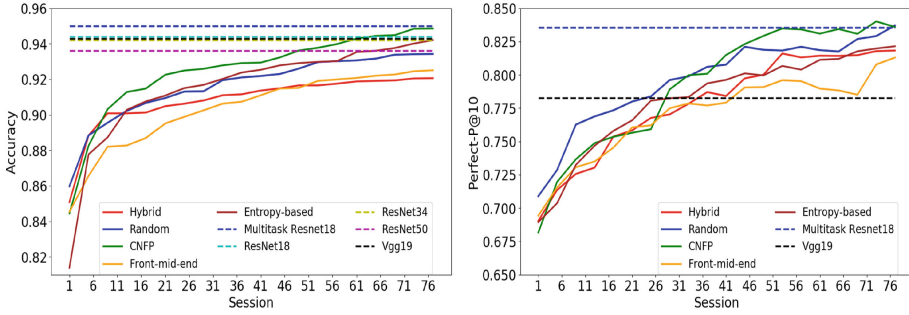
Fig. 2. Held out test set: performance of proposed interactive learning technique across feedback sessions on the heldout test set of the CRC dataset [9] in comparison to baseline methods [17] that were trained using all the annotated data. The image on the left is the accuracy, the image on the right is the macro averaged perfect P@10 score.

technique, a commonly used sampling strategy in the active learning literature. This demonstrates the efficiency and relevance of the proposed interactive learning over learning from a fully annotated dataset available upfront.

4.2   Evaluation on ICIAR Dataset

The segmentation masks are obtained from the patch label predictions of the classifier by assigning the colour code (green represents tumor and black represents normal patch). Patch locations lost during the patch extraction process are assigned black by default. Figure 3 indicates the performance of the proposed technique in the segmentation annotation task.

We notice that the annotation performance using CNFP sampling strategy (Ref. 2.2) increases with prolonged interaction, whereas the performance
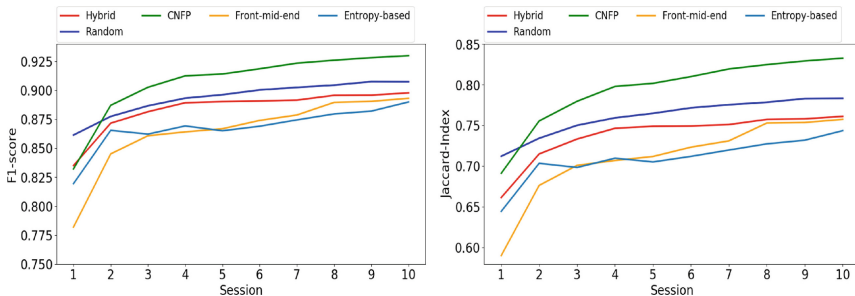

Fig. 3. Annotation performance: across feedback sessions using different sampling strategies on the ICIAR dataset. The image on the left is the slide wise macro averaged patch classification F1 score and the image on the right is the slide wise macro averaged Jaccard index score. (Color figure online)
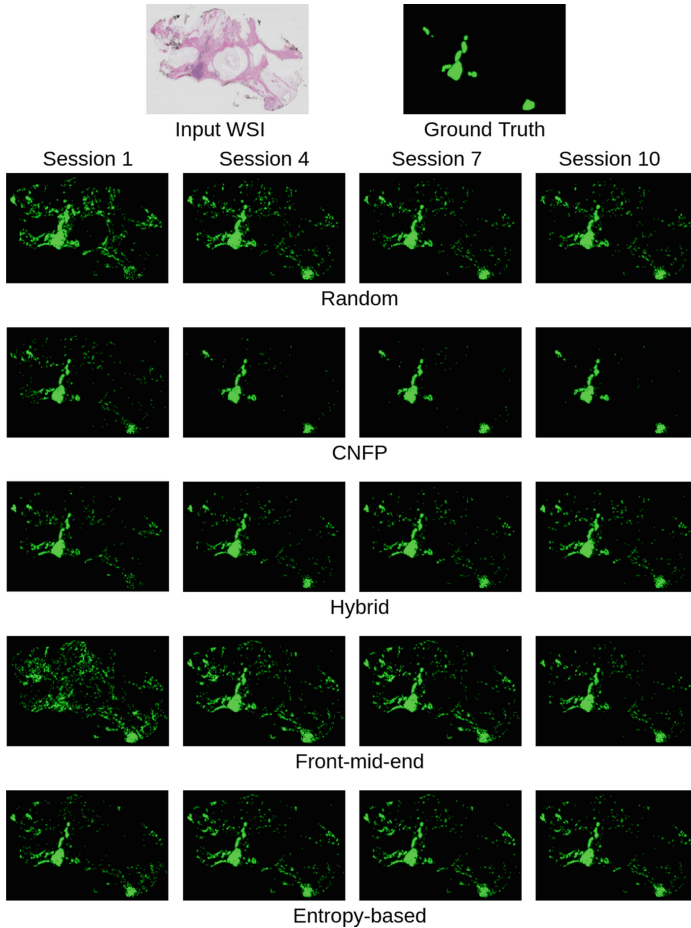
Fig. 4. Annotation output of a sample slide across feedback sessions using different sampling strategies on the ICIAR challenge dataset. The CNFP sampling strategy performs the best with segmentation output being less noisy and closest to the groundtruth

using other sampling strategies saturates. The CNFP sampling strategy (Ref. 2.2) achieves an average F1 score of 0.94 and an average Jaccard index score (represents the amount of overlap with the ground truth segmentation mask) of 0.85 within 10 sessions per slide. This includes providing 10 random tumor query patches and reviewing 250 patches ($\approx$ 2% of total patches per slide). It is important to note that though the entropy-based sampling strategy shows a similar trend, its performance is poor during the initial sessions. Figure 4 shows the segmentation output of a sample slide using the proposed interactive
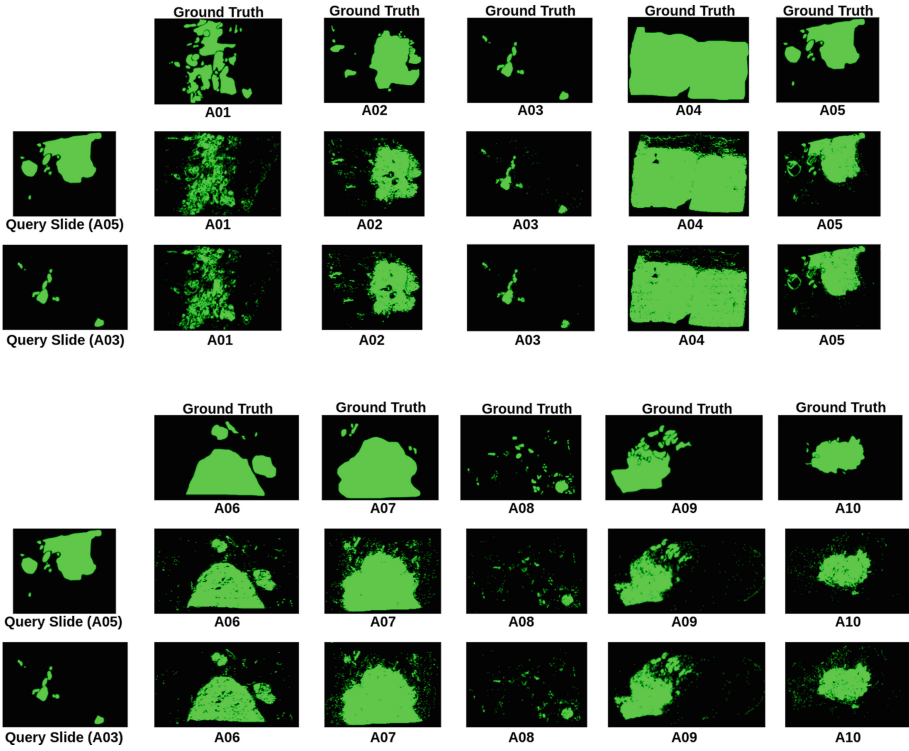
Fig. 5. Inter slide annotation: annotation results across slides with a single slide under review using the CNFP sampling strategy. The segmentation output results show that the proposed method is robust to the queried slide and is able to find anomalous patterns existing across multiple slides

learning technique. The improvement of segmented output using different sampling strategies over feedback sessions is shown. We can observe that the CNFP sampling strategy (Ref. 2.2) performs better than the other sampling strategies by obtaining a segmentation output with a Jaccard index of (≈ 0.98) at the end of 10 sessions.

Figure 5 shows the inter slide annotation results. Here the patches sampled for feedback belong to a slide different from the one understudy. The technique can be useful to help annotate multiple slides concurrently using patches queried from one slide. The segmentation results shown here are the results across 10 feedback sessions. From a diagnostic viewpoint, this could be helpful to search across multiple slides for anomalous patterns present in a slide that was already diagnosed.
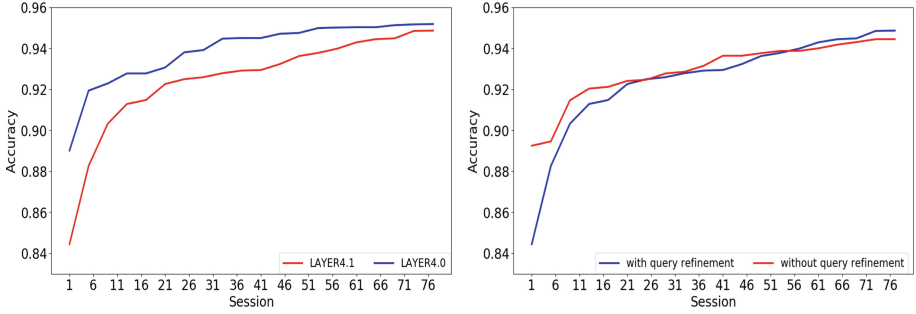
Fig. 6. Ablation studies: showing the effect of layers used for finetuning the model and query refinement on the final results using the CRC dataset [9]

4.3   Ablation Studies

We analysed two aspects for the ablation purpose on the final results. The layers used for finetuning and the effect of query refinement. Figure 6 shows the results of the ablation study performed on the CRC dataset [9].

For the finetuning experiment, we compared the performance obtained by finetuning layer4.0 of ResNet-18 and finetuning layer4.1 of ResNet-18. The model finetuned from layer4.0 reached the desired accuracy faster (20 sessions earlier) than the one finetuned from layer4.1. However, the number of trainable parameters resulting in finetuning from layer4.0 is about 78% more than layer4.1. This would increase both the time duration between every feedback session and memory for training. Hence we chose layer4.1 by prioritising the efficiency and time over the number of feedback sessions taken to achieve the desired accuracy.

We also analysed the effect of query refinement on the final results. It was observed that experiments without query refinement initially had good accuracy, but it saturates within 50 feedback sessions. The experiments performed using query refinement had a low initial accuracy but continued to increase over feedback sessions.

5   Conclusion

We propose expert feedback-driven interactive learning, which effectively reduces the requirement of large amounts of patch-level annotated images for a deep-learning-based WSI analysis. Our technique also assists WSI tumor(anomalous) region annotations and WSI diagnosis with minimal expert involvement. The proposed technique requires multiple rounds of expert feedback on informative patches selected using a novel method of sampling from a ranked list of patches. Experimentation results show that the proposed technique required only 5% of the total annotated patches to achieve state-of-the-art results. Experimentation on WSI annotation shows that the proposed technique reduces the expert efforts by requiring feedback of only 250 patches which is about 2% of total patches

per slide, to obtain segmentation outputs with 85% IOU. Inter slide annotation results also show that the proposed technique can annotate anomalous regions across multiple slides using a single slide under review. These annotated data can further be refined and used to build AI models aimed at being used as potential prognostic and diagnostic tools.

References

1. Aeffner, F., et al.: Commentary: roles for pathologists in a high-throughput image analysis team. Toxicologic Pathol. **44**(6), 825–34 (2016)
2. Bándi, P., et al.: From detection of individual metastases to classification of Lymph node status at the patient level: The CAMELYON17 challenge. IEEE Trans. Med. Imaging **38**, 550–560 (2019)
3. Bejnordi, B.E., et al.: Diagnostic assessment of deep learning algorithms for detection of Lymph node metastases in women with Breast Cancer. JAMA **318**, 2199–2210 (2017)
4. Cho, S., et al.: DeepScribble: interactive pathology image segmentation using deep neural networks with scribbles. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp. 761–765 (2021)
5. Coudray, N., et al.: Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat. Med. **24**(10), 1559–1567 (2018)
6. Jeelani, S., et al.: Histopathological examination of nail clippings using PAS staining (HPE-PAS): gold standard in diagnosis of Onychomycosis. Mycoses **58**, 27–32 (2015)
7. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. arXiv preprint arXiv:1702.08734 (2017)
8. Kather, J.N., et al.: Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. PLoS Med. **16**(1), e1002730 (2019)
9. Kather, J.N., Halama, N., Marx, A.: 100,000 histological images of human colorectal cancer and healthy tissue. Version v0.1, April 2018. https://doi.org/10.5281/zenodo.1214456
10. Kather, D.J.N., et al.: Collection of textures in colorectal cancer histology, May 2016. https://doi.org/10.5281/zenodo.53169
11. Li, H., Yin, Z.: Attention, suggestion and annotation: a deep active learning framework for biomedical image segmentation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 3–13. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_1
12. Li, W., et al.: Path R-CNN for prostate cancer diagnosis and Gleason grading of histological images. IEEE Trans. Med. Imaging **38**, 945–954 (2019)
13. Liao, H., et al.: Deep learning-based classification and mutation prediction from histopathological images of hepatocellular carcinoma. Clin. Transl. Med. **10**, e102 (2020)
14. Lindvall, M., et al.: TissueWand, a rapid histopathology annotation tool. J. Pathol. Inform. **11**, 27 (2020)

15. Musgrave, K., Belongie, S., Lim, S.-N.: PyTorch metric learning. arXiv: 2008.09164 [cs.CV] (2020)
16. Nalisnik, M., et al.: Interactive phenotyping of large-scale histology imaging data with HistomicsML. Sci. Rep. **7**, 14588 (2017)
17. Peng, T., Boxberg, M., Weichert, W., Navab, N., Marr, C.: Multi-task learning of a deep K-nearest neighbour network for histopathological image classification and retrieval. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 676–684. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_75
18. Polónia, A., Eloy, C., Aguiar, P.: BACH dataset: grand challenge on breast cancer histology images. Med. Image Anal. **56**, 122–139 (2019). https://doi.org/10.5281/zenodo.3632035
19. Putzu, L., Piras, L., Giacinto, G.: Convolutional neural networks for relevance feedback in content based image retrieval. Multimedia Tools Appl. **79**, 26995–27021 (2020)
20. Raczkowski, Ł, et al.: ARA: accurate, reliable and active histopathological image classification framework with Bayesian deep learning. Sci. Rep. **9**, 14347 (2019)
21. Sardanelli, F., et al.: Sensitivity of MRI versus mammography for detecting foci of multifocal, multicentric breast cancer in Fatty and dense breasts using the whole-breast pathologic examination as a gold standard. AJR Am J. Roentgenol. **183**(4), 1149–57 (2004)
22. Shaban, M., et al.: Context-aware convolutional neural network for grading of colorectal cancer histology images. IEEE Trans. Med. Imaging **39**, 2395–2405 (2020)
23. Shen, Y., Ke, J.: Representative region based active learning for histological classification of colorectal cancer. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp. 1730–1733 (2021)
24. Tabibu, S., Vinod, P.K., Jawahar, C.: Pan-Renal Cell Carcinoma classification and survival prediction from histopathology images using deep learning. Sci. Rep. **9**, 10509 (2019)
25. Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z.: Suggestive annotation: a deep active learning framework for biomedical image segmentation. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 399–407. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_46