IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 5 April 2023, accepted 7 May 2023, date of publication 22 May 2023, date of current version 30 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3278275

RESEARCH ARTICLE

Semantic Segmentation With Multiple Contradictory Annotations Using a Dynamic Score Function

POOYA ESMAEIL AKHOONDI AND MAHDIEH SOLEYMANI BAGHSHAH
Department of Computer Engineering, Sharif University of Technology, Tehran 11155-9517, Iran
Corresponding author: Mahdieh Soleymani Baghshah (soleymani@sharif.edu)

ABSTRACT Semantic Segmentation aims to partition an image into separate regions where each region conveys certain valuable information. In recent years, deep learning models have achieved high performance in this task. However, when several ground truth segmentations are available, aggregating the information of these segmentations into a single ground truth becomes a crucial pre-processing step. This task becomes challenging when the segmentations are contradictory and the existing classes in the segmentations are imbalanced. An elegant example is the grading of Prostate Cancer in the Gleason 2019 Challenge dataset. This dataset provides six annotations of relatively high contrast from expert pathologists for each image. Additionally, the low number of images for Gleason grade 5 has also resulted in an imbalanced dataset. The Majority Voting and the Simultaneous Truth And Performance Level Estimation (STAPLE) algorithm are the most popular algorithms for combining annotations. In this paper, we visually show that the outputs of these algorithms discard the semantics of the image in regions of high inconsistency and point out that they demote low-frequency patterns, making the final segmentations even more imbalanced. We claim these drawbacks highly decrease the performance of deep learning models trained on these ground truths. To solve this problem, we propose a dynamic score function that selects one of the six annotations for each image while balancing the Gleason grading among the annotations in terms of variability and quantity. Finally, we train and evaluate a Pyramid Scene Parsing network on the final ground truths to validate our claims.

INDEX TERMS Convolutional neural networks, Gleason 2019 challenge, Gleason grading, Multi-expert annotations, Prostate cancer, Semantic segmentation.

I. INTRODUCTION

Semantic segmentation refers to the task of identifying several regions in an image where all the pixels of each region belong to the same class. In other words, semantic segmentation assigns a categorical label to every pixel in an image, which is also regarded as a pixel-wise classification problem. Due to its nature of dealing with images and recognizing local features, utilizing the convolution operator is an appropriate approach for solving this problem [1], [2]. In recent years, various convolutional neural networks have been introduced for the task of semantic segmentation that have shown promising results on several datasets [3], [4], [5]. However, the great performance of these networks is partly due to the ground truths in the dataset being semantically valid.

In some cases, several segmentations are provided for an image and it is required to aggregate these segmentations into a single ground truth before feeding them to convolutional neural networks. One of the most well-known medical image analysis datasets containing several segmentation for an image is the dataset of the Gleason 2019 Challenge [6], which was part of the MICCAI 2019 Conference. This dataset includes 244 TMA images and their corresponding pixel-level annotations prepared by six pathologists [7], [8]. Specifically, six expert pathologists were instructed to draw closed contours around different regions of each of the TMA

The associate editor coordinating the review of this manuscript and approving it for publication was Essam A. Rashed.

images and mark all pixels inside each specified region as benign or Gleason grade of 3 to 5 [9]. Some images were not annotated by all of the six pathologists, but for every TMA image, four to six pixel-level annotations were available in the dataset. Furthermore, we observed that few annotations provided by the pathologists contained Gleason grade 5. Therefore, the Gleason 2019 Challenge dataset is regarded as an imbalanced dataset.

Previously, Majority Voting [10] and the Simultaneous Truth And Performance Level Estimation (STAPLE) [11] algorithm have been introduced as methods for combining multiple segmentations. Majority Voting refers to assigning to each pixel a class that is agreed by the highest number of segmentations available [10]. On the other hand, the STAPLE algorithm constructs the final ground truth by an EM [12] algorithm. i.e., iteratively updates its constructed segmentation until convergence. Specifically, the STAPLE algorithm takes the probabilities of the category labels for each pixel and updates these probabilities by maximizing an objective function through the use of an EM setting [13].

These methods have shown promising results on many problems and have been widely used in recent years. However, in this paper, we indicate that the application of these algorithms to segmentations of images with high diversity and imbalanced classes results in two shortcomings: 1) creating semantically incorrect segmentations in regions of high contrast; 2) Making the dataset imbalanced even further by demoting the minority class and promoting the majority class. Therefore, taking these issues into account is necessary for improving the performance in the semantic segmentation task.

In this paper, we propose a dynamic score function (DSF) to solve these two problems in contradictory segmentations. Specifically, we assign scores to each available segmentation map of an image and aim to select one of these segmentations as the final ground truth for that image. The score assigned to each segmentation includes a static part and a dynamic part. To value the class with the highest probability, the static part of our function assigns higher values to segmentations closer to the majority voting segmentation, which can be done as a pre-processing step and therefore, provides scores with static values. On the other hand, to consider the classes with lower probabilities and promote a more balanced segmentation set, the dynamic part of our function assigns higher values to segmentations which can efficiently improve the dataset balancing and lower values to other segmentations. Since the whole dataset is required to understand which classes are scarce and which classes are abundant, this part of the function is dependent on the current dataset selection and is, therefore, considered dynamic. To settle with an appropriate selection set, we implement a hill-climbing [13] method to iteratively improve our selection set by choosing segmentations with higher scores and finalize the ground truth selections afterward.

The main contributions of our paper are described as follows:

- We provide an algorithm for selecting a segmentation map from the several segmentations of a single image. This algorithm has three properties: 1) It maintains the semantic information available in the segmentation maps. We claim that this is an important factor to be considered for the appropriate training of neural networks. 2) Our algorithm prioritizes segmentations containing classes that are common in most of the maps. This imitates the practical result of assigning higher probabilities to commonly agreed regions. 3) Our algorithm prioritizes segmentations containing rare classes. This imitates the practical result of assigning high variance for the rare classes and the common ones.
- We graphically illustrate the difference between our method, the STAPLE, and the Majority Voting algorithm on the segmentation maps to visually show that STAPLE and Majority Voting algorithms do not necessarily maintain the semantic information in the final results.
- We train and evaluate a PSPNet on [14] different settings, including our segmentation results and the STAPLE segmentation maps to verify our claim that segmentation neural networks train better on our proposed method. Moreover, we provide an ablation study to verify our claim that data augmentation and resampling methods [15] are not enough to improve the model's performance, and the third property of our algorithm is fundamental for good performance.

This paper is structured as follows: In the Related Work section, we outline the previous methods implemented for the Gleason grading task of the Gleason 2019 Challenge dataset and provide a general overview of the task of prostate cancer detection. In the Preliminaries section, we first introduce the notations that will be used throughout this paper and formulate the semantic segmentation problem. Next, we revisit the application of Majority Voting and the STAPLE algorithm on the Gleason 2019 Challenge dataset with additional detail. In the Proposed Method section, we explicitly propose our method based on the notations provided in the paper and then introduce the general network structure of the PSPNet and the details of the training and evaluation of our network. In the Results And Discussion section, we compare our results with the top participating teams in the challenge based on several metrics and different tasks. In the Conclusion section, we conclude our contributions.

II. RELATED WORK

In this section, we first review the general approaches for aggregating maps and then present the recent methods for histopathological Gleason grading.

A. AGGREGATION APPROACHES

The task of aggregating several mapping functions into a single function has been extensively studied both directly and

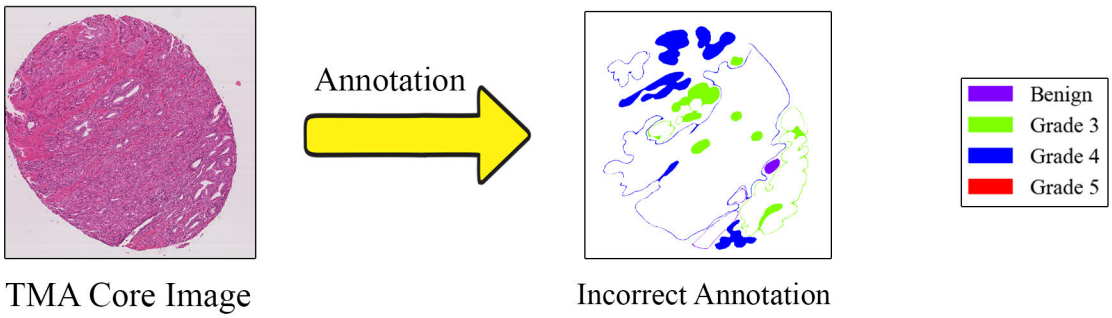

FIGURE 1. An example of an incorrect annotation provided by one of the pathologists for a TMA image in the Gleason 2019 Challenge dataset [6].

indirectly by researchers focusing on other tasks with similar ideas. Two of the algorithms that can be directly applied to sets of image inputs include the Majority Voting and the STAPLE algorithm. In addition, it is possible to view the segmentation maps as different domains or views of a single entity and the goal is to combine the information of these separate domains into a single global domain. In this way, it is possible to approach the problem from different perspectives.

One of these viewpoints is the domain adaptation problem, in which several domains correspond to different segmentation maps, and the goal is to propose a semantic segmentation approach that adapts one domain to another, eventually concluding with a final ground truth. Xie et al. [16] proposed a self-training method for making predictions about a target domain given labeled source domains. In their work, they proposed a centroid-aware pixel contrast method that employs the class centroids of the source domain for learning discriminative features. To compensate for the imbalance in the classes, they proposed a distribution-aware pixel contrast method, in which they approximated the true distribution of classes from the statistics of the source domain. Specifically, to calculate the statistics of the source domain, they considered the mean and the covariance matrix of the pixel features corresponding to a particular class for each image. Next, they stacked the mean features for several images as a bank feature and aggregated the mean and the covariance matrix features on the total dataset to obtain global features for the corresponding class.

In a dataset such as the Gleason 2019 Challenge dataset where several contradictory segmentations are available for each image, we could consider six source domains for each image corresponding to the maps according to this paper. However, since several different classes could be assigned to the same region of the input image, this has the drawback that the mean feature vectors and the covariance matrices contain similar vectors that come from different classes, making these features difficult to distinguish among different classes. In other words, the features of different classes collide with one another in the feature space, making the mean vectors and the covariance matrices inappropriate prototypes for the features. As opposed to our work, we endeavor to pick a subset of these features, i.e. one segmentation from all the available segmentations of each image as our final ground truth.

Another point of view is the image registration problem, in which the spatial relationship between several images in the same location is found to obtain the maximum image information. Ban et al. [17] proposed a weighted spatial histogram algorithm to extract statistical features. In their method, they computed the mean and the covariance matrix of the image pixels with a specific value. They further formulated the similarity of two histograms using a weighted sum on all the different possible values of pixels. In the Gleason 2019 Challenge dataset, we could consider the segmentation maps as the inputs of image registration problem. However, since every input has contradictory segmentations available, this implementation has the disadvantage that the mean and the covariance matrices contain similar content from the classes in regions of high contrast which decreases the discriminative power of these features among the different classes. On the contrary, our work aims to select a portion of these features and assign as the final ground truth.

B. DEEP LEARNING FOR GLEASON GRADING

Computer-aided detection and grading of prostate cancer [18] can be used as a powerful tool for enhancing the histopathological Gleason grading and the treatment selection of prostate cancer [19]. In recent years, deep learning [20] models have achieved state-of-the-art performance in the semantic segmentation of histopathology images for medical analysis [21], [22]. Several papers have presented the superiority of deep neural networks in prostate cancer diagnoses.

Regarding the Gleason 2019 Challenge, Khani et al. [23] implemented a DeepLabV3+ [24], [25] with a pre-trained MobileNetV2 [26] backbone for the semantic segmentation task. To solve the imbalance in the Gleason 2019 Challenge dataset, they applied data augmentation [16] on the less frequent Gleason grades. To prevent overfitting and achieve more generalization, they also applied data augmentation for the second time on the whole dataset. They further indicated that some of the annotations provided by the pathologists in the dataset contained improperly closed contours which

led to poor training of deep learning models. Fig.1 shows an example of an annotation with improperly closed regions. To solve this problem, they manually corrected the annotations by filling in the unclosed regions. For the final ground truth, they implemented the STAPLE algorithm using the SimpleITK [27], [28] library.

Qiu et al. [29] utilized the PSPNet with an auxiliary branch for the semantic segmentation of TMA images. The PSPNet architecture is composed of the ResNet–101 [30], [31], followed by pooling layers and finally, the head of the Fully Convolutional Network (FCN) [32]. Qiu et al. added an auxiliary branch to their network by using the auxiliary loss of the ResNet–101 and passing the features through an FCN. They trained their final network by taking a linear combination of the losses of the two branches and backpropagating through the parameters. For obtaining the ground truth labels for the dataset, they utilized the gold-standard STAPLE algorithm on the annotations provided by the pathologists. Qiu et al. submitted their code to the Gleason 2019 Challenge and achieved first place in the challenge.

Zhang et al. [33] implemented a UNet [34], [35] by exploiting convolutional blocks for feature extraction. They connected the encoder of the network to the decoder by concatenating the feature maps of the encoder to the corresponding convolutional blocks in the decoder. To obtain the ground truths of the dataset, they encoded the annotated values onto six channels and merged annotations by different pathologists for each channel to reduce the inter-variance among the annotations. Zhang et al. participated in the Gleason 2019 Challenge and finished fourth place in the challenge.

Iqbal et al. [36] conducted a thorough analysis and compared the performance of traditional machine learning models such as SVM [37], Decision Tree [38], and Kernel Naive Bayes [39], [40] to deep learning models such as ResNet and LSTM [41] on the task of prostate cancer detection. They concluded that ResNet–101 and LSTM achieve optimal results in the feature extraction of prostate cancer images.

The ISIC dataset provides a large set of images containing skin lesions [42]. Similar to the Gleason 2019 Challenge dataset, the ISIC dataset provides several skin lesion segmentation boundaries for each image. One of the tasks defined on this dataset is the segmentation of skin lesions, in which the problem contains two classes, the malignant class, and the benign class. Ribeiro et al. provided an analysis of the different methods used for the segmentation maps of the ISIC dataset [43]. They also indicated that up to 2018, the ISIC dataset was the only dataset that provided several segmentation maps for images. However, we report that the 71670 images publicly made available in the ISIC archive and the ones in the challenge websites only contain the input images and do not include the several corresponding segmentation masks [44]. Moreover, the ground truth images provided in the challenge website only contain one mask for each image. Therefore, we only provide our experimental results on the Gleason 2019 Challenge dataset which contains four classes with imbalanced frequencies.

III. PRELIMINARIES

First, we provide the notations that will be used throughout this paper and demonstrate a mathematical formulation for our problem. Then, we provide a thorough analysis of the advantages and shortcomings of the Majority Voting and STAPLE algorithms.

A. NOTATIONS AND PROBLEM FORMULATION

We formulate the problem of semantic segmentation of TMA core images on the Gleason 2019 Challenge dataset as follows:

Define $g = \{0, 3, 4, 5\}$ as the set of all possible Gleason grades assigned to the pixels of an image, h as the height and w as the width of the image input. Also, define $p = \{0, 1, \ldots, 255\}$ as the set of all possible values of a pixel in an RGB image input channel. Therefore, $p^{3\times h\times w}$ is the space of all input images with RGB channels, $g^{h\times w}$ is the space of all possible segmentation results and $f : p^{3\times h\times w} \rightarrow g^{h\times w}$ is defined as the function that takes an image input and outputs a corresponding segmentation using the Gleason grades 0, 3, 4, and 5. Let the set $\{Map_i\}_{i=1}^{6}$ contain instances of the mentioned functions, where Map_i takes an image input from the dataset and outputs the segmentation provided in the *i-th* map. With these definitions, we wish to devise a function $f : p^{3\times h\times w} \rightarrow g^{h\times w}$ such that $f(x)$ provides us with a suitable semantic segmentation result where $x \in p^{3\times h\times w}$ is an image given as input.

B. MAJORITY VOTING

The majority voting algorithm proposes an objective function as the classification error among the six maps and formulates the problem as obtaining a function f that minimizes the value of the objective function [10]. The objective function is defined as follows:

$$\sum_{i=1}^{h}\sum_{j=1}^{w}\sum_{k=1}^{6} I\left(f(x)_{ij} \neq Map_k(x)_{ij}\right). \tag{1}$$

We can rewrite the classification error in the following way:

$$\begin{aligned}
&\sum_{i=1}^{h}\sum_{j=1}^{w}\sum_{k=1}^{6} I\left(f(x)_{ij} \neq Map_k(x)_{ij}\right) \\
&\quad = \sum_{i=1}^{h}\sum_{j=1}^{w}\sum_{k=1}^{6}\left(1 - I\left(f(x)_{ij} = Map_k(x)_{ij}\right)\right) \\
&\quad = 6hw - \sum_{i=1}^{h}\sum_{j=1}^{w}\sum_{k=1}^{6} I\left(f(x)_{ij} = Map_k(x)_{ij}\right).
\end{aligned} \tag{2}$$

In other words, minimizing the classification error in (2) is equivalent to maximizing the following expression:

$$\sum_{i=1}^{h}\sum_{j=1}^{w}\sum_{k=1}^{6} I\left(f(x)_{ij} = Map_k(x)_{ij}\right). \tag{3}$$

It is straightforward to see that the objective function in (3) is maximized when we take the majority vote of all six maps for each pixel of an image. Therefore, the majority voting algorithm defines $MajVote(x)\colon p^{3\times h\times w} \rightarrow g^{h\times w}$ as the ground truth segmentation of an image x as follows:

$$MajVote(x) := \arg\max_{f} \sum_{i=1}^{h}\sum_{j=1}^{w}\sum_{k=1}^{6} I\left(f(x)_{ij} = Map_k(x)_{ij}\right). \tag{4}$$

Despite the algorithm's simplicity and its success in finding a solution that globally optimizes the proposed objective function (3), it has shortcomings described as follows:

- First, in the semantic segmentation task, the ground truth maps contain semantically different regions where all the pixels within the same region share the same class. This implies a correlation between the pixels within each semantic region. In addition, deep learning [20] models try to capture the correlation between the contiguous pixels using convolutional layers and fully connected layers. Therefore, maintaining the correlation between pixels of each semantic region in the final suggested ground truth is fundamental for good performance. However, the objective function in (3) indicates that the class of each pixel depends only on the corresponding pixel among the different maps. This fact implies that the majority voting equation in (4) cannot guarantee that the distinct pixels within each semantic region share the same class.
- Second, the majority voting algorithm is predicated on the idea that segmentations containing low-frequency classes are fallacious, and should therefore be removed from the final representation, and only segmentations with which most annotators agree should be considered as ground truth. Thus, there is a bias or preference of selecting high-frequency classes which may not be suitable for problems such as the grading of prostate cancer in TMA images. For this reason, providing a ground truth that does not necessarily eradicate low-frequency segmentation is crucial for a successful implementation.
- Third, the majority voting algorithm's supremacy dissipates when applied to an imbalanced dataset, which is the case with the Gleason 2019 Challenge dataset. Our study on this dataset reveals that for many of the images in the dataset, at most one expert assigned grade 5 to some region of the image, while the other experts mostly agreed with one another on assigning other grades. Furthermore, we observed that there are few images where most of the experts agreed with one another on assigning grade 5. As explained in the previous paragraph, discarding low-frequency segmentations for proposing a final ground truth eliminates the data of the classes that were available in the low-frequency segmentations. In the case of the Gleason 2019 Challenge dataset, this has the consequence that many of the segmentations which contain the Gleason grade 5 are eliminated from the final ground truth, making the Gleason grade 5 even more scarce than the original set of segmentations. Ultimately, this algorithm decreases the variability of the Gleason grade 5 patterns in the dataset, which results in poor training of any deep learning model.

To justify our point, we have implemented the majority voting algorithm on the segmentation maps of various images from the Gleason 2019 Challenge dataset and showed the results in Fig.5. In this figure, the Image column refers to the original image of the dataset, and the columns Map_1, Map_3, Map_4, Map_5, and Map_6 refer to the segmentation maps in the dataset. The entries in which the particular segmentation maps were unavailable are empty. The MajVot column refers to the majority voting algorithm applied to the segmentation maps. It is evident that in regions of high contrast, unusual shapes have been created by the algorithm which eliminates the required semantic properties for regions of those classes. Moreover, we can observe that low-frequency classes such as Gleason grade 5 were completely discarded in the output of this algorithm. Additionally, Fig.2(a) shows the distribution of the classes among the original segmentations provided by the experts in the dataset. Fig.2(b) shows the distribution of the classes among the segmentations resulting from applying the majority voting algorithm to the original segmentation maps.

C. STAPLE

The objective function defined in (1) suggests that taking into account high-frequency classes is beneficial for appropriately obtaining the final ground truth. This implication has led deep learning researchers to devise more elaborate algorithms, the most successful one, to the knowledge of the authors, is the STAPLE [13] algorithm for the segmentation task. It is noteworthy to mention that the participating teams in the Gleason 2019 Challenge were evaluated on the test dataset with the STAPLE algorithm implemented as the aggregation method [45]. Moreover, several papers [23], [29] have implemented the STAPLE algorithm as a preprocessing step in the Gleason 2019 Challenge to obtain the ground truth annotations for the training dataset and achieved state-of-the-art performance with their deep learning models.

The STAPLE algorithm constructs the final ground truth as an EM [14] algorithm, iteratively updating its constructed segmentation to maximize the incomplete data log-likelihood function. Specifically, the algorithm initially constructs a test segmentation by simple voting on each pixel. Further, the algorithm calculates the performance level parameters of each of the annotators compared to the test segmentation,

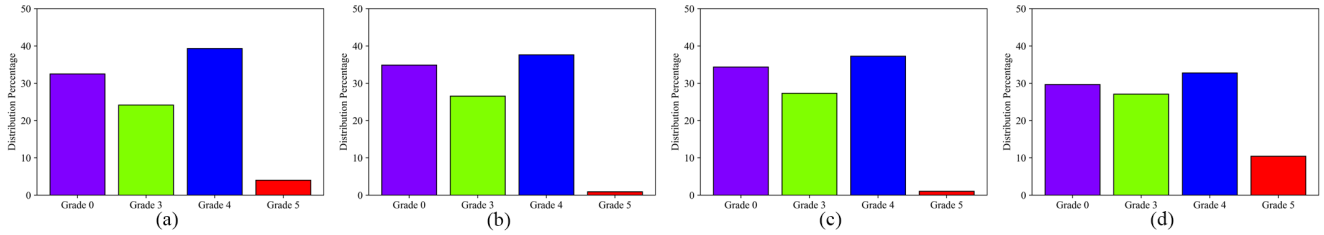

FIGURE 2. The Gleason grade distribution in Gleason 2019 Challenge dataset [6]. (a) the distribution in the set of segmentation maps of the Gleason 2019 Challenge dataset. (b) the distribution after applying the majority voting algorithm to the semantically correct segmentations. (c) the initial set of ground truths in the hill-climbing algorithm. (d) the final set of ground truths in the hill-climbing algorithm.

as the E-step, and constructs a new test segmentation using the weights previously calculated as the M-step. This process is repeated until the test segmentation converges and the converged test segmentation is considered as the final ground truth [13].

Despite its widespread application on several datasets as a novel aggregation method, its performance on the semantic segmentation task of the Gleason 2019 Challenge dataset has two shortcomings. First, the EM procedure implemented in the STAPLE algorithm does not consider the class assigned to a pixel depending on the classes of its surrounding pixels. In other words, the same semantic properties of an acceptable segmentation result that were discussed in the previous subsection are not taken into account in the result of the STAPLE algorithm. Second, the STAPLE algorithm performs poorly on datasets with a high imbalance in the class categories and full of contradictory segmentations, which is the case in our Gleason 2019 Challenge. Similar to the majority voting algorithm, the STAPLE algorithm usually discards the class with the least frequency, Gleason grade 5, and severely decreases the frequency of grade 5 in the final ground truth segmentations.

To justify our point, we have implementated the STAPLE algorithm on the segmentation maps of various images from the Gleason 2019 Challenge dataset and showed the results in Fig.5. In this figure, the STAPLE column refers to the output of the STAPLE algorithm applied to the segmentation maps. Similar to the output of the majority voting algorithm, it can be seen that in regions of high contrast, unusual shapes have been created that discard the required semantic properties for regions of those classes. Moreover, it can be seen that low-frequency classes such as Gleason grade 5 were completely discarded in the output of this algorithm.

IV. PROPOSED METHOD

In this paper, we shall propose a method that maintains the necessary semantic properties mentioned in the Preliminaries section for a successful implementation, and simultaneously solve the imbalance of classes in the dataset. As explained in the previous section, it is necessary to consider a ground truth that identifies groups of glandular cells that share the same Gleason grade, without containing any peculiarly shaped regions.

More precisely, if we define D to be the subspace of $g^{h\times w}$ that harbors the necessary semantic properties, we wish to propose a function $f: p^{3\times h\times w} \rightarrow D$ as the ground truth. As described in the Majority Voting subsection of the previous section, it is not guaranteed that the maximization of the objective function in (3) necessarily lies in the subspace D. However, we explained in the Introduction section that the pathologists were instructed to draw the contours of regions belonging to a particular Gleason grade for preparing their annotation. In other words, the pathologists were instructed to provide annotations lying in the subspace D.

A. COMPATIBILITY OBJECTIVE

First, we focus on proposing an objective function that helps the output of our method become compatible with the correct segmentations provided by the pathologists. Fig.1 shows an example of an annotation with improperly closed regions. Various solutions such as manually correcting these mistakes have been pointed out by researchers [23]. However, in our work, we simply manually removed these incorrect segmentations from the whole set of segmentations provided by the pathologists. After removing fallacious segmentations, we are left with only segmentations that lie inside the domain D in our study. Since we wish to make sure to present ground truth segmentations that lie inside subspace D, we opt to skillfully select one segmentation from the segmentation maps available for each image in the dataset and designate it as the ground truth.

To address the importance of considering segmentations with high agreement among the annotators, we propose a function as a measurement of the segmentation's similarity to the majority voting result as follows:

$$F_1(a,x) := \frac{1}{hw}\left(\sum_{i=1}^{h}\sum_{j=1}^{w} I\left(a_{ij} = MajVote(x)_{ij}\right)\right) \quad (5)$$

where $x \in p^{3\times h\times w}$ is a TMA image from the Gleason 2019 Challenge dataset and $a \in g^{h\times w}$ is an arbitrary segmentation presented for our image input x.

It is evident from (5) that a higher value of $F_1(a,x)$ indicates higher similarity of the segmentation to the majority voting segmentation. Concretely, the output of the function

$F_1(a, x)$ always lies in the interval [0, 1], where the function claims the maximum value 1 when $a = MajVote(x)$.

B. BALANCING OBJECTIVE

In this subsection, we focus on proposing an objective function that balances the selected segmentations. Initially, we define a function with which we can identify an arbitrary segmentation's ability to contribute to the abundance of a particular class as follows:

$$J_k(a) := I\left(\frac{1}{hw}\left(\sum_{i=1}^{h}\sum_{j=1}^{w} I\left(a_{ij} = k\right)\right) \geq \alpha_k\right) \tag{6}$$

where $a \in g^{h\times w}$ is an arbitrary annotation, $k \in g$ is the particular class in our study, and α_k is defined as the median of the objective function on the left-hand side among all the images in the dataset with a positive contribution to the class k. In other words, we define α_k as follows:

$$\alpha_k := Median\left(\left\{\frac{1}{hw}\left(\sum_{i=1}^{h}\sum_{j=1}^{w} I\left(Map_m(x)_{ij} = k\right)\right)\right| I\left(\sum_{i=1}^{h}\sum_{j=1}^{w} I\left(Map_m(x)_{ij} = k\right) > 0\right)\right\}\Bigg) \tag{7}$$

where $m \in \{1, 2, \ldots, 6\}$ is any map index and x is an arbitrary image input from the dataset. Therefore, considering a feasible set is essential for correct implementation. With this definition, approximately half of the segmentation maps with a positive contribution to the class k would have $J_k(a) = 1$ and the other half would have $J_k(a) = 0$.

In particular, we argue that segmentations containing a low number of a particular class cannot contribute to the balancing of the dataset, and there should be a lower bound on the number of pixels that are marked as a particular class. When the indicator function in (6) takes the value of 1, it is determined that the particular segmentation at hand can contribute to the balancing of the dataset. Otherwise, the indicator function takes the value of 0, implying that the number of pixels of the particular class in the segmentation is not sufficient for balancing purposes. In other words, the function $J_k(a)$ determines whether the chosen segmentation can effectively contribute to the balancing of the dataset.

Using our defined function $J_k(a)$, we can further extend our notion to the whole dataset as follows:

$$N_k := \sum_{n=1}^{N} J_k(a_n) \tag{8}$$

where $k \in g$ indicates a class, and for every $1 \leq n \leq N$, $a_n \in g^{h\times w}$ is the segmentation map chosen for the n-th image in the dataset and N is the size of the dataset.

Now, we propose a function that determines whether the dataset is deficient in a particular class as follows:

$$Q(k) := I\left(\frac{N_k}{\sum_{i\in g} N_i} \leq \frac{1}{2|g| - 1}\right) \tag{9}$$

where $k \in g$ is the class under our study.

The function Q claims the value of 1 when the number of segmentations among the whole set of segmentations containing a reasonable amount of a particular class falls below a threshold, indicating a need for an increase in the segmentations containing this scarce class. When the function claims the value of 0, it implies that enough segmentations are containing the particular class at hand.

The intuition of selecting $\frac{1}{2|g|-1}$ as a threshold is as follows: In our method, to ensure that the variables N_i, where $i \in g$, are not very distant from one another, we wish to maintain the following inequality:

$$\max_{i\in g} N_i \leq 2\min_{i\in g} N_i \tag{10}$$

With this idea in mind, an acceptable situation in which the expression on the left-hand side of (9) claims its minimum value for class k is when $N_i = 2N_k$ for all $i \in g$ and $i \neq k$, in which case the value of $\frac{1}{2|g|-1}$ is achieved. In other words, when the expression on the left-hand side of (9) falls below the threshold, there exists some class N_i where $N_i > 2N_k$, therefore, class k needs to be increased among the segmentation maps.

To address the value of an arbitrary segmentation's contribution to the balancing of the dataset, we propose the following function:

$$F_2(a) := \sum_{k\in g} Q(k) J_k(a) w_k \tag{11}$$

where $a \in g^{h\times w}$ is an arbitrary segmentation, and w_k are positive real numbers defined as follows:

$$w_k := \frac{\frac{1}{N_k}}{\sum_{i\in g}\frac{1}{N_i}}. \tag{12}$$

The expression $Q(k) J_k(a)$ in each summand equals 1 when there is not only a need for increasing the frequency of class k in the set of segmentation maps but also the segmentation at hand contains class k enough to be considered suitable for contributing to the balancing of the dataset. In this case, a reward is given to the segmentation with a value of w_k. Otherwise, the expression $Q(k) J_k(a)$ equals 0, and consequently, the corresponding summand equals 0. Concretely, the output of the function $F_2(a)$ always lies in the interval [0, 1] where the function claims the maximum value 1 when for all $k \in g$, $Q(k) J_k(a) = 1$. Therefore, the function $F_2(a)$ can be used as a measurement with which we can express the value of the contribution of a particular segmentation to the balancing of the dataset.

C. OPTIMIZATION

Using the mentioned functions, we define a score for each chosen segmentation as follows:

$$F(a, x) := F_1(a, x) + CF_2(a) \tag{13}$$

where C is a positive real number that plays the trade-off between the two functions.

To achieve the goal of choosing semantically appropriate segmentations that are simultaneously close to the majority voting segmentations and balance the final chosen segmentations to the best, we adopt a hill-climbing [15] algorithm as follows:

First, we introduce the general setting of our hill-climbing algorithm. As mentioned before, we aim to select one map among all the maps provided for each image as the ground truth. Therefore, the states in our hill-climbing method can be regarded as segmentation map indices $(i_1, \ldots, i_N)$ where i_k indicates that the i_k-*th* segmentation map should be chosen as the ground truth for the k-*th* image in the dataset. In this setting, assuming that there are exactly six maps available for each image, there would be 6^N states. Next, we define the neighbors of each hill-climbing state. For each state $(i_1, \ldots, i_N)$, the state $(j_1, \ldots, j_N)$ is its neighbor if and only if $i_k = j_k$ for all $1 \leq k \leq N$ except for one index. In this setting, there are exactly six maps available for each image, i.e., there are $5N$ neighbors for each state in our hill-climbing method. For conciseness, we have provided the pseudocode of our algorithm in Fig.3.

For implementing our algorithm, we first apply the majority voting algorithm on all the segmentation maps for each image in the dataset, and further, choose the map among the available maps that is closest to the corresponding majority voting result, according to the objective function in (5), and choose this set as the initial situation for our hill-climbing algorithm. In another sense, we choose a segmentation map among the available maps that maximizes the function $F_1(a, x)$ for the corresponding image x, regardless of what value of $F_2(a)$ is obtained, and choose these segmentations as the initial setting. Afterward, we calculate the values N_i for each $i \in g$ according to (8).

Next, we randomly select an image, investigate among the available segmentation maps, and identify a segmentation with the maximum score, according to (13). If the observed annotation achieved a score greater than the score of the previously chosen segmentation for the corresponding image, we replace the previous segmentation with the new one, and update the values N_i for each $i \in g$ accordingly.

We repeat this process for many iterations, and finally, terminate the hill-climbing algorithm. We present the final ground truth segmentation for each image in the dataset as the corresponding segmentation found in the terminating state of the hill-climbing method. In our implementation, we set the number of iterations in our algorithm to $5N$. Since there are six segmentation maps available in the Gleason 2019 Challenge dataset, this number of iterations gives roughly 5 chances for each image to change its segmentation map through the process. The segmentation map chosen by our DSF algorithm can be seen in the DSF column in Fig.5.

We indicate that F_1 in (13) is simply a function of the segmentation maps of a given image (according to (5)), which are available in our dataset and do not change over time. Therefore, this function provides us with a static score. However, F_2 (according to (11)) is a function of N_i for $i \in g$ and the values of these variables change through the hill-climbing process, and F_2 becomes a dynamic score. In the special case where $C = 0$, the score in (11) is removed from (13). In this case, we can refer to this method of segmentation selection as the Static Score Function (SSF) method. We shall study this case as an ablation study in the Results And Discussion section.

Algorithm 1: Pseudocode of DSF

Input: Segmentation Maps: Map_1, Map_2, ..., Map_6
Dataset Images: $X = \{x_1, ..., x_N\}$
Output: Segmentation Selections: $hill\_state$

$MajVot\_array \leftarrow [None, None, ..., None]$
for $i \in \{1, 2, ..., N\}$ **do**
  $MajVot\_array[i] \leftarrow MajVot(X[i])$
end for
Compute α_k using Eq. 7
$hill\_state \leftarrow [None, None, ..., None]$
$J\_array_k \leftarrow [None, None, ..., None]$
for $i \in \{1, 2, ..., N\}$ **do**
  $hill\_state[i] \leftarrow \arg\max_j F_1(Map_j(X[i]), X[i])$
  $J\_array_k[i] \leftarrow J_k(Map_{hill\_state[i]}(X[i]))$
end for
Compute N_k using Eq. 8
$n\_iter \leftarrow 5N$
for $iter \in \{1, 2, ..., n\_iter\}$ **do**
  $index \leftarrow random\{1, 2, ..., N\}$
  Compute Q_k using Eq. 9
  Compute w_k using Eq. 12
  $hill\_state[index] \leftarrow \arg\max_j F(Map_j(X[index]), X[index])$
  $temp_k \leftarrow J_k(Map_{hill\_state[index]}(X[index]))$
  $N_k \leftarrow N_k - J\_array_k[index] + temp_k$
  $J\_array_k[index] \leftarrow temp_k$
end for
return $hill\_state$

FIGURE 3. The pseudocode of our proposed method. The parameter α_k is calculated initially and the parameters J_k, N_k, Q_k, and w_k is updated in each iteration. In this pseudocode, all lines containing the subscript k are executed for all $k \in g = \{0, 3, 4, 5\}$. Since N_k is the sum of J_k and only one J_k changes in each iteration, instead of recalculating N_k anew, we simply add the new term, $temp_k$, and subtract the previous term from N_k for additional speedup for our algorithm.

D. FINAL DATASET PREPARATION

Fig.2(b) shows the distribution of the classes among the segmentations resulting from the implementation of majority voting on the original set of annotations. Fig.2(c) shows the distribution of the classes among the chosen segmentations at the initial situation of the hill-climbing method. The fact that the distribution of the classes in Fig.2(b) and Fig.2(c) are similar to each other, implies that there are segmentations among the six maps for each image that achieve a high $F_1(a, x)$ score, which is reassuring since we would like to start our hill-climbing algorithm from the state that is agreed by most experts. Fig.2(d) shows the distribution of the classes among the chosen annotations by the hill-climbing method. It is more balanced than the distribution of classes in all of the previous situations, i.e., Fig.2(a), Fig.2(b), and Fig.2(c). Furthermore, it can be seen that the distribution of benign and Gleason grades 3 and 4 have become closer to one another compared to the previous distributions. This indicates that

our method was successful in presenting a set of annotations as the final ground truth that is close to the ideally balanced distribution. However, it is still possible to bring our dataset closer to an ideally balanced set in the following ways:

- To further balance the dataset, we randomly select images that contain the low-frequency classes in their corresponding ground truth and replicate them until their frequency reaches the higher-frequency classes.
- Since data augmentation has proven as a powerful method for increasing the dataset size for training machine learning models by representing data in different forms [16], we also applied augmentation methods such as Random Rotation, Random Horizontal Flip, Random Vertical Flip, and Random Crop [16], [46] on the whole train dataset for further generalization of the model and to reduce overfitting.

There is also another important note about the significance of our proposed hill-climbing method on increasing the classes with lower frequency for balancing the final segmentation set. Previous works [23] propose data augmentation of the minority class as a method for balancing the dataset, whereas, in our paper, we emphasize optimizing the score function in (11) and replicating the data afterward to obtain the ideally balanced dataset.

The difference between our hill-climbing method and data augmentation methods is that in data augmentation, images with similar semantic contexts are created through the application of various transformations. In other words, data augmentation methods do not create genuinely new semantic patterns from their input. However, in our proposed method, since we generally assume that different images from the dataset contain different patterns, switching from one segmentation to another, results in adding a new pattern to the minority class and removing a pattern from another class. In the Results And Discussion section, we shall present the difference in the performance of deep learning models trained on a dataset with sufficient pattern variability and trained on a dataset with less pattern variability but with more data augmentation.

V. EXPERIMENTS

Among the convolutional networks, the PSPNet [12] has shown superior performance in the semantic segmentation of medical images. The PSPNet is mainly composed of three parts: ResNet, the pyramid pooling module, and the FCN head, each of which plays key roles in the performance of the network. Fig.4 illustrates the internal structure of the PSPNet architecture implemented in our work.

In PSPNet, the ResNet–101 extracts features from the image input through its built-in residual blocks. The long skip connections in the residual blocks help harness category information of the image input from a global point of view. Furthermore, as discussed in [36], ResNet–101 is an excellent network for extracting features of medical images and outperforms traditional machine learning methods.

The feature outputs of ResNet–101, which are weights of size 96×96, are given as input to the pyramid pooling module, inside of which are several adaptive average pooling layers [47], [48], [49], [50]. To capture global information from the ResNet features, different kernel sizes are chosen for the adaptive average pooling layers. In PSPNet, four different kernel sizes of 1×1, 2×2, 3×3 and 6×6 are applied on the ResNet–101 output features. After passing these pooling layers through convolutional layers, these features are upsampled to the original 96×96 weight sizes and concatenated together with the original ResNet–101 output features to represent features that capture both local information from the ResNet–101 outputs, and global information from the upsampling applied on the pooling layers.

Afterward, these concatenated features are given as input to the head of the FCN. The FCN classifies the image pixel by pixel from upsampling and ignores the adjacent information during the downsampling of low-resolution feature images. Since the concatenated features in the input of the FCN head appropriately exploit the local and global features of the image input, the output of the FCN head, which is the output of the PSPNet, becomes dependent on both local and global information, resulting in a good performance [29]. In this way, the weakness of the FCN in making good use of category information from a global point of view is compensated by the pyramid pooling module deployed after the ResNet–101 features.

As explained in [29], it is also possible to add another branch to the network by exploiting another FCN on the output features of the ResNet backbone, referred to as the auxiliary branch. However, in our work, we did not opt to consider the auxiliary branch since our experiments from training both versions reveal no significant improvement in either case.

Additionally, we noticed a subtle mistake made by the pathologists in the annotations. Most annotations were not accurate on the surrounding borders of the tissue, and some background pixels were mistakenly assigned a Gleason grade 3 or above. To fix this problem, we applied a background filtering method [51] on the final ground truth, where the loss of this corrected ground truth and the model output would be backpropagated for the training procedure, to ensure that the model can learn the border features appropriately.

We implemented our code using the Python [52] programming language and ran the code on the Google Colab Pro service [53]. We utilized the numpy [54] and Pytorch [46] framework for the training and evaluation of the PSPNet, and used the matplotlib [55], scikit-learn [56] and SimpleITK [27], [28] libraries for the figures of this paper.

The preprocessing procedure which includes the calculation of the initial set of annotations for the hill-climbing algorithm and the iterations afterward took twenty hours to complete. The appropriate value for the parameter C defined in (13) can be found in the following way: The static part of our score function defined in (5) can be calculated as a

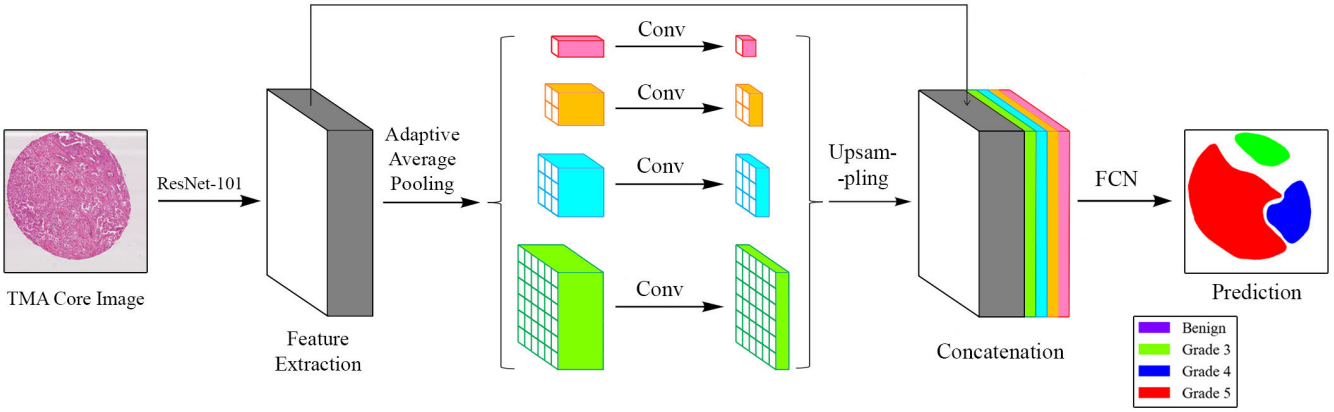

FIGURE 4. The network structure of PSPNet. The internal structure of PSPNet was first introduced and illustrated in [12].

preprocessing step and an average is taken over the whole segmentation maps. Next, the possible values that the dynamic part of our score function defined in (11) can be estimated. Afterward, the parameter C is set to a value that can dominate the dynamic part of the score function when the dataset requires major balancing. With this methodology, we used the value of $C = 5$ in our experiments.

For training the PSPNet, we used a Stochastic Gradient Descent optimizer [57]. For better convergence of model parameters, we deployed a learning rate decay [58] with the strategy

$$\eta = \eta_0 \left(1 - \frac{mB}{NE}\right)^{0.9} \tag{14}$$

where $B = 2$ is the batch size, N is the size of the training dataset, $E = 50$ is the total number of epochs for our training procedure, m is an iterator that goes from 0 to $\frac{NE}{B}$, and $\eta_0 = 0.001$.

Due to the limitations in the available RAM, we restricted our implementation to a size of 2 for the batch size and reduced the original high-resolution images to the size of 1024×1024 before the execution of data augmentation [16]. We believe that considering different batch sizes and training the network with the original resolution improves the performance results.

VI. RESULTS AND DISCUSSION

According to [45], the aggregation method applied to the expert annotations by the organizing committee of the Gleason 2019 Challenge was the STAPLE algorithm. However, the STAPLE results obtained by the organizing committee were not publicly made available [45]. The organizers provided the TMA images of the training dataset and test dataset, but the six annotations of the experts were made publicly available only for the training dataset [6]. The participating teams trained their models on the 244 TMA images from the training dataset, without having access to the annotation maps of the 87 TMA images in the test dataset, submitted their programs to the official challenge, and were ranked according to an evaluation metric which includes a combination of the Cohen's kappa [59] score and F1-Score [60] calculated on the test dataset [6]. Furthermore, the confusion matrices of the top participating teams were publicly made available on the official website of the challenge [6].

Since the annotations of the experts in the test dataset were unavailable, we divided our training dataset into a new training dataset of size 200, and a test dataset of size 44. In the separation of the original training dataset from the latter datasets, we carefully maintained the relative distribution of classes between the two datasets, to ensure all patterns of a particular class would not completely fall in the training dataset or the test dataset. To make a fair comparison between our proposed method and some of the participating teams in the challenge, we trained and evaluated the models of the 1'st and 4'th ranked teams on the same 200 and 44 partitions we introduced.

The results of our performance and two of the other teams in the challenge are shown in Table 1, Table 2, Table 3, and Table 4. We refer to Table 1 and Table 2 to demonstrate the superiority of our method in the task of identifying benign and malignant tumors. Table 3 and Table 4 show the superiority of our method in the task of distinguishing between the three distinct malignant classes 3, 4, and 5. Since the organizing committee evaluated the output of models to the STAPLE algorithm applied on the whole dataset for providing the confusion matrices, we provide the performance of our model evaluated both on our proposed DSF method and the STAPLE method in all the tables even though we mentioned the shortcomings of choosing STAPLE algorithm as an aggregation method.

Moreover, we provide an ablation study to emphasize the importance of our DSF method in all of the tables. To reveal the importance of including (11) in the final score of (13), we trained and evaluated our model with the same setting but only changed the parameter C by choosing $C = 0$. In this case, the frequency of Gleason grade 5 has not been increased by (11). Therefore, we solely relied on using data augmentation and resampling methods to the extent that the frequency of Grade grade 5 reaches the

TABLE 1. The performance of different methods on the task of Benign vs. Malignant tumor detection on the Gleason 2019 Challenge dataset. We compared the model outputs using the method applied to the expert segmentation maps written in parentheses next to the titles.

Benign vs. Malignant Performance					
Method	Accuracy	Precision	Recall	F1-Score	Cohen's Kappa
DSF (DSF)	0.9519	0.9481	0.9502	0.9491	0.8982
DSF (STAPLE)	0.9390	0.9425	0.9313	0.9360	0.8721
SSF (STAPLE)	0.9319	0.9354	0.9235	0.9285	0.8570
SSF w.o. any balancing (STAPLE)	0.9335	0.9364	0.9256	0.9302	0.8605
1'st Ranked Team (STAPLE)	0.8822	0.8841	0.8982	0.8814	0.7653
4'th Ranked Team (STAPLE)	0.6534	0.7399	0.5779	0.5323	0.1781

TABLE 2. This table shows the performance of different methods on the task of Benign vs. Malignant tumor detection on an ideally balanced dataset. We compared the model outputs using the method applied to the expert segmentation maps written in parentheses next to the titles.

Benign vs. Malignant Performance				
Method	Accuracy	Precision	F1-Score	Cohen's Kappa
DSF (DSF)	0.9479	0.9480	0.9479	0.8957
DSF (STAPLE)	0.9232	0.9271	0.9230	0.8464
SSF (STAPLE)	0.9033	0.9100	0.9029	0.8066
SSF w.o. any balancing (STAPLE)	0.9053	0.9116	0.9050	0.8107
1'st Ranked Team (STAPLE)	0.8971	0.9086	0.8964	0.7942
4'th Ranked Team (STAPLE)	0.5761	0.7125	0.4951	0.1523

TABLE 3. The performance of different methods on the detection of Gleason grades 3, 4, and 5 on the Gleason 2019 Challenge dataset. We compared the model outputs using the method applied to the expert segmentation maps written in parentheses next to the titles.

Grade 3, 4, and 5 Performance					
Method	Accuracy	Precision	Recall	F1-Score	Cohen's Kappa
DSF (DSF)	0.9697	0.9662	0.9695	0.9678	0.9471
DSF (STAPLE)	0.8754	0.7090	0.9180	0.7451	0.7692
SSF (STAPLE)	0.8691	0.6867	0.5962	0.5851	0.7363
SSF w.o. any balancing (STAPLE)	0.8587	0.6802	0.5941	0.5786	0.7194
1'st Ranked Team (STAPLE)	0.6340	0.5384	0.3838	0.3636	0.1603
4'th Ranked Team (STAPLE)	0.6102	0.5107	0.3600	0.3237	0.0869

TABLE 4. This table shows the performance of different methods for the detection of Gleason grades 3, 4, and 5 on an ideally balanced dataset. We compared the model outputs using the method applied to the expert segmentation maps written in parentheses next to the titles.

Grade 3, 4, and 5 Performance				
Method	Accuracy	Precision	F1-Score	Cohen's Kappa
DSF (DSF)	0.9693	0.9699	0.9695	0.9540
DSF (STAPLE)	0.9180	0.9185	0.9167	0.8770
SSF (STAPLE)	0.5960	0.5528	0.4962	0.3940
SSF w.o. any balancing (STAPLE)	0.5940	0.5423	0.4922	0.3910
1'st Ranked Team (STAPLE)	0.3840	0.4723	0.2848	0.0760
4'th Ranked Team (STAPLE)	0.3600	0.4046	0.2432	0.0400

higher frequency classes, as discussed in the Final Dataset Preparation subsection of our proposed method. This ablated version of the proposed method is called *SSF*. As it can be seen in the tables, the ablation study does not reveal a significant decrease in the metrics of benign vs malignant diagnosis. However, it has shown a significant decrease in

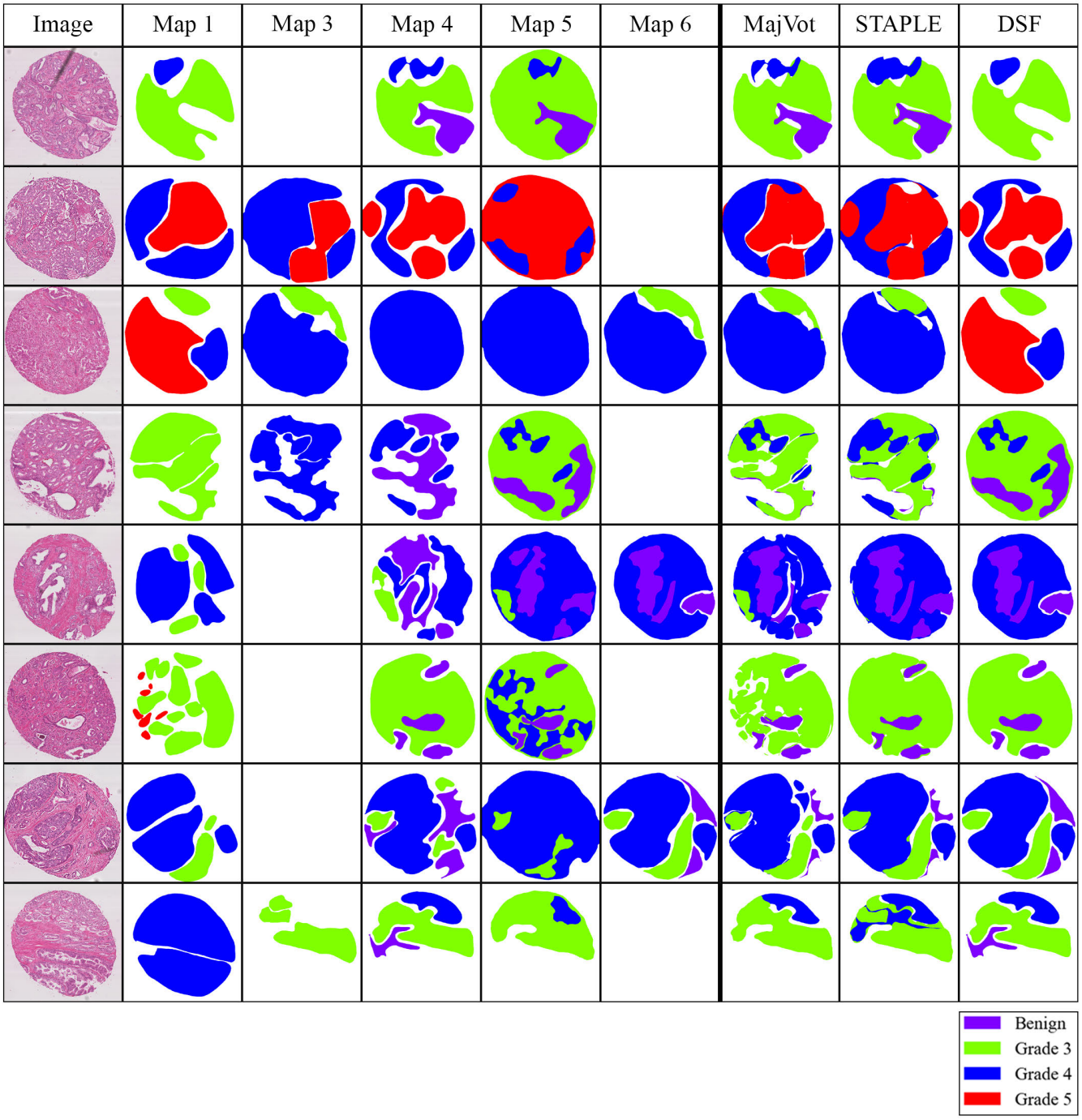

FIGURE 5. Several images from the Gleason 2019 Challenge dataset and their corresponding segmentation maps. The MajVot column refers to the output of the majority voting algorithm when applied to the segmentation maps. The STAPLE column refers to the output of the STAPLE algorithm when applied to the segmentation maps. The DSF column refers to the segmentation map chosen by our proposed method. In this figure, unavailable maps or fallacious maps were left empty. Since none of these images had an acceptable segmentation from the second map, column Map2 has been omitted. According to our proposed method, it is reasonable that in the third row, Map1 is chosen due to containing a decent amount of Gleason grade 5 and the other rows indicate selections similar to the MajVot and STAPLE columns. Moreover, the odd-shaped curves and regions mentioned in the paper are clear in the MajVot and STAPLE column entries [6], [7], and [8].

Table 3 and Table 4, especially the recall metric and the accuracy metric in the balanced case. This is in agreement with our previous claim that data augmentation alone is not enough to improve a model's performance by increasing the samples of the low-frequency classes when the dataset lacks variability.

Additionally, we have implemented our ablation study in the case where no data augmentation and replication methods

are used as well. The result of this study is entitled *SSF w.o. any balancing* in the tables. As it can be seen from the tables, the evaluation metrics of the ablated versions are close to each other. This finding reveals that when the pattern variety in a particular class falls below some threshold, complex models tend to focus on accurately predicting the other classes and forget the class with the least variability, regardless of whether data augmentation and replication methods have been used or not. Consequently, maintaining a relation similar to (10) is mandatory for good performance.

Since we trained our PSPNet similarly to the 1'st ranked team, by comparing the results of our proposed method and the 1'st ranked team's model in the tables, we validate our claim that segmentation models cannot learn properly when their ground truth segmentations do not maintain semantic properties, and that choosing semantically correct ground truth segmentations contribute a lot to the model performance. In another sense, the first and the second rows in the tables reveal that training models based on our DSF method but evaluating them based on other methods still greatly improves the performance. Additionally, by observing some of the confusion matrices in the challenge's website, we observe that some of the methods had difficulty correctly classifying Gleason grade 5, whereas our proposed method can achieve good accuracy in all the classes.

As a final note, in obtaining the results of Table 1 and Table 2, the Gleason grade 0 is considered as the benign class, and the grades 3, 4, and 5 as malignant. Moreover, in obtaining the results of Table 2 and Table 4, the balanced dataset was achieved by normalizing the number of input pixels in each class.

VII. CONCLUSION

In this paper, we thoroughly explained the main advantages and shortcomings of the majority voting and the STAPLE algorithm on the Gleason 2019 Challenge dataset and highlighted the importance of semantic properties of ground truths on the performance of segmentation models. We argued that using the segmentations from the original segmentation maps is a good choice for ensuring that the semantic properties are maintained. To deal with the high dissimilarity among the segmentations, we proposed a dynamic score function for each segmentation that highlights the importance of similarity to the majority voting result and balancing the dataset. To settle with a decent set of segmentations, we ran the hill-climbing algorithm for many iterations.

In future work, some extensions can be made to our method for other machine learning problems as well. We highlight that the score in (5) emphasizes choosing semantically correct annotations which are close to the majority voting annotation, and the score in (11) emphasizes balancing the annotations. We anticipate that our SSF method performs well on ideally balanced datasets which do not require the balancing score function in (11). We further believe that the scores defined in our paper can be generalized to advanced versions that perform well on a broad set of datasets.

ACKNOWLEDGMENT

The authors acknowledge Dr. Mokhtari for their medical opinions and suggestions for this article. They further express their thanks to Sadeghi and Ahmadi from Chakavak Yekta (Dade Chi) Company for providing them with a Google Colab Pro account for implementing their code.

REFERENCES

[1] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Lake Tahoe, NV, USA, Mar. 2018, pp. 1451–1460, doi: 10.1109/WACV.2018.00163.

[2] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017, doi: 10.1109/TPAMI.2016.2572683.

[3] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 3213–3223, doi: 10.1109/CVPR.2016.350.

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010, doi: 10.1007/s11263-009-0275-4.

[5] J. Fritsch, T. Kuhnl, and A. Geiger, "A new performance measure and evaluation benchmark for road detection algorithms," in *Proc. 16th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, The Hague, The Netherlands, Oct. 2013, pp. 1693–1700, doi: 10.1109/ITSC.2013.6728473.

[6] *The Gleason 2019 Challenge*. Accessed: May 15, 2023. [Online]. Available: https://gleason2019.grand-challenge.org/.

[7] D. Karimi, G. Nir, L. Fazli, P. C. Black, L. Goldenberg, and S. E. Salcudean, "Deep learning-based Gleason grading of prostate cancer from histopathology images—Role of multiscale decision aggregation and data augmentation," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 5, pp. 1413–1426, May 2020, doi: 10.1109/JBHI.2019.2944643.

[8] G. Nir, S. Hor, D. Karimi, L. Fazli, B. F. Skinnider, P. Tavassoli, D. Turbin, C. F. Villamil, G. Wang, R. S. Wilson, K. A. Iczkowski, M. S. Lucia, P. C. Black, P. Abolmaesumi, S. L. Goldenberg, and S. E. Salcudean, "Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts," *Med. Image Anal.*, vol. 50, pp. 167–180, Dec. 2018, doi: 10.1016/j.media.2018.09.005.

[9] J. I. Epstein, L. Egevad, M. B. Amin, B. Delahunt, J. R. Srigley, and P. A. Humphrey, "The 2014 international society of urological pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma: Definition of grading patterns and proposal for a new grading system," *Amer. J. Surgical Pathol.*, vol. 40, no. 2, pp. 244–252, Feb. 2016, doi: 10.1097/PAS.0000000000000530.

[10] L. Lam and S. Y. Suen, "Application of majority voting to pattern recognition: An analysis of its behavior and performance," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 27, no. 5, pp. 553–568, Sep. 1997, doi: 10.1109/3468.618255.

[11] P. Rawla, "Epidemiology of prostate cancer," *World J. Oncol.*, vol. 10, no. 2, pp. 63–89, 2019, doi: 10.14740/wjon1191.

[12] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2881–2890, doi: 10.1109/CVPR.2017.660.

[13] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Trans. Med. Imag.*, vol. 23, no. 7, pp. 903–921, Jul. 2004, doi: 10.1109/TMI.2004.828354.

[14] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Process. Mag.*, vol. 13, no. 6, pp. 47–60, Nov. 1996, doi: 10.1109/79.543975.

[15] L. Hernando, A. Mendiburu, and J. A. Lozano, "Hill-climbing algorithm: Let's go for a walk before finding the optimum," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Rio de Janeiro, Brazil, Jul. 2018, pp. 1–7, doi: 10.1109/CEC.2018.8477836.

[16] A. Mikolajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," in *Proc. Int. Interdiscipl. PhD Workshop (IIPhDW)*, Winóu, Poland, May 2018, pp. 117–122, doi: 10.1109/IIPHDW.2018.8388338.

[17] B. Xie, Sh. Li, M. Li, Ch. H. Liu, G. Huang, and G. Wang, "SePiCo: Semantic-guided pixel contrast for domain adaptive semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jan. 17, 2023, doi: 10.1109/TPAMI.2023.3237740.

[18] Y. Ban, Y. Wang, S. Liu, B. Yang, M. Liu, L. Yin, and W. Zheng, "2D/3D multimode medical image alignment based on spatial histograms," *Appl. Sci.*, vol. 12, no. 16, p. 8261, Aug. 2022, doi: 10.3390/app12168261.

[19] W. Bulten et al., "Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists," *Modern Pathol.*, vol. 34, no. 3, pp. 660–671, Mar. 2021, doi: 10.1038/s41379-020-0640-y.

[20] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[21] Y. Wang, H. Dou, X. Hu, L. Zhu, X. Yang, M. Xu, J. Qin, P. Heng, T. Wang, and D. Ni, "Deep attentive features for prostate segmentation in 3D transrectal ultrasound," *IEEE Trans. Med. Imag.*, vol. 38, no. 12, pp. 2768–2778, Dec. 2019, doi: 10.1109/TMI.2019.2913184.

[22] G. Litjens, C. I. Sánchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, C. H. van de Kaa, P. Bult, B. van Ginneken, and J. van der Laak, "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," *Sci. Rep.*, vol. 6, no. 1, pp. 1–11, May 2016, doi: 10.1038/srep26286.

[23] A. A. Khani, S. A. F. Jahromi, H. O. Shahreza, H. Behroozi, and M. S. Baghshah, "Towards automatic prostate gleason grading via deep convolutional neural networks," presented at the 5th Iranian Conf. Signal Process. Intell. Syst. (ICSPIS), Shahrood, Iran, Dec. 2019, doi: 10.1109/ICSPIS48872.2019.9066019.

[24] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," Dec. 2019, *arXiv:1706.05587*, doi: 10.48550/arXiv.1706.05587.

[25] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. 15th Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 833–851, doi: 10.1007/978-3-030-01234-2_49.

[26] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 4510–4520, doi: 10.1109/CVPR.2018.00474.

[27] Z. Yaniv, B. C. Lowekamp, H. J. Johnson, and R. Beare, "SimpleITK image-analysis notebooks: A collaborative environment for education and reproducible research," *J. Digit. Imag.*, vol. 31, no. 3, pp. 290–303, Jun. 2018, doi: 10.1007/s10278-017-0037-8.

[28] B. C. Lowekamp, D. T. Chen, L. Ibáñez, and D. Blezek, "The design of SimpleITK," *Frontiers Neuroinform.*, vol. 7, Dec. 2013, Art. no. 45, doi: 10.3389/fninf.2013.00045.

[29] Y. Qiu, Y. Hu, P. Kong, H. Xie, X. Zhang, J. Cao, T. Wang, and B. Lei, "Automatic prostate Gleason grading using pyramid semantic parsing network in digital histopathology," *Frontiers Oncol.*, vol. 12, Apr. 2022, doi: 10.3389/fonc.2022.772403.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[32] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 3431–3440, doi: 10.1109/CVPR.2015.7298965.

[33] Y.-h. Zhang, J. Zhang, Y. Song, C. Shen, and G. Yang, "Gleason score prediction using deep learning in tissue microarray image," May 2020, *arXiv:2005.04886*, doi: 10.48550/arXiv.2005.04886.

[34] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Munich, Germany, Oct. 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[35] W. Bulten, H. Pinckaers, H. van Boven, R. Vink, T. de Bel, B. van Ginneken, J. van der Laak, C. H.-van. de Kaa, and G. Litjens, "Automated Gleason grading of prostate biopsies using deep learning," *The Lancet. Oncol.*, vol. 21, no. 2, pp. 233–241, Jan. 2020, doi: 10.1016/S1470-2045(19)30739-9.

[36] S. Iqbal, G. F. Siddiqui, A. Rehman, L. Hussain, T. Saba, U. Tariq, and A. A. Abbasi, "Prostate cancer detection using deep learning and traditional techniques," *IEEE Access*, vol. 9, pp. 27085–27100, Feb. 2021, doi: 10.1109/ACCESS.2021.3057654.

[37] A. Sohail and F. Arif, "Supervised and unsupervised algorithms for bioinformatics and data science," *Prog. Biophys. Mol. Biol.*, vol. 151, pp. 14–22, Mar. 2020, doi: 10.1016/j.pbiomolbio.2019.11.012.

[38] S. R. Mounce, K. Ellis, J. M. Edwards, V. L. Speight, N. Jakomis, and J. B. Boxall, "Ensemble decision tree models using RUSBoost for estimating risk of iron failure in drinking water distribution systems," *Water Resour. Manage.*, vol. 31, no. 5, pp. 1575–1589, Mar. 2017, doi: 10.1007/s11269-017-1595-8.

[39] R. Al-Khurayji and A. Sameh, "An effective Arabic text classification approach based on kernel naive Bayes classifier," *Int. J. Artif. Intell. Appl.*, vol. 8, no. 6, pp. 1–10, Nov. 2017, doi: 10.5121/ijaia.2017.8601.

[40] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, Dec. 2007, doi: 10.1007/s10115-007-0114-2.

[41] A. Graves, "Long short-term memory," in *Supervised Sequence Labelling With Recurrent Neural Networks*, vol. 385, Berlin, Germany: Springer, 2012, pp. 37–45, doi: 10.1007/978-3-642-24797-2_4.

[42] *International Skin Imaging Collaboration: Melanoma Project*. Accessed: May 15, 2023. [Online]. Available: https://isic-archive.com

[43] V. Ribeiro, S. Avila, and E. Valle, "Handling inter-annotator agreement for automated skin lesion segmentation," Jun. 2019, *arXiv:1906.02415*, doi: 10.48550/arXiv.1906.02415.

[44] *ISIC 2018 Challenge*. Accessed: May 15, 2023. [Online]. Available: https://challenge.isic-archive.com/data/#2018

[45] A. Foucart, O. Debeir, and C. Decaestecker, "Processing multi-expert annotations in digital pathology: A study of the Gleason 2019 challenge," presented at the 17th Int. Symp. Med. Inf. Process. Anal., Campinas, Brazil, Dec. 2021, doi: 10.1117/12.2604307.

[46] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Vancouver, BC, Canada, Dec. 2019, pp. 8026–8037, Art. no. 721.

[47] C.-Y. Lee, P. Gallagher, and Z. Tu, "Generalizing pooling functions in CNNs: Mixed, gated, and tree," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 863–875, Apr. 2018, doi: 10.1109/TPAMI.2017.2703082.

[48] B. Betro and A. Guglielmi, "Methods for global prior robustness under generalized moment conditions," in *Robust Bayesian Analysis*, vol. 152. New York, NY, USA: Springer, 2000, pp. 273–293, doi: 10.1007/978-1-4612-1306-2_15.

[49] N. G. Polson and J. G. Scott, "On the half-cauchy prior for a global scale parameter," *Bayesian Anal.*, vol. 7, no. 4, pp. 887–902, Dec. 2012, doi: 10.1214/12-BA730.

[50] B. Yu, L. Yang, and F. Chen, "Semantic segmentation for high spatial resolution remote sensing images based on convolution neural network and pyramid pooling module," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 9, pp. 3252–3261, Sep. 2018, doi: 10.1109/JSTARS.2018.2860989.

[51] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979, doi: 10.1109/TSMC.1979.4310076.

[52] G. van Rossum, "Python tutorial," Centrum voor Wiskunde en Informatica (CWI), Amsterdam, The Netherlands, Tech. Rep. CS-R9526, May 1995. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8591001/#CR16

[53] *Google Colaboratory Workspace*. Accessed: May 15, 2023. [Online]. Available: https://workspace.google.com/marketplace/app/colaboratory/1014160490159

[54] C. R. Harris et al., "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020, doi: 10.1038/s41586-020-2649-2.

[55] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, 2007, doi: 10.1109/MCSE.2007.55.

[56] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, and D. Cournapeau, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Research.*, vol. 12, no. 85, pp. 2825–2830, Oct. 2011.

[57] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, no. 3, pp. 400–407, Sep. 1951, doi: 10.1214/aoms/1177729586.
[58] K. You, M. Long, J. Wang, and M. I. Jordan, "How does learning rate decay help modern neural networks?" Sep. 2019, *arXiv:1908.01878*, doi: 10.48550/arXiv.1908.01878.
[59] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit," *Psychol. Bull.*, vol. 70, no. 4, pp. 213–220, 1968, doi: 10.1037/h0026256.
[60] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: An overview," Aug. 2020, *arXiv:2008.05756*, doi: 10.48550/arXiv.2008.05756.

POOYA ESMAEIL AKHOONDI was born in Tehran, Iran, in 2002. He is currently pursuing the B.Sc. degree in computer engineering with the Sharif University of Technology, Tehran. His research interests include computer vision, deep learning, image and video processing, and machine learning. His awards and honors include the Silver Medal in the Iranian National Mathematical Olympiad, in 2018, the Gold Medal in the Iranian National Mathematical Olympiad, in 2019, the Gold Medal in the Iranian Combinatorics Olympiad, in 2020, and a number of medals at international mathematics Olympiads.

MAHDIEH SOLEYMANI BAGHSHAH received the B.S., M.S., and Ph.D. degrees from the Department of Computer Engineering, Sharif University of Technology, Iran, in 2003, 2005, and 2010, respectively. Her M.S. and Ph.D. theses were in the field of machine learning. She joined the Sharif University of Technology as an Assistant Professor of computer engineering, where she has founded the Machine Learning Laboratory (MLL), in 2012. She is currently an Associate Professor with the Sharif University of Technology and the Director of MLL. Her research interests include machine learning and deep learning.