Class Probability-guided Ensemble Learning-based Semantic Segmentation to Predict Cancerous Regions on Hematoxylin and Eosin-stained Images

Sanghoon Lee
Department of Computer Science
Kennesaw State University
Marrietta, GA
slee297@kennesaw.edu

Sai Chandana Koganti
Department of Computer Science
Kennesaw State University
Marrietta, GA
skogant3@students.kennesaw.edu

InChan Hwang
School of Data Science and Analytics
Kennesaw State University
Marrietta, GA
ihwang@students.kennesaw.edu

MinJae Woo
School of Data Science and Analytics
Kennesaw State University
Marrietta, GA
mwoo1@kennesaw.edu

***Abstract*—Ensemble-based machine learning aims to obtain generalized prediction outcomes improving predictive performance by combining multiple neural network models. The success of ensemble-based machine learning in providing accurate predicted regions has been widely presented in many research areas, but it is rarely investigated in the field of histopathology image analysis. In this paper, we propose a class probability-guided ensemble learning method that aims to enhance the effectiveness of semantic segmentation on cancer-related regions in histopathological images. The proposed method combines the prediction probabilities of ensemble-based semantic segmentation with the prediction probabilities of weighted ensemble-based image classification. Ensemble-based semantic segmentation is conducted by averaging from five backbone networks, ResNet18, ResNet50, Mobilenetv2, Xception, and InceptionResNetv2 from the DeepLabv3 model. Weighted ensemble-based image classification is conducted by computing the weighted averages from five different semantic segmentation models: GoogLeNet, AlexNet, InceptionResNet, VGG16, and ResNet50. The performance of the proposed method was evaluated and comprehensively analyzed using the BCSS dataset including cancer-related hematoxylin and eosin images and immune-related hematoxylin and eosin images. The results show that the proposed methods outperform the state-of-the-art semantic segmentation models in terms of accuracy, F1 measure, and IoU evaluation metrics.**

***Index Terms*—deep learning, ensemble learning, semantic image segmentation, image classification, histopathology**

I. Introduction

The recent success of artificial intelligence (AI) provides evidence based on the effectiveness of data-driven approaches in a variety of research societies [1]. AI-guided data-driven approaches have been a crucial rule for both accurate identification and rapid classification of target subjects in a large amount of data [2]. Among AI-guided data-driven approaches, deep learning models have received much attention because of their ability to capture specific characteristics defining the target subjects' presence by learning features from the data, introducing new evidence-based methods. Deep learning models and their applications have impacted our society positively and delivered significant portions of information leading to the creation of new technologies [3].

Traditional deep learning-based models have mainly used multiple layers to create a seamless architecture and repeat an iterative process to obtain optimized ones. LeNet [4], a multilayer neural network, has received a lot of attention because this 7-layered neural network applies to various situations such as audio recognition and visual analysis. Compared with LeNet, AlexNet [5] using five convolutional layers and three fully connected layers achieved the top-five error rate in an ImageNet challenge which is a large consecutive visual recognition project containing millions of images and 20,000 categories. A very deep convolutional network for large-scale image recognition called VGG [6] investigated the deep neural network by increasing the network depth more, showing a significant improvement on the ImageNet challenge. A Residual Neural Network called ResNet [7] presented an alternative way to avoid the network burden by skipping the network connections so that training is easier than the previous network even though the ResNet layers are deeper than VGG. These traditional deep learning models have been separately adopted in different fields of studies leading to better performance results [8]. Our paper aims to address the combination of different deep-learning models in image classification problems.

While the traditional deep learning-based models categorize one image into one single label, the semantic deep learning-based models focus on predicting each pixel of an image region as a class label. Fully convolutional networks called FCN [9] are a semantic deep learning algorithm performing a semantic segmentation transforming a pixel to a class label. FCN runs on the forward and backward learning tasks

2472-9671/24/$31.00 ©2024 IEEE
DOI 10.1109/ICSC59802.2024.00014

following existing image classification networks, but it takes advantage of adding skip connections in a decoder module of the upper sampling layer providing fine-grain details of image shapes. U-Net [10] extended the FCN's decoder module by enabling more accurate localization through its symmetric network shape. Over the years, it has undergone several iterations, with DeepLabV3 [11] representing one of the latest and most advanced versions. The primary objective of DeepLab is to perform semantic segmentation with exceptional precision. It achieves this by accurately assigning class labels to each pixel in an image, thereby enabling precise object localization and segmentation. A key feature of DeepLab is its extensive use of dilated convolution, often referred to as dilated convolution [12]. Our paper adopts the advantage of DeepLab's backbone networks for semantic image segmentation.

Deep learning-based models often suffer from the generalization problem; deep learning fundamentally does not provide general explanations equally because there is a gap between the losses of the training set and the test set. The ensemble-based approaches have been introduced as an alternative way to avoid the generalization problem by combining multiple models in one problem. Although these ensemble-based approaches have already matured in various fields, they are not fully investigated in digital pathology because domain experts target specific clinical and pathologic features by taking a special deep-learning algorithm assisting their works.

In this paper, we propose a class probability-guided ensemble learning method that aims to not only increase the effectiveness of semantic image segmentation but also avoid the problem of generalization in predicting cancer-related regions in the histopathology image data. The proposed method integrates the predictions of ensemble-based semantic segmentation with the predictions of weighted ensemble-based image classification by multiplying the class probabilities by the pixel probabilities. Ensemble-based semantic segmentation is conducted by averaging from five backbone networks, ResNet18, ResNet50, Mobilenetv2, Xception, and Inception-ResNetv2 from the DeepLabv3 model. Weighted ensemble-based image classification is conducted by computing the weighted averages from five different semantic segmentation models: GoogLeNet, AlexNet, InceptionResNet, VGG16, and ResNet50.

Out contribution of the paper can be summarized as follows:

- Accuracy: The proposed method was evaluated on two different cancerous regions: tumor and tumor-infiltrate lymphocytes of hematoxylin and eosin images across five state-of-the-art deep learning architectures and two traditional ensemble learning methods. The prediction results of the proposed method outperform those results predicted by the traditional learning methods.
- Generalization: The framework of the proposed method is based on ensemble-based approaches and can provide a technical way to avoid the generalization problem by taking prediction probabilities from different models.
- Validity: The use of the proposed ensemble-based methods on both image classification and semantic image segmentation was validated by comparison with state-of-the-art deep learning models. Our Ensemble-based methods can provide more accurate predictions of the cancerous regions.
- Explainability: The visual representation of Grad-CAM was compared with the probability map of the proposed method on cancer-related images and immune-related images. We observed that the predictions of our method well follow the explainability of the Grad-CAM.

II. Related Works

Ensemble learning is a valuable technique in semantic image segmentation with diverse applications, notably in medical image analysis. It plays a pivotal role in accurately detecting anomalies within medical images, benefiting both patients and healthcare professionals. Ensemble methods offer a promising strategy to enhance the precision and reliability of semantic image segmentation models [13]. By thoughtfully aggregating predictions from multiple models, ensemble techniques effectively address critical challenges like overfitting and sensitivity to initial conditions. In the context of semantic image segmentation, ensemble learning can combine different backbone architectures (e.g., ResNet, DenseNet) and models trained with various loss functions, augmentations, and hyperparameters [14]. The degree of performance improvement through ensembling depends on the diversity of the individual models, with models exhibiting uncorrelated errors typically yielding better results when combined. For even more precise results, weighted ensembling, based on model uncertainty or confidence, can be employed to improve performance compared to simple averaging. In this approach, models receive higher weights in regions where they demonstrate greater certainty. It's important to recognize that ensemble learning significantly boosts segmentation accuracy by harnessing the strengths of diverse models.

Ensemble learning in histopathology image analysis is a valuable approach that enhances diagnostic and prognostic models' accuracy and reliability, particularly in the context of pathology and cancer diagnosis [15]. Histopathology, which involves the microscopic examination of tissue samples to detect abnormalities, relies on ensemble techniques to ensure dependable and consistent results. In histopathology image analysis, ensemble learning entails combining multiple machine learning or deep learning models to create a more accurate diagnostic system [16]. These models may be trained using distinct data subsets, techniques, or architectures to reduce diagnostic errors and improve predictions. Ensemble learning is commonly applied to tissue classification tasks, where models identify various tissue structures, cellular components, or pathologies in microscopic images. By aggregating multiple models' outputs, ensemble techniques provide more precise and robust classification results. The majority voting or weighted averaging determines the final class label for a given image region. Ensemble learning is particularly valuable in reducing false positives and false negatives in histopathology image analysis, ensuring that different models complement

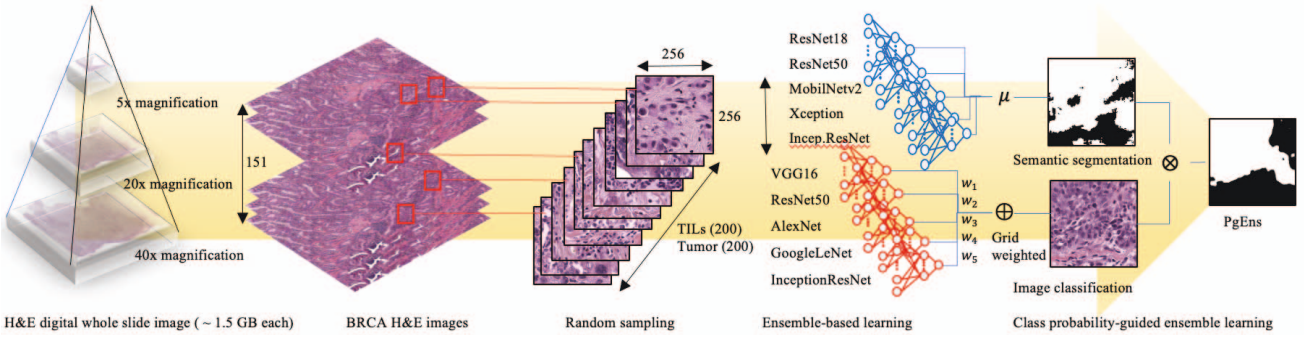

Fig. 1. The overall process of the class probability-guided ensemble learning-based semantic image segmentation for H&E digital whole slide images. Breast invasive carcinoma (BRCA) images are obtained from 20x magnification whole slide images (1.5 GB each). Fixed-sized images including 400 cancer-related images and immune-related images are randomly selected for training five deep learning models (VGG16, ResNet50, AlexNet, GoogleLetNet, and InceptionResNet) for image classification as well as for training DeepLabv3 with five backbones (ReNet18, ResNet50, MibilNetv2, Xception, and InceptionResnet) for semantic image segmentation. Ensemble-based learnings are performed for both image classification and semantic image segmentation. The binary mask for either cancer-related image or immune-related image is created by using the class probabilty-guided ensemble learning method (PgEns).

each other's strengths and compensate for their weaknesses. This contributes to a balanced and reliable diagnostic outcome, minimizing unnecessary medical interventions and missed diagnoses. Ensemble learning's utility extends to tasks such as tumor detection and grading. By combining models trained to recognize specific tumor characteristics, ensemble techniques provide comprehensive and accurate tumor assessments. This approach finds promising applications in pathology, cancer research, and other critical areas of medical diagnostics reliant on accurate image-based assessments.

III. Class Probability-guided Ensemble Learning for Semantic Image Segmentation

A. DeepLabV3 and its five backbone networks

The proposed class probability-guided ensemble learning method uses five backbone networks of DeepLabV3 and a weighted ensemble learning model generated by five deep neural network models: GoogLeNet, AlexNet, InceptionResNet, VGG16, and ResNet50. In this section, we describe DeepLabV3 and its five backbone networks on the semantic image segmentation problem. DeepLabV3 is one of the latest iterations in the DeepLab series and has demonstrated that it provides notable improvements to enhance its performance and applicability in the realm of semantic segmentation. For H&E image segmentation, we extend DeepLab's versatility to ensemble learning, allowing the combination of predictions from multiple backbone networks for heightened segmentation accuracy and overall robustness, particularly valuable in precision-demanding scenarios [17]. This multi-class semantic segmentation model is expected to excel beyond binary categorization (i.e., tumor vs. nontumor). In this paper, we adopt the advantages of DeepLab and its backbone networks to predict both cancer-related images and immune-related images. The first backbone network is ResNet18 renowned for its compact yet powerful design. Comprising 18 layers, it has made significant contributions to various computer vision tasks, including semantic segmentation. This model's strength lies in its capacity to capture intricate image features, primarily through the utilization of skip connections and residual blocks. These innovative connections allow us to mitigate vanishing gradient issues and effectively facilitate the training of exceedingly deep networks. In DeepLabV3, ResNet18 assumes a foundational role by serving as the initial feature extractor. The feature extraction capabilities of ResNet18 can significantly contribute to this complex task by providing a rich and comprehensive representation of the input data. Its design is adept at capturing both low-level and high-level features, ensuring that crucial image details are meticulously preserved throughout the entire segmentation process. We used ResNet50 as another backbone network of DeepLabV3, bringing deep residual power to the forefront for precise semantic segmentation. ResNet50 contains a 50-layer deep residual neural network architecture for capturing complex patterns and intricate details within the input data. With its depth, ResNet50 is expected to excel in extracting high-level features that contribute to the model's nuanced understanding of the content within images. Its role within the DeepLabV3 framework extends beyond foundational feature extraction, as it provides a deep and intricate foundation upon which the segmentation model is built. In scenarios where in-depth analysis and segmentation precision are paramount, We used Mobilenetv2 as another backbone network of DeepLabV3. Its lightweight architecture is tailored to provide an optimal balance between efficient feature extraction and resource optimization [18]. This feature makes Mobilenetv2 an excellent choice in scenarios where computational resources are limited. Despite its focus on efficiency, Mobilenetv2 excels in feature extraction, capturing valuable insights into the input data's content. This contribution enhances DeepLabV3's flexibility, allowing it to address semantic segmentation tasks in resource-constrained environments. With Mobilenetv2 as one of its backbone networks, DeepLabV3 gains an added dimension of adaptability and accessibility. We also used Xception as a backbone network of DeepLabV3. Xception's primary role is

to capture intricate and fine-grained details present in the input data, making it an invaluable asset in semantic segmentation tasks that demand comprehensive feature extraction. Xception's extensive architecture is designed to excel at extracting high-level features and intricate patterns from images. This capability enriches DeepLabV3's understanding of image [19] content, enabling it to recognize and delineate objects and scenes with a high degree of precision. Xception's role extends beyond mere feature extraction; it provides DeepLabV3 with the depth required to handle complex visual scenarios and objects. We assume that Xception's depth and complexity can ensure that DeepLabV3 performs exceptionally well in semantic segmentation tasks. Xception's contribution to the DeepLabV3 architecture is fundamental, enhancing the model's capability to deliver precise and detailed segmentation results, thus establishing its status as a critical and indispensable component within the DeepLabV3 framework. InceptionResNetV2 is the last backbone network used on DeepLabV3. InceptionResNetV2 can perform both intricate feature extraction and capture multi-scale contextual information, a combination that significantly enhances the overall performance of DeepLabV3. With its deep neural network architecture and inception-style modules, InceptionResNetV2 [20] contributes to the precise semantic segmentation capabilities of DeepLabV3. Its role as one of the five backbone networks further reinforces DeepLabV3's position as a leading model for semantic segmentation. InceptionResNetV2's combination of depth and inception modules makes it a valuable asset in the model's architecture, ensuring accurate and detailed segmentation, and marking it as a critical contributor to the success of DeepLabV3.

B. Weighted ensemble learning

In this section, we define the proposed weighted ensemble learning on the image classification problem. Let $B = \{b_1, b_2, \ldots, b_n\}$ be the set of base classifiers, $W = \{w_1, w_2, \ldots, w_n\}$ be the set of weights assigned to the classifiers, and $L = \{l_1, l_2, \ldots, l_m\}$ be the set of class labels, where n is the number of base classifiers and m is the number class labels. Given image I, the probability distribution for each base classifier c_i is defined as $p(b_i(I) = l_k)$, where l_k is a kth class label for ith classifier. The proposed weighted ensemble learning method combines the predictions of seven base classifiers for pixel-based image segmentation using the weighted method. The probability of the weighted ensemble learning classifier is computed by the weighted sum of class probabilities and its decision is made as below:

$$\arg\max_{l_k} \sum_{n=1}^{n} w_i p(b_i(I) = l_k) \tag{1}$$

The weight w_i is determined by using greed search optimization to maximize the performance of the weighted ensemble learning. In this paper, a grid search algorithm is used to find the optimized weights.

C. Class probability-guided ensemble learning

In this section, we define the proposed overlay-guided weighted ensemble learning. Let I be the input image, C be the class probability of I, and S be the segmentation map where each pixel belongs to a class label. The weighted ensemble learning model for image segmentation can provide S as a probability map for I, while another weighted ensemble learning model for image classification can provide C as a class probability for I. The new semantic map $\hat{S}$ through overlay-guided weighted ensemble learning is defined as:

$$\hat{S} = C_i S_{x,y} \tag{2}$$

, where x and y are the pixel coordinates of an image I, and C_i is the ith class probability on the image classification. The overall process of the proposed method is described in Fig. 1.

D. Data augmentation

Our work aims to assign a semantic label to each pixel in a histopathology image to tackle the problem of generalization of each deep neural network model by taking advantage of ensemble learning. We define an augmented image as $T(I)$, where I is the input image and $T(\cdot)$ is a transformation function corresponding to the augmentation. To strengthen the proposed method in terms of generalization, we extended our training on augmented images for each classifier (i.e., rotation angle between -5 and 5, x and y reflection, and shearing in the range of -0.05 and 0.05 degrees).

IV. Experiment

A. Tumor image classification on H&E images

Datasets: The dataset used in the tumor classification problem is collected from the Breast Cancer Semantic Segmentation (BCSS) dataset [21]. BCSS dataset consists of 151 H&E breast cancer slides obtained from the Digital Slide Archive [22]. A total of 25 participants including senior residents, junior residents, and medical students of pathology annotated 151 images following the annotation review process. We used 200 tumor-related images sized 256x256 randomly extracted from the 151 images. Tumor images are labeled as 'tumor' if the percentage of tumor regions is greater than or equal to 70% in the annotated images, while others are labeled as 'non-tumor'. The images are split into training, validation, and test sets with 60%, 20%, and 20% of images respectively.

Baselines and Metrics: We compare three ensemble methods, average-based ensemble (AvgEns), maximum-based ensemble (MaxEns), and weighted-based ensemble (WeightedEns), to five deep convolutional neural network models including AlexNet [5], GoogLeNet [23], InceptionResNet [20], VGG16 [6], and ResNet50 [7]. We use well-known evaluation metrics such as Precision (*PREC.*), Recall (*REC.*), Specificity (*SPEC.*), Accuracy (*ACC.*), and F1 score (*F1*) to evaluate the performance of the classification methods. *PREC.* is the ratio of correctly classified images to the total number of images. *REC.* is the ratio of correctly classified images to the total

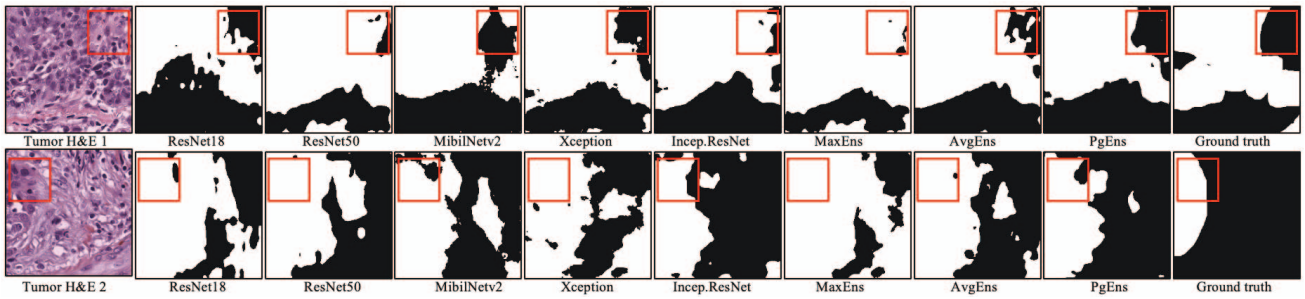

Fig. 2. Comparison of the proposed method (PgEns) on the BCSS dataset using tumor-predicted images. The proposed PgEns shows better results than the baseline and other ensemble learning methods. The details are shown in the red boxes.

number of images correctly classified. *SPEC.* is the ratio of incorrectly classified images to the total number of incorrectly classified images. F1 is the combination of *PREC.* and *REC.* We created a confusion matrix to compute the evaluation matrix; *PREC.* = TP/(TP+FP), *REC.*= TP/(TP+FN), *SPEC.*= TN/(TN+FP), *F1* = 2* *PREC.** *REC.*/(*PREC.*+ *REC.*), where TP is the number of true positive images, FP is the number of false positive images, FN is the number of false negative images, and TN is the number of true negative images. We separate the images into two groups: tumor and non-tumor and set the threshold to 0.5 classifying a binary outcome.

Results. Table 1 shows that WeightedEns outperforms both traditional deep convolutional neural network models and traditional ensemble learning in terms of *PREC.*, *SPEC.*, *ACC.*, and *F1*. We found that the results of GoogLeNet are identical to the ones from InceptionResNet and ResNet50. This is because the results of the confusion matrix of the models are the same: TP(16), FP(1), FN(4), and TN(19). WeightedEns shows an average precision of 0.8636, an average specificity of 0.8500, and an average F1 score of 0.9048. MaxEns, instead, leads to improved results on recall of 1.0000. The overall performance is measured by using *F1*, indicating that the WeightedEns produces the best performance results on the tumor classification.

TABLE I
RESULTS OF TUMOR IMAGE CLASSIFICATION

Model	*PREC.*	*REC.*	*SPEC.*	*ACC.*	*F1*
AlexNet [5]	0.8182	0.9000	0.8000	0.8500	0.8571
GoogLeNet [23]	0.8261	0.9500	0.8000	0.8750	0.8837
InceptionResNet [20]	0.8261	0.9500	0.8000	0.8750	0.8837
ResNet50 [7]	0.8261	0.9500	0.8000	0.8750	0.8837
VGG16 [6]	0.7917	0.9500	0.7500	0.8500	0.8636
AvgEns	0.7600	0.9500	0.7000	0.8250	0.8444
MaxEns	0.7143	**1.0000**	0.6000	0.8000	0.8333
WeightedEns	**0.8636**	0.9500	**0.8500**	**0.9000**	**0.9048**

B. Tumor semantic image segmentation on H&E images

Datasets: The dataset used in the tumor image semantic segmentation problem is collected from the same BCSS dataset. We used the same images (200) as the tumor image classification, but the entire annotated regions were used rather than classifying the images into tumor or non-tumor. Every pixel is labeled as either 'tumor' or 'non-tumor'. The images are split into training, validation, and test sets with 60%, 20%, and 20% of images respectively.

Baselines and Metrics: We compare three ensemble methods, average-based ensemble (AvgEns), maximum-based ensemble (MaxEns), and class probability-guided ensemble (PgEns), to five DeepLabv3-backboned models: ResNet18 [7], ResNet50 [7], MobilNetv2 [24], Xception [25], and Inception-ResNet [20]. We use *PREC.*, *REC.*, *ACC.*, *F1*, and Intersection Over Union (IoU) to evaluate the performance of the image segmentation methods.

Results: Table 2 shows that the PgEns outperforms both traditional deep convolutional neural network models and traditional ensemble learning in terms of *REC.*, *ACC.*, *F1*, and *IoU*. In particular, the MaxEns shows the best on *PREC.*, but it shows the worst on *REC.* These results concrete the reason why the *F1* is necessary to evaluate the performance of models. Likewise, AvgEns leads to improved results compared with other deep convolutional neural network models, but PgEns shows the best overall performance. Assuming that the overall performance is measured by using *F1* and *IoU*, we conclude that the PgEns produces the best performance results on the tumor image semantic segmentation. The prediction comparisons between semantic segmentation models are shown in Fig. 2.

TABLE II
RESULTS OF TUMOR IMAGE SEMANTIC SEGMENTATION

Model	*PREC.*	*REC.*	*ACC.*	*F1*	*IoU*
ResNet18 [7]	0.9471	0.8761	0.9045	0.9102	0.8352
ResNet50 [7]	0.9685	0.8614	0.9042	0.9118	0.8380
MobilNetv2 [24]	0.9503	0.8862	0.9122	0.9171	0.8469
Xception [25]	0.9638	0.8606	0.9016	0.9092	0.8336
Incep.ResNet [20]	0.9245	0.8958	0.9064	0.9099	0.8348
MaxEns	**0.9929**	0.8077	0.8755	0.8908	0.8031
AvgEns	0.9673	0.8847	0.9188	0.9242	0.8590
PgEns	0.9626	**0.8924**	**0.9215**	**0.9261**	**0.8625**

C. TILs image classification on H&E images

Datasets: The dataset used in the TILs image semantic segmentation problem is also collected from the same BCSS

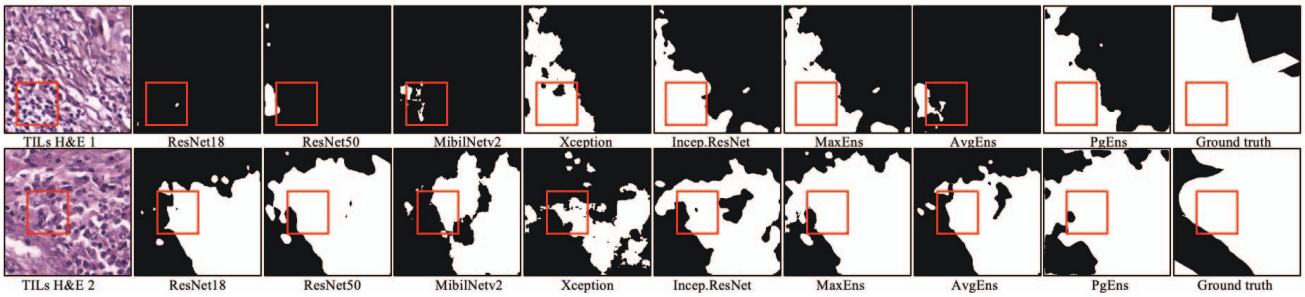

Fig. 3. Comparison of the proposed method (PgEns) on the BCSS dataset using TILs predicted images. The proposed PgEns shows better results than the baseline and other ensemble learning methods. The details are shown in the red boxes.

dataset. However, we used 200 tumor-infiltrated lymphocytes (TILs) images sized 256x256 randomly extracted from the 151 images. The images are labeled as 'tils' if the percentage of tumor regions is greater than or equal to 70% in the annotated images, while others are labeled as 'non-tils'. The images are split into training, validation, and test sets with 60%, 20%, and 20% of images respectively.

Baselines and Metrics: The baselines and the evaluation metrics are the same as the tumor image classification.

Results Table 3 shows that both VGG16 and WeightedEns outperform traditional deep convolutional neural network models as well as traditional ensemble learning in terms of *PREC.*, *SPEC.*, *ACC.*, and *F1*. Unlike WeightedEns only showed the best results on the tumor image classification, both VGG16 and WeightedEns showed the best results on TILs image classification. These results create an open question 'Is ensemble learning always producing the best results?' for image classification. From this experiment, we are not able to determine whether the ensemble-based learning in TILs image classification provides the best performance results or not. In this paper, we further experimented on TILs image semantic segmentation using the class probability-guided ensemble learning.

TABLE III
RESULTS OF TILS IMAGE CLASSIFICATION

Model	*PREC.*	*REC.*	*SPEC.*	*ACC.*	*F1*
AlexNet [5]	0.8571	0.9000	0.8500	0.8750	0.8780
GoogLeNet [23]	0.8947	0.8500	0.9000	0.8750	0.8718
Incep.ResNet [20]	0.7895	0.7500	0.8000	0.7750	0.7692
ResNet50 [7]	0.9375	0.7500	**0.9500**	0.8500	0.8333
VGG16 [6]	**0.9474**	0.9000	**0.9500**	**0.9250**	**0.9231**
AvgEns	0.9412	0.8000	0.9500	0.8750	0.8649
MaxEns	0.7308	**0.9500**	0.6500	0.8000	0.8261
WeightedEns	**0.9474**	0.9000	**0.9500**	**0.9250**	**0.9231**

D. TILs semantic image segmentation on H&E images

Datasets: The dataset used in the TILs image semantic segmentation problem is collected from the same BCSS dataset. The images are the same as the TILs image classification. However, the entire annotated regions were used rather than classifying the images into tils or non-tils. Every pixel is labeled as either 'tils' or 'non-tils'. The images are split into training, validation, and test sets with 60%, 20%, and 20% of images respectively.

Baselines and Metrics: The baselines and the evaluation metrics are the same as the tumor image semantic segmentation classification.

Results: Table 4 shows that the PgEns outperforms both traditional deep convolutional neural network models and traditional ensemble learning in terms of *ACC.*, *F1*, and *IoU*. The MaxEns showed the best results on *PREC.*, but it showed the worst results on *REC.* These results also concrete the reason why the *F1* is necessary to evaluate the performance of models. The Xception showed the best results on *REC.*, but the model showed the lowest results on *IoU*. Assuming that the overall performance is measured by using *F1* and *IoU*, we conclude that the PgEns produces the best performance results on the TILs image semantic segmentation. The prediction comparisons between semantic segmentation models are shown in Fig. 3.

TABLE IV
RESULTS OF TILS IMAGE SEMANTIC SEGMENTATION

Model	*PREC.*	*REC.*	*ACC.*	*F1*	*IoU*
ResNet18 [7]	0.8397	0.8779	0.8736	0.8583	0.7518
ResNet50 [7]	0.8397	0.8779	0.8736	0.8583	0.7518
MobilNetv2 [24]	0.8126	0.8964	0.8717	0.8524	0.7428
Xception [25]	0.7933	**0.9067**	0.8685	0.8462	0.7334
Incep.ResNet [20]	0.8371	0.8919	0.8794	0.8636	0.7600
MaxEns	**0.9389**	0.8202	0.8782	0.8755	0.7786
AvgEns	0.8547	0.9030	0.8918	0.8782	0.7828
PgEns	0.9256	0.8508	**0.8920**	**0.8866**	**0.7964**

E. Explainability of the proposed method on cancer-related cells and immune-related cells

We further investigated the explainability of the proposed method on cancer-related cells (Tumor) and immune-related cells (TILs) by using an explainable artificial intelligence (XAI). Gradient-weighted class activation mapping (Grad-CAM) is an XAI technique that represents an enhanced algorithm of class activation mapping (CAM) [26], an approach used to generate heat maps highlighting critical image regions influencing the classification decisions of Convolutional

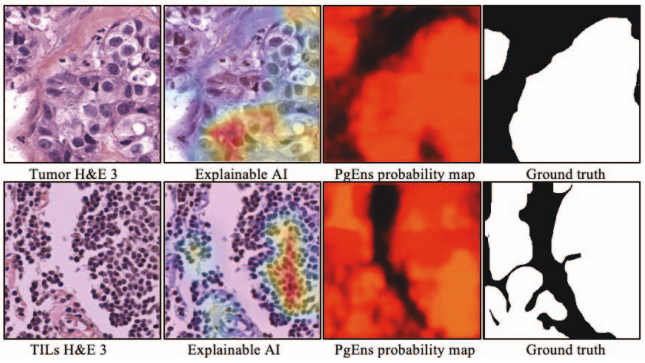

Fig. 4. Explainable AI is compared with the probability map of the proposed method (PgEns). The heatmaps of the explainable AI show the visual explanation of the model, while the probability maps generated by the proposed method show the class probabilities of tumor and TILs respectively.

Neural Networks (CNNs). Unlike its predecessors, such as CAM, saliency maps, and Grad-CAM++, Grad-CAM reveals its superiority in the localization of abnormalities in medical images [27]. It accomplishes such excellence by producing more focused and concentrated heatmaps that effectively delineate the boundaries between CNN-identified abnormalities and unaffected normal tissue, thereby enhancing classification accuracy. These heatmaps provide a comprehensive visualization of the collective topological importance of image features while analyzing the final CNN layer. We followed the Grad-CAM process that entails computing gradient scores for each label for feature map activations from a CNN layer. Subsequently, these gradient scores are averaged along both width and height dimensions to derive numeric importance scores. These importance scores are then multiplied by the feature map activation function of the last CNN layer. Then, it is followed by the application of the Rectified Linear Unit (ReLU) function to filter out and emphasize significant influences with image classification. Fig. 4 shows the visual comparison with the Grad-CAM generated predictions and the predictions generated by the proposed method (PgEns) on cancer-related image and immune-related images. We observed a robust concurrence between the prediction probability maps generated by the proposed method and Grad-CAM visualizations, further enhancing the explainability of the process of our ensemble-based semantic segmentation.

V. Conclusion

In this paper, we proposed a class probability-guided ensemble learning method for histopathology image segmentation. The proposed method combines the predictions of ensemble-based semantic segmentation with the predictions of weighted ensemble-based image classification by multiplying the class probabilities by the pixel probabilities. Five backbone networks, ResNet18, ResNet50, Mobilenetv2, Xception, and InceptionResNetv2, were ensembled in the DeepLabv3 model to perform semantic image segmentation, while five deep neural networks: GoogLeNet, AlexNet, InceptionResNet, VGG16, and ResNet50, were used for performing a weighted ensemble-based image classification. Our experiment results show that the proposed method outperforms the traditional deep learning models as well as the traditional ensemble learning methods, avoiding the generalization problem by taking prediction probabilities from different models. We also compared the prediction probability maps with the visual representation of Grad-CAM on both cancer-related images and immune-related images, observing that the predictions of our method follow the explainability of the Grad-CAM.

Acknowledgment

This research was supported by the National Science Foundation under Grant No. 2138260 and Grant No. 2153063.

References

[1] P. Liu, L. Wang, R. Ranjan, G. He, and L. Zhao, "A survey on active deep learning: From model driven to data driven," *ACM Comput. Surv.*, vol. 54, no. 10s, sep 2022. [Online]. Available: https://doi.org/10.1145/3510414

[2] M. Ng, J. Zhao, Q. Yan, G. J. Conduit, and Z. W. Seh, "Predicting the state of charge and health of batteries using data-driven machine learning," *Nature Machine Intelligence*, vol. 2, pp. 161–170, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:215947098

[3] J. yun He, S. L. Baxter, J. Xu, J. Xu, X. Zhou, and K. Zhang, "The practical implementation of artificial intelligence technologies in medicine," *Nature Medicine*, vol. 25, pp. 30 – 36, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:57574622

[4] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, pp. 2278–2324, 1998. [Online]. Available: https://api.semanticscholar.org/CorpusID:14542261

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, pp. 84 – 90, 2012. [Online]. Available: https://api.semanticscholar.org/CorpusID:195908774

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:14124313

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:206594692

[8] S. Lee, M. Amgad, P. Mobadersany, M. McCormick, B. Pollack, H. Elfandy, H. Hussein, D. A. Gutman, and L. A. D. Cooper, "Interactive classification of whole-slide imaging data for cancer researchers," *Cancer Research*, vol. 81, pp. 1171 – 1177, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:229687633

[9] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:1629541

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *ArXiv*, vol. abs/1505.04597, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:3719281

[11] Y. Zhao, D. Cai, X. Lyu, and D. Cheng, "Terraced field extraction in uav imagery using improved deeplabv3+ network," *2023 8th International Conference on Intelligent Computing and Signal Processing (ICSP)*, pp. 854–859, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:262076777

[12] L.-C. Chen, G. Papandreou, I. Kokkinos, K. P. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 834–848, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:3429309

[13] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *European Conference on Computer Vision*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:202734362

[14] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Eds., vol. 24. Curran Associates, Inc., 2011.

[15] J. Xie, R. Liu, J. Luttrell IV, and C. Zhang, "Deep learning based analysis of histopathological images of breast cancer," *Frontiers in genetics*, vol. 10, p. 80, 2019.

[16] S. Lee, M. Amgad, M. E. M. Masoud, R. Subramanian, D. A. Gutman, and L. A. D. Cooper, "An ensemble-based active learning for breast cancer classification," *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2549–2553, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:211058552

[17] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *CoRR*, vol. abs/1511.07122, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:17127188

[18] M. Akay, Y. Du, C. L. Sershen, M. Wu, T. Y. Chen, S. Assassi, C. Mohan, and Y. M. Akay, "Deep learning classification of systemic sclerosis skin using the mobilenetv2 model," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 2, pp. 104 – 110, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233262213

[19] Z. Liao, H. Hu, J. Zhang, and C. Yin, "Residual attention unit for action recognition," *Comput. Vis. Image Underst.*, vol. 189, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:204188111

[20] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," *ArXiv*, vol. abs/1602.07261, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:1023605

[21] M. Amgad, H. Elfandy, H. Hussein, L. A. Atteya, M. A. T. Elsebaie, L. S. A. Elnasr, R. A. Sakr, H. S. E. Salem, A. F. Ismail, A. M. Saad, J. Ahmed, M. A. T. Elsebaie, M. Rahman, I. A. Ruhban, N. M. Elgazar, Y. Alagha, M. H. Osman, A. M. Alhusseiny, M. M. Khalaf, A.-A. F. Younes, A. Abdulkarim, D. M. Younes, A. M. Gadallah, A. M. Elkashash, S. Y. Fala, B. M. Zaki, J. D. Beezley, D. R. Chittajallu, D. Manthey, D. A. Gutman, and L. A. D. Cooper, "Structured crowdsourcing enables convolutional segmentation of histology images," *Bioinformatics*, vol. 35, pp. 3461 – 3467, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:73424650

[22] D. A. Gutman, M. Khalilia, S. Lee, M. Nalisnik, Z. Mullen, J. D. Beezley, D. R. Chittajallu, D. Manthey, and L. A. D. Cooper, "The digital slide archive: A software platform for management, integration, and analysis of histology for cancer research." *Cancer research*, vol. 77 21, pp. e75–e78, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:3894618

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:206592484

[24] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:4555207

[25] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1800–1807, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:2375110

[26] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, pp. 336 – 359, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:15019293

[27] J.-C. Chien, J.-D. Lee, C.-S. Hu, and C.-T. Wu, "The usefulness of gradient-weighted cam in assisting medical diagnoses," *Applied Sciences*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:251298176