M-VAAL: Multimodal Variational Adversarial Active Learning for Downstream Medical Image Analysis Tasks

Bidur Khanal[1(✉)], Binod Bhattarai[4], Bishesh Khanal[3], Danail Stoyanov[5], and Cristian A. Linte[1,2]

[1] Center for Imaging Science, RIT, Rochester, NY, USA
bk9618@rit.edu
[2] Biomedical Engineering, RIT, Rochester, NY, USA
[3] NepAl Applied Mathematics and Informatics Institute for Research (NAAMII), Lalitpur, Nepal
[4] University of Aberdeen, Aberdeen, UK
[5] University College London, London, UK

Abstract. Acquiring properly annotated data is expensive in the medical field as it requires experts, time-consuming protocols, and rigorous validation. Active learning attempts to minimize the need for large annotated samples by actively sampling the most informative examples for annotation. These examples contribute significantly to improving the performance of supervised machine learning models, and thus, active learning can play an essential role in selecting the most appropriate information in deep learning-based diagnosis, clinical assessments, and treatment planning. Although some existing works have proposed methods for sampling the best examples for annotation in medical image analysis, they are not task-agnostic and do not use multimodal auxiliary information in the sampler, which has the potential to increase robustness. Therefore, in this work, we propose a Multimodal Variational Adversarial Active Learning (M-VAAL) method that uses auxiliary information from additional modalities to enhance the active sampling. We applied our method to two datasets: i) brain tumor segmentation and multi-label classification using the BraTS2018 dataset, and ii) chest X-ray image classification using the COVID-QU-Ex dataset. Our results show a promising direction toward data-efficient learning under limited annotations.

Keywords: multimodal active learning · annotation budget · brain tumor segmentation and classification · chest X-ray classification

1 Introduction

Automated medical image analysis tasks, such as feature segmentation and disease classification play an important role in assisting with clinical diagno-

G. Waiter et al. (Eds.): MIUA 2023, LNCS 14122, pp. 48–63, 2024.
https://doi.org/10.1007/978-3-031-48593-0_4

sis, as well as appropriate planning of non-invasive therapies, including surgical interventions [2,8]. In recent years, supervised deep learning-based methods have shown promising results in clinical settings. However, clinical translation of supervised deep learning-based models is still limited for most applications, due to the lack of access to a large pool of annotated data and generalization capability.

As medical data is expensive to annotate, some methods have explored the generation of synthetic images with ground truth annotation [18]. However, generative models for synthesizing high-quality medical data have not yet achieved a state where their distribution perfectly matches the distribution of real medical data, especially those containing rare cases [1,20]. This limitation can often result in biased models and poor performance. Another approach with a growing interest is semi-supervised learning where very few labeled data along with a large number of unlabeled data is used to train deep learning models [6,13,23]. Nevertheless, in a real-world scenario, semi-supervised learning still requires the selection of a certain number of image samples to be annotated by experts.

Active learning attempts to sample the best subset of examples for annotation from a pool of unlabeled examples to maximize the downstream task performance. Methods to sample examples that provide the best improvement with a limited budget have a long history [12], and several approaches have been proposed in recent years [25]. Active learning (AL) has experienced increased traction in the medical imaging domain, as it enables a human-in-the-loop approach to improve deployed AI models [5]. Recent works have proposed various sampling methods in AL specific to tasks such as classification, segmentation, and depth estimation in medical imaging. Shao *et al.* [16] used active sampling to minimize annotation for nucleus classification in pathological images. Yang *et al.* [24] proposed a framework to improve lymph node segmentation from ultrasound images using only 50% training data. Laradji *et al.* [11] adopted an entropy-based method for a region-based active sampler for COVID-19 lesion segmentation from CT images. Thapa *et al.* [22] proposed a task-aware AL for depth estimation and segmentation from endoscopic images. Related to our work, Sharma *et al.* [17] used uncertainty and representativeness to actively sample training examples for 2D brain tumor segmentation. Lastly, Kim *et al.* [10] studied various AL approaches for 3D brain tumor segmentation task.

Task Agnostic: The methods described earlier require training of downstream task-specific models precluding the possibility of training a task-agnostic model from a given pool of unannotated input images when the intended downstream task is not known a priori. On the other hand, task-agnostic methods could enable the use of the same model for new tasks that may arise in real-world continuous deployment settings. VAAL [19] is a promising task-agnostic method that trains a variational auto-encoder (VAE) with adversarial learning. VAAL produces a low-dimensional latent space that aims to capture both uncertainty and representativeness from the input data, enabling effective sampling for AL.

Sampling from multimodal images: Clinicians rarely reach a clinical decision by looking into a single medical image of a patient. They rely on other informa-

tion, such as clinical symptoms, patient reports, multimodal images, and auxiliary device information. We believe that sampling methods in AL could benefit from using multimodal information. Although some recent works have explored multimodal medical images in the context of AL [5], none of the existing methods directly use the multimodal image as auxiliary information to learn the sampler.

In this paper, we propose a task-agnostic method that exploits multimodal imaging information to improve the AL sampling process. We modify the existing task-agnostic VAAL framework to enable exploiting multimodal images and evaluate its performance on two widely used publicly available datasets: the multimodal BraTS dataset [14] and COVID-QU-Ex dataset [21] containing lung segmentation maps as additional information.

The contributions of this work are as follows: **1)** We propose a novel multimodal variational adversarial AL method (M-VAAL) that uses additional information from auxiliary modalities to select informative and diverse samples for annotation; **2)** Using the BraTS2018 and COVID-QU-Ex datasets, we show the effectiveness of our method in actively selecting samples for segmentation, multi-label classification, and multi-class classification tasks. **3)** We show that the sample selection in AL can potentially benefit from the use of auxiliary information.

2 Methodology

In active learning, a subset of most informative samples X_s is selected from a large pool of unlabelled set X^* to query labels Y_s. The number of samples selected for label annotation is dictated by the set budget b of the sampler. Let us consider $(x, y) \sim (X, Y)$ as a labeled data pair initially present in the dataset. After sampling b examples from an unlabelled pool X^*, they are labeled and added to the labeled pool (X, Y). The sampling process is iterative, such that b examples are queried at each active sampling round and added to the labeled pool. The labeled samples are used to train a task network at each round by minimizing the task objective function. In our study, we have considered three different downstream tasks: a) semantic segmentation of brain tumors, b) multi-label classification of tumor types, and c) multi-class classification of chest conditions. The overall workflow of our proposed method is shown in Fig. 1.

2.1 Task Learner (A)

Our objective is to enhance the performance of downstream tasks by utilizing the smallest number of labeled samples. The task network is defined by the parameter θ_T. For the segmentation task, we train a U-Net model [15] to generate pixel-wise segmentation maps y', given a labeled sample $(x, y) \sim (X, Y)$, where y is a binary image. We chose the U-Net model due to its simplicity, as our focus is on evaluating the effectiveness of AL, not on developing a state-of-the-art segmentation technique.

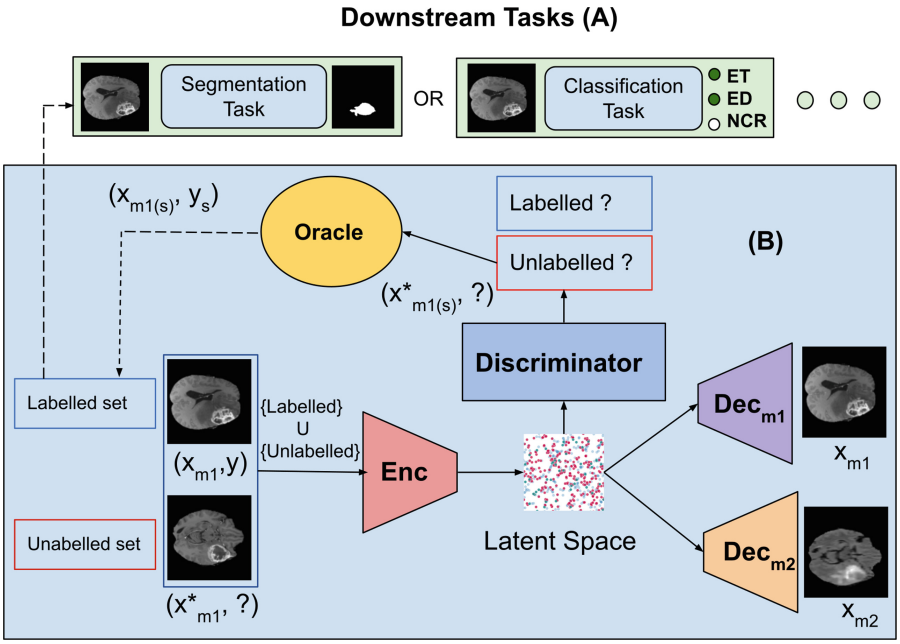


Fig. 1. M-VAAL Pipeline: Our active learning method uses multimodal information ($m1$ and $m2$) to improve VAAL. M-VAAL samples the unlabelled images, and selects samples complementary to the already annotated data, which are passed to Oracle for annotation. Incorporating auxiliary information from the second modality ($m2$) produces a more generalized latent representation for sampling. Our method learns task-agnostic representations, therefore the latent space can be used for both classification and segmentation tasks to sample the best-unlabelled images for annotation. (Refer to 2.2 for the meaning of each notation)

For the classification task, we use a ResNet-18 [9] classifier to predict y', given $(x, y) \sim (X, Y)$ where $y \subseteq \{0, 1, 2, ..\}$. In a multi-label classification setting, y' can contain more than one class, while in a multi-class setting, y' belongs to only one class from the set. Once a task is trained using labeled samples, we move on to the sampler (B) stage to select the best samples for annotation. After annotating the selected samples, we add them to the labeled sample pool and retrain the task learner using the updated sample set. Thus, the task learner is completely retrained for multiple rounds, with an increased training budget each time.

2.2 M-VAAL Sampler (B)

Our proposed M-VAAL method extends the VAAL approach by using an encoder-decoder architecture combined with adversarial learning to generate a lower dimensional latent space that captures sample uncertainty and representativeness. Unlike other AL methods that directly estimate uncertainty or diversity

using task representation, VAAL learns a separate task-agnostic representation, providing the freedom to model uncertainty independent of task representation. To further strengthen the representation learning capability, we introduce a second modality as auxiliary information. We modify the VAE to reconstruct the original input image from modality $m1$ and the auxiliary image of modality $m2$ using the same latent representation generated from the $m1$ images. The objective function of the VAE is formulated as:

$$\mathcal{L}_{\text{VAE}}^{m1} = \mathbb{E}\left[\log p_{\theta_{m1}}\left(x_{m1} \mid z\right)\right] + \mathbb{E}\left[\log p_{\theta_{m1}}\left(x_{m1}^{*} \mid z^{*}\right)\right] \tag{1}$$

$$\mathcal{L}_{\text{VAE}}^{m2} = \mathbb{E}\left[\log p_{\theta_{m2}}\left(x_{m2} \mid z\right)\right] + \mathbb{E}\left[\log p_{\theta_{m2}}\left(x_{m2}^{*} \mid z^{*}\right)\right] \tag{2}$$

where $p_{\theta_{m1}}$ and $p_{\theta_{m2}}$ are the decoders for modality $m1$ and $m2$, respectively, parameterized by θ_{m1} and θ_{m2}. Likewise, z, x_{m1}, and x_{m2} represent a latent representation, an image of modality $m1$, and an image of modality $m2$, respectively, of the samples belonging to the labelled set; similarly, z^{*}, x_{m1}^{*}, and x_{m2}^{*} represent a latent representation, an image of modality $m1$, and an image of modality $m2$, respectively, of the samples belonging to the unlabelled set. The VAE also uses a Kullback-Leibler (KL) distance loss to regularize the latent representation, formulated as:

$$\mathcal{L}_{\text{VAE}}^{KL} = -\,\mathrm{D}_{\text{KL}}\left(q_{\phi_{m1}}\left(z \mid x_{m1}\right) || p(z)\right) - \mathrm{D}_{KL}\left(q_{\phi_{m1}}\left(z^{*} \mid x_{m1}^{*}\right) || p(z)\right) \tag{3}$$

where $q_{\phi_{m1}}$ is the encoder of modality $m1$ parameterized by ϕ_{m1} and $p(z)$ is the prior chosen as a unit Gaussian. It should be noted that only a single encoder is used to generate the latent representation, while there are two decoders.

Finally, an adversarial loss is added to encourage the VAE towards generating the latent representation that can fool the discriminator in distinguishing labeled from unlabelled examples (shown in Eq. 4). We train both the VAE and the discriminator in an adversarial fashion. The discriminator (D) is trained on the latent representation of an image to distinguish if an image comes from a labeled set or an unlabeled set. The loss function for the discriminator is shown in Eq. 5. We used Wasserstein GAN loss with gradient penalty [7].

$$\mathcal{L}_{\text{VAE}}^{adv} = \mathbb{E}\left[D\left(q_{\phi_{m1}}\left(z \mid x_{m1}\right)\right)\right] - \mathbb{E}\left[D\left(q_{\phi_{m1}}\left(z^{*} \mid x_{m1}^{*}\right)\right)\right] \tag{4}$$

$$\mathcal{L}_{D} = -\mathbb{E}\left[D\left(q_{\phi_{m1}}\left(z \mid x_{m1}\right)\right)\right] + \mathbb{E}\left[D\left(q_{\phi_{m1}}\left(z^{*} \mid x_{m1}^{*}\right)\right)\right] + \lambda\mathbb{E}\left[\left(\|\nabla_{\hat{x}} D(\hat{x})\| - 1\right)^{2}\right] \tag{5}$$

where the third term in Eq. 5 is the gradient penalty and $\hat{x} = \epsilon x_{m1} + (1-\epsilon)x_{m1}^{*}$ is a randomly weighted average between x_{m1} and x_{m1}^{*}, such that $0 \leq \epsilon \leq 1$.

During the sampling process, the top samples that the discriminator votes as belonging to the unlabelled set are chosen; our intuition is that these samples contain information most different from the ones carried by the samples already belonging to the labeled set. Figure 2 shows that our method selects the examples that are far from the distribution of labeled examples, thus capturing diverse samples. The overall training and sampling sequence of M-VAAL is shown in Algorithm 1. In the first round, the task network is trained only

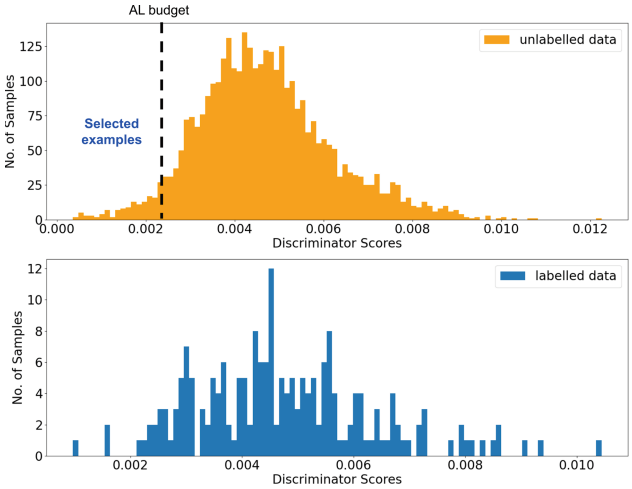


Fig. 2. Histogram of the discriminator scores for unlabeled and labeled data at third AL round. The discriminator is adversarially trained to push unlabeled samples toward lower values and labeled samples toward higher values. Our method involves selecting unlabelled instances that are far from the peak distribution of the labeled data. The number of samples to select is dictated by the AL budget.

Algorithm 1. Multimodal Variational Adversarial Active Learning

Given: Hyperparameters: epochs, γ_1, γ_2, γ_3, δ_1, δ_2, δ_3
Input: Labeled data (x_{m1}, y, x_{m2}), Unlabeled data (x^*_{m1}, x^*_{m2})
Initialize: Model parameters θ_T, $\theta_{VAE} = \{\theta_{m1}, \theta_{m2}, \phi\}$, and θ_D

for $e = 1$ to epochs **do**
  sample $(x_{m1}, y, x_{m2}) \sim (X_{m1}, Y, X_{m2})$
  sample $(x^*_{m1}, x^*_{m2}) \sim (X^*_{m1}, X^*_{m2})$
  $\mathcal{L}_{\text{VAE}} \leftarrow \gamma_1 \mathcal{L}^{adv}_{\text{VAE}} + \gamma_2 \mathcal{L}^{m1}_{\text{VAE}} + \gamma_3 \mathcal{L}^{m2}_{\text{VAE}} + \mathcal{L}^{KL}_{\text{VAE}}$
  Update VAE: $\theta'_{VAE} \leftarrow \theta_{VAE} - \delta_1 \nabla \mathcal{L}_{\text{VAE}}$
  Update D: $\theta'_D \leftarrow \theta_D - \delta_2 \nabla \mathcal{L}_{\text{D}}$
  Train and update T: $\theta'_T \leftarrow \theta_T - \delta_3 \nabla \mathcal{L}_{\text{T}}$

Sampling Phase
Input: b, X_{m1}, X^*_{m1}
Output: X_{m1}, X^*_{m1}

Select samples $X^*_{m1(s)}$ with $\min_b\{\theta_D(z^*)\}$
$y_s \leftarrow \mathcal{ORACLE}(X^*_{m1(s)})$
$(X_{m1}, Y) \leftarrow (X_{m1}, Y) \cup (X^*_{m1(s)}, Y_s)$
$X^*_{m1} \leftarrow X^*_{m1} - X^*_{m1(s)}$
return X_{m1}, X^*_{m1}

with the initially labeled samples to minimize the task objective function $\mathcal{L}_T$. Simultaneously, the M-VAAL sampler is also trained independently using all the labeled and unlabelled examples, with the end goal of improving the discrimi-

nator at distinguishing between labeled and unlabelled pairs. After the initial round of training, the trained discriminator is used to select the top b unlabelled samples identified as belonging to the unlabelled set. These selected samples are sent to the Oracle for annotation. Finally, the selected samples are removed from the unlabelled set and then added to the pool of labeled sets, along with their respective labels. The next round of AL proceeds in a similar fashion, but with the updated labeled and unlabelled sets.

3 Experiments

3.1 Dataset

BraTS2018: We used the BraTS2018 dataset [14], which includes co-registered 3D MR brain volumes acquired using different acquisition protocols, yielding slices of T1, T2, and T2-Flair images. To sample informative examples from unlabelled brain MR images, we employed the M-VAAL algorithm using contrast-enhanced T1 sequences as the main modality, and T2-Flair as auxiliary information. Contrast-enhanced T1 images are preferred for tumor detection as the tumor border is more visible [4]. In addition, T2-Flair, which captures cerebral edema (fluid-filled region due to swelling), can also be utilized for diagnosis [14]. Our focus was on the provided 210 High-Grade Gliomas (HGG) cases, which included manual segmentation verified by experienced board-certified neuroradiologists. There are three foreground classes: Enhancing Tumor (ET), Edematous Tissue (ED), and Non-enhancing Tumor Core (NCR). In practice, the given foreground classes can be merged to create different sub-regions such as whole tumor and tumor core for evaluations [14]. The whole tumor entails all foreground classes, while the tumor core only entails ET and NCR.

Before extracting 2D slices from the provided 3D volumes, we randomly split the 210 volumes into training and test cases with an 80:20 ratio. The training set was further split into training and validation cases using the same ratio of 80:20, resulting in 135 training, 33 validation, and 42 test cases. These splits were created before extracting 2D slices to avoid any patient information leakage in the test splits. Each 3D volume had 155 (240×240) transverse slices in axial view with a spacing of 1 mm. However, not all transverse slices contained tumor regions, so we extracted only those containing at least one of the foreground classes. Some slices contained only a few pixels of the foreground segmentation classes, so we ensured that each extracted slice had at least 1000 pixels representing the foreground class; any slice not meeting this threshold was discarded. Consequently, the curated dataset comprised 3673 training images, 1009 validation images, and 1164 test images of contrast-enhanced T1, T2-Flair, and the segmentation map.

We evaluated our method on two downstream tasks: whole tumor segmentation and multi-label classification task. In the multi-label classification task, our prediction classes consisted of ET, ED, and NCR, with each image having either one, two, or all three classes. It's worth noting that for the downstream

task, only contrast-enhanced T1 images were used, while M-VAAL made use of both contrast-enhanced T1 and T2-Flair images.

COVID-QU-Ex: The COVID-QU-Ex dataset is composed of 256×256 chest X-ray images from different patients that have been categorized into one of three groups: COVID infection, Non-COVID infection, and Normal. The dataset has two subsets: lung segmentation data and COVID-19 infection segmentation data, and the latter was chosen for our experiment. In addition to the X-ray images, the dataset also provides a segmentation mask for each image, which was utilized as an auxiliary modality during training M-VAAL. The dataset has a total of 5,826 images, consisting of 1,456 Normal, 1,457 Non-COVID-19, and 2,913 COVID-19 cases. These images are split into three sets: training, validation, and test, with the training set containing 3,728 images, the validation set containing 932 images, and the test set containing 1,166 images. The downstream task was to classify the input X-ray image into one of three classes.

3.2 Implementation Details

BraTS2018: All the input images have a single channel of size 240×240. To preprocess the images, we removed 1% of the top and bottom intensities and normalized them linearly by dividing each pixel with the maximum intensity, bringing the pixel intensity to a range of 0 to 1. We also normalized the images by subtracting and dividing them by the mean and standard deviation, respectively, of the training data. For VAAL and M-VAAL, the images were center-cropped to 210×210 pixels and resized to 128×128 pixels. For the downstream task, we used the original image size. Furthermore, to stabilize the training of VAAL and M-VAAL under similar hyperparameters as the original VAAL [19], we converted the single-channel input image to a three-channel RGB with the same grayscale value across all channels.

For M-VAAL, we used the same β-VAE and discriminator as in original VAAL [19], but added a batch normalization layer after each linear layer in the discriminator. Additionally, instead of using vanilla GAN with binary-cross entropy loss, we used WGAN with a gradient penalty to stabilize the adversarial training [7], with $\lambda = 1$. The latent dimension of VAE was set to 64, and the initial learning rate for both the VAE and discriminator was $1e^{-4}$. We tested γ_3 in the range $M = 0.2, 0.4, 0.8, 1$ via an ablation study reported in Sec 5, while γ_1 and γ_2 were set to 1. Both VAAL and M-VAAL used a mini-batch size of 16 and were trained for 100 epochs using the same hyperparameters, except for γ_3, which was only present in M-VAAL. For consistency, we initialized the model at each stage with the same random seed and repeated three trials with three different seeds for each experiment, recording the average score.

For the segmentation task, we used a U-Net [15] architecture with four down-sampling and four up-sampling layers, trained with an initial learning rate of $1e^{-5}$ using the RMSprop optimizer for up to 30 epochs in a mini-batch size of 32. The loss function was the sum of the pixel-wise cross-entropy loss and Dice

coefficient loss. The best model was identified by monitoring the best validation Dice score and used on the test dataset. For the multi-label classification, we used a pre-trained ResNet18 architecture trained on the ImageNet dataset. Instead of normalizing the input images with our training set's mean and standard deviation, we used the ImageNet mean and standard deviation to make them compatible with the pre-trained ResNet18 architecture. We used an initial learning rate of $1e^{-5}$ with the Adam optimizer and trained up to 50 epochs in a mini-batch size of 32. The best model was identified by monitoring the best validation mAP score and used on the test set. We used AL to sample the best examples for annotation for up to six rounds and seven rounds for segmentation and classification tasks, respectively, starting with 200 samples and adding 100 examples (budget – b) at each round.

COVID-QU-Ex: All the input images were loaded as RGB, with all channels having the same gray-scale value. To bring the pixel values within the range of 0 to 1, we normalized each image by dividing all the pixels by 255. Additionally, we normalized the images by subtracting the mean and dividing by the standard deviation, both with values of 0.5. The same hyperparameters were used for both VAAL and M-VAAL as those used in BraTS, and we also downsampled the original 256×256 images to 128×128 for both models. For the downstream multi-class classification task, we utilized a pre-trained ResNet18 architecture that was trained on the ImageNet dataset. We trained the model with an initial learning rate of $1e^{-5}$ using the Adam optimizer and a mini-batch size of 32 for up to 50 epochs. The best model was determined based on the highest validation overall accuracy score, and we evaluated this model on the test set. We employed an AL method that involved sampling up to seven rounds using an AL budget of 100 samples. The initial budget for classification before starting active sampling was 100 samples.

We have released the GitHub repository for our source code implementation[1]. We implemented our method using the standard Pytorch 12.1.1 framework in Python 3.8 and trained in a single A100 GPU (40 GB).

3.3 Benchmarks and Evaluation Metrics

For our study, we employed two baseline control methods: random sampling and VAAL – a state-of-the-art task agnostic learning method [19] that doesn't use any auxiliary information while training. To quantitatively evaluate the segmentation, multi-label classification, and multi-class classification, we measured the Dice score, mean Average Precision (mAP), and overall accuracy, respectively.

[1] https://github.com/Bidur-Khanal/MVAAL-medical-images.

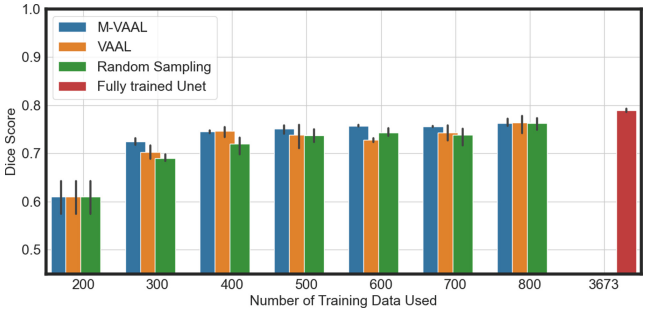


Fig. 3. Whole tumor segmentation performance comparison of proposed (M-VAAL) method against the VAAL and random sampling baselines according to the Dice score.

4 Results

4.1 Segmentation

Figure 3 compares the whole tumor segmentation performance (in terms of Dice score) between our proposed method (M-VAAL), the two baselines (random sampling and VAAL), and a U-Net trained on the entire fully labeled dataset, serving as an upper bound with an average Dice score of 0.789.

As shown, with only 800 labeled samples, the segmentation performance starts to saturate and approaches that of the fully trained U-Net on 100% labeled data. Moreover, M-VAAL performs better than baselines in the early phase and gradually saturates as the number of training samples increase.

Figure 4 illustrates a qualitative, visual assessment of the segmentation masks yielded by U-Net models trained with 400 samples selected by M-VAAL against ground truth, VAAL, and random sampling, at the AL second round. White denotes regions identified by both segmentation methods. Blue denotes regions missed by the test method (i.e., method listed first), but identified by the reference method (i.e., method listed second). Red denotes regions that were identified by the test method (i.e., method listed first), but not identified by the reference method (i.e., method listed second). As such, an optimal segmentation method will maximize the white regions, while minimizing the blue and red regions. Furthermore, Fig. 4 clearly shows that the segmentation masks yielded by M-VAAL are more consistent with the ground truth segmentation masks than those generated by VAAL or Random Sampling.

4.2 Multi-label Classification

In Fig. 6, a comparison is presented of the multi-label classification performance, measured by mean average precision, between the M-VAAL framework and two baseline methods (VAAL and random sampling). The upper bound is represented by a fully fine-tuned ResNet18 network. It should be noted that, on average, M-VAAL performs better than the two baseline methods, particularly when fewer

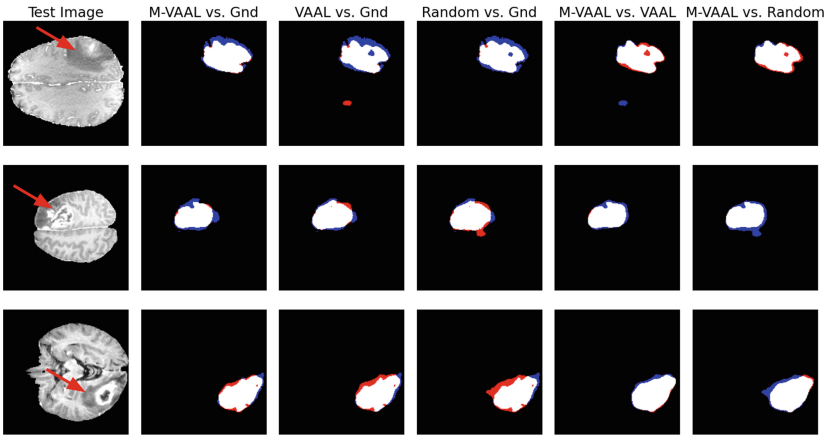


Fig. 4. Qualitative comparison of test segmentation masks generated by U-Nets trained on 400 samples using M-VAAL's selection method at the second round of AL phase. Columns 2–4 compare M-VAAl, VAAL, and random sampling against ground truth. Columns 5–6, compare M-VAAL against VAAL and random sampling. White denotes regions identified by both segmentation methods. Blue denotes regions missed by the test method (i.e., method listed first), but identified by the reference method (i.e., method listed second). Red denotes regions that were identified by the test method (i.e., method listed first), but not identified by the reference method (i.e., method listed second). Additionally, an arrow indicates the features that were segmented.

training data samples are used (i.e. 300 to 600 samples). The comparison models achieve a mean average precision close to that of a fully trained model even with a relatively small number of training samples, suggesting that not all training samples are equally informative. The maximum performance, which represents the upper bound, is achieved when using all samples and is 96.56%.

4.3 Multi-class Classification

The classification performance of ResNet18 models in diagnosing a patient's condition using X-ray images selected by M-VAAL was compared with baselines. Figure 5 illustrates that M-VAAL consistently outperforms the other methods, albeit by a small margin. The maximum performance, which represents the upper bound, is achieved when using all samples and is 95.10%.

5 Ablation Study: M-VAAL

We conducted an ablation study to assess the effect of the multimodal loss component, which involved varying the hyperparameter γ_3 across a range of values ($M = 0.2, 0.4, 0.8, 1.0$). The results suggested that both VAAL and M-VAAL were sensitive to hyperparameters, which depended on the nature of the downstream

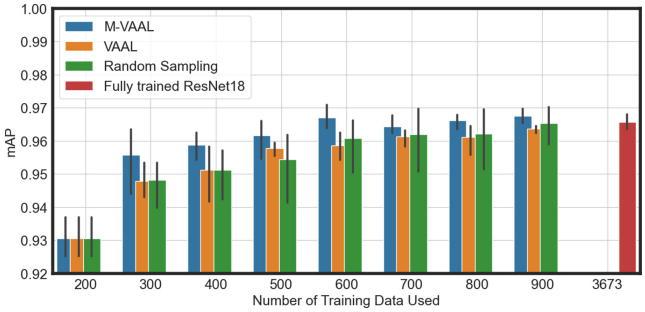


Fig. 5. Comparison of the mean average precision (mAP) for multi-label classification of tumor types between the proposed M-VAAL method and two baseline methods (VAAL and random sampling).

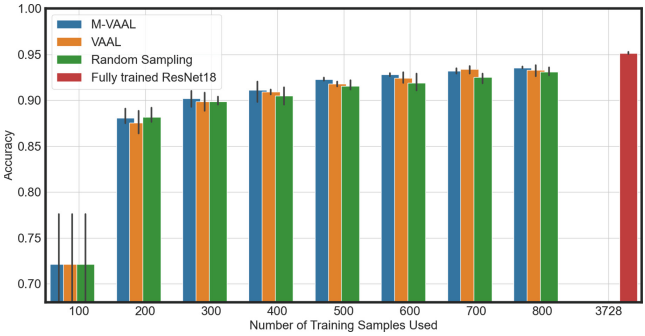
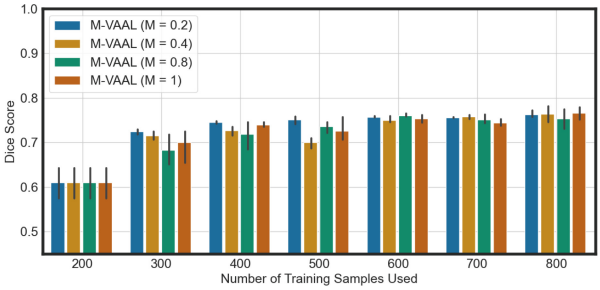
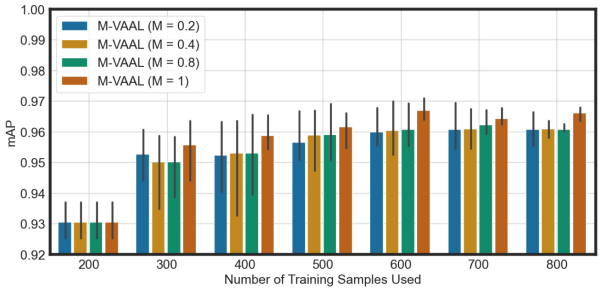


Fig. 6. Comparison of the overall accuracy for multi-class classification of COVID, Non-COVID infections, and normal cases between the proposed M-VAAL method and two baseline methods (VAAL and random sampling).

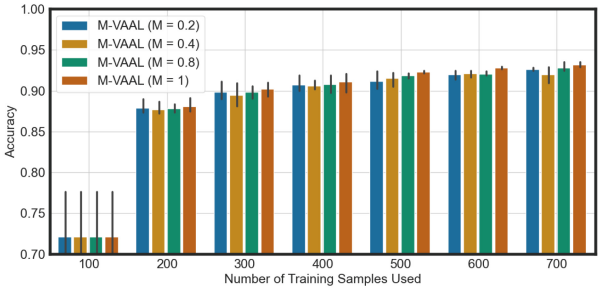
task (see Fig. 7). For example, M = 0.2 performed optimally for the whole tumor segmentation task, while M = 1 was optimal for tumor-type multi-label classification and Chest X-ray image classification. We speculate that the multimodal loss components (see 1 and 2) in the VAE loss serve as additional regularizers during the learning of latent representations. The adversary loss component, on the other hand, contributes more towards guiding the discriminator in sampling unlabelled samples that are most distinct from the distribution of the labeled dataset.



(a)



(b)



(c)

Fig. 7. Comparing the performance of (a) whole tumor segmentation, (b) tumor type multi-label classification, (c) chest X-ray infection multi-class classification with different training budgets, using the samples selected by M-VAAL trained with different values of M.

6 Discussion and Conclusion

M-VAAL, while being sensitive to hyperparameters, samples more informative samples than the baselines. Our method is task-agnostic, but we observed that the optimal value of M differs for different tasks, as shown in Sec. 5. The type of dataset also plays an important role in the effectiveness of AL. While AL is usually most effective with large pools of unlabelled data that exhibit high diversity and uncertainty, the small dataset used in this study, coupled with the

nature of the labels, led to high-performance variance across each random initialization of the task network. For instance, in the whole tumor segmentation task, as tumors do not have a specific shape, the prediction was based on texture and contrast information. Similarly, in the Chest X-ray image classification task, the images looked relatively similar, with only subtle minute features distinguishing between classes. In addition, the evaluation test also plays a crucial role in accessing the AL sampler. If the test distribution is biased and does not contain diverse test cases, the effective evaluation of AL will be undermined. We conducted several Student's t-tests to evaluate the statistical significance of the performance results; nevertheless, we observed high variance across runs in downstream tasks when smaller AL budgets are used as shown by the error bars in Fig. 6, 3, and 5, as a result, the scores are not always statistically significant, given the large variance.

In the future, we plan to investigate this further by evaluating the benefit of multimodal AL on a larger pool of unannotated medical data on diverse multimodal datasets that guarantee a diversified distribution. Additionally, we aim to explore the potential of replacing the discriminator with metric learning [3] to contrast the labeled and unlabelled sets in the latent space. There is also a possibility of extending M-VAAL to other modalities. For instance, depth information can serve as auxiliary multimodal information to improve an AL sampler for image segmentation and label classification on surgical scenes with endoscopic images.

In this work, we proposed a task-agnostic sampling method in AL that can leverage multimodal image information. Our results on the BraTS2018 and COVID-QU-Ex datasets show initial promise in the direction of using multimodal information in AL. M-VAAL can consistently improve AL performance, but the hyperparameters need to be properly tuned.

Acknowledgements.. Research reported in this publication was supported by the National Institute of General Medical Sciences Award No. R35GM128877 of the National Institutes of Health, and the Office of Advanced Cyber Infrastructure Award No. 1808530 of the National Science Foundation. BB and DS are supported by the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) [203145Z/16/Z]; Engineering and Physical Sciences Research Council (EPSRC) [EP/P027938/1, EP/R004080/1, EP/P012841/1]; The Royal Academy of Engineering Chair in Emerging Technologies scheme; and the EndoMapper project by Horizon 2020 FET (GA 863146).

References

1. Al Khalil, Y., et al.: On the usability of synthetic data for improving the robustness of deep learning-based segmentation of cardiac magnetic resonance images. Med. Image Anal. **84**, 102688 (2023)
2. Ansari, M.Y., et al.: Practical utility of liver segmentation methods in clinical surgeries and interventions. BMC Med. Imaging **22**, 1–17 (2022)

3. Bellet, A., Habrard, A., Sebban, M.: A survey on metric learning for feature vectors and structured data. arXiv preprint arXiv:1306.6709 (2013)
4. Bouget, D., et al.: Meningioma segmentation in T1-weighted MRI leveraging global context and attention mechanisms. Front. Radiol. **1**, 711514 (2021)
5. Budd, S., et al.: A survey on active learning and human-in-the-loop deep learning for medical image analysis. Med. Image Anal. **71**, 102062 (2021)
6. Chen, X., et al.: Semi-supervised semantic segmentation with cross pseudo supervision. In: Proceedings IEEE Computer Vision and Pattern Recognition, pp. 2613–2622 (2021)
7. Gulrajani, I., et al.: Improved training of Wasserstein GANs. In: Advances in Neural Information Processing Systems 30 (2017)
8. Hamamci, A., et al.: Tumor-cut: segmentation of brain tumors on contrast-enhanced MR images for radiosurgery applications. IEEE Trans. Med. Imaging **31**, 790–804 (2011)
9. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 630–645. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_38
10. Kim, D.D., et al.: Active learning in brain tumor segmentation with uncertainty sampling, annotation redundancy restriction, and data initialization. arXiv preprint arXiv:2302.10185 (2023)
11. Laradji, I., et al.: A weakly supervised region-based active learning method for COVID-19 segmentation in CT images. arXiv:2007.07012 (2020)
12. Lewis, D.D.: A sequential algorithm for training text classifiers: corrigendum and additional data. In: ACM SIGIR Forum, vol. 29, pp. 13–19. ACM New York, NY, USA (1995)
13. Luo, X., et al.: Semi-supervised medical image segmentation through dual-task consistency. In: AAAI Conference on Artificial Intelligence, pp. 8801–8809 (2021)
14. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (BraTS). IEEE Trans. Med. Imaging **34**, 1993–2024 (2015)
15. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
16. Shao, W., et al.: Deep active learning for nucleus classification in pathology images. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 199–202 (2018)
17. Sharma, D., et al.: Active learning technique for multimodal brain tumor segmentation using limited labeled images. In: Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data: MICCAI Workshop 2019, pp. 148–156 (2019)
18. Singh, N.K., Raza, K.: Medical image generation using generative adversarial networks: a review. In: Patgiri, R., Biswas, A., Roy, P. (eds.) Health Informatics: A Computational Perspective in Healthcare. SCI, vol. 932, pp. 77–96. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-9735-0_5
19. Sinha, S., et al.: Variational adversarial active learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5972–5981 (2019)
20. Skandarani, Y., et al.: GANs for medical image synthesis: an empirical study. J. Imaging **9**(3), 69 (2023)
21. Tahir, A.M., et al.: COVID-QU-Ex Dataset (2022), https://www.kaggle.com/dsv/3122958

22. Thapa, S.K., et al.: Task-aware active learning for endoscopic image analysis. arXiv:2204.03440 (2022)
23. Verma, V., et al.: Interpolation consistency training for semi-supervised learning. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, pp. 3635–3641 (7 2019)
24. Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z.: Suggestive annotation: a deep active learning framework for biomedical image segmentation. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 399–407. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_46
25. Zhan, X., et al.: A comparative survey of deep active learning. arXiv:2203.13450 (2022)