National Taiwan Ocean University

Department of Computer Science & Engineering

A Novel Ensemble Learning Method for Medical Image Segmentation

Jiunn wei Chiang 姜竣維 **00957056@mail.ntou.edu.tw**
Bao Wei Chiam 詹堡崴 **00957058@mail.ntou.edu.tw**
Dwayne Reinaldy 許漢強 **00957059@mail.ntou.edu.tw**
Vincensius Hobart Wijaya 何天勝 **00957060@mail.ntou.edu.tw**
Hanson Gabriel Cendana 曾漢盛 **00957062@mail.ntou.edu.tw**

Advisor： Dr. Kuan Y. Chang 張光遠

2023 / 09 / 01

Table of Contents

Abstract

Medical image segmentation plays a big and vital role especially in diagnosing disease accurately and for treatment planning. High performance medical image segmentation is vastly employing ensemble learning techniques. Ensemble learning combines multiple models or predictions, allowing for enhanced accuracy and boost performance.

This study aims to compare deep learning-based approaches and ensemble learning-based approaches for the segmentation of brain tumors in MRI scans, segmentation of polyp images, and Pulmonary Chest X-Ray Abnormalities. In this study, we use multiple deep learning-based architectures, namely SegNet, U-Net, and DeepLab.

In this paper, we proposed three different ensemble learning technique for analyzing the performance impact of ensemble learning, which is bootstrap stacking which combine bootstrapping technique with stacking, bootstrap meta-model stacking classifier which use stacking classifier method to combine prediction, and bootstrap multiple meta-model stacking which use multiple layer of meta-model to combine prediction.

Overall, our best ensemble learning technique gives a performance boost of +23.26% in IoU score, +6.31% in F1-score, +6.73% in Precision, +14.59% in Recall, +2.34% in Accuracy compared to the base model.

Introduction

Medical image segmentation serves a critical purpose of enhancing the visibility of anatomical structures within images, which enables more precise analysis and diagnosis [1]. By partitioning the image into different regions, it allows for extraction of relevant features and improves diagnostic accuracy. With the advancements in imaging technologies, like magnetic resonance imaging (MRI) scans and computed tomography (CT) scans, medical image segmentation has garnered significant interest and attention from researchers and practitioners alike in recent years.

The goal of medical image segmentation is to accurately describe the boundaries and contours of organs, tissues, or other structures of interest within medical images. The purpose of this process is for extracting quantitative information, volumetric measurements, and spatial relationships, which are vital for clinical decision-making and precise intervention.

A wide range of segmentation techniques have been developed and employed in the field of medical imaging. Traditional methods include thresholding, region-based methods, clustering, and edge detection [2,3]. Nevertheless these approaches often struggle with the complexity and variability of medical images and it leads to non optimal results.

Medical image segmentation has been revolutionized in recent years by the emergence of deep learning-based techniques, with convolutional neural networks (CNNs) playing a pivotal role in this transformation. CNNs have shown exceptional performance in various segmentation tasks by

automatically learning hierarchical representations and capturing contextual information from medical images. Models such as U-Net [4], SegNet [5], DeepLab [6], TransUNet [7], U-Net++ [8], and V-net [9] have been extensively used and adapted for medical image segmentation. U-Net [4], SegNet [5], DeepLab [6] and U-Net++ [8] are all deep CNN models for semantic segmentation, featuring an encoder (compression) path to capture context and a decoder (expansion) path for precise localization. From the encoder to the decoder, U-Net transfers the entire feature maps, while SegNet transfers only the pooling indices. U-Net uses skip connections to fuse low-level and high-level features, and up-convolutions to upsample the feature maps; DeepLab uses atrous convolutions to enlarge the receptive field of the feature maps, and atrous spatial pyramid pooling to capture multi-scale information. U-Net++ is an improved version of U-Net, which uses nested and dense skip connections to enhance the feature fusion between the encoder and the decoder. Instead TransUNet is a transformer-based neural network that combines the strengths of both transformers and U-Net [7]. V-Net is a 3D CNN designed for volumetric image segmentation, known for its effectiveness in tasks like organ and tumor segmentation in medical imaging [9].

Ensemble learning is a machine learning technique that integrates the predictions of multiple models to improve a single model by reducing variance and exploiting the advantages of each model [10]. Some popular ensemble learning methods are Bagging and Stacking. Bagging trains multiple models independently on random subsets of data and then combines their predictions through a majority vote [11]. Stacking trains multiple models (usually of different types) on the same data and then uses another model, which is called a meta-learner, to learn how to best combine their predictions [12].

In this study, we explored whether integrating ensemble learning techniques could substantially improve the performance of medical image segmentation models. To choose our base model, we carefully reviewed the existing literature, focusing on benchmark results such as those found in the paperswithcode Kvasir-SEG Benchmark [13]. We did not select models like U-Net++, TransUnet, or DeepLabv3 because these models consistently demonstrated superior segmentation capabilities compared to the model we ultimately chose: SegNet, U-Net, and DeepLab. Although SegNet, U-Net, and DeepLab are not considered state-of-the-art in medical image segmentation, they can still provide reasonable segmentation results. Their simplicity in architecture and computational efficiency make them suitable candidates for ensemble learning. Furthermore, their relative weaknesses compared to models like U-Net++, TransUnet, or DeepLabv3 align with the principle of leveraging diversity in ensemble methods, as their biases and error patterns differ from those of stronger learners. This diversity can enhance the ensemble's ability to generalize well to a wide range of medical images and potentially improve overall segmentation performance.

Here we proposed three techniques called Bootstrap Stacking, Bootstrap Multiple Meta-Model Stacking, and Bootstrap Stacking with Stacking Classifier. These techniques were evaluated on three distinct datasets: LGG MRI Segmentation, Kvasir-SEG, and Pulmonary Chest X-Ray Abnormalities.

Motivation & Purpose

The main challenge in medical image segmentation is the varying performance of models across different datasets. This variability is often due to differences in data characteristics and model architecture. Precise prediction is crucial in this field, making it important to understand and address this variability.

Our study is motivated by the critical need to confront the inherent variability in model predictions. This variability is not merely an inconvenience, but a barrier to achieving consistent and reliable results across diverse datasets, particularly when models are constructed using different architectures or weak learners. The subtle complexities within these models can exert a profound influence on their performance. We aim to unravel this issue and determine whether ensemble learning techniques can emerge as a robust remedy. Ensembles have the capability to combine multiple models and harness the collective wisdom of diverse architectures or weak learners, holding the potential to mitigate the nuanced peculiarities that compromise prediction stability.

Related Works

DivergentNets

DivergentNets is a medical image segmentation technique that uses an ensemble of multiple high-performing image segmentation architectures [14]. The model combines the TriUNet segmentation model with an ensemble of the well-known segmentation models, namely UNet++, FPN, DeepLabv3, and DeepLabv3+ [14]. The TriUNet model takes a single image as input, which is then passed into two distinct U-Net models [14]. The outputs of these two models are combined and then processed through a third U-Net model to generate the final segmentation masks [14].

In the EndoCV2021 challenge, the TriUNet architecture used in DivergentNets was the winning model in terms of segmentation accuracy and generalization [14]. In contrast to the original TriUNet method, DivergentNets employs an ensemble of five intermediate models (TriUnet, U-Net++, FPN, DeepLabv3, and DeepLabv3+), which are trained separately and then combined by averaging the pixels between each mask[14] .

In comparison, DivergentNets employs five different medical image segmentation models, while MingleNets uses three (SegNet, U-Net, DeepLab). DivergentNets combines the predictions of these models using an averaging method, whereas MingleNets aims to employ a meta-model that is distinct from averaging to achieve improved results. The meta-models used in our technique are more complex compared to simple averaging techniques.

Materials and Methods

Datasets

1. **Low-Grade Glioma (LGG) MRI Segmentation**

   LGG MRI Segmentation dataset obtained from The Cancer Imaging Archive (TCIA) [15] contains 3929 brain MR images of 256 x 256 pixels with manual Fluid-Attenuated Inversion Recovery (FLAIR) abnormality segmentation masks [15]. We split the dataset into training, validation and test sets by a ratio of 8:1:1, resulting in 3143, 393 and 393 images respectively. The total infected tumor case is 1373 and non-infected tumor case is 2556.

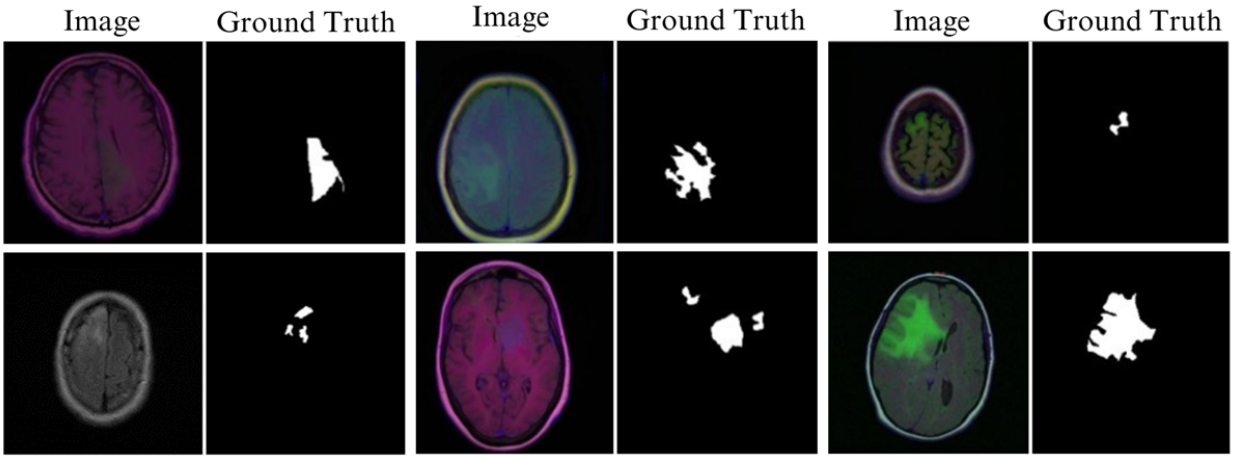

Fig. 1. Example of Images and Masks from LGG MRI Segmentation.
Showcasing six distinct images paired with their corresponding masks positioned on the right side of each image.

2. **Kvasir-SEG**

   Kvasir-SEG v2 dataset, which consists of 1000 gastrointestinal segmented polyp images and corresponding segmentation masks, were manually annotated and verified by medical experts [16]. The image resolution ranges from 332 x 487 to 1920 x 1072 pixels. We split the dataset into training, validation and test sets by a ratio of 8:1:1, resulting in 800, 100 and 100 images respectively. We resized and masked every image into 256 x 256 pixels.

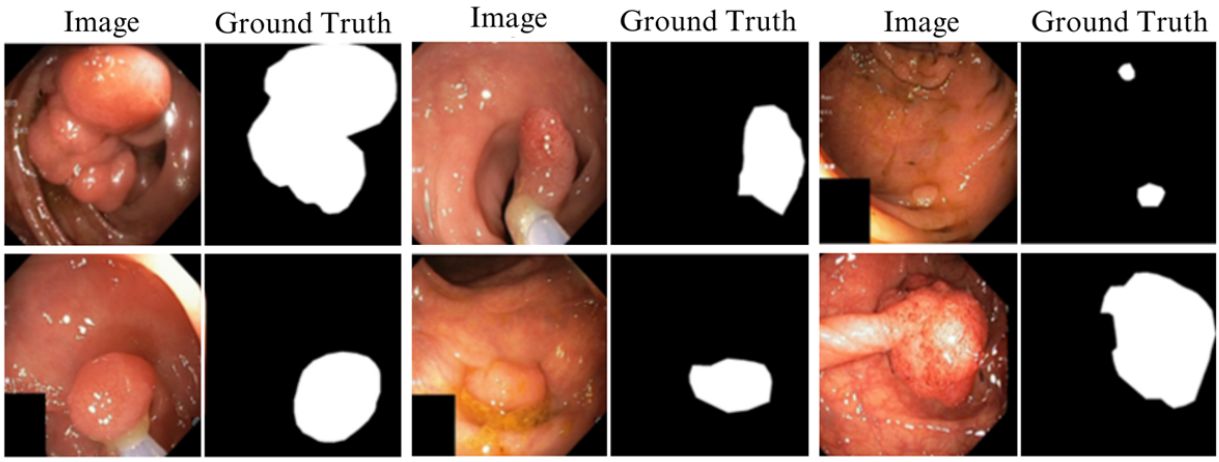

Fig. 2. Example of Images and Masks from Kvasir-SEG.
Showcasing six distinct images paired with their corresponding masks positioned on the right side of each image.

3. Pulmonary Chest X-Ray Abnormalities

Pulmonary Chest X-Ray Abnormalities dataset, which contains 662 chest X-ray images and clinical labels for Tuberculosis (TB) diagnosis, was collected from two sources [17]. The first source is the China Shenzhen set, with 336 images showing TB signs. The second source is the Montgomery County set, with 80 normal images and 58 images with TB signs. Every image and mask has a resolution of 4020 x 4892 pixels. We split the dataset into training, validation and test sets by a ratio of 8:1:1, resulting in 530, 66 and 66 images respectively. We resized every image into 256 x 256 pixels.

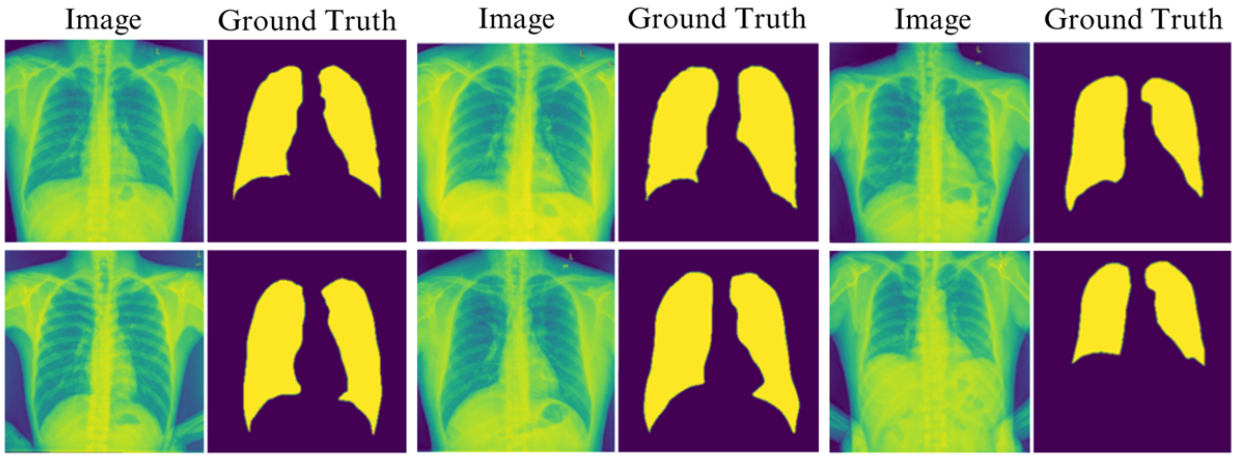

Fig. 3. Example of Images and Masks from Pulmonary Chest X-Ray Abnormalities.
Showcasing six distinct images paired with their corresponding masks positioned on the right side of each image.

Deep Learning Models for Image Segmentation

1. SegNet

SegNet is a deep learning architecture specifically designed for semantic segmentation tasks in computer vision.[5] It utilizes a fully convolutional neural network (CNN) with an encoder-decoder structure. In the encoder stage, it progressively downsampled the input image to capture hierarchical feature representations [5]. The decoder stage then performs upsampling and pixel-wise classification to produce a segmentation mask, assigning each pixel in the input image to a specific class label.[5] SegNet is known for its efficiency and ability to achieve accurate pixel-level segmentation, making it a valuable tool for tasks such as object detection and scene understanding in applications like autonomous driving and medical image analysis [5].

2. U-Net

The U-Net architecture developed in 2015 has become a widely adopted deep-learning model for image segmentation [4]. U-Net architecture has a U-shaped network structure, which consists of a contracting path and an expanding path. The contracting path in U-Net serves as a feature extractor, gradually decreasing spatial dimensions while capturing significant contextual information, the expanding path performs upsampling of the extracted features to match the original image size, facilitating precise localization [4]. U-Net can handle limited training data and produce detailed and accurate segmentation masks efficiently [4]. It is a successful model for various biomedical applications, such as cell segmentation [18], tumor detection [19], and organ segmentation [20].

3. DeepLab

DeepLab was made to overcome challenges in Deep Convolutional Neural Network (DCNN), these challenges are reduced feature resolution, the existence of objects at multiple scales, and reduced localization accuracy[6]. Reduced feature resolution is caused by a repeated combination of max-pooling and downsampling that resulted in reduced spatial resolution. To fix this problem, we can use *atrous convolution*, which is to sparsely sample input feature maps by inserting zeros in filters which resulted in denser feature maps[6]. The solution to fix the issue of the existence of objects at multiple scales is Atrous Spatial Pyramid Pooling (ASPP), which is to use multiple parallel atrous convolutional layers with different sampling rates[6]. ASPP can help to detect objects at different scales and can improve accuracy.
The solution to fix the issue of reduced localization accuracy is to use fully-connected Conditional Random Field (CRF), which is to help refine the coarse outputs. CRF uses Gaussian Convolutions and the model is efficient and also can significantly speed up the computation[21]. In CRF, every pair of nodes in the graph is connected by an edge and the weights on these edges are learned from data[21].

Concept of Ensemble Learning

Ensemble learning is a powerful machine learning approach that involves combining multiple individual models to achieve more accurate predictions or classifications than what a single model can achieve on its own. The fundamental idea behind ensemble learning is to leverage the diversity of these models to overcome the limitations and biases inherent in individual models, thereby enhancing overall performance and generalization capabilities[22].

1. **Bagging**

Bagging improves model predictions by following three steps: Bootstrapping, training, and aggregation. It randomly draws and repeats data points from the training dataset to create diverse samples. These samples are trained separately and simultaneously using weak learners. The diagram below shows how the dataset was split into three subsets, and different models were used for each subset. Finally, the predictions are averaged or majority-voted for regression or classification tasks, resulting in a more accurate and robust estimate [23].

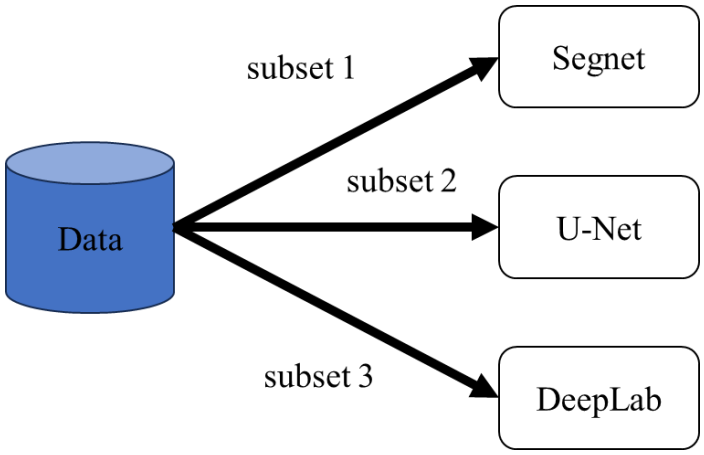

Fig. 4. Structure diagram of ensemble learning by bagging.

2. **Stacking**

Stacking is an ensemble learning technique that improves predictive performance by combining multiple models. It uses their predictions as inputs to a meta-model, which makes the final prediction. The meta-model compensates for the weaknesses of individual base models, leading to better performance than using a single base model [24].

Meta-Model

Meta-Model is a machine learning algorithm that learns from and combines the output of other learners. A Meta-Model is a meta-learner that improves its own performance by using other learners. Meta-Model can increase the accuracy, reduce the data, and make the model more robust and efficient. The machine learning classification models such as Gaussian Naïve Bayes, Decision Tree, Random Forest Classifier, and Logistic Regression can be used as the meta-model [25].

Our Ensemble Learning Approaches

1. **Bootstrap Stacking**

   Bootstrap Stacking merges the principles of bagging and stacking, by employing the bootstrapping technique from bagging and subsequently consolidating predictions through stacking[26]. Here, the approach involves combining predictions from individual models using a

meta-model. The meta-model is trained and subsequently utilized to make predictions on the test dataset.

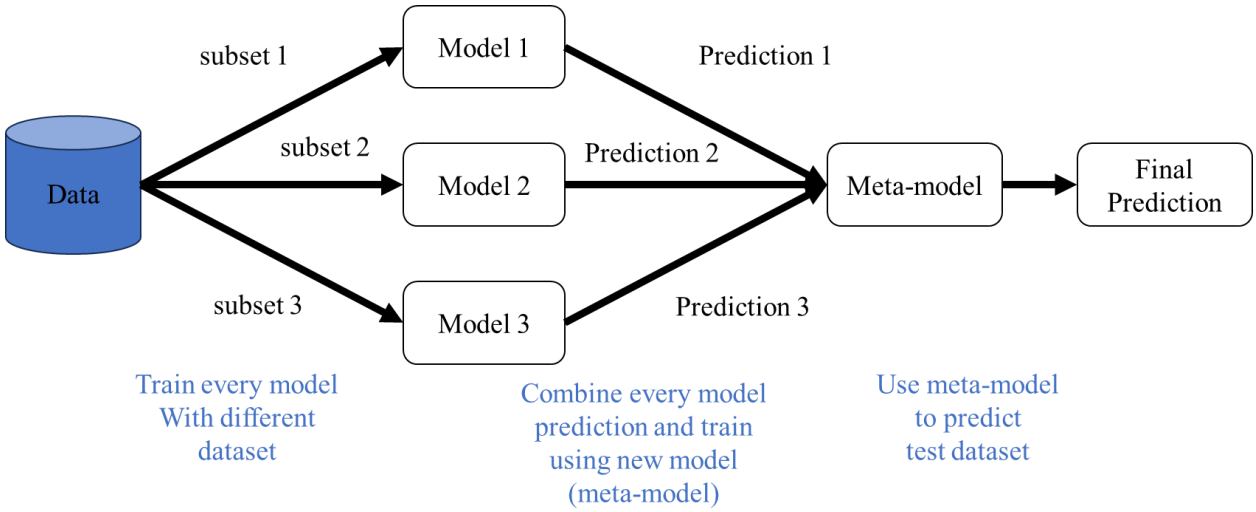

Fig. 5. Structure diagram of Bootstrap Stacking.
The Meta-Model used here are Gaussian Naïve Bayes, Random Forest, and Decision Tree.

2. Bootstrap Meta-Model Stacking Classifier

Bootstrap Meta-Model Stacking Classifier combines bootstrapping technique with Meta-Model Stacking Classifier. Meta-Model Stacking Classifier itself means combining multiple meta-models by using a Stacking Classifier from scikit-learn. Stacking Classifier combines predictions from previous models (model 1,2 and 3) and uses k-fold cross-validation. With k-fold cross-validation, the combined prediction produced by model 1,2, and 3 is divided into k subsets, and the meta-model is trained and evaluated k times using different subsets as training and validation sets.

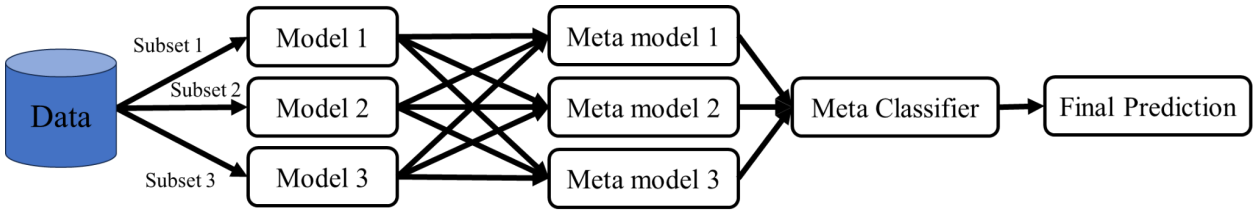

Fig. 6. Structure diagram of Bootstrap Meta-Model Stacking Classifier

We implemented three variations of a Bootstrap Meta-Model Stacking-Classifier (MMSC). The first one, MMSC 1, used three different meta-models (Gaussian Naïve Bayes, Decision Tree Classifier, and Linear Discriminant Analysis) with Logistic Regression as the Meta Classifier, employing 5-fold cross-validation. The second, MMSC 2, was identical to MMSC 1 but used 10-fold cross-validation. The third, MMSC 3, used four meta-models (Gaussian Naïve Bayes, Decision Tree Classifier, Linear Discriminant Analysis, and Random Forest Classifier with 32 trees) with Logistic Regression as the Meta Classifier and 5-fold cross-validation.

3. **Bootstrap Multiple Meta-Model Stacking**

Bootstrap Multiple Stacking is a combination between bagging and multiple stacking. It uses a bootstrapping method from bagging and then combines the prediction by using multiple meta-model stacking. Multiple meta-model stacking itself is a method to use multiple layers of stacking and the method is as follows. Each layer consists of three meta-models and first we train the combined prediction on each meta-model on the first layer. After that, we combine two meta-models predictions from the first layer and train it on different meta-models predictions on the next layer and this works until the last layer. Lastly, combine the last layer predictions and use a meta-classifier for final prediction.

As an illustration, we train the combined predictions using meta-models f11, f12, and f13. Subsequently, we merge predictions from meta-models f11 and f12, followed by training on meta-model f23. This process is mirrored across the other meta-models: combining predictions from meta-models f12 and f13, then training on meta-model f21, and similarly combining predictions from meta-models f11 and f13, and training on meta-model f22.

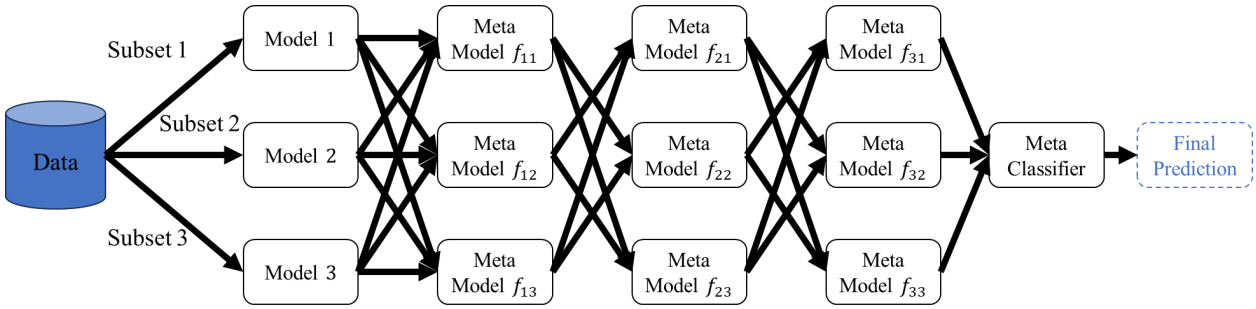

Fig. 7. Structure diagram of Bootstrap Multiple Meta-Model Stacking

In this case, we are using three different Meta-Models for each layer, including Gaussian Naïve Bayes, Random Forest with 32 numbers of trees, and Decision Tree.

Implementation Details

In this study, we use Google Colab Pro with V100 GPU and High-RAM. The environments that we used are Tensorflow 2.13.0 and Python 3.10.12.

1. **LGG MRI Segmentation**

In LGG MRI Segmentation Dataset, we choose 100 epoch with a batch size of 16, with Learning Rate 0.0001 in the first 5 epoch , after 5 epoch use learning rate scheduler to change learning rate we employ the formula of new_learning_rate = current_learning_rate * tf.math.exp(-0.1) and we also employ steps per epoch of the length of training sets divided by batch size. The optimizer that we use is Adam with binary_crossentropy loss as the objective function. Additionally, we implement an early stopping mechanism that triggers when the model fails to show improvement for ten consecutive epochs.

2. **Kvasir-SEG**

In this Kvasir-SEG Dataset, we choose 100 epochs with a batch size of 16, with a Learning Rate of 0.001 and steps per epoch of the length of training sets divided by batch size. The optimizer that we use is

Adam with binary_crossentropy loss as the objective function. Furthermore, we choose not to apply Early Stopping.

3. Pulmonary Chest X-Ray Abnormalities

In this Pulmonary Chest X-Ray Abnormalities Dataset, we choose 100 epoch with a batch size of 16, with Learning Rate 0.0001 in the first 5 epoch, after 5 epoch, use learning rate scheduler to change learning rate with this formula lr * tf.math.exp(-0.1). We use Adam as an optimizer with binary_crossentropy loss as the objective function. We implement an early stopping mechanism when the model fails to show improvement for 10 consecutive epochs.

Evaluation

For evaluation we use the following performance metrics: F1-score, IoU (Intersection over Union), Precision, Recall, and Accuracy. All these metrics are calculated with reference to the true mask that we possess. The formula are presented below:

$$F1\text{–}score = \frac{2TP}{2\,TP+FP+FN}$$

$$IoU = \frac{TP}{TP+FP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

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

All the metrics utilized in this study rely on the binary classification's confusion matrix, where TP, FP, TN, and FN correspond to the true positive, false positive, true negative, and false negative rates, respectively [27].This approach ensures that our evaluation process is grounded in the comparison between the predicted results and the ground truth, represented by the true mask.

In this study, we have chosen not to employ Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) as the primary metrics for evaluating our image segmentation model. The rationale behind this decision stems from the nature of our specific task, which involves delineating objects and regions of interest within images with the availability of true masks (ground truth). ROC and AUC metrics are typically well-suited for assessing the trade-offs between sensitivity and specificity when making decisions based on class probabilities or scores. In contrast, image segmentation focuses on pixel-level delineation, and our primary objective is to measure the spatial overlap and accuracy of our model's predictions with respect to the ground truth. Due to this reason, metrics like F1 Score, IoU, Precision, Recall, and Accuracy are better aligned with our objectives, providing a more meaningful and detailed assessment of the quality and accuracy of our segmentation results. These metrics offer insights into how effectively our model captures the spatial distribution of objects within images, making them the preferred choice for our evaluation.

Results

Table 1 presents the base models with and without bootstrapping, along with three ensemble learning techniques on the LGG MRI segmentation dataset. The first ensemble learning technique is bootstrap stacking, which incorporates bagging and stacking. The second technique, the bootstrap meta-model stacking classifier (MMSC), wherein it uses a stacking classifier to blend models with various meta-models to perform predictions. The last technique is called bootstrap multiple meta-model stacking, utilizing a bootstrapping approach from bagging to consolidate predictions through multiple meta-models.

The bootstrap MMSC techniques classifier 1 or 2 performs the best in LGG MRI Segmentation Dataset, with an average F1 score of 95.97%, an IoU of 85.21%, a Precision of 90.85%, and an Accuracy of 99.85%. By the way, the bootstrap stacking model with Gaussian NB achieved the highest Recall of 99.27%.

Table 1. Segmentation Accuracy on the LGG MRI Segmentation.

Low-Grade-Glioma MRI Segmentation					
Method	**Evaluation**				
	F1-Score	**IoU**	**Precision**	**Recall**	**Accuracy**
SegNet	92.29%	73.50%	85.27%	84.20%	99.73%
U-Net	93.52%	77.25%	87.14%	87.21%	99.76%
DeepLab	89.79%	66.32%	81.66%	77.97%	99.65%
SegNet (with bootstrapping)	94.60%	80.67%	89.51%	89.09%	99.81%
U-Net (with bootstrapping)	95.45%	83.45%	92.10%	89.88%	99.84%
DeepLab (with bootstrapping)	92.67%	74.63%	85.16%	85.78%	99.74%
Bootstrap Stacking with Gaussian Naive Bayes	91.20%	70.33%	70.70%	**99.27%**	99.62%
Bootstrap Stacking with Random Forest	95.59%	83.93%	91.36%	91.16%	99.84%
Bootstrap Stacking with Decision Tree	93.96%	78.63%	87.28%	88.80%	99.78%
Bootstrap Meta-Model Stacking Classifier 1	**95.97%**	**85.21%**	**90.85%**	93.20%	**99.85%**
Bootstrap Meta-Model Stacking Classifier 2	**95.97%**	**85.21%**	**90.85%**	93.20%	**99.85%**
Bootstrap Meta-Model Stacking Classifier 3	95.84%	84.77%	90.40%	93.15%	99.85%
Bootstrap Multiple Meta-Model Stacking (Final Layer Result)	94.51%	80.38%	86.54%	91.87%	99.80%

The best results are in **bold**.

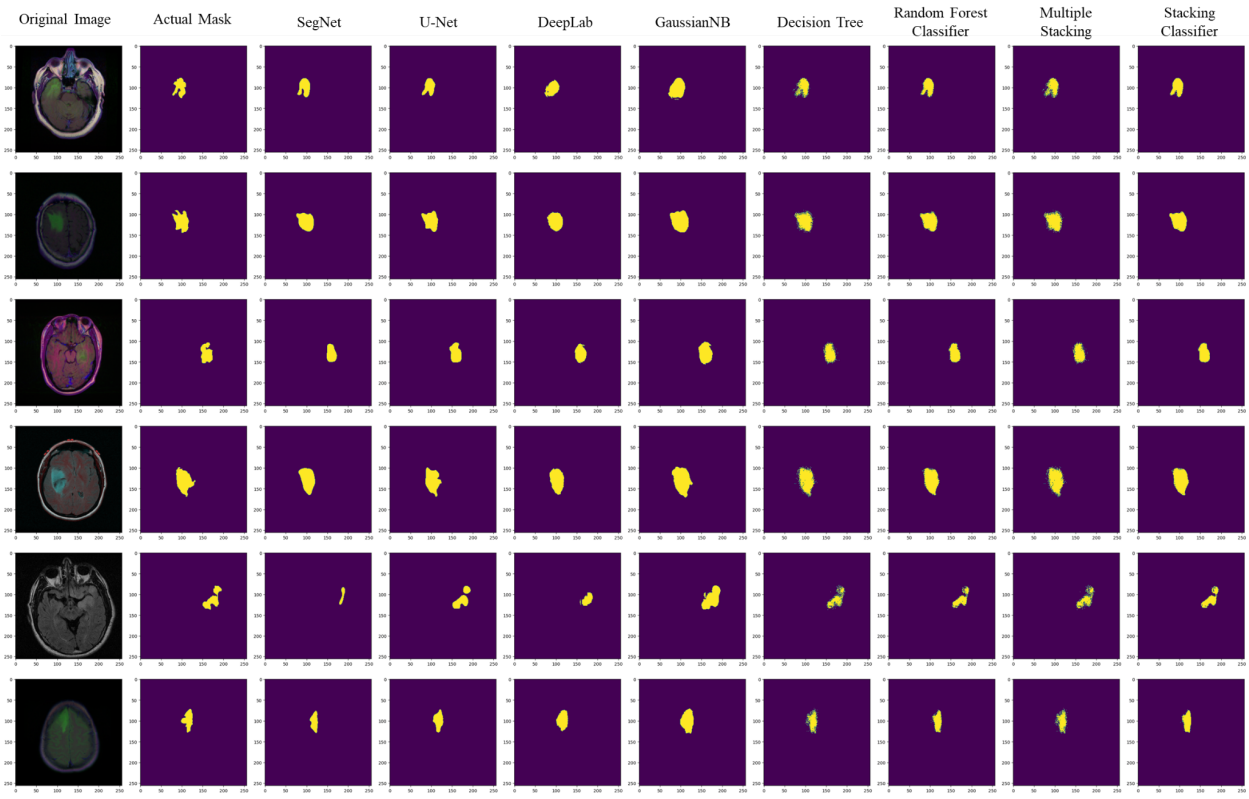

Fig. 8. Comparison of predicted tumor masks on LGG MRI dataset.

The same segmentation methods were also tested on the Kvasir-SEG dataset as seen in Table 2. Among these methods, the bootstrap MMSC classifier 3 stands out as the top performer. It achieves remarkable results with a F1 score of 96.99%, an IoU of 90.58%, a Precision of 94.51%, and an Accuracy of 98.23%. Interestingly, for Recall, the best-performing model is Bootstrap Stacking with Gaussian Naive Bayes, achieving an impressive Recall rate of 99.00%.

Table 2. Segmentation Accuracy on Kvasir-SEG.

Kvasir-SEG					
Method	**Evaluation**				
	F1 Score	**IoU**	**Precision**	**Recall**	**Accuracy**
SegNet	86.31%	62.57%	85.75%	69.89%	92.69%
U-Net	87.28%	64.88%	86.32%	72.64%	93.13%
DeepLab	86.51%	63.38%	82.21%	73.53%	92.56%
SegNet (with bootstrapping)	93.34%	80.13%	93.34%	85.83%	96.22%
U-Net (with bootstrapping)	93.12%	79.54%	93.16%	84.46%	96.11%
DeepLab (with bootstrapping)	92.76%	78.62%	91.77%	84.58%	95.90%
Bootstrap Stacking with Gaussian Naive Bayes	94.75%	84.38%	85.10%	**99.00%**	96.74%
Bootstrap Stacking with Random Forest Classifier	96.91%	90.33%	95.62%	94.23%	98.21%
Bootstrap Stacking with Decision Tree	95.68%	86.73%	93.38%	92.42%	97.49%
Bootstrap Meta-Model Stacking Classifier 1	96.66%	89.62%	93.71%	95.37%	98.04%
Bootstrap Meta-Model Stacking Classifier 2	96.66%	89.63%	93.70%	95.37%	98.04%
Bootstrap Meta-Model Stacking Classifier 3	**96.99%**	**90.58%**	**94.51%**	95.62%	**98.23%**
Bootstrap Multiple Meta-Model Stacking (Final Layer Result)	95.97%	87.55%	94.47%	92.35%	97.67%

The best results are in **bold**.

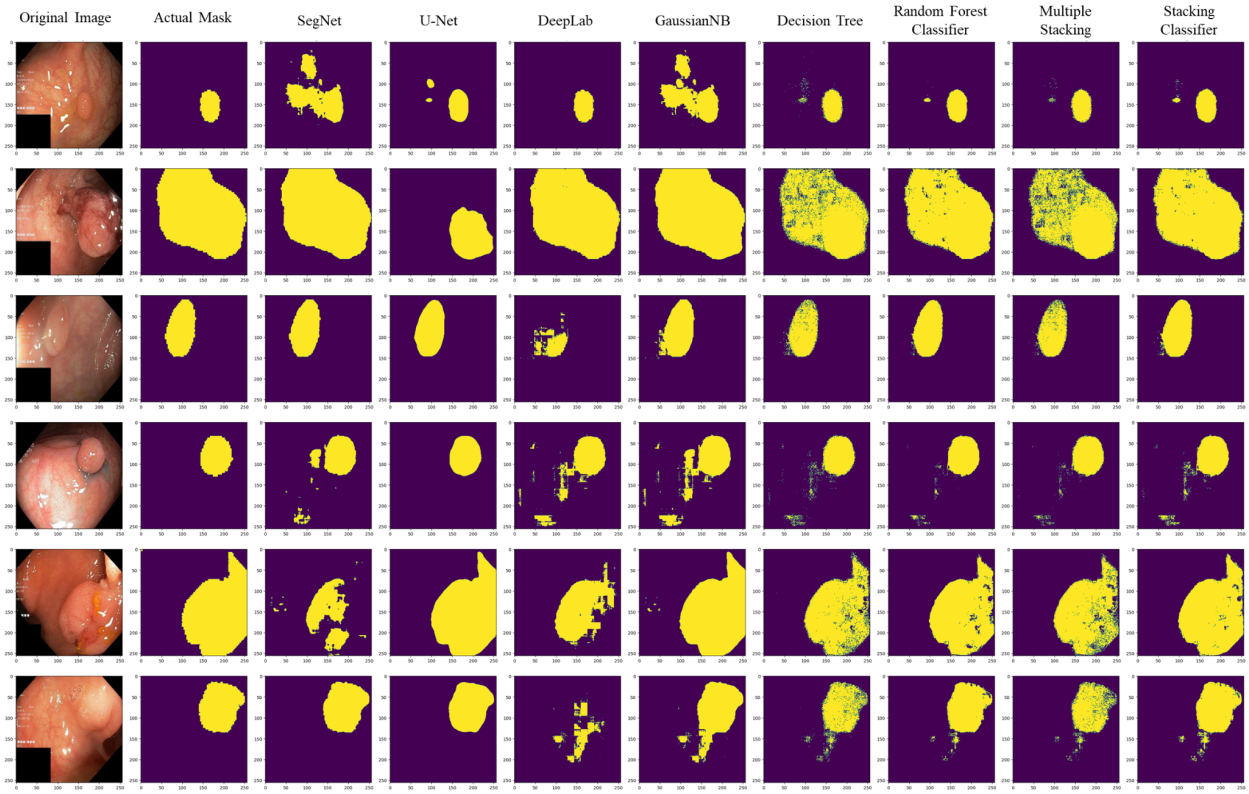

Fig. 9. Comparison of predicted polyp masks on Kvasir-SEG.

In addition, the proposed segmentation methods were examined on the Pulmonary Chest X-Ray Abnormalities dataset. Among these techniques, the Bootstrap Meta-Model Stacking Classifier achieves the highest performance, with a F1 score of 98.69%, an IoU of 96.15%, a Precision of 97.82%, and an Accuracy of 99.02%. To our surprise, the model with best recall is bootstrap stacking with Gaussian NB with a Recall of 99.35%.

Table 3. Segmentation Accuracy on Pulmonary Chest X-Ray Abnormalities.

Pulmonary Chest X-Ray Abnormalities					
Method	**Evaluation**				
	F1 Score	**IoU**	**Precision**	**Recall**	**Accuracy**
SegNet	96.08%	78.35%	95.36%	96.80%	98.02%
U-Net	95.92%	88.45%	96.25%	95.60%	97.96%
DeepLab	94.91%	86.21%	95.19%	94.62%	97.45%
SegNet (with bootstrapping)	98.41%	95.34%	97.73%	97.49%	98.81%
U-Net (with bootstrapping)	98.17%	94.67%	97.43%	97.09%	98.64%
DeepLab (with bootstrapping)	98.15%	94.61%	97.64%	96.82%	98.62%
Bootstrap Stacking with Gaussian Naive Bayes	98.47%	95.54%	96.13%	**99.35%**	98.84%
Bootstrap Stacking with Random Forest	98.67%	96.09%	98.39%	97.61%	99.01%
Bootstrap Stacking with Decision Tree	98.17%	94.66%	97.56%	96.95%	98.63%
Bootstrap Meta-Model Stacking Classifier 1	98.68%	96.12%	97.74%	98.31%	99.01%
Bootstrap Meta-Model Stacking Classifier 2	98.68%	96.12%	97.74%	98.31%	99.01%
Bootstrap Meta-Model Stacking Classifier 3	**98.71%**	**96.21%**	**97.99%**	98.15%	**99.03%**
Bootstrap Multiple Meta-Model Stacking (Final Layer Result)	98.47%	95.50%	98.03%	97.37%	98.85%

The best results are in **bold**.

Fig. 10. Comparison of predicted lung masks on pulmonary chest X-Ray abnormalities dataset.

As we can see from Table 1, 2, 3, our ensemble learning technique excels in terms of IoU, F1-Score, Precision, Recall, and Accuracy when compared to both base model and base model with bootstrapping. Our top performing ensemble learning technique is MMSC with decision tree (classifier 3) in average with an improvement of +23.26% in IoU score, +6.31% in F1-score, +6.73% in Precision, +14.59% in Recall, +2.34% in Accuracy compare to base model without bootstrapping. While our Bootstrap MMSC outperforms other techniques, bootstrap stacking with Random Forest Classifier yields nearly equivalent performance to the bootstrap MMSC with an improvement of +22.34% in IoU score, +6.18% in F1-score, +7.66% in Precision, +12.82% in Recall, +2.32% in Accuracy compared to the base models without bootstrapping.

Discussion

Our research outcomes provide strong evidence of the substantial performance enhancements attainable through the application of ensemble learning compared to individual base models. Ensemble learning capitalizes on the collaborative strength of these base models, where each model contributes its unique insights. This synergistic approach results in significantly improved accuracy and robustness. This phenomenon was consistently observed in our study, as our ensemble learning models consistently outperformed each individual base model. While the individual base models excelled in specific aspects of the task, they demonstrated limitations in handling the full complexity of our dataset. Ensemble learning effectively bridged these gaps by leveraging the diverse patterns captured by each model, leading to a more comprehensive understanding of the data.

Additionally, we can see that bootstrapping techniques can improve the performance of the based models. Bootstrapping technique involves the random selection of multiple subsamples, each of the same size as the original dataset, with replacement. These samples, characterized by their inherent variation due to the random sampling process, contribute to the creation of diverse datasets for training. The introduction of this diversity in the training data facilitates the learning process of the model and helps improve performance.

Even though our ensemble learning method gives an improvement over base models, our Bootstrap Multiple Meta-Model stacking doesn't consistently improve performance when compared to other proposed ensemble learning methods. The reason for this is that in many cases, one layer of meta-model stacking is already effective enough in capturing the underlying patterns and relationships in the data. Using too many layers of stacking can have a diminishing return and even reduce performance, as it introduces more complexity without necessarily enhancing predictive power. However, it's important to note that having multiple layers can be beneficial in certain scenarios, particularly when dealing with complex datasets prone to overfitting. Multiple layers can help reduce overfitting and increase model diversity, potentially leading to better generalization. Nonetheless, the actual impact on performance can vary depending on factors like dataset size; larger datasets may make it more challenging for models to generalize with excessive stacking. Therefore, the decision to use multiple layers of meta-model stacking should be based on careful experimentation and consideration of the specific characteristics of the data at hand.

Based on the medical image segmentation results on Kvasir-SEG [13], our best model bootstrap MMSC outperforms all other models except for Meta-Polyp [28]. Meta-Polyp[28] achieves a highest IoU performance score of 0.921 compared to our score of 0.9058. However, our model still slightly beats **DUCK-Net [29] which is the second place, which has an IoU score of 0.9051.** This evidence supports the conclusion that our approach enhances overall model performance.

When we see the aforementioned results, it becomes apparent that the Random Forest Classifier excels as the preferred choice for the stacking technique. Despite Bootstrap Meta-Model Stacking Classifier (MMSC) 3 delivering a superior score when compared to the Random Forest Classifier, the consideration of computational efficiency suggests that opting for the Random Forest Classifier is a more practical

choice, as it requires less processing time, in which MMSC 3 runs 6 times slower than Random Forest Classifier.

Our results indicate that Gaussian Naive Bayes performs exceptionally well in achieving high Recall, which has left many of us wondering about the reasons behind this unexpected result.. Through some research, we found out that it is primarily due to its ability to handle continuous data and noise robustness, both of which are prevalent in segmentation tasks. Its Gaussian distribution assumption aligns well with the data distributions commonly found in segmentation, allowing it to effectively capture the nuances of continuous features such as pixel values in images.

Encounter Problems

1. Configuring the necessary Tensorflow environment and installing the required drivers on our computer systems proved to be a time-consuming and intricate task, leading us to adopt Google Colab as a practical workaround.

2. Learning about ensemble learning and its code has taken a while. It involves understanding how to combine different models, choose the right techniques, and adjust settings, which can be a bit tricky.

3. When using the Colab environment, we encountered recurrent issues with system memory leaks, attributed to the inherent complexity of our base models. This happens because of the complexity of our models.

4. Google Colab provides access to GPU, but the availability and type of hardware can be limited. For instance, we only have access to a V100 GPU. This research involves running multiple experiments with different parameters and data preprocessing techniques. Since our resources are limited, it slows down the process of optimizing.

Future Works

In future research, we aim to further develop data processing and augmenting techniques, as we believe they will enable the model to learn more effectively. Additionally, we plan to investigate why our meta model's multiple layers did not lead to a significant improvement in performance. By gaining insights into the specific challenges posed by medical images, we can refine our ensemble strategy for this domain.

We plan to investigate the characteristics of each individual meta-model within our ensemble and assess their respective impacts on the final performance. This analysis will enable us to optimize the composition of our ensemble by selecting and weighting meta-models strategically. Additionally, we will explore the possibility of using alternative base models to further enhance the versatility and adaptability of our ensemble approach, tailoring it to the unique requirements of different tasks and datasets.

Lastly, we plan to explore the use of strong learners in place of weak learners to assess their influence on ensemble learning techniques and to compare their performance. This will help us determine the most effective approach for our specific applications, as well as provide valuable insights into the strengths and weaknesses of different learner types in the context of ensemble learning.

References

[1]R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi, “Medical image segmentation using deep learning: A survey,” IET Image Processing, vol 16 , no. 5 , pp. 1243-1267 Jan. 2022, doi: https://doi.org/10.1049/ipr2.12419.

[2] R. M. Haralick and L. G. Shapiro, “Image segmentation techniques” Computer Vision, Graphics, and Image Processing, vol. 29, no. 1, pp. 100–132, Jan. 1985, doi: https://doi.org/10.1016/S0734-189X(85)90153-7.

[3] D. L. Pham, C. Xu, and J. L. Prince, “Current Methods in Medical Image Segmentation,” Annual Review of Biomedical Engineering, vol. 2, no. 1, pp. 315–337, Aug. 2000, doi: https://doi.org/10.1146/annurev.bioeng.2.1.315.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," Medical Image Computing and Computer-Assisted Intervention, vol. 9351, pp. 234–241, May. 2015. doi: https://doi.org/10.1007/978-3-319-24574-4_28

[5] V. Badrinarayanan, A. Kendall and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 12, pp. 2481-2495, 1 Dec. 2017, doi: 10.1109/TPAMI.2016.2644615.

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. 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, no. 4, pp. 834–848, Apr. 2018, doi: https://doi.org/10.1109/tpami.2017.2699184.

[7] Jieneng Chen, Yongyi Lu, and Chenyang Xu, "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," Feb. 2021, doi: https://doi.org/10.48550/arXiv.2102.04306.

[8] Z. Zhou, Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and J. Liang, “UNet++: A Nested U-Net Architecture for Medical Image Segmentation,”Springer International Publishing, vol. 11045, pp. 3-11 ,Sep 2018,doi:https://doi.org/10.1007/978-3-030-00889-5_1

[9] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation,”2016 fourth international conference on 3D vision (3DV),pp. 565-571,Oct 2016,doi:https://doi.ieeecomputersociety.org/10.1109/3DV.2016.79

[10]D. Opitz and R. Maclin, “Popular Ensemble Methods: An Empirical Study,” Journal of Artificial Intelligence Research, vol. 11, no. 1, pp. 169–198, Aug 1999, doi: https://doi.org/10.1613/jair.614.

[11] L. Breiman, “Bagging predictors”. Springer Link Machine Learning., vol. 24, no. 2, pp. 123–140, Aug. 1996. doi: https://doi.org/10.1007/BF00058655

[12]D. H. Wolpert, “Stacked generalization,” Neural Networks, vol. 5, no. 2, pp. 241–259, Jan. 1992, doi: https://doi.org/10.1016/s0893-6080(05)80023-1.

[13] Papers with Code. Kvasir-SEG Benchmark (Medical Image Segmentation). Retrieved September 14, 2023, Available: https://paperswithcode.com/sota/medical-image-segmentation-on-Kvasir-SEG.

[14] V. Thambawita, S. A. Hicks, P. Halvorsen, and M. A. Riegler, "DivergentNets: Medical Image Segmentation by Network Ensemble," EndoCV@ISBI, July 2021, Available:https://api.semanticscholar.org/CorpusID:235614237

[15] Mateusz Buda, Ashirbani Saha, Maciej A. Mazurowski “Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm” Computers in Biology and Medicine, vol. 109, pp. 218-255, May. 2019, doi:https://doi.org/10.1016/j.compbiomed.2019.05.002

[16] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen “Kvasir-SEG: A segmented polyp dataset.” In International Conference on Multimedia Modeling, pp. 451 - 462, Jan. 2020, doi:https://doi.org/10.1007/978-3-030-37734-2_37

[17] Stefan Jaeger ,Sema Candemir ,Sameer Antani , Yì-Xiáng J. Wáng, Pu-Xuan Lu, and George Thoma. “Two public chest X-ray datasets for computer-aided screening of pulmonary diseases.” Quant Imaging Med Surg, vol.4 , no. 6, pp. 475 - 477, Dec. 2014, doi: https://doi.org/10.3978/j.issn.2223-4292.2014.11.20

[18] S. Alagu, Ahana Priyanka N, Kavitha G and Bhoopathy Bagan K, “Automatic Detection of Acute Lymphoblastic Leukemia Using UNET Based Segmentation and Statistical Analysis of Fused Deep Features,” Applied Artificial Intelligence,vol. 35, no. 15, pp. 1952–1969, Oct. 2021, doi: https://doi.org/10.1080/08839514.2021.1995974.

[19]N. Phani Bindu and P. Narahari Sastry, “Automated brain tumor detection and segmentation using modified UNet and ResNet model”, vol 27, no. 13, pp. 9179–9189, Jul. 2023, Available: https://doi.org/10.1007/s00500-023-08420-5

[20]U. Javaid, D. Dasnoy, and J. A. Lee, “Multi-organ Segmentation of Chest CT Images in Radiation Oncology: Comparison of Standard and Dilated UNet,” Lecture Notes in Computer Science, vol 11182, pp. 188–199, Jan. 2018, doi: https://doi.org/10.1007/978-3-030-01449-0_16.

[21]P. Krähenbühl and V. Koltun, “Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials.”Advances in neural information processing systems,pp. 109-117,Dec 2011,doi:https://dl.acm.org/doi/10.5555/2986459.2986472

[22]D. Opitz and R. Maclin, “Popular Ensemble Methods: An Empirical Study,” Journal of Artificial Intelligence Research, vol. 11, pp. 169–198, Aug. 1999, doi: https://doi.org/10.1613/jair.614.

[23]G. Ngo, R. Beard, and R. Chandra, “Evolutionary bagged ensemble learning.” Neurocomputing, vol 510, pp. 1-14, Oct. 2022, doi: https://doi.org/10.1016/j.neucom.2022.08.055

[24]P. Proscura and A. Zaytsev, “Effective training-time stacking for ensembling of deep neural networks,”Proceedings of the 2022 5th International Conference on Artificial Intelligence and Pattern Recognition,pp. 78-82 , May 2023,doi: https://doi.org/10.1145/3573942.3573954

[25]Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay, “Scikit-learn: Machine Learning in Python” the Journal of machine Learning research,vol 12,pp. 2825-2830, July. 2011, doi:https://dl.acm.org/doi/10.5555/1953048.2078195

[26]Matheus Henrique Dal Molin Ribeiro,Ramon Gomes da Silva ,Sinvaldo Rodrigues Moreno,Viviana Cocco Mariani and Leandro dos Santos Coelho“Efficient bootstrap stacking ensemble learning model applied to wind power generation forecasting,” International Journal of Electrical Power & Energy Systems, vol. 136, pp. 107712, Mar. 2022, doi: https://doi.org/10.1016/j.ijepes.2021.107712.

[27]Lever Jake,Krzywinski Martin and Altman Naomi, “Classification evaluation,” Nature Methods, vol. 13, no. 8, pp. 603–604, Jul. 2016, doi: https://doi.org/10.1038/nmeth.3945.

[28] Quoc-Huy Trinh, “Meta-Polyp: a baseline for efficient Polyp segmentation.”2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS),pp. 742-747,June 2023,doi:http://doi.ieeecomputersociety.org/10.1109/CBMS58004.2023.00312

[29]Razvan-Gabriel Dumitru, Darius Peteleaza, and Catalin Craciun, “Using DUCK-Net for polyp image segmentation,”Scientific Reports, vol. 13, no. 1, Jun. 2023, doi: https://doi.org/10.1038/s41598-023-36940-5.

Self-Assessment of Summer Project

The content of our study remains virtually unchanged from the original plan. We have successfully achieved our project objectives. Our research results contribute to the current state of medical imaging to a certain extent and may be suitable for publication in journals. Our study focuses on a less mature technology, and any research in this area is valuable. We discovered that combining ensemble learning methods with a few datasets leads to some improvement. However, the extent of improvement is highly dependent on the dataset. Interestingly, we observed that increasing the number of ensembles used results in diminishing returns, with some ensembles performing worse than others.

Comments from Instructor

I am truly impressed with the teamwork, dedication and achievements of the five international students who participated in our summer research project. Their unwavering work ethic and self-discipline throughout the project were exemplary, highlighting their ability to excel in active research. Their hard work has paid off, as their research project is on the verge of publication, which is a testament to its scientific value.

In their self-assessment, the students demonstrate a clear understanding of the significance of their work. They've diligently adhered to their original research plan and successfully achieved their objectives. Their contributions to the field of medical imaging are noteworthy, especially considering the relatively less mature technology they explored. Their findings regarding the combination of ensemble learning methods and dataset dependency provide valuable insights, shedding light on the nuances of this area. I have no doubt that their research will make a meaningful contribution to the scientific community, and I look forward to seeing their work published in journals.