RESEARCH ARTICLE

Gradual Self-Training via Confidence and Volume Based Domain Adaptation for Multi Dataset Deep Learning-Based Brain Metastases Detection Using Nonlocal Networks on MRI Images

Andrea Liew, MEng,[1] Chun Cheng Lee, MD, MRAD,[2] Valarmathy Subramaniam, MD, MMED,[2] Boon Leong Lan, PhD,[1,3] and Maxine Tan, PhD[1,4*]

Background: Research suggests that treatment of multiple brain metastases (BMs) with stereotactic radiosurgery shows improvement when metastases are detected early, providing a case for BM detection capabilities on small lesions.
Purpose: To demonstrate automatic detection of BM on three MRI datasets using a deep learning-based approach. To improve the performance of the network is iteratively co-trained with datasets from different domains. A systematic approach is proposed to prevent catastrophic forgetting during co-training.
Study Type: Retrospective.
Population: A total of 156 patients (105 ground truth and 51 pseudo labels) with 1502 BM (BrainMetShare); 121 patients with 722 BM (local); 400 patients with 447 primary gliomas (BrATS). Training/pseudo labels/validation data were distributed 84/51/21 (BrainMetShare). Training/validation data were split: 121/23 (local) and 375/25 (BrATS).
Field Strength/Sequence: A 5 T and 3 T/T1 spin-echo postcontrast (T1-gradient echo) (BrainMetShare), 3 T/T1 magnetization prepared rapid acquisition gradient echo postcontrast (T1-MPRAGE) (local), 0.5 T, 1 T, and 1.16 T/T1-weighted-fluid-attenuated inversion recovery (T1-FLAIR) (BrATS).
Assessment: The ground truth was manually segmented by two (BrainMetShare) and four (BrATS) radiologists and manually annotated by one (local) radiologist. Confidence and volume based domain adaptation (CAVEAT) method of co-training the three datasets on a 3D nonlocal convolutional neural network (CNN) architecture was implemented to detect BM.
Statistical Tests: The performance was evaluated using sensitivity and false positive rates per patient (FP/patient) and free receiver operating characteristic (FROC) analysis at seven predefined (1/8, 1/4, 1/2, 1, 2, 4, and 8) FPs per scan.
Results: The sensitivity and FP/patient from a held-out set registered 0.811 at 2.952 FP/patient (BrainMetShare), 0.74 at 3.130 (local), and 0.723 at 2.240 (BrATS) using the CAVEAT approach with lesions as small as 1 mm being detected.
Data Conclusion: Improved sensitivities at lower FP can be achieved by co-training datasets via the CAVEAT paradigm to address the problem of data sparsity.
Level of Evidence: 3
Technical Efficacy Stage: 2

J. MAGN. RESON. IMAGING 2023;57:1728–1740.

View this article online at wileyonlinelibrary.com. DOI: 10.1002/jmri.28456

Received Jul 13, 2022, Accepted for publication Sep 20, 2022.

*Address reprint requests to: M.T., Electrical and Computer Systems Engineering Discipline, School of Engineering, Monash University Malaysia, Bandar Sunway 47500, Malaysia. E-mail: maxine.tan@monash.edu

From the [1]Electrical and Computer Systems Engineering Discipline, School of Engineering, Monash University Malaysia, Bandar Sunway, Malaysia; [2]Radiology Department, Sunway Medical Centre, Bandar Sunway, Malaysia; [3]Advanced Engineering Platform, School of Engineering, Monash University Malaysia, Bandar Sunway, Malaysia; and [4]School of Electrical and Computer Engineering, The University of Oklahoma, Norman, Oklahoma, USA

Brain metastasis (BM) occurs in 10%–26% of patients with cancer, making it the most common brain tumor. Given the high-contrast resolution and thin voxel capability, MRI with intravenous gadolinium enhancement is the imaging modality of choice for BM detection.[1] BM can be treated with radiation therapy, performed by the radiation oncologists. This includes whole brain radiation therapy (WBRT) for widespread BM; and stereotactic radiosurgery (SRS), which provides targeted treatment to individual metastasis up to submillimeter accuracy.[2]

As awareness increases, early brain MRI screening has been carried out to detect BM for cancer patients, especially for those with neurological symptoms like headache, paranesthesia, or weakness. Early brain MRI screening enables detection of BM at the early stages, usually presented in smaller size (in the range of millimeter) and number. SRS is particularly useful to treat the BM at the early stages.[3]

Brain MRI is analyzed by a neuroradiologist and each BM is detected and labeled on a workstation, so that the radiation oncologist can subsequently use to treat the BM with SRS (Fig. 1). Hence, the accuracy of BM detection, mostly in the range of millimeter, heavily relies on the experience of the neuroradiologist. Observer error is a known factor that contributes to missed BM, which will lead to delayed treatment with SRS.

To date, there is no clinically accepted autonomous computer aided detection (CAD) feature for BM detection. Recent state-of-the-art results achieved in brain glioma segmentation,[4] lung nodule detection[5] and ischemic stroke lesion segmentation[6] are based on deep learning methods using convolutional neural networks (CNNs). The speed of progress in primary brain cancer detection methods compared to BM detection can be attributed to the widely available BrATS dataset.[7–9] A handful of work using deep learning on BM is found in the literature.[10–13] To improve sensitivity and reduce false positive (FP) detections of BM, two-stage pipelines were proposed that use Laplacian of Gaussian operators in generating regions of interest[10] and using FP reduction methods as a postprocessing step.[11] While two-step methods proved effective, CNN-based single-pass methods such as modified GoogLeNet[12] and DeepMedic[13] are preferred because of its simplicity. To achieve performances on par with two-step methods, our work explores enhancements to the network from an architecture and training strategy standpoint.

Convolutional operations are building blocks of CNNs, which are primed to extract information from their local neighborhood. In 3D medical detection, convolutions may be insufficient to capture information beyond the scope of their convolutional kernel. In this regard, attention and squeeze and excitation methods highlight regions of interests through saliency maps.[14,15] However, these methods evaluate an overall map from feature maps, which is less precise than nonlocal blocks as the latter model long-range dependencies at each pixel position.[16] The nonlocal block's approach to model a pixel's response in relation with the entire volume is advantageous in detecting BM that are observable across a few slices.[16]

In practice, we are not able to dictate the input modalities for the CAD system—this requirement varies in each

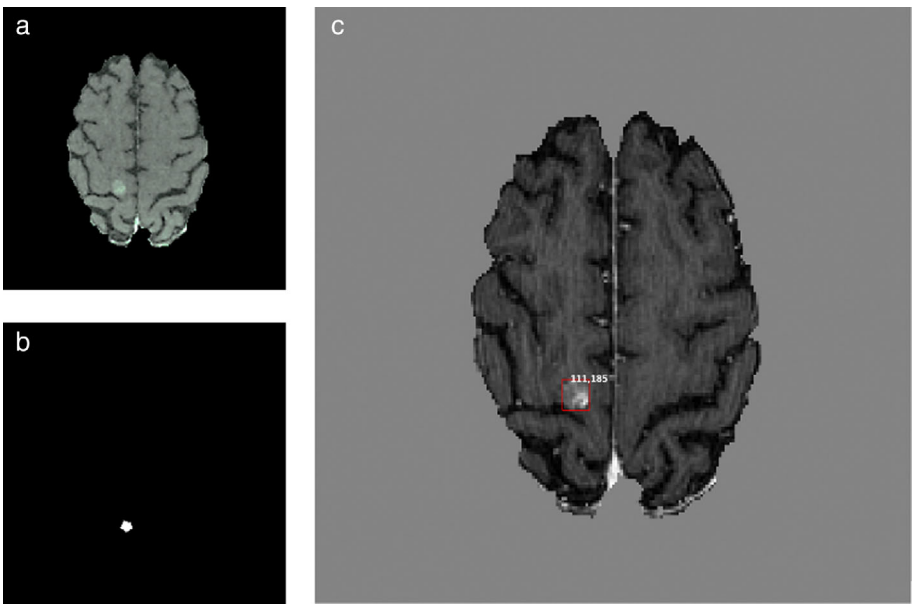

FIGURE 1: Representative axial slice of a brain MRI study. (a) Original image from T1 spin-echo postcontrast sequence (b) corresponding ground truth provided by BrainMetShare and (c) resulting bounding box with coordinate locations of the tumor centroid generated by connected component analysis and considered as ground truth for the detector network on the postprocessed image

radiology department acquisition protocols. By training multimodal data on an input dropout layer, previous work has shown negligible effect of missing modalities on the network.[17] However, this limits the network to having a fixed type of multimodal data with some missing modalities.

It is widely accepted that to sufficiently train the neural network, large volumes of annotated data are required to achieve good results. The original Faster-RCNN paper was benchmarked against the PASCAL VOC 2007, which contained 5000 training and 5000 test images.[18] In contrast, a Faster-RCNN used in BM detection used 73 patients for training and 48 for testing.[11] The scarcity of data is more pronounced in medicine, from concerns of ethics in data collection to getting a team of radiologists to dedicate time on top of their daily workload to process the MRI images in the dataset. Models Genesis demonstrated that a model trained on chest CT images contained domain-transferable qualities for other different domain medical datasets.[19] However, same-domain training is favored over training on different domains because of the smaller knowledge gap in learned representations.

Model performance is measured by its ability to generalize on unseen test data. Oftentimes, a model trained on a homogenous data source will not perform well when presented with data from different acquisition protocols or different hospitals.[20] This phenomenon is common in medical data where different contrast, settings, machines, and modalities could cause domain shifts. Self-training paradigms that incorporate unlabeled data and auxiliary datasets have been shown to narrow these domain gaps.[19,21]

In this article, we aim to:

1. detect BM using a 3D CNN-based object detection method through efficient use of nonlocal networks through a single-channel network using data from various MRI sequences;
2. develop a method to leverage-specific characteristics in data for domain adaptation via gradual self-training to prevent catastrophic forgetting; and
3. improve the performance (increase sensitivity and reduce FP) of the BM detection network by incorporating a multi dataset strategy and leveraging out-of-domain data from public datasets.

Materials and Methods

The internal ethics review board approved the use of the local dataset in this study (reference number: 2019-19668-35727). Since the study is retrospective, the requirement for written informed consent was waived.

Local MRI Dataset

The authors obtained the head MRIs database of patients with metastatic brain cancers, which were candidates for Gamma Knife therapy collected from 2017 to 2020. One hundred and twenty-one patients were included in the dataset with the following inclusion criteria: 1) at least one lesion in the scan and 2) a 3D post-gadolinium T1-MPRAGE sequence (three-dimensional T1 magnetization prepared rapid acquisition gradient echo).

All images were obtained from a 3 T MRI scanner (SIEMENS, Erlangen, Germany). The acquisition parameters for 3D T1-MPRAGE sequence were as follows: matrix = 256 × 256 × 192; slice thickness = 0.90 mm; pixel spacing = 0.90–0.93 mm; TE = 2.32–2.49 msec; TR = 1900–2000 msec; TI = 900 msec; flip angle = 8°–9°.

Public MRI Datasets

The authors used two publicly available datasets for multi-institutional and domain-shifted data, named BrainMetShare and BrATS 2021. Both datasets consist of ground truth masks, which delineate the entire tumor volume for BM segmentation tasks. In BM detection, a tumor volume is identified by the coordinate of its centroid.

BrainMetShare DATASET. This dataset provided by Stanford University contained 156 patients (105 for the training and 51 for the validation set) with at least one cerebral metastasis.[12] From the four sequences presented in this dataset, only the T1 spin-echo post-contrast sequence was selected as training data, to conform with the local dataset which utilizes a single sequence (see Table 1 for demographic data). The entire ground truth segmentation mask was used in determining the centroid of the tumor volume.

BrATS 2021 DATASET. The BrATS 2021 dataset is arguably the largest and most dominant dataset in brain tumor segmentation containing 1251 patients with four multiparametric MRI (mpMRI) scans with ground truth annotations by four radiologists.[7–9] Lesions originate from primary brain cancers and the MRI sequence used for training are T1-FLAIR images. For the evaluation of centroids, only the tumor core regions of the ground truth segmentation masks with labels 1 and 2 were considered; the edema regions (label 4) were omitted.

Image Ground Truths of BM Detection

BrainMetShare dataset was annotated by two neuroradiologists with 8- and 2-years' experience by manually delineating regions of interest on each slice of post-Gd 3D T1 weighted IR-FSPGR sequence with guidance from 3D FLAIR and post-Gd T1-weighted spin echo on OsiriXMD software package (v. 8.0; Geneva, Switzerland).[12]

BrATS data were manually annotated into four regions (namely edema, tumor core, enhancing core and nonenhancing core) by up to four experts using a standardized annotation protocol given in the literature[8] using co-registered T1, T1c, T2 and FLAIR MRI contrasts.

On the local dataset, postgadolinium T_1-MPRAGE sequences were interpreted by a clinical interventional radiologist with 10 years of experience (C.C.L), on a Carestream workstation (New York, USA). The lesions were manually marked using "Response Evaluation Criteria in Solid Tumors" (RECIST) guidelines on key slices and saved as DICOM Key Images for future reference (22). Depending on the size and shape of the metastases, every metastasis

TABLE 1. Patient Demographics of BrainMetShare, Local Dataset, and BrATS Dataset*

	BrainMetShare Dataset	Local Dataset	BrATS Dataset
Gender	Male: 51 Female: 105	Male: 56 Female: 65	Male: N/A Female: N/A Total: 1251
Dataset type	BM	BM	Glioblastoma multiforme
Primary cancer	Lung: 99 Breast: 33 Skin/melanoma: 7 Genitourinary: 7 Gastrointestinal: 5 Miscellaneous: 5	Breast: 3 Pituitary: 3 Lung: 12 Skin: 1 Spinal cord: 2 Colon: 1 Unknown: 99	Glioblastoma multiforme: 1251
Number of lesions	≤3: 64 (41%) 4–10: 47 (30%) >10: 45 (29%)	≤3: 85 (71%) 4–10: 27 (21%) >10: 9 (8%)	≤3: 1251 (100%)

*Numbers indicate the number of patients.

was described by one or two of RECIST diameters accompanied with a length (in mm or cm) to measure the lesion extent.

Convolutional Neural Network Details

NLMET: NONLOCAL BM DETECTION NETWORK. The model was based on N-Net, a lung cancer detector for computed tomography (CT) images that won the Data Science Bowl Lung Cancer Detection Competition.[5] Due to the parallels of size and multiplicity in lung nodules and BM, our work used the competition winning lung nodule detector architecture as the basis for BM detection.[5] The proposed model was implemented on Pytorch framework (v1.4.0; Meta AI, New York City, New York), Python (v3.8.5; Python Software Foundation, Delaware, USA), on a standard PC with 2 NVIDIA GeForce RTX 2080 Ti (Santa Clara, California, USA).

The backbone of N-Net contained a UNET-like structure with four down-sampling blocks and two up-sampling blocks. The last two blocks (see Fig. 2b) were replaced by a region proposal network (RPN), consisting of two CNN layers with $1 \times 1 \times 1$ kernels, which produced an output of $24 \times 24 \times 24 \times 5 \times 3$. This output tensor represents every location in $24 \times 24 \times 24$ with three anchors of sizes 10, 15, and 20 and the five regression values, corresponding to the probability, coordinates in x, y, z and the length of the bounding box. The probability is an objectness score provided by the RPN, which represents the degree of certainty which the location contains a metastases. This objectness score will be used in pseudo-labeling of the BrainMetShare dataset in our gradual domain adaptation method.

To increase depth in the UNET, residual units from[23] were introduced in the down-sampling path (see Fig. 2c). Each convolution layer was followed by batch normalization. The original N-Net architecture used for lung nodule detection consisted of an input block with coordinate locations of the input crop in relation to the position of the lung.[5] This location information is on the pretext that malignant lung nodules are typically found at the edge of the lungs. In NLMET, this we omit the location block because BMs can be found anywhere in the brain.[24]

Due to GPU memory limitations, we reduced the crop size from 128 to 96 with a batch size of 4. However, smaller crop sizes tend to demonstrate lower accuracy in various CNN architectures.[25] To overcome this limitation, we integrated a nonlocal block at the skipped connection. In the UNET, these skipped connections compensated for the loss of spatial information during deconvolution. However, skipped connections utilize feature maps in early layers, which do not have sufficient contextual information for differentiation of background and regions of interest within the volume. Nonlocal blocks were retrofitted into the skipped connections to evaluate attention in volumetric MRI data (see Fig. 2a). For computation efficacy, our method placed the nonlocal block only at the first skipped connection to provide contextual information to the up-sampling path (see Fig. 2b). A subsampling trick was also applied on the nonlocal blocks via max-pooling layers to reduce computational complexity. In Fig. 2d, we have proposed another placement of the nonlocal block at the skipped connection to compare the performance of the two configurations. Our model is called nonlocal BM network (NLMET) for short.

TRAINING DETAILS. The previously defined three datasets were used for training the network. Prior to training, all images were skull stripped by applying the brain mask generated by the BrainMaGe package.[26] All non-zero intensities below 0.2% and above 99.8% quantiles were clipped. All images in the datasets were then normalized to a [0–1] range via min–max normalization. To transform the BM segmentation datasets for use in BM detection, the coordinate of the centroid of the tumor is obtained through 3D connected component analysis of the ground truth mask (see Fig. 1). A lesion

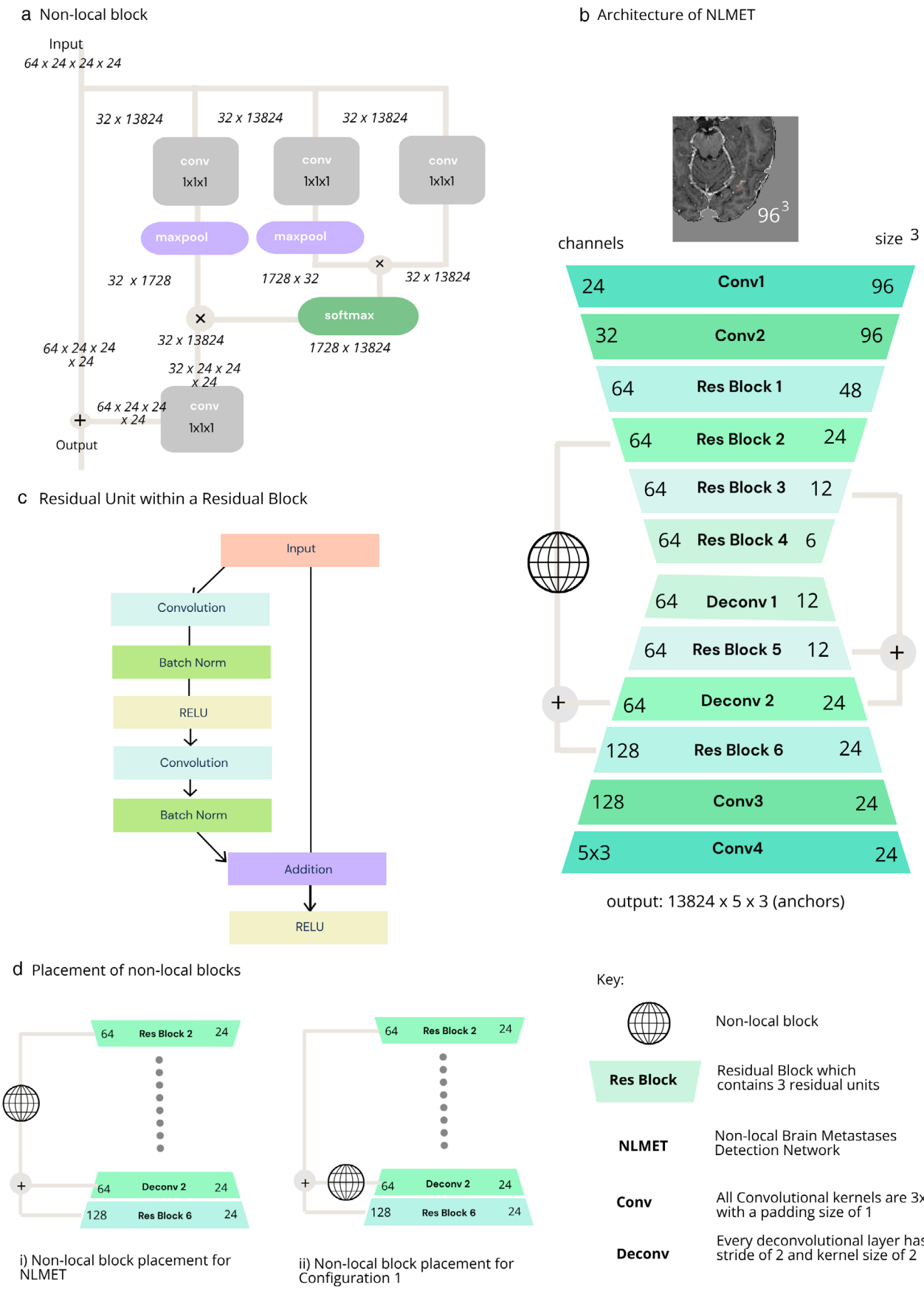

FIGURE 2: Architecture diagram of major components. (a) Nonlocal block, (b) NLMET, (c) residual units within a residual block, of the NLMET BM detection network, (d) placement of nonlocal blocks in NLMET and configuration 1.

was described as a cube with length equivalent to the longest lesion size.

During training, patches of size 96 × 96 × 96 were sampled randomly within the centroids of the lesions. We omitted sampling negative patches as the background area (negative samples) outnumber the lesions found in a training patch. Additionally, negative hard mining was implemented to only select the top k negative samples to be considered in the loss function. In our work, we used $k = 2$*batch_size.

An adaptive learning rate was used with an initial rate of 0.001 for the first half of the epochs, reduced to a factor 0.1 of the initial rate till the 80th percentile of epochs and finally to a factor 0.01 for the remaining epochs.

LOSS FUNCTION. Let the ground truth of a training input patch be defined as $T = [t_x, t_y, t_z, t_d]$ where there are N instances of ground truth lesions. The output from NLMET describes a bounding box for an anchor, $B = [b_p, b_x, b_y, b_z, b_d]$ where x, y, z are coordinate locations and d is the bounding box length (or lesion diameter) and its corresponding probability, b_p.The first part of the loss function L_{MSE} is a mean-squared error loss, only the lesions are considered. L_{MSE} is the L2-norm between the bounding box coordinates and the ground truth, where N is the number of ground truth lesions, defined as:

$$L_{\text{MSE}} = \frac{1}{N}\sum_{1}^{N}(t_x - b_x)^2$$

The second loss, L_{BCE} is the binary cross entropy of b_p and its ground truth g_p. Note that $g_p \in \{0, 1\}$ where 0 is the background samples and 1 is for lesion candidates.

$$L_{\text{BCE}} = g_p \log(b_p) + (1 - g_p)\log(b_p)$$

The loss function L_{box} for each anchor box is the sum of both losses defined as:

$$L_{\text{box}} = L_{\text{MSE}} + L_{\text{BCE}}$$

INFERENCE DURING TESTING. Due to GPU memory constraints, it was infeasible to obtain the output using a single pass of the network. Instead, the image was split into 96 × 96 × 96 patches in a sliding window fashion—with a window of 32 pixels. The output (x, y, z, r, p) from all the patches will result in proposals. The combined patches are processed with non-maximum suppression (NMS) to remove overlapping proposals.

STATISTICAL ANALYSIS. The model's capacity to detect metastases was determined by measuring the detection sensitivity and FP rate per patient (FP/patient). Free receiver operating characteristic (FROC) curves were generated at seven predefined FP rates per patient: 1/8, 1/4, 1/2, 1, 2, 4, and 8. To ensure consistency, scripts were adapted from those provided to evaluate lung cancer detection in the LUNA Challenge.[27] This script was implemented on the Python package (v3.8.5; Python Software Foundation, Delaware, USA), with Scikit-learn (v1.0.1), Matplotlib (v3.5.1) libraries.

To determine the impact of the nonlocal block on the original model's performance, we compare the sensitivity and FP/patient of NLMET with the original N-Net. To compare the sensitivities of both networks, we use a FP/patient of approximately 4.

In this work, we are presented with datasets with limited sample size (the ground truth from the BrainMetShare Validation dataset is not accessible); therefore, we employ a 5-fold cross validation (CV) method to evaluate the overall performance of NLMET. The data are partitioned into five equal folds; 4-fold are used to train the network while the remaining fold is retained as the test set. Training is performed on four folds, until each fold has been used once as the test set (in our case, training is repeated five times). The average sensitivity and FP rates are taken from the five test folds. The train-test data split is generated using Scikit-learn (scikit-learn v 0.24.2) k-fold function to eliminate bias.

The CAVEAT method is tested on the test fold with the median sensitivity and FP rates within the five test folds, instead of re-running the experiments in a 5-fold method for every iteration due to time constraints.

Consider a candidate lesion l in a 3D volume described by coordinates (x_l, y_l) at slice z_l within a cube with length d_l. Consider an annotated metastasis from the ground truth g given by $[x_g, y_g, z_g, d_g]$ where (x_g, y_g) are the coordinates of the centroid location of bounding box with length d_g at slice z_g. A candidate lesion was a true positive with respect to ground truth g when the condition in Eq[1] was satisfied:

$$[x_g - x_l]^2 + [y_g - y_l]^2 + [z_g - z_l]^2 < d_g^2 \therefore l \text{ is a true positive} \quad (1)$$

CAVEAT: Gradual Self Training Through Confidence and Volume Based Domain Adaptation

Here, the methods used to co-train the datasets will be described. We present three datasets for training NLMET, consisting of two BM datasets and an auxiliary dataset containing primary brain cancer patients. The first BM dataset contains labeled data from the local dataset. In the second BM dataset from BrainMetShare, we utilize labeled data from the training set and also unlabeled data from its validation set. Lastly, we utilize a primary brain cancer dataset containing labeled data from BraTS.

In every batch, a 1:2:2 ratio of BrATS to BrainMetShare and local samples were used. Among the three datasets, BrainMetShare and the local dataset contained BM that were closer in domain compared to BraTS. The latter dataset consisted of primary gliomas, which are larger in volume than BM. Here, we assumed that lesion volumes of primary gliomas correlated with the appearance of metastases. The lesion volumes are determined by the counting the labels 1 and 2 (edema marked with label 4 was excluded) from the segmentation mask ground truth. A total of 375 BrATS samples with lesions of comparable volumes to BrainMetShare and the local dataset were selected as auxiliary data to generate more datapoints for training NLMET. To train on the multiple datasets, we

employed a two-prong strategy, one aimed at generating higher confidence pseudo labels in the BrainMetShare validation set and the other to partition the BrATS data through lesion volume to create mini datasets. For higher confidence pseudo labels, we utilized the objectness score from the output of NLMET at successively higher probabilities in every iteration. For BrATS data, we gradually increased the volume of the tumors in the training set. We call this method confidence and volume based domain adaptation (CAVEAT) (see Fig. 3 Flowchart description of CAVEAT. Pseudo labels were generated with successive levels of confidence and lesions were selected based on volume).

Results

We presented a local dataset with 121 patients with BMs, where most patients had cancers of unknown primary (CUP) as they are symptomatic patients who have secondary metastases discovered on their first scan (Table 1). In both BrainMetShare and the local dataset, we found that most of the patients had between one and three cancers.

Total training time for 150 epochs with batch size of 5 was around 29 hours on 2 NVIDIA RTX 2080 Ti (Santa Clara, California) GPUs. The inference time for a full MRI volume was 850 msec. In Table 2, by comparing the N-Net model with NLMET and Configuration 1, we found that the addition of the nonlocal block improved sensitivity as well as FP/patient for models trained independently on the BrainMetShare and Local dataset. We found that placement of the nonlocal block at the output of the downsampling path yields better results than placing it at the upsampling path. The sensitivity and FP evaluated on the five test folds of the 5-fold cross validation (CV) method were used to determine overall performance of NLMET in our dataset (Table 3).

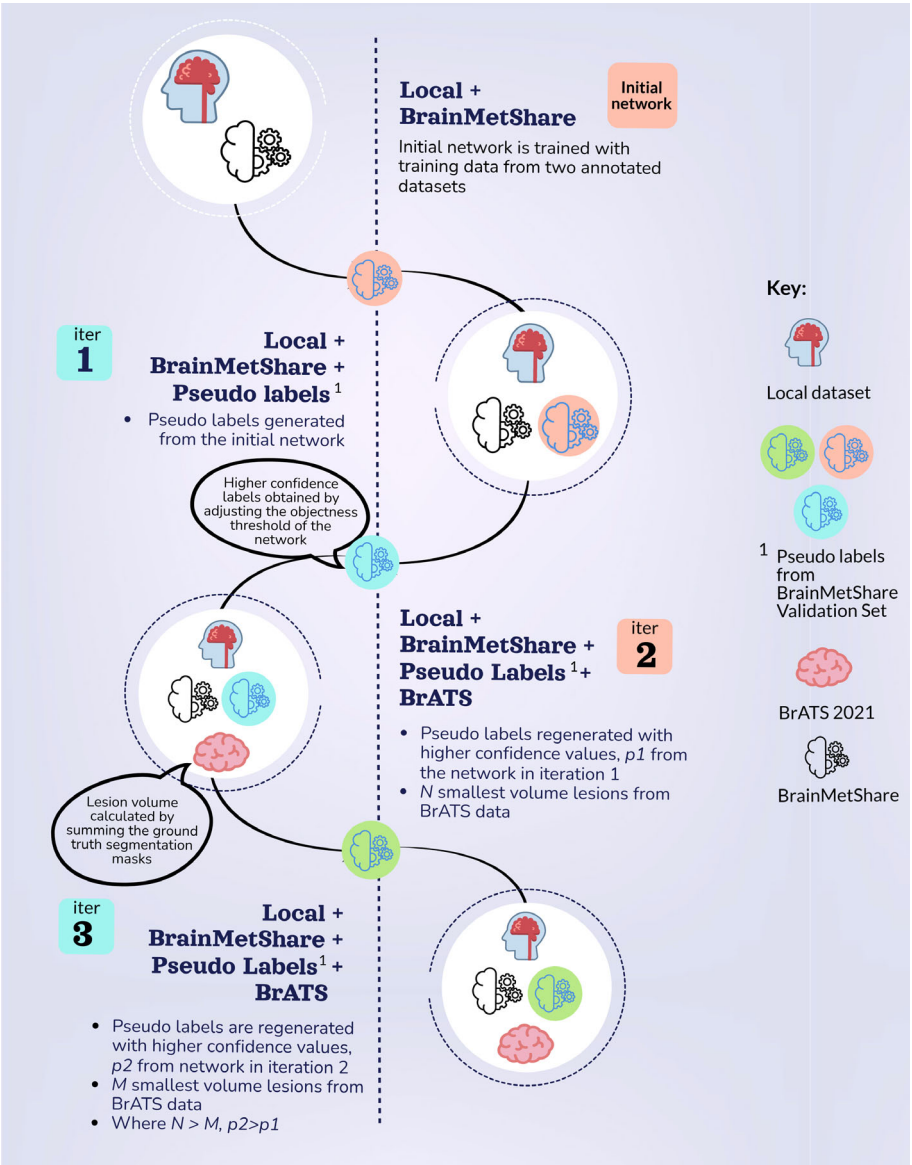

FIGURE 3: Flowchart description of CAVEAT. Pseudo labels were generated with successive levels of confidence and lesions were selected based on volume.

TABLE 2. Summary of the Detection Metrics for Validation Data on BrainMetShare and Local Dataset for the Base Model, Configuration 1, and NLMET

	BrainMetShare				Local Dataset			
Model Description	**TP**	**FP**	**FP/Patient**	**Sensitivity**	**TP**	**FP**	**FP/Patient**	**Sensitivity**
N-Net (base)	145	86	4.095	0.609	47	83	3.609	0.610
Configuration 1	166	83	3.952	0.697	50	78	3.391	0.649
NLMET (ours)	170	74	**3.523**[a]	**0.718**	50	76	**3.304**	**0.649**

[a]Numbers in bold denote the better model performance, higher sensitivity at lower FP/patient.
TP = true positives; FP = false positives.

It was found that BrainMetShare on the fifth test fold (CV5) and third test fold (CV3) on Local dataset was the median sensitivity and FP among the five test folds, and were selected as the folds to be used for training and testing the CAVEAT method.

Table 4 tabulates the results for the CAVEAT method of gradual domain adaptation incorporating all three datasets. We obtained BrATS results from iteration 0 and 1 that was trained only on BrainMetShare and the local dataset, where the first iteration registered 0.638 at 2.200 FP/patient. In the second iteration, we found a drop in sensitivity, but at a lower FP/patient. It can be observed that the method generalized well to BrATS, with a sensitivity of 0.723 and a FP/patient of 2.240 at iteration 3 without any fine tuning. BrainMetShare's performance also benefited from the CAVEAT method—with marked improvements of ~17% in Sensitivity with improved FP/patient. The CAVEAT iterations showed steady improvement from the fully supervised baseline model to the next three iterations where more confident pseudo labels and small volume BrATS data were introduced (see Fig. 4).

Correctly inferred lesions and false positives from the final iteration of CAVEAT can be found in Fig. 5. These lesion candidates showed that CAVEAT helped the model detect different sizes of lesions. The FPs found in the representative subject of Fig. 5c with motion artifacts were found to be blood vessels mistaken for metastases. A qualitative review of MRIs in the validation dataset found that similar FPs were also present in MRIs without motion artifacts (Fig. 5d,e). Figure 5f is an example of the lateral ventricle mistaken as a FP. Missed detections were found to be common in samples with multiple tumors in close proximity, which were inadvertently removed by the NMS algorithm typically used in RPNs.

Figure 6 depicted heatmaps of three arbitrary points, with the point labeled 1 located at the tumor site, while points 2 and 3 are structures unrelated to the tumor. The purpose of the nonlocal attention maps (NLAM) was to compare the different nonlocal responses within different structures in the test image. While every point generated a volumetric NLAM, only a single slice was selected to be displayed. As observed from the NLAM in (Fig. 6c–e), the response of each point in the test volume was different, and in the proximity of the tumor area (Fig. 6c), regions of saliency are highlighted in red, whereas there is less response in areas that do not correspond with any tumors (Fig. 6d–e). Artifacts in the form of horizontal lines observed in NLAM are attributed to the subsampling trick described in the literature[16] to save computational resources. In our case, the subsampling reduced the patch from $32 \times 32 \times 32$ to $32 \times 32 \times 8$ and hence reshaping the image to its original size of $128 \times 128 \times 128$ may have caused some distortions.

Discussion

In this study, we developed a deep learning method, NLMET, which detects BM reliably on different postcontrast MRI sequences (T1-MPRAGE postcontrast, T1-gradient echo poscontrast and T1-FLAIR datasets). The various sequences demonstrate the potential of training with a mixed bag of MRI sequences, to simulate multicenter scenarios.

Also, the appearance of tumors on the MRI are dependent on the origin of the primary malignancies.[24] Since 82% of the data in the local dataset have unknown primary tumors, we are not able to ascertain which tumor morphologies are under- or over-represented in the dataset.

We utilized a lung nodule detection network as there are parallels between metastases in brain and lung nodules. For instance, nodules are comparable in size to metastases, and often mistaken for vessels and other structures, similar to how metastases are misrepresented as blood vessels in the brain. Additionally, using CAVEAT iterative gradual self-training framework for co-training of the three datasets further improved the quality of the detection network. Our findings were validated on held out data with samples of different sequences originating from three separate datasets.

TABLE 3. Detection Performance 5-Fold CV on BrainMetShare and Local Dataset on the NLMET Model

	BrainMetShare						Local Dataset					
	CV1	CV2	CV3	CV4	CV5	Average	CV1	CV2	CV3	CV4	CV5	Average
Sensitivity	0.737	0.688	0.525	0.766	0.727	0.689	0.678	0.683	0.610	0.543	0.585	0.620
TP	182	258	202	196	173	202.200	61	127	47	44	110	77.800
Missed BM	65	117	183	60	65	98.000	29	59	30	37	78	46.600
FP	93	92	93	96	97	94.200	105	107	93	115	107	105.400
FP/patient	4.429	4.381	4.619	4.571	4.429	4.486	4.375	4.458	4.429	4.792	4.458	4.502

BM = brain metastases; TP = true positives; FP = false positives; CV = trained models from the 5-fold cross-validation on the training set.

TABLE 4. Brain Metastasis Detection Performance of NLMET When Trained With the CAVEAT Method

	Datasets				BrainMetShare Validation				Local Validation				BRATS Validation			
	Local	BrainMetShare														
Iter	TR	TR	PS	BraTS[a]	TP	FP	FP/p	SENS	TP	FP	FP/p	SENS	TP	FP	FP/p	SENS
0	✓	✓	×	×	151	97	4.619	0.634	54	115	5.000	0.630	30	55	2.200	0.638
1	✓	✓	✓	×	171	73	3.476	0.718	60	89	3.870	0.779	27	44	1.760	0.574
2	✓	✓	✓	350	183	63	3.000	0.769	57	75	3.261	0.740	32	51	2.040	0.681
3	✓	✓	✓	375	193	62	**2.952**	**0.811**	57	72	**3.131**	**0.740**	34	44	**1.760**	**0.723**
Baselines					170	74	3.523	0.718	50	76	3.304	0.649	26	69	2.760	0.553

Iter = iteration of the CAVEAT method; TP = true positives; FP = false positives; TR = training data; PS = pseudo-labeled data; FP/p = false positives per patient; SENS = sensitivity, baselines provide values when trained on the training set of its corresponding dataset; ✓ = dataset not used in training; × = dataset used in training.
Numbers in bold denotes the highest sensitivity and lowest FP/patient achieved by CAVEAT method.
[a]Indicates number of samples used for training.

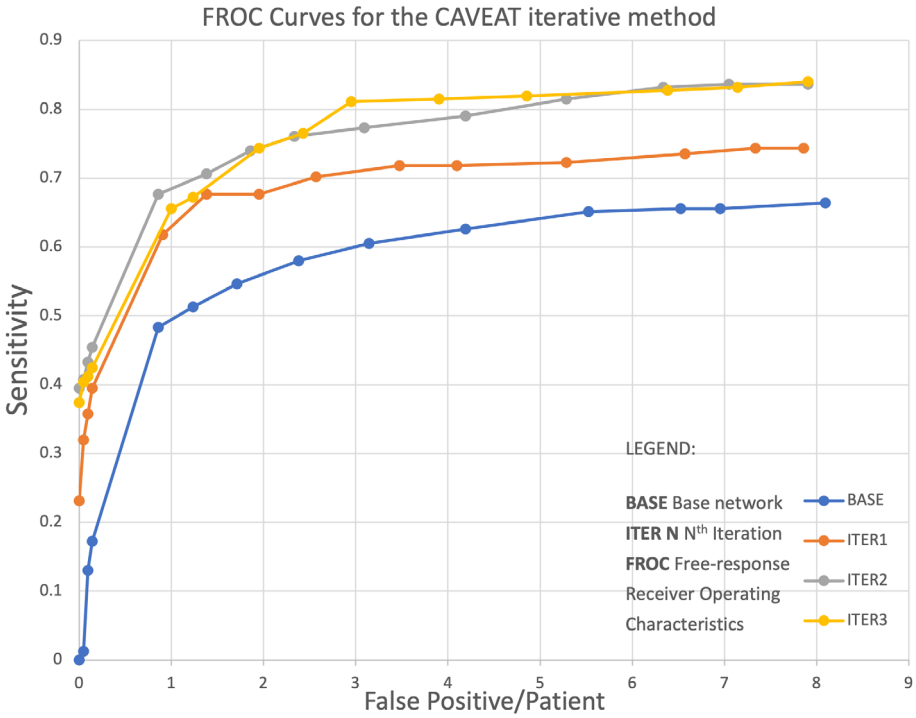

FIGURE 4: FROC curves for the CAVEAT method at different iterations evaluated at 1/4, 1/2, 1, 2, 4, 8 FP/patient.

We found that the addition of a single nonlocal block improved the sensitivity and FP/patient with BrainMetShare and local datasets by running ablation studies between the NLMET and N-Net on both datasets. While it would be intuitive to increase the number of nonlocal blocks in the object detection network, diminishing returns upon increasing the number of blocks have been found.[16] This result may be explained by the fact that BM cause subtle morphological distortions and midline shifts within the brain structure, which may not be adequately described by convolutional operations. These distortions serve as visual cues radiologists look out for in locating BM. As the goal of any CAD system is to model intrinsic knowledge, we included a nonlocal block to evaluate information, which may not be within range of a convolutional operation. The NLAM suggests that the response for different regions in an image is different and

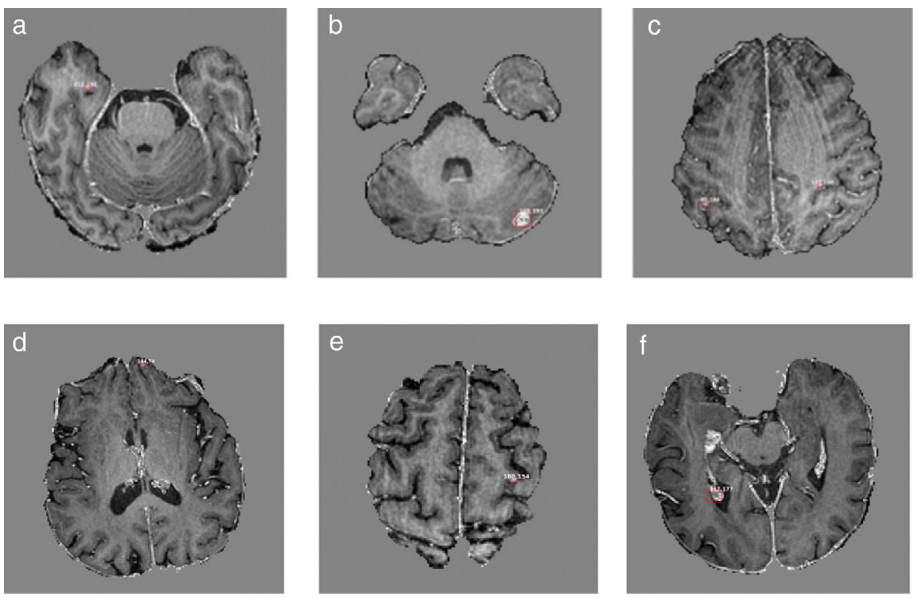

FIGURE 5: Detected lesions from the BrainMetShare dataset in six representative patients with lesions of varying sizes (a) small metastasis 2 mm in size, (b) large metastasis, measuring 9 mm (c) two false positives (d) one false positive (e) one false positive (f) one false positive.

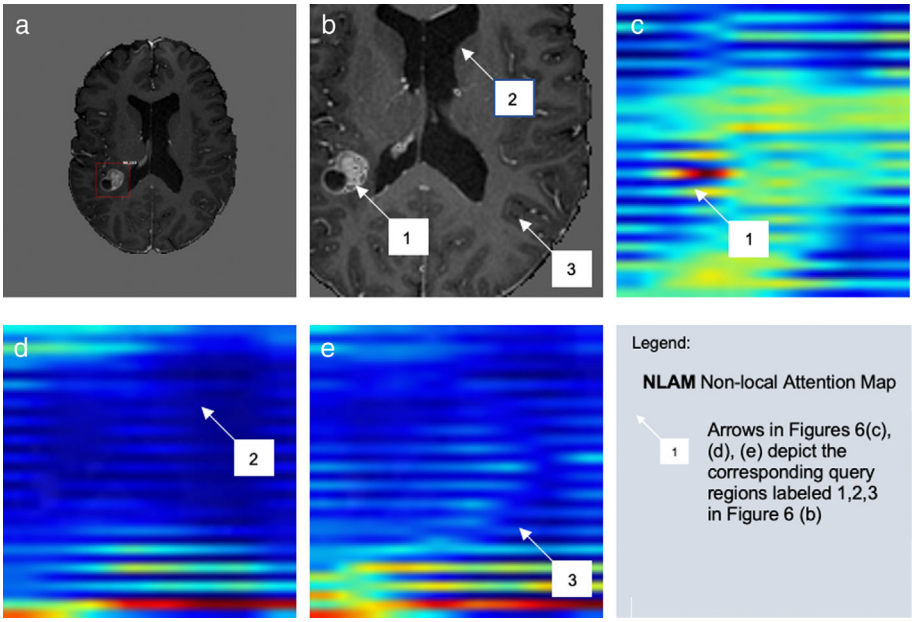

FIGURE 6: Depiction of the nonlocal response at three predetermined locations using NLAM. (a) Ground truth of tumor in a representative patient with 240 × 240 resolution. (b) Patch of 128 × 28, which denotes the query regions of NLAM. (c) NLAM near tumor region 1. (d) NLAM near nontumor region 2. (e) NLAM near nontumor region 3.

methods that generate single attention maps for the entire volume[28,29] may provide insufficient context to the network.

Prior work on BM was done on private datasets, which does not allow comparison of the various neural networks.[11–13,30,31] We utilized results from the publicly available BrainMetShare dataset to benchmark the performance of our method with other methods that used the same dataset. With supervised learning on 3D NLMET, we achieved an average of sensitivity of 0.69 at 4.5 FP per patient, compared to the original BrainMetShare architecture using a modified 2.5D GoogLeNet with 0.53 sensitivity at 8.3 FP/patient.[12] The benefits of 3D NLMET were 2-fold, it corroborated the effectiveness of nonlocal blocks over increased network depth (of the GoogLeNet) in capturing long range dependencies.[16] Second, it demonstrated the benefits of a 3D over 2.5D scheme—despite reduction in crop size and batch size. The results were obtained by 5-fold CV because the ground truth was unavailable for the validation set of BrainMetShare.

While we have shown that a nonlocal block can provide more global context, the networks were trained on their respective datasets in a fully supervised manner. This meant that each model had to be trained individually and there was no knowledge shared between datasets. Since all three datasets represented a domain shift from each other, training the model with all labeled data in one go may lead to model instability. It has been shown that systematic training such as curriculum learning proposed in[26] makes the model impervious to label noise. However, in our case, there is no objective way to distinguish "easy" cases from their more "difficult" counterparts. Therefore, it is more advantageous to narrow the domain gap between datasets. An incremental domain adaptation approach was shown to yield higher accuracies at the target dataset compared to direct domain adaptation.[27] One challenge in our multi dataset system is in determining the intermediate domains to which we would incrementally train the dataset.

To achieve a more generalized model and to leverage other datasets, we co-trained the datasets in a gradual self-training paradigm known as CAVEAT, which is inspired by incremental learning and curriculum learning.[32,33] With CAVEAT, the aim was to create small shifts in domain so that the network can adapt to the target domain. Therefore, in the first iteration of CAVEAT, we applied fully supervised learning on the local and BrainMetShare dataset as it had the smallest shift in domain. Next, we applied pseudo labels from the unlabeled data in BrainMetShare. As the BrATS data were a primary glioma dataset, it had a larger domain shift, hence was only present from the second iteration. In each batch, one BrATS sample was presented to two BrainMetShare and two local samples to prevent catastrophic forgetting as we were dealing with three different datasets. We found that a marginal increase in the amount of data in the next iteration was enough to illicit a better sensitivity at lower FP per scan. Also, a sudden increase in BrATS sample size may cause an imbalance in the number of samples being trained in an epoch, potentially causing overfitting.

In terms of stability of the CAVEAT approach, we evaluate this based on the performance of three datasets, where the improvements are consistent across the board. Without training on BrATS data (iteration 0 and 1) we can attain a relatively acceptable result on the BrATS dataset. We do find that after the third iteration, there are depreciating returns as

we find that we no longer have enough small volume tumors to train from and the confidence of the pseudo labels has reached its peak.

Empirical work on larger BrATS ratios (2:2:2) have been shown to be detrimental to the performance of metastases detection. As the main goal of our detection was to identify the smallest metastases and the trickiest to determine manually—we did not apply lesion size thresholds when evaluating the results. Validation of results on CAVEAT was performed with a single fold on the datasets to save on computation time.

After three iterations on CAVEAT, we found that the BrainMetShare and local dataset improved at least 10% sensitivity and decreased in FP/patient down to 2.9. Another approach of synthesizing large numbers of lesions using generative adversarial networks (GANs) achieved similar increase of 4 to 7 FP per slice.[34] Simply put, a radiologist must contend with three extra false lesions on each slice, whereby a single patient's MRI would have 450 incorrectly marked lesions (for an MRI with 150 slices). In the two-step methods, we have reviewed, "CropNet"[10] gives 85% sensitivity at 5.85 FP/patient and the Faster RCNN with RUSBoost[11] registers 96% sensitivity at 20 FP/patient. In comparison with CAVEAT, the two-step methods achieved higher sensitivities, albeit at higher FPs. The reason for lower sensitivities could be attributed to the data distribution of the datasets and MRI sequences selected for use during training. In BrainMetShare, the authors contend that the diversity in data makes it a more challenging dataset to work with.[12] Similarly, the local dataset has a range of different BMs ranging from 2 mm to 40 mm, with different distribution of primary cancers. One downside of higher FPs is that it creates additional overhead and is the main reason for poor adoption of CAD in clinical settings.[35]

The presence of motion artifacts in a representative subject from BrainMetShare is expected as motion artifacts were not an exclusion criterion.[12] We have also ascertained that FPs were not only limited to MRIs with motion artifacts, so we rule out the possibility that there is limited training data with motion artifacts. Another instance of FPs, are from high-contrast structures in the MRI such as lateral ventricles. To reduce the prevalence of these FPs, these FPs can be incorporated in the training set for further negative sample training.

The CAVEAT approach of introducing increasingly confident labels and out of domain samples in a controlled manner provided regularization to prevent overconfidence in the network during self-training, evident by the 2.9 FP/patient result. However, no marked improvements were observed from the second to third iteration as the pseudo labels were sufficiently confident after three iterations, and the number of small lesions BrATS data was insufficient to perform further iterations. This is typical in iterative learning where improvements saturate after the third iteration.[36]

Limitations

Three imaging sequences were selected (T1-MPRAGE, T1-gradient echo and T1-FLAIR) for training with the aim of demonstrating generalizability to various sequences. Further studies on other combinations of multiparametric sequences to improve detection performance should be performed. Also, the local dataset that we curated does not have enough data points for a training and validation set as performed in BrainMetShare. Second, there is a concern that the ground truth of the local dataset was annotated by a single radiologist—we try to mitigate this by introducing two datasets, BrainMetShare, which is publicly available and annotated by two specialists and BrATS, which is annotated by four radiologists. The other limitation is that the two metastases' datasets, may not be similarly distributed as most of the primary cancers in the local dataset are CUP—and therefore domain adaptation methods and self-training with out of domain BrATS dataset were perused.

Conclusion

The CAVEAT deep learning detection ecosystem alleviated the scarcity of training data in a single dataset with a small FP rate and improved sensitivity. In addition, the proposed method is MRI sequence-agnostic, which is advantageous for clinical applications.

Acknowledgments

This study was funded by Sunway Medical Centre, under grant ENG/SUNMED/11-2019/008. Open access publishing facilitated by Monash University, as part of the Wiley - Monash University agreement via the Council of Australian University Librarians.

References

1. Tong E, McCullagh KL, Iv M. Advanced imaging of brain metastases: From augmenting visualization and improving diagnosis to evaluating treatment response. Front Neurol 2020;11:270. https://doi.org/10.3389/FNEUR.2020.00270/BIBTEX.

2. Ma L, Fogh S, Gupta N, Hwang A, et al. A technique for achieving submillimeter accuracy of volume-staged stereotactic radiosurgery. J Radiosurgery SBRT 2012;2:11.

3. Wolf A, Kvint S, Chachoua A, et al. Toward the complete control of brain metastases using surveillance screening and stereotactic radiosurgery. J Neurosurg 2017;128:23-31.

4. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2020;18(2):203-211. https://doi.org/10.1038/s41592-020-01008-z.

5. Liao F, Liang M, Li Z, Hu X, Song S. Evaluate the malignancy of pulmonary nodules using the 3D deep leaky noisy-or network. arXiv 2017;11:3484-3495. https://doi.org/10.1109/tnnls.2019.2892409

6. Clèrigues A, Valverde S, Bernal J, Freixenet J, Oliver A, Lladó X. Acute ischemic stroke lesion core segmentation in CT perfusion images using fully convolutional neural networks. Comput Biol Med 2019;115:103487.

7. Baid U, Ghodasara S, Mohan S, et al. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv:2107.02314. https://doi.org/10.48550/arxiv.2107.02314.
8. Menze BH, Jakab A, Bauer S, et al. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging 2015;34:1993-2024. https://doi.org/10.1109/TMI.2014.2377694.
9. Bakas S, Akbari H, Sotiras A, et al. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci Data 2017;4:170117. https://doi.org/10.1038/sdata.2017.117.
10. Dikici E, Ryu JL, Demirer M, et al. Automated brain metastases detection framework for T1-weighted contrast-enhanced 3D MRI. IEEE J Biomed Heal Informatics 2019;24:2883-2893.
11. Zhang M, Young GS, Chen H, et al. Deep-learning detection of cancer metastases to the brain on MRI. J Magn Reson Imaging 2020;52:1227-1236.
12. Grøvik E, Yi D, Iv M, Tong E, Rubin D, Zaharchuk G. Deep learning enables automatic detection and segmentation of brain metastases on multisequence MRI. J Magn Reson Imaging 2020;51:175-182.
13. Jünger ST, Hoyer UCI, Schaufler D, et al. Fully automated MR detection and segmentation of brain metastases in non-small cell lung cancer using deep learning. J Magn Reson Imaging 2021;54(5):1608-1622.
14. Vaswani A, Shazeer N, Parmar N, et al. (2017). Attention is all you need. *Advances in neural information processing systems* (pp. 5998-6008).
15. Hu J, Shen L, Sun G. "Squeeze-and-Excitation Networks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132-7141, https://doi.org/10.1109/CVPR.2018.00745.
16. Wang X, Girshick R, Gupta A, He K. Non-local neural networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol 14 Salt Lake City, UT, USA: IEEE; 2018. p 7794-7803.
17. Grøvik E, Yi D, Iv M, et al. Handling missing MRI sequences in deep learning segmentation of brain metastases: A multicenter study. npj Digit Med 2021;41:1-7.
18. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. arXiv 2015;1:91-99.
19. Zhou Z, Sodha V, Pang J, Gotway MB, Liang J. Models Genesis. Med Image Anal 2021;67:101840.
20. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med 2018;24(249):1342-1350.
21. Zhu Y, Zhang Z, Wu C, et al. Improving semantic segmentation via self-training. arXiv 2020. https://doi.org/10.1109/TPAMI.2021.3138337
22. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). Eur J Cancer 2009;45:228-247.
23. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *2016 IEEE conference on computer vision and pattern recognition*, Las Vegas, NV, USA: IEEE; 2015. p 770-778. https://doi.org/10.48550/arxiv.1512.03385.
24. Ellenbogen RG, Abdulrauf SI, Sekhar LN. *Principles of neurological surgery*, Philadelphia, PA: Elsevier Health Sciences; 2012. p 1-820.
25. Hamwood J, Alonso-Caneiro D, Read SA, Vincent SJ, Collins MJ. Effect of patch size and network architecture on a convolutional neural network approach for automatic segmentation of OCT retinal layers. Biomed Opt Express 2018;9:3049.
26. Thakur S, Doshi J, Pati S, et al. Brain extraction on MRI scans in presence of diffuse glioma: Multi-institutional performance evaluation of deep learning methods and robust modality-agnostic training. Neuroimage 2020;220:117081.
27. Setio AAA, Traverso A, de Bel T, Berens MSN, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. Med Image Anal 2017;42:1-13.
28. Tao A, Sapra K, & Catanzaro B. Hierarchical Multi-Scale Attention for Semantic Segmentation. 2020;ArXiv, abs/2005.10821.
29. Cao Y, Xu J, Lin S, Wei F, Hu H. GCNet: Non-local networks meet squeeze-excitation networks and beyond. *Proc - 2019 Int Conf Comput Vis work ICCVW 2019 [internet]*; IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019. p 1971-1980.
30. Zhou Z, Sanders JW, Johnson JM, et al. Computer-aided detection of brain metastases in T1-weighted MRI for stereotactic radiosurgery using deep learning single-shot detectors. Radiology 2020;295:407-415. https://doi.org/10.1148/radiol.2020191479.
31. Xue J, Wang B, Ming Y, et al. Deep learning-based detection and segmentation-assisted management of brain metastases. Neuro Oncol 2020;22:505-514.
32. Mittal S, Galesso S, Brox T. Essentials for class incremental learning. *2021 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)*. Nashville, TN, USA: IEEE 2021.
33. Bengio Y, Louradour J, Collobert R, Weston J. Curriculum Learning. *Proceedings of the 26th annual international conference on machine learning*, New York, NY, USA: Association for Computing Machinery, 2009; p 41-48.
34. Han C, Murao K, Noguchi T, et al. Learning more with less: Conditional PGGAN-based data augmentation for brain metastases detection using highly-rough annotation on MR images. *The 28th ACM international conference on information and knowledge management*. New York, NY: Association for Computing Machinery; 2019. p 119-127. https://doi.org/10.1145/3357384.3357890.
35. Strohm L, Hehakaya C, Ranschaert ER, Boon WPC, Moors EHM. Implementation of artificial intelligence (AI) applications in radiology: Hindering and facilitating factors. Eur Radiol [Internet] 2020;30:5525-5532.
36. Xie Q, Hovy EH, Luong M, & Le QV. Self-training with noisy student improves ImageNet classification. arXiv 2020;10684-10695.