Brain Attention Regularized Networks: Infection Diagnosis in Hydrocephalus CT Images

This paper was downloaded from TechRxiv (<https://www.techrxiv.org>).

LICENSE

CC BY 4.0

SUBMISSION DATE / POSTED DATE

10-02-2022 / 13-02-2022

CITATION

Yu, Mingzhao; Peterson, Mallory; Cherukuri, Venkateswararao; Hehnly, Christine; Mbabazi-Kabachelor, Edith; Mulondo, Ronnie; et al. (2022): Brain Attention Regularized Networks: Infection Diagnosis in Hydrocephalus CT Images. TechRxiv. Preprint. <https://doi.org/10.36227/techrxiv.19146362.v1>

DOI

[10.36227/techrxiv.19146362.v1](https://doi.org/10.36227/techrxiv.19146362.v1)

Brain Attention Regularized Networks: Infection Diagnosis in Hydrocephalus CT Images

Mingzhao Yu, Mallory R. Peterson, Venkateswararao Cherukuri, Christine Hehnly, Edith Mbabazi-Kabachelor, Ronnie Mulondo, Brian Nsubuga Kaaya, James R. Broach, Steven J. Schiff, and Vishal Monga

Abstract—Objective: Hydrocephalus is a medical condition characterized by an abnormal accumulation of cerebrospinal fluid (CSF) in the brain. Identification of postinfectious hydrocephalus (PIH) versus non-postinfectious hydrocephalus (NPIH), as well as the pathogen involved in PIH is crucial for developing an appropriate treatment plan. Unlike conventional classification tasks, this problem is particularly challenging as there is a great deal of overlap between the visual patterns that guide the classification of hydrocephalus images into PIH and NPIH on computed tomography (CT) scans. Moreover, the size and shape of the head vary significantly across different hydrocephalic patients, making it even more difficult to identify consistent features for classification. These challenges are only exacerbated for the task of identifying the pathogen within a PIH scan.

Method: state-of-the-art (SOTA) classification performance is achieved via deep convolutional neural networks (CNNs). However, deep learning often relies on generous training data and may produce class activations that are not physically meaningful. To address the aforementioned challenges, we first introduce a novel brain attention regularizer on a 2D DenseNet that forces the CNN to put more focus inside brain regions that are crucial for classification. Information from only 2D slices may not be sufficient to obtain reasonable performance. Therefore, we add a 3D CNN branch to the existing 2D CNN branch for capturing additional inter-slice information. Then a mutual attention regularization loss term is introduced to the training of the network which enables the two CNN branches to share information and puts more attention on important regions that are distributed between slices in a given CT stack. To incorporate this regularizer effectively, an alternative optimization strategy is employed to handle the stability issues that are common in training a 3D CNN. Since we introduce attention regularizers to brain image classification, we call our 2D CNN the brain attention regularized network (BAR-Net) and we refer to the hybrid 2D/3D CNN as the mutual brain attention regularized network (MBAR-Net).

Results: extensive evaluation of our method is carried out on a challenging real-world dataset obtained from the CURE Children's Hospital of Uganda (CCHU) in Mbale, Uganda. BAR/MBAR-Net demonstrates that it outperforms SOTA alternatives and exhibits graceful degradation as the number of training images is reduced, which is an important practical benefit. Furthermore, the class activation maps (CAMs) obtained from our method are more interpretable and help us understand the characteristics of PIH, NPIH, and pathogen positive images.

I. INTRODUCTION

A. Introduction to the Problem

Hydrocephalus is the leading indication for pediatric neuro-surgical care worldwide, with the majority of the pediatric hydrocephalus burden falling on the developing world [1]. Over half of the pediatric hydrocephalus cases in regions such as sub-Saharan Africa are PIH in nature [2], while the NPIH cases have etiologies such as hemorrhage or congenital malformations [1]. Etiologies such as congenital malformations in particular lead to distinct anatomic pathology in brain imaging. Infectious pathology can be more diverse in hydrocephalic patients, and is characterized by calcifications, abscesses, loculations, and debris in the ventricles (CSF filled cavities within the brain) [3]. Distinguishing PIH from NPIH is critical since a child with an active infection within the brain may not have clear signs of infection such as fever, yet performing surgery to divert the excess fluid of hydrocephalus should wait until the infection is properly treated with antimicrobials. Despite the different underlying pathophysiology behind these types of hydrocephalus, the major imaging pathology seen is dilation of the ventricles, and there are many other shared features such as edema within the brain tissue and enlarged overall head size.

In the developed world, magnetic resonance imaging (MRI) is used to diagnose the etiology of hydrocephalus and plan surgical treatment, but in the developing world, brain computerized tomography (CT) scans are the main imaging technology used for these purposes. The expertise of a medical professional is necessary to diagnose the underlying hydrocephalus etiology, so that the treatment plan can be developed appropriately. There are not sufficient medical professionals to manually diagnose in some developing regions that can hinder the process of diagnosis and treatment. Even in a setting with medical professionals who can diagnose PIH/NPIH, the infectious pathogen underlying PIH is extremely difficult to determine. A recent publication found that *Paenibacillus* bacteria most commonly leads to severe PIH in Uganda, making

This work was supported by NIH Grants 2R01HD085853, 1R01HD096693, 1U01NS107486 and NIH Director's Transformative Award 1R01AI145057.

M. Yu is with the Department of Electrical Engineering and the Center for Neural Engineering, The Pennsylvania State University, University Park, PA 16801 USA (email: mv5241@psu.edu).

M. R. Peterson is with the Center for Neural Engineering, The Pennsylvania State University.

V. Cherukuri is with the Department of Electrical Engineering and the Center for Neural Engineering, The Pennsylvania State University.

C. Hehnly is with Hershey Medical Center and College of Medicine. The Pennsylvania State University.

E. Mbabazi-Kabachelor, R. Mulondo, and B. N. Kaaya are with CURE Children's Hospital of Uganda.

J. R. Broach is with Department of Biochemistry and Molecular Biology, The Pennsylvania State University.

S. J. Schiff is with the Center for Neural Engineering and Infectious Disease Dynamics, the Departments of Neurosurgery, Engineering Science and Mechanics, and Physics. The Pennsylvania State University.

V. Monga is with the Department of Electrical Engineering, The Pennsylvania State University.

it a crucial pathogen to diagnose [3], which has been found to be correlated with more severe imaging pathology in terms of abscesses, calcifications, and other infectious processes [3]. Hence, this work aims to develop an automatic brain CT diagnosis system for PIH/NPIH with a *Paenibacillus* detection model which would alleviate the demand on medical experts and laboratories and improve the management of pediatric hydrocephalus. The flow chart in Fig. 1 illustrates the objective of our work: given CT images of a patient, we would like our classifier 1 to classify whether a patient has PIH or NPIH. CAMs are also desired for clinicians to know which regions contribute to the model's decision and which regions do not. If the outcome of the classification is PIH, the same CT images are sent to model 2 for pathogen classification i.e. *Paenibacillus*/non-*Paenibacillus*, with CAMs output for this decision as well. A few examples of CT images from NPIH

Fig. 1: Flow chart of classification framework for the hydrocephalus and pathogen diagnosis problem.

patients and PIH patients with and without *Paenibacillus* are shown in Fig. 2.

(a) Examples of CT scans for patients with PIH and *Paenibacillus*

(b) Examples of CT scans for patients with PIH and non-*Paenibacillus*

(c) Examples of CT scans for patients with NPIH

Fig. 2: Examples of CT images from the CURE dataset.

B. Related Work

Traditional computer-aided diagnosis systems (CADs) involve manually extracting features from the images and feeding them to subsequent classifiers such as support-vector machine (SVM) or logistic classifiers. In [4], the authors developed a thresholding technique to extract features from brain MRI images and applied a SVM to classify cerebral microbleeds. In [5], a rule-based classification method is employed on features extracted from images. These techniques have obvious drawbacks and are only compatible with very well studied pathogens or features. In recent years, deep learning-based methods have shown remarkable performance in computer vision tasks such as scene recognition [6,7], object detection [8,9], and image super-resolution [10,11]. Among these, deep CNNs are the most popular and successful architecture, which are trained to extract discriminative features and determine their mappings to the corresponding class from labeled training data in an end-to-end manner. CNNs have also been very successful in medical imaging applications, such as medical image classification [12,13], segmentation [14–16], super-resolution [17,18], and synthetic image generation [19,20]. In this paper, we leverage the advancement of CNNs for identifying PIH or NPIH in infant patients, as well as the underlying infectious pathogen by using CT brain images.

For problems related to medical image classification or detection, there are many existing works that can be approximately categorized as 2D, 3D, or hybrid approaches, respectively. As for 2D approaches, individual 2D slices are used as inputs to 2D CNN-based models. In [21], the authors proposed a 2D CNN for classifying brain tumors on MRI slices. In [22], the authors proposed a DenseNet interleaved with SENet [23] to classify breast cancer histopathological images. In [24], whole x-ray images of the chest are fed to a global branch, outputting attention maps which guide the extraction of local informative patches. A local branch is trained to classify patches. The features from the global and local branches are fused for the final prediction. Some works focus on utilizing information from more dimensions. In [25], a multi-plane model is proposed, in which a stack of images are decomposed into sagittal, coronal and axial plane slices in order to train multiple CNNs to detect lymph nodules from different planes. The decomposition of 3D volumetric images into 2D images enables the information along the z-axis to be learned, but there still may be a loss of information, which prompted the exploration of 3D CNN-based approaches. In [13], a plain and residual 3D CNN is developed to diagnose Alzheimers disease from MRI scans. Singh *et al.* [26] developed a shallow 3D CNN for detecting acute brain hemorrhage from CT scans. Training 3D CNNs is challenging due to stability issues. Using pre-trained models or reducing the input size are approaches that have been adopted to address this issue. The authors of [27] reshaped a pre-trained 2D CNN to a 3D corresponding one for diagnosis of Covid-19 from x-ray images. In [28], a 2D CNN and a 3D CNN are trained to classify 2D MRI slices and 3D patches separately and combine their predictions. Similarly, in [29], the authors proposed a 2-stage patch-wise 3D fully convolution

Fig. 3: Deep network CAMs, activations outside brain in (a).

network (FCN). In the first stage, a probability map of the 3D image is generated from the classification results of 3D patches extracted from the whole 3D image. In the second stage, the probability map is fed to a multi-layer perceptron for final predictions. 3D patch base methods still has drawbacks: 1.) network [29] is not trained in an end-to-end manner and 2.) global 3D context is not mined.

Motivation and Challenges: automated classification of hydrocephalic imagery faces the following open challenges:

1) The condition of hydrocephalus varies greatly between different patients in terms of imaging pathology, which makes it difficult for 2D or 3D CNNs to discriminate which feature is informative and which is not. In the absence of any domain knowledge, decision making is guided by training imagery and their geometric variations alone. A key drawback of such black-box learning can be observed by the phenomena of activated areas in the CAMs that occur far outside of the region of interest (the brain in this particular case) for deep models like ResNet [30] or DenseNet [31]. A few examples are shown in Fig. 3a. As we can see, large areas besides the brain are activated. Further, when there is limited training data available, the deep models may over-fit. A possible explanation for this phenomena is that the model “memorizes” the mapping from the training sample to its label during the training process to a certain degree, such that it learns to activate non-diagnostic details in the image rather than the actual pathological patterns.

2) Different etiologies of hydrocephalus often have many features in common, which hinders the ability of a CNN to find strong discriminative features between classes. While 2D CNNs can be trained more efficiently, the lack of knowledge from inter-slice information limits the CNN performance, especially in a setting with less discriminate pathology. On the other hand, 3D approaches, while capable of extracting inter-slice information from 3D volume data, are difficult to train [25] and computationally prohibitive for larger volumes.

To address the aforementioned challenges, we propose a novel brain attention regularized (BAR)-Net that guarantees activations to be inside the brain region. This is achieved by

a skull-stripped mask of the brain region, which in turn is used to generate a regularization term that encourages within brain activations during the training (learning) phase. Our design naturally exhibits inter-patient robustness since masks of different sizes and shapes get represented in the training. Furthermore, to incorporate 3D information, a 3D branch and a fusion block are added to the BAR-Net, which leads to a hybrid of 2D and 3D models. CT stacks and slices from same patient are fed into the 3D and 2D branch respectively, outputting feature vectors which are fused in the fusion block for final prediction. To enhance the learning of the 3D branch, we propose a mutual brain attention regularizer, which is inspired by the concept of knowledge distillation (KD) [32]. In KD, a student model learns knowledge from the extra-supervision provided by a teacher model in the form of prediction scores [32], intermediate feature maps, [33] or activation maps [34]. The motivation behind this technique is that the final or intermediate outputs of a capable model will contain useful information which can help a student model find and converge to good local minima. Particularly, a mutual learning strategy proposed by [35] allows two student models to learn from each other, which inspired us to develop the mutual brain attention regularization (MBAR). Conceptually, in MBAR two branches act as teacher model and student model alternatively and transfer specific aspects of the attention information to the other side via an alternating optimization scheme. Our hybrid 2D/3D model is named mutual brain attention regularized network (MBAR-Net).

The contributions of our work are summarized as follows:

- 1) **A novel 2D brain attention regularized network (BAR-Net) is proposed**, which encourages the network to focus inside the brain region for classification decisions. The newly proposed regularizer yields more interpretable class activations and experimental results confirm that the issue of over-fitting is mitigated.
- 2) **Extending the 2D BAR-Net, a hybrid 2-D/3-D CNN is proposed, that facilitates interactions between the 2D and 3D branches. Furthermore, a mutual attention regularization term is introduced** in order to enable branches to exchange knowledge learned from the other by transferring attention maps for regularization.
- 3) Extensive evaluation is carried out on a challenging real dataset obtained from the CURE Children’s Hospital of Uganda (CCHU).
- 4) The results obtained in this paper are fully reproducible and the code for BAR-Net and MBAR-Net is made publicly available at: <https://scholarsphere.psu.edu/resources/59e7b646-ae41-4d54-a38f-1d7a790bd92e>

II. BRAIN ATTENTION REGULARIZED NETWORK

A. Brain Attention Regularized Network

In this section, we will introduce the proposed brain attention regularized network (BAR-Net) for hydrocephalus and pathogen classification. We first explain the motivation, and then illustrate the architecture of the BAR-Net in section II-A. Finally, we will introduce the novel loss function and brain attention regularizer that are used to optimize the model.

Fig. 4: Architecture of brain attention regularized network, which consists of a main classification network and a branch for brain attention regularization.

1) **Motivation:** when standard deep CNN models like ResNet [30] or DenseNet [31] are used, there is always area outside the region of interest (i.e. the brain in our case) with higher activations, as demonstrated in Fig. 3a. Fig. 3a consists of superposition of CAMs (red area is activated most) over the corresponding input brain CT images. It is concerned that the activations outside of the brain region are uninterpretable, which could cause confusion for clinicians who are interested in understanding the decision-making process of deep networks. To address this issue, we propose a brain attention regularizer that suppresses the activations outside the brain region with the help of the brain region mask generated by a skull stripping algorithm [36].

2) **Structure of BAR-Net:** our proposed brain attention regularized network is illustrated in Fig. 4. The backbone of the BAR-Net is a DenseNet [31], which is composed of a 2D convolution layer, 4 2D dense blocks, 3 convolution block attention modules (CBAM) [37], and 3 transition blocks. Each dense block is comprised of several densely connected convolutional layers, a Rectified Linear Unit (ReLU) function, and a batch normalization layer. The CBAM consists of a channel attention branch and a spatial attention branch, which learn to pay more attention on important channel-wise and spatial-wise features. The transition block aims at adjusting the number of convolution filter channels in each block and reducing the spatial size of feature maps to increase the network's robustness to variation within the input images. Lastly, each of the final feature map is represented by its average value, forming a feature vector which is subsequently fed to a fully connected layer for classification. Meanwhile, CAMs [38] are generated by weighted averaging the model's final feature maps using the fully connected layer's parameters as the weights. The corresponding equation for the CAMs is shown in Eqn. (1)

$$CAM^i = \sum_{j=1}^C \mathbf{f}_j^m \cdot V_j^i, \quad (1)$$

where $\mathbf{f}^m$ represents the final convolution layer's feature maps in shape of $C \times H \times W$ and V^i are parameters in the fully connected layer corresponding to class i . The main objective

of this branch is to learn the final prediction scores via binary cross-entropy (BCE) loss $\mathcal{L}_{CE}$ given by:

$$\mathcal{L}_{CE}(W_{barn}) = - \sum_c y_c \log(\hat{\mathbf{p}}_c) \quad (2)$$

$$\hat{\mathbf{p}} = f_{barn}(I_{2D}), \quad (3)$$

where $f_{barn}(\cdot)$ is BAR-Net with parameters W_{barn} , I_{2D} is the 2D input image, and y_c is the ground truth label of the input image. $y_c = 1$ if c is the class of ground truth and $y_c = 0$ otherwise. $\hat{\mathbf{p}}_c$ is the predicted probability for the class c .

3) **Brain Attention Regularization:** since our goal is to prevent the model from over-fitting and to increase the model's robustness, it is important to restrict the model's activations that are not directly related to the discriminative patterns and therefore prompt the model to focus more attention on information inside the brain. First, we generate the brain region mask M from a given input image by skull stripping [36] and CAMs via Eqn. (1). Next, the mask M is down-sampled to M_d of size $H \times W$ in order to align with the CAM and the regularization loss is calculated as follows:

$$\mathcal{L}_{bar} = \sum_{i=1}^H \sum_{j=1}^W (\|(1 - M_d) \odot CAM\|_2^2)_{ij}, \quad (4)$$

where $\odot$ is (element-wise) Hadamard product.

The overall loss function for the BAR-Net is given by

$$\mathcal{L}_{barn} = \mathcal{L}_{CE}(W_{bar}) + \alpha \mathcal{L}_{bar}, \quad (5)$$

where α is a hyper-parameter balancing the weight of the regularization term. Note that our proposed $\mathcal{L}_{bar}$ is the summation of the activations in the non-brain region. By minimizing this regularization term in the learning, the activation outside brain is therefore suppressed.

In essence, BAR-Net is learning to classify 2D slices into PIH/NPIH or *Paenibacillus*/non-*Paenibacillus* with its activation map constrained inside the brain region. Fig. 3b illustrates the effect of this regularizer. As can be observed, the region outside the brain is not activated and the activation inside the brain becomes more focused, which results in more informative patterns and reduces the risk of over-fitting.

4) **Ensemble of Predictions:** note that BAR-Net is a 2D CNN model and each CT brain scan set contains multiple slices spanning from the base to the top of the head. Hence, the model is trained on individual 2D slices and an ensemble of multiple predictions determine the final label for that patient.

In the training phase, N_i slices from the i^{th} patient from the CT scan of the patient are extracted. In the testing phase, the probability scores of the BAR-Net on N_i slices from the i^{th} patient are averaged as the ensemble of the predictions. By doing so, in a statistical sense, the predictions will have less variance and hence have a better generalization ability. However, even though multiple slices are used, an obvious shortcoming of this 2D ensemble strategy is the lack of inter-slice correlation. To overcome this, we further propose an extension structure including 3D information extraction, which is illustrated in Sec. II-B.

B. Mutual Brain Attention Regularized Network

1) **Motivation:** BAR-Net is capable of extracting 2D features efficiently but it is limited by the lack of ability to learn 3D contextual information, which is crucial for the classification of hydrocephalus and pathogens. 3D CNNs [13,39,40] are known to address this issue, but they are also known to suffer from stability issues, particularly in the face of limited training data.

In short, a 2D model is efficient in training with a lack of 3D information while a 3D model possesses inter-slice information but suffers from training inefficiency. Motivated by the complementary properties of 2D and 3D models and the fact that 2D slices are subsets of the 3D stack, a fusion of 2D and 3D models, and a mutual regularization between 2D and 3D models are proposed to address these issues. The fusion model, called mutual brain attention regularized network (MBAR-Net), aims at leveraging both the 2D and 3D information by concatenating features extracted through the 2D and 3D branches. At the same time, inspired by mutual learning [35], we propose to prompt each branch model to mutually transfer their knowledge alternatively via a mutual brain attention regularizer during the training process. The details are reported in Sec. II-B4

2) **Structure of 2D and 3D branch:** The architecture of MBAR-Net consists of a 2D branch, 3D branch, and a fusion block, which is shown in Fig. 5.

The 2D branch is BAR-Net, while the 3D branch is composed of four 3D dense blocks, three 3D transition blocks followed by a global average pooling, a fully connected layer, and a softmax layer. The 3D dense blocks and 3D transition blocks are in the same structure as those in the 2D branch with the 2D convolution layer, 2D batch normalization and 2D pooling operations replaced by 3D counterparts. The 3D branch has the same number of layers as the 2D branch in order to ensure that they have the same receptive field in the activation maps.

In the feed-forwarding process, a stack of 16 slices I_{3D} is fed into the 3D branch, which outputs probability scores, feature vectors, and CAMs. The feature vectors are the output of a global average pooling layer following the final convolutional layer. Note that the 3D branch outputs 2-D CAMs due to several 3D pooling operations. In spite of this, they activate discriminative information of a 3D flavor, i.e. across 16 input slices. At the same time, 2D slices, I_{2D} , extracted from the 3D cube, I_{3D} , are fed into the 2D branch, outputting probability scores, CAMs and feature vectors. The details of the receptive field relationship between 2D and 3D branches are illustrated in Sec. II-B4. Both branches are optimized by minimizing a binary cross-entropy loss between the prediction and ground truth label

$$\mathcal{L}_{CE}(W_{2D}) = - \sum_c y_c \log((\hat{\mathbf{p}}_{2D})_c) \quad (6)$$

$$\mathcal{L}_{CE}(W_{3D}) = - \sum_c y_c \log((\hat{\mathbf{p}}_{3D})_c) \quad (7)$$

$$\hat{\mathbf{p}}_{2D} = f_{2D}(I_{2D}) \quad (8)$$

$$\hat{\mathbf{p}}_{3D} = f_{3D}(I_{3D}), \quad (9)$$

where I_{2D} and I_{3D} are 2D image and 3D image respectively. $y_c = 1$ if c is the class of ground truth, and $y_c = 0$ otherwise. $f_{2D}(\cdot)$ and $f_{3D}(\cdot)$ are 2D and 3D branches with parameters W_{2D} and W_{3D} .

3) **Fusion Block:** The input to the fusion block is a feature vector $\mathbf{f}^v$ which is a concatenation of $\mathbf{f}_{2D}^v$ and $\mathbf{f}_{3D}^v$. Both are generated via global averaging pooling (GAP) feature maps $\mathbf{f}_{2D}^m, \mathbf{f}_{3D}^m$ from 2D/3D branches. The steps for generating $\mathbf{f}^v$ are

$$\mathbf{f}^v = \text{concatenate}(\mathbf{f}_{2D}^v, \mathbf{f}_{3D}^v) \quad (10)$$

$$\mathbf{f}_{2D}^v = \text{GAP}(\mathbf{f}_{2D}^m), \quad \mathbf{f}_{3D}^v = \text{GAP}(\mathbf{f}_{3D}^m), \quad (11)$$

where

$$\text{GAP}(\mathbf{f}^m) = \frac{1}{H \times W} \sum_i \sum_j \mathbf{f}^m(:, i, j), \quad (12)$$

and $\mathbf{f}^m$ is the set of feature maps in shape of $C \times H \times W$.

The mapping from concatenated features to ground truth are learned through two fully connected layers, which are optimized by minimizing the loss function Eqn. (14)

$$\mathcal{L}_{CE}(W_{fuse}) = - \sum_c y_c \log(\hat{\mathbf{p}}_c) \quad (13)$$

$$\hat{\mathbf{p}} = f_{fuse}(\mathbf{f}^v), \quad (14)$$

where $f_{fuse}(\cdot)$ is the fusion module with parameters W_{fuse} .

4) **Mutual brain attention regularization:** The 2-D BAR-Net can leverage pre-trained networks and regularized learning is feasible using brain attention masks applied to 2-D class activations. However, the 2-D BAR-Net is trained to learn patterns that only exist in the 2D plane, which inevitably lacks 3D context information. The alternative is to learn 3-D networks which naturally process volumetric image data but they suffer from convergence issues during the training process when the input volume is large in size and no specific pre-training or extra supervision is provided. Hence, a mechanism that is able to combine the advantages of both of these methods and overcome their respective disadvantages is desirable. The relationship between the receptive fields of the 2D branch and the 3D branch is demonstrated in Fig. 6. The stack of slices in the right represents the I_{3D} , which is the input to the 3D branch, while the slice in the left represents a 2D slice I_{2D} extracted from the stack. Two grids above the slices are original CAMs from 2D and 3D branches.

When specific discriminative features exist in the 2D slice, the corresponding regions in the 2D branch's CAM are activated. Since the 2-D slice is a subset of the 3-D stack, the 3D branch will be activated as well. Hence, both the $unit_a^{3D}$ and $unit_a^{2D}$ are activated. *The 2D branch therefore guides the 3D branch about which regions to activate.* However, the 2D branch is not supposed to guide the 3D branch about the non-activation, since information from adjacent slices (parallelograms in slices in Fig. 6) is not available in the 2D branch, but is within 3D branch's receptive field. *Conversely, the 3D branch can guide the 2D branch about where not to activate.* This interplay of activation and non-activation is achieved via masks M_a and M_{na} as illustrated in Fig. 6.

Ideally, the 2D and 3D branch should reach an agreement about their corresponding activations after sufficient training

Fig. 5: Architecture of mutual brain attention regularized network. It consists of a 2D branch (i.e. BAR-Net), a 3D branch and a fusion block. The CAMs of two branches are mutually regularized during training.

Fig. 6: Demonstration of the relationship between receptive fields of the 2D and 3D branches. The slice in the left represents input slice of 2D branch extracted from the stack of slices in the right which is input to the 3D branch. The parallelograms in different color on the slice and stack represent discriminative features. The grids above the parallelogram are activation maps of the 2D branch and the 3D branch. $Unit_a$ and $Unit_b$ are units in the activation maps.

so that the 3D model should activate the regions where the 2D model activates, and the 2D model should not activate any units where the 3D model does not activate. Based on this

logic, we propose a mutual attention learning scheme where two branches play the role of teacher model alternatively via transferring the specific information of each activation map

through M_a and M_{na} to the other branch. At each learning stage, one branch will regularize the other to be consistent with its advantageous perspective. Next, the details of optimization step are described.

Algorithm 1: Optimization for Mutual Brain Attention Regularized network (MBAR-Net)

Input: input set I , label set Y , brain region mask M_a , number of epochs N_e , number of iteration for one optimization step N_{iter}
Output: 2D branch $f_{2D}(\cdot)$, 3D branch $f_{3D}(\cdot)$, fusion block $f_{fuse}(\cdot)$
while $Epochs \leq N_e$ **do**
 0.0 obtain a batch of 3D images I_{3D}^b ;
 0.1 obtain a batch of 2D images I_{2D}^b by extracting slices from each 3D image in I_{3D}^b ;
 while iterating all batches **do**
 if iteration % $2N_{iter} \leq N_{iter}$ **then**
 1.0 output CAM_{3D} , $\mathbf{f}_{3D}^v$, $\hat{p}_{3D}$ from $f_{3D}(I_{3D}^b)$;
 1.1 output CAM_{2D} , $\mathbf{f}_{2D}^v$ from $f_{2D}(I_{2D}^b)$;
 1.2 update $f_{3D}(\cdot)$ by optimizing (21) with 2D branch fixed;
 1.3 update $f_{fuse}(\cdot)$ by minimizing (14);
 else
 2.1 output CAM_{2D} , $\mathbf{f}_{2D}^v$, $\hat{p}_{2D}$ from $f_{2D}(I_{2D}^b)$;
 2.2 output CAM_{3D} , $\mathbf{f}_{3D}^v$ from $f_{3D}(I_{3D}^b)$;
 2.3 update $f_{2D}(\cdot)$ by optimizing (17);
 2.4 update $f_{fuse}(\cdot)$ by minimizing (14);
 end
 end
end

Optimization step of 2D branch: in this step, the 3D branch is temporally considered as the teacher model, which transfers its knowledge about non-activated regions to the 2D branch because there should be no activation in those particular regions for the 2D model either if it finds no useful information over the entire 3D receptive area. Therefore, a non-activation mask is first generated via a soft gating function, which is shown in Eqn. (15)

$$M_{na}(i, j) = 1 - \frac{1}{1 + \exp(-w_{3D} \cdot CAM_{3D}^{gt}(i, j))}, \quad (15)$$

where $M_{na}(i, j)$ is the pixel in the non-activation mask at position (i, j) , CAM_{3D}^{gt} is the activation map of ground truth class from the 3D branch, and w_{3D} is the weight controlling the slope of the gating function. The non-activation mask M_{na} is therefore designed to achieve higher values (closer to 1) where no activations are desired. Then, with CAM_{3D} fixed, the mask M_{na} is element-wise multiplied by the absolute value of difference between CAM_{3D} and CAM_{2D} . The sum is used as the loss term for regularization, which is

$$\mathcal{L}_{mbar,2D} = \sum_{i=1}^H \sum_{j=1}^W (M_{na} \odot \|(CAM_{3D} - CAM_{2D})\|_1)_{ij}, \quad (16)$$

where CAM_{3D} is fixed.

The overall loss function for the 2D branch is

$$\mathcal{L}_{2D} = \mathcal{L}_{CE}(W_{2D}) + \alpha_1 \mathcal{L}_{mbar,2D} + \alpha_2 \mathcal{L}_{bar,2D}, \quad (17)$$

where $\mathcal{L}_{bar,2D}$ is the brain attention regularizer in Eqn. (4).

Optimization step of 3D branch: in a similar manner, when the 2D branch acts as the teacher model, the region activated by the 2D branch should be activated by the 3D region as the 2D region is a subset of 3D region.

An activation mask indicating the region activated by the 2D branch is first generated as

$$M_a(i, j) = \frac{1}{1 + \exp(-w_{2D} \cdot CAM_{2D}^{gt}(i, j))}, \quad (18)$$

where CAM_{2D}^{gt} is the activation map of the ground truth class of the 2D branch and w_{2D} is the parameter controlling the slope of the function Eqn. (18). A larger positive value of w_{2D} will make the output of the masking function change rapidly around the threshold.

Regularization of the CAM_{3D} is constructed as

$$\mathcal{L}_{mbar,3D} = \sum_{i=1}^H \sum_{j=1}^W (M_a \odot \|CAM_{3D} - CAM_{2D}\|_1)_{ij}, \quad (19)$$

where CAM_{2D} is fixed.

In addition, the brain attention regularizer is applied here for constraining the over-fitting issue and enhancement of the attention inside the brain. With the constraint by $\mathcal{L}_{bar,2D}$ and $\mathcal{L}_{bar,3D}$, the knowledge transferred between the 2 branches is guaranteed to contain valuable information rather than meaningless activations. The $\mathcal{L}_{bar,2D}$ is defined in Eqn. (4) and $\mathcal{L}_{bar,3D}$ is defined as

$$\mathcal{L}_{bar,3D} = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W ((1 - \bigcup_{k=1}^{16} M_d^k) \odot \|CAM_{3D}\|_1)_{ij}, \quad (20)$$

where $\bigcup_{k=1}^{16} M_d^k = 1 - \prod_{k=1}^{16} (1 - M_d^k)$, which is the union of masks of 16 CT slices in the I_{3D} .

The overall loss function of the 3D branch is

$$\mathcal{L}_{3D} = \mathcal{L}_{CE}(W_{3D}) + \alpha_3 \mathcal{L}_{mbar,3D} + \alpha_4 \mathcal{L}_{bar,3D}, \quad (21)$$

where α_1 and α_2 are parameters manually selected for the purpose of balancing the significance of the regularization term's influence on the entire optimization process. The optimization steps are summarized in Algorithm 1.

stack of CT images from one patient

Fig. 7: Demonstration of ensemble mechanism for inference.

5) *Ensemble of Predictions*: Because input to the 3D branch is a stack of 16 CT slices, for each patient's CT image set, N stacks are extracted for training and inference. In the inference stage, N stacks (and corresponding slices for the 2-D branch) are input to the MBAR-Net, which in turn outputs N sets of probability scores from the fusion block, which are averaged for the final prediction. The demonstration of the overall inference process is illustrated in Fig. 7.

III. EXPERIMENTAL RESULTS

A. Dataset

1) *Dataset*: we collected challenging real-world images from the CURE Children's Hospital of Uganda [3] that include CT scans from 383 infants labeled as PIH and NPIH, *Paenibacillus* and non-*Paenibacillus* by medical experts and genomic testing [41]. Each scan consists of multiple axial plane slices (34 slices in average) in the shape of 512×512 . All scans are stored in DICOM file format. Furthermore, the proportion of patients with PIH and those with NPIH is shown in Table I(a). The proportion of *Paenibacillus* within the PIH cohort is shown in Table I(b).

PIH	NPIH	<i>Paenibacillus</i>	non- <i>Paenibacillus</i>
202/383	181/383	96/202	106/202
(a)		(b)	

TABLE I: (a) Proportion of PIH verses NPIH (b) proportion of *Paenibacillus* verses non-*Paenibacillus* within the PIH cohort

2) *Pre-processing*: masks of the brain region are generated via a skull stripping algorithm [36], which sets the intensity of pixels outside the brain to be zero. These masks are utilized for the brain attention regularizer in the Eqn. (4). Example masks are shown in Fig. 8.

Fig. 8: Illustration of CT images and their corresponding brain region masks generated from skull stripping algorithm.

The slices of each patient are stacked to form a 3D image in the shape of $D \times 512 \times 512$ in which D is variable for different patient. Each 3D image is resized to $32 \times 512 \times 512$ via nearest neighbor interpolation.

B. Experimental Setting

1) *Parameter Selection*: some parameters in Eqn. (3), Eqn. (18), Eqn. (15), Eqn. (9) and Alg. 1 are selected carefully via cross-validation [42]. The best values are found to be $\alpha = 0.00001$, $\alpha_1 = 0.000005$, $\alpha_2 = 0.000005$, $\alpha_3 = 0.00005$, $\alpha_4 = 0.00005$, $w_{2D} = 4$, $w_{3D} = 4$, $N_i = 50$.

2) *Training-validation Configuration*: For classification of PIH vs. NPIH, 5 fold cross-validation is conducted on 383 patients. At each fold, 80% of the scans are assigned for training and 20% are assigned for validation. As for classification of *Paenibacillus* vs. non-*Paenibacillus*, 202 scans with PIH are used for the same 80%/20% 5 fold cross-validation scheme.

3) *Data Augmentation*: Several data augmentation operations are adopted, including flipping along each axis, randomly rotating -20° to 20° in the axial plane, cropping and resizing, and adding Gaussian noise. The operation is randomly selected at each iteration's batch loading phase. The 2D slice is extracted from the augmented 3D sample. Thus, the spatial information in the 2D and 3D samples are in correspondence.

4) *Implementation Setting*: We optimize the models using Adam [43] with a mini-batch size of 32, 8, and 8 for the 2D branch, 3D branch, and fusion block, respectively. We train each branch for 100 epochs. The initial learning rate is set to be 0.0001 and divided by 2 after every 30 epochs. We use weight decay of 0.000001 and a momentum of 0.9. The models and training procedures are implemented within the Pytorch framework [44] and all of the computation is performed on a NVIDIA TITAN X GPU.

C. Evaluation Metrics

We use accuracy, sensitivity and specificity to evaluate the performance of the models, which are defined as:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (22)$$

$$sensitivity = \frac{TP}{TP + FN} \quad (23)$$

$$specificity = \frac{TN}{TN + FP} \quad (24)$$

where TP , TN , FP , FN represent the number of true positive samples, true negative samples, false positive samples, and false negative samples.

In addition, the receiver operating characteristic (ROC) curve and area under the curve (AUC) are presented.

We evaluate all competing methods on the CURE dataset 3 times independently with 5-fold cross validation. First, we analyze the effectiveness of the components in BAR-Net and MBAR-Net. Second, we evaluate BAR-Net with MBAR-Net on classification of PIH/NPIH and *Paenibacillus*/non-*Paenibacillus* and compare the results against the state-of-the-art (SOTA) methods. Finally, we illustrate the impact of MBAR on the models by presenting and analyzing activation maps of the 2D and 3D branches in MBAR-Net.

D. Ablation Study

Here we study the effectiveness of the components in both BAR/MBAR-Net. The results are shown in Tables II and III.

First, we study the impact of the brain attention regularizer $\mathcal{L}_{bar}$ in BAR-Net by evaluating the performance of using $\mathcal{L}_{CE}(W_{bar})$ alone against that of applying the regularizer under the training-validation configuration in Section III-B2 and another low-training setup where 40% of the available training is used. The low training setting is particularly important to examine generalizeability, i.e. the ability of adapting to new test samples and datasets beyond the training data. Table II shows that brain attention regularization ($\mathcal{L}_{bar}$) in BAR-Net offers accuracy improvements that are particularly pronounced when the training set is smaller. A comprehensive evaluation of BAR/MBAR-Net against SOTA methods for varying training regimes is discussed in Sec. III-E3.

Loss function		Size of training set		Accuracy	Gain
$\mathcal{L}_{CE}(W_{bar})$	$\mathcal{L}_{bar}$	100%	40%		
✓		✓		93.5 %	
✓	✓	✓		93.9 %	+ 0.4%
✓			✓	88.1 %	
✓	✓		✓	91.7 %	+3.6%

TABLE II: Brain attention regularization’s impact on BAR-Net’s performance on classification of PIH/NPIH.

Second, we evaluate the impact of the number of input slices to the 3D branch in the MBAR-Net. It is shown in Table III that inputting 16 slices to the 3D branch and making predictions based on ensembles of each prediction from the extracted 16 slices achieves better performance. Using 16 slices has two obvious advantages: first, avoiding using a large number of slices for the 3D input mitigates the training difficulty for 3D models; second, the dimension of the z-axis will be downsized to 1 if 16 slices are used as an input to most of the SOTA models. For our case, it is convenient to implement mutual attention regularization if CAMs from both branches are in the same size.

Third, we study the effectiveness of mutual attention regularization in the hybrid model. We first evaluate the performance without using any regularization in MBAR-Net. Then, we evaluate the performance using $\mathcal{L}_{mbar,2D}$, and $\mathcal{L}_{mbar,3D}$ respectively and jointly. The results are shown in Table III, which show that either using one of the attention regularizers or both of them boosts the performance.

MBAR-Net			regularizer		
2D branch	3D branch		$\mathcal{L}_{mbar,2D}$	$\mathcal{L}_{mbar,3D}$	Accuracy
	16 slices	32 slices			
	✓				90.9%
		✓			90.1%
✓	✓				94.2%
✓	✓		✓		94.2%
✓	✓			✓	94.2%
✓	✓		✓	✓	94.7%

TABLE III: Ablation study of MBAR-Net’s performance on classification of PIH/NPIH. Effect of size of input to the 3D branch and mutual brain attention regularizer are studied.

E. Comparisons Against SOTA methods

We compare against carefully selected SOTA methods, which are introduced here:

a. Korolev *et al.* [13]: SOTA method that uses 3D networks for brain MRI images. A mix of plain and residual 3D convolutional networks are used.

b. Gao *et al.* [28]: Deep network for CT brain image classification that separately designs and fuse 2D and 3D networks.

c. Qiu *et al.* [29]: This method also trains a 3D CNN to classify 3D patches, which are extracted from the whole 3D brain MRI volume. The 3D CNN outputs a probability map; a multi-layer perceptron is constructed, which assigns the 3D probability map to the ground truth label.

d. Li *et al.* [22]: DenseNet [31] based method for classification of histopathology images but is broadly applicable.

1) Evaluation of classification of PIH/NPIH: We first evaluate the performance of our proposed methods on PIH verses NPIH classification against state-of-the-art alternatives. The mean and standard deviation of accuracy, sensitivity, specificity results are shown in Table IV. The receiver operating characteristic curve (ROC curve) is shown in Fig. 9 (a) and corresponding area under the curve (AUC) is reported in the last column of Table IV.

Since no extra-information or pre-training is provided for the training of the 3D model, the 3D model proposed by [13] in Table IV does not achieve compelling results.

Overall, our proposed BAR-Net and MBAR-Net outperform state-of-the-art methods on classification of PIH/NPIH in terms of accuracy, sensitivity, specificity, and AUC. First, the results showed that the 3D method may not necessarily lead to better performance primarily due to over-fitting and training stability issues. 2D methods with a powerful architecture and careful regularization can achieve high accuracy. Second, the results confirm that a joint 2-D/3-D design with a composite loss function is an effective strategy.

In addition, the 5-fold cross validation experiments are conducted independently 3 times for an Analysis of Variance (ANOVA) study [45]¹ over two groups of classification accuracy observations - the proposed MBAR-Net and the results of Li *et al.* [22]. The 2-group ANOVA results are shown in Fig. 10. The F-value and p-value calculated in the ANOVA study for methods on classification of PIH/NPIH are 9.881 and 0.0056 respectively. The F-value and p-value combined with Fig. 10 confirm that the difference in mean accuracy of MBAR-Net and Li *et al.* [22] is statistically significant.

2) Evaluation of detection of *Paenibacillus* within PIH cohort: The evaluation results of classification of *Paenibacillus* verses non-*Paenibacillus* are reported in Table V. We can observe from Table V that the task of identification of the pathogen is much more challenging as revealed by relatively lower classification accuracy compared to Table IV. This is to be expected because this task involves a more fine-grained classification within the PIH cohort of patients.

3) Performance under different percentages of training samples: A well-designed and robust machine learning method can not only exploit abundant training but its performance should drop gracefully as the number of training images is reduced. To examine training robustness, we report

¹Prior to application of ANOVA, we rigorously verified that the observations (accuracy) satisfy ANOVA assumptions [46].

Methods	Accuracy(%)	Sensitivity(%)	Specificity(%)	AUC
Korolev <i>et al.</i> [13](3D CNN)	85.8 $\pm$ 4.37	85.4 $\pm$ 5.43	86.2 $\pm$ 6.21	0.90 $\pm$ 0.018
Gao <i>et al.</i> [28](fusion of 2D/3D CNN)	89.5 $\pm$ 4.41	87.2 $\pm$ 4.35	90.2 $\pm$ 6.23	0.92 $\pm$ 0.016
Qiu <i>et al.</i> [29](patch 3D CNN+MLP)	91.4 $\pm$ 2.26	89.6 $\pm$ 2.41	93.4 $\pm$ 3.12	0.94 $\pm$ 0.012
Li <i>et al.</i> [22](2D DenseNet+SE)	92.2 $\pm$ 1.99	92.1 $\pm$ 2.56	92.4 $\pm$ 2.98	0.94 $\pm$ 0.013
BAR-Net	93.9 $\pm$ 1.55	92.4 $\pm$ 1.98	95.1 $\pm$ 2.54	0.96 $\pm$ 0.012
MBAR-Net	94.7 $\pm$ 1.67	94.1 $\pm$ 2.03	95.2 $\pm$ 2.67	0.97 $\pm$ 0.011

TABLE IV: Classification result of PIH verses NPIH on CURE dataset.

Methods	Accuracy(%)	Sensitivity(%)	Specificity(%)	AUC
Korolev <i>et al.</i> [13](3D CNN)	76.5 $\pm$ 4.57	68.2 $\pm$ 6.67	84.3 $\pm$ 6.42	0.72 $\pm$ 0.020
Gao <i>et al.</i> [28](fusion of 2D/3D CNN)	77.6 $\pm$ 4.10	72.3 $\pm$ 6.32	83 $\pm$ 7.65	0.75 $\pm$ 0.021
Qiu <i>et al.</i> [29](patch 3D CNN+MLP)	78.2 $\pm$ 3.01	72.4 $\pm$ 5.14	83.6 $\pm$ 6.87	0.79 $\pm$ 0.019
Li <i>et al.</i> [22](2D DenseNet+SE)	80 $\pm$ 3.48	73 $\pm$ 5.45	86.1 $\pm$ 6.99	0.80 $\pm$ 0.016
BAR-Net	81.8 $\pm$ 4.21	73.8 $\pm$ 6.18	87.9 $\pm$ 8.03	0.83 $\pm$ 0.015
MBAR-Net	82.3 $\pm$ 5.94	74.5 $\pm$ 6.39	89.1 $\pm$ 8.55	0.83 $\pm$ 0.015

TABLE V: Classification result of *Paenibacillus* verses non-*Paenibacillus* on CURE dataset.

Fig. 9: Comparisons of ROC curves output from BAR/MBAR-Net against state-of-the-art methods.

Fig. 10: Result of ANOVA study for Li *et al.* [22] and MBAR-Net on classification of PIH/NPIH. The F-value and p-value calculated in the study are 9.881 and 0.0056

accuracy results for various percentages of available training samples. The training data utilized in this experiment varies from 40% - 100% of the original training size. The results are

shown in Fig. 11. As can be observed, BAR-Net and MBAR-Net exhibit graceful degradation with respect to training size compared to other SOTA methods. This is because our proposed regularizers $\mathcal{L}_{bar}$ in Eqns. (4) and (5) exploit expected output characteristic of CAMs, which are independent of the choice of training samples thereby improving robustness to the quantity and quality of available training images.

4) Impact of MBAR on CAMs: The impact of MBAR on the output activation maps is demonstrated in Fig. 12. The results from the classification of PIH versus NPIH are shown in the left 2 columns while the right 2 columns are results from the classification of *Paenibacillus* versus non-*Paenibacillus*. The first row in Fig. 12 shows the input CT images. The second row shows the CAMs of Li *et al.*'s method [22]. The third row shows the results of BAR-Net. The fourth and the last row are results from the 2D branch and 3D branch of MBAR-Net respectively. Note that even though the activation map from the 3D branch is a 2D map, its receptive field is actually in 3D which covers the local space along the z-axis. Thus, it is possible that the activation shown in the 3D branch's CAM is not only influenced by the present slice but instead by adjacent slices in its vicinity. For BAR and MBAR-Net, first we notice that all the activations are concentrated inside the brain region with the help of regularization. Since no mask is provided as

Fig. 11: Illustration of BAR/MBAR-Nets’ and SOTA methods’ classification performance under variable training size. The proportion of training data in the whole dataset ranges from 0.32 to 0.8. Note that BAR/MBAR-Net offer significantly better generalizability owing to our custom designed regularizers, as evidenced by a greater than 5% accuracy gain over SOTA in the low-training regime.

input in the inference phase, it indicates that the model learns to distinguish the brain from non-brain and inactivate the units outside the region of interest in activation maps. By doing so, the model will learn to focus on reliable discriminative regions. Secondly, we can observe that the CAMs from BAR-Net and MBAR-Net are significantly more interpretable than those from SOTA methods but non-domain enriched approach such as Li *et al.* [22]. In particular, BAR/MBAR-Net are often able to highlight the pathological features that neuroscience experts may look for in their decision making. As an example in Fig. 12 a loculation is highlighted by the BAR-Net as shown in column 1, row 3 and calcifications are highlighted by MBAR-Net as shown in row 5. Finally, when comparing activation maps of 2D and 3D branch, we observe that there are some overlapping regions indicating that the two branches reach agreement on certain features. However, there are also some non-overlapping activations, which indicates that the two branches still capture diverse information due to varied inputs.

IV. DISCUSSIONS AND CONCLUSION

In this paper, we present two deep learning-based methods for classification of hydrocephalic brain images into PIH/NPIH and identify the responsible pathogen in PIH scans. In particular, an attention regularizer is introduced which ensures that the network focuses its attention inside the brain region. Moreover, to utilize 3D context information and to reduce the risk of over-fitting in the 2D BAR-Net, we propose a mutual attention regularized network that fuses a 2D branch and a 3D branch and mutually regularizes two branches in an alternative fashion to improve the stability of the network. The performance of the proposed BAR-Net and MBAR-Net is

Fig. 12: The CAMs of Li *et al.*’s method [22], BAR-Net, 2D branch and 3D branch of MBAR-Net are shown. The CT image of each column is from PIH, NPIH, *Paenibacillus*, and non-*Paenibacillus* from left to right.

validated on real images against several SOTA methods. Currently, the classification of PIH/NPIH and *Paenibacillus*/non-*Paenibacillus* are conducted separately by two networks. In the future, combining two models into one via multi-task learning or transfer learning is worth further investigation. Finally, while our *domain enriched* approach produces more reliable activation maps than SOTA methods, obtaining ‘optimal’ and fully interpretable activations remains an open challenge. A particularly interesting direction is where activations predicted by a deep network may be treated as feedback to the expert neuroscientist, such that collaborative interactions between the algorithm and human expert can be facilitated. Achieving the above may involve newer learning frameworks and forms an exciting future research frontier.

REFERENCES

- [1] M. C. Dewan, B. C. Warf *et al.*, “Global hydrocephalus epidemiology and incidence: systematic review and meta-analysis,” *Journal of Neurosurgery*, vol. 130, no. 4, pp. 1065 – 1079, 2019.
- [2] B. C. Warf, “Comparison of endoscopic third ventriculostomy alone and combined with choroid plexus cauterization in infants younger than 1 year of age: a prospective study in 550 african children,” *Journal of Neurosurgery: Pediatrics*, vol. 103, no. 6, pp. 475–481, 2005.
- [3] J. N. Paulson, B. L. Williams, C. Hehny, N. Mishra, S. A. Sinnar, L. Zhang, P. Ssentongo, E. Mbabazi-Kabachelor, D. S. Wijetunge, B. Von Bredow *et al.*, “*Paenibacillus* infection with frequent viral coinfection contributes to postinfectious hydrocephalus in ugandan infants,” *Science translational medicine*, vol. 12, no. 563, 2020.

- [4] S. R. Barnes, E. M. Haacke, M. Ayaz, A. S. Boikov, W. Kirsch, and D. Kido, "Semiautomated detection of cerebral microbleeds in magnetic resonance images," *Magnetic resonance imaging*, vol. 29, no. 6, pp. 844–852, 2011.
- [5] T. Chan, "Computer aided detection of small acute intracranial hemorrhage on computer tomography of brain," *Computerized Medical Imaging and Graphics*, vol. 31, no. 4-5, pp. 285–298, 2007.
- [6] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," 2014.
- [7] L. Wang, S. Guo, W. Huang, Y. Xiong, and Y. Qiao, "Knowledge guided disambiguation for large-scale scene classification with multi-resolution cnns," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 2055–2068, 2017.
- [8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.
- [9] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 11, pp. 3212–3232, 2019.
- [10] T. Guo, H. S. Mousavi, and V. Monga, "Adaptive transform domain image super-resolution via orthogonally regularized deep networks," *IEEE Transactions on Image Processing*, vol. 28, no. 9, 2019.
- [11] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional neural networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015.
- [12] J. Zhang, Y. Xie, Q. Wu, and Y. Xia, "Medical image classification using synergic deep learning," *Medical image analysis*, vol. 54, 2019.
- [13] S. Korolev, A. Safiullin, M. Belyaev, and Y. Dodonova, "Residual and plain convolutional neural networks for 3d brain mri classification," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE, 2017, pp. 835–838.
- [14] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 2018, pp. 3–11.
- [15] V. Cherukuri, V. K. Bg, R. Bala, and V. Monga, "Deep retinal image segmentation with regularization under geometric priors," *IEEE Transactions on Image Processing*, vol. 29, pp. 2552–2567, 2019.
- [16] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: achievements and challenges," *Journal of digital imaging*, vol. 32, no. 4, pp. 582–596, 2019.
- [17] V. Cherukuri, T. Guo, S. J. Schiff, and V. Monga, "Deep mr brain image super-resolution using spatio-structural priors," *IEEE Transactions on Image Processing*, vol. 29, pp. 1368–1383, 2019.
- [18] S. Zhang, G. Liang, S. Pan, and L. Zheng, "A fast medical image super resolution method based on deep learning network," *IEEE Access*, vol. 7, pp. 12 319–12 327, 2018.
- [19] K. Armanious, C. Jiang, M. Fischer, T. Küstner, T. Hepp, K. Nikolaou, S. Gatidis, and B. Yang, "Medgan: Medical image translation using gans," *Computerized Medical Imaging and Graphics*, vol. 79, 2020.
- [20] D. Nie, R. Trullo, J. Lian, L. Wang, C. Petitjean, S. Ruan, Q. Wang, and D. Shen, "Medical image synthesis with deep convolutional adversarial networks," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 12, pp. 2720–2730, 2018.
- [21] N. M. Balasooriya and R. D. Nawarathna, "A sophisticated convolutional neural network model for brain tumor classification," in *2017 IEEE International Conference on Industrial and Information Systems (ICIIS)*. IEEE, 2017, pp. 1–5.
- [22] X. Li, X. Shen, Y. Zhou, X. Wang, and T.-Q. Li, "Classification of breast cancer histopathological images using interleaved densenet with senet (idsnet)," *PloS one*, vol. 15, no. 5, p. e0232127, 2020.
- [23] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
- [24] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang, "Thorax disease classification with attention guided convolutional neural network," *Pattern Recognition Letters*, vol. 131, pp. 38–45, 2020.
- [25] H. R. Roth, L. Lu, A. Seff, K. M. Cherry, J. Hoffman, S. Wang, J. Liu, E. Turkbey, and R. M. Summers, "A new 2.5d representation for lymph node detection using random sets of deep convolutional neural network observations," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014*, 2014.
- [26] S. P. Singh, L. Wang, S. Gupta, B. Gulyás, and P. Padmanabhan, "Shallow 3d cnn for detecting acute brain hemorrhage from medical imaging sensors," *IEEE Sensors Journal*, 2020.
- [27] K. Bayoudh, F. Hamdaoui, and A. Mtibaa, "Hybrid-covid: a novel hybrid 2d/3d cnn based on cross-domain adaptation approach for covid-19 screening from chest x-ray images," *Physical and engineering sciences in medicine*, vol. 43, no. 4, pp. 1415–1431, 2020.
- [28] X. W. Gao, R. Hui, and Z. Tian, "Classification of ct brain images based on deep learning networks," *Computer methods and programs in biomedicine*, vol. 138, pp. 49–56, 2017.
- [29] S. Qiu, P. S. Joshi, M. I. Miller, C. Xue, X. Zhou, C. Karjadi, G. H. Chang, A. S. Joshi, B. Dwyer, S. Zhu *et al.*, "Development and validation of an interpretable deep learning framework for alzheimer's disease classification," *Brain*, vol. 143, no. 6, pp. 1920–1933, 2020.
- [30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [31] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
- [32] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.
- [33] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1365–1374.
- [34] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *arXiv preprint arXiv:1612.03928*, 2016.
- [35] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4320–4328.
- [36] S. Roy and P. Maji, "A simple skull stripping algorithm for brain mri," in *2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR)*. IEEE, 2015, pp. 1–6.
- [37] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.
- [38] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.
- [39] K. Jnawali, M. R. Arbabshirani, N. Rao, and A. A. Patel, "Deep 3d convolution neural network for ct brain hemorrhage classification," in *Medical Imaging 2018: Computer-Aided Diagnosis*, vol. 10575. International Society for Optics and Photonics, 2018, p. 105751C.
- [40] D. Wang, Y. Zhang, K. Zhang, and L. Wang, "Focalmix: Semi-supervised learning for 3d medical image detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3951–3960.
- [41] S. U. Morton, C. Hehnly *et al.*, "Paenibacillus infection causes neonatal sepsis and subsequent postinfectious hydrocephalus in ugandan infants." [Online]. Available: <https://ssrn.com/abstract=4016548>
- [42] V. Monga, *Handbook of Convex Optimization Methods in Imaging Science*. Springer, 2017, vol. 1.
- [43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [44] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *arXiv preprint arXiv:1912.01703*, 2019.
- [45] C. J. Wu and M. S. Hamada, *Experiments: planning, analysis, and optimization*. John Wiley & Sons, 2011, vol. 552.
- [46] J. H. McDonald, *Handbook of biological statistics*. sparky house publishing Baltimore, MD, 2009, vol. 2.