A multi-path 2.5 dimensional convolutional neural network system for segmenting stroke lesions in brain MRI images

Yunzhe Xue[a], Fadi G. Farhat[a], Olga Boukrina[b,c], A. M. Barrett[b,c], Jeffrey R. Binder[d], Usman W. Roshan[a,*], William W. Graves[e]

[a] *Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102, USA*
[b] *Stroke Rehabilitation Research, Kessler Foundation, West Orange, NJ, USA*
[c] *Department of Physical Medicine and Rehabilitation, Rutgers – New Jersey Medical School, Newark, NJ, USA*
[d] *Department of Neurology, Medical College of Wisconsin, Milwaukee, WI, USA*
[e] *Department of Psychology, Rutgers University – Newark, Newark, NJ, USA*

Abstract

Automatic identification of brain lesions from magnetic resonance imaging (MRI) scans of stroke survivors would be a useful aid in patient diagnosis and treatment planning. It would also greatly facilitate the study of brain-behavior relationships by eliminating the laborious step of having a human expert manually segment the lesion on each brain scan. We propose a multi-modal multi-path convolutional neural network system for automating stroke lesion segmentation. Our system has nine end-to-end UNets that take as input 2-dimensional (2D) slices and examines all three planes with three different normalizations. Outputs from these nine total paths are concatenated into a 3D volume that is then passed to a 3D convolutional neural network to output a final lesion mask. We trained and tested our method on datasets from three sources: Medical College of Wisconsin (MCW), Kessler Foundation (KF), and the publicly available Anatomical Tracings of Lesions After Stroke (ATLAS) dataset. To promote wide applicability, lesions were included from both subacute (< 5 weeks) and chronic (> 3 months) phases post stroke, and were of both hemorrhagic and ischemic etiology. Cross-study validation results (with independent training and validation datasets) were obtained to compare with previous methods based on naive Bayes, random forests, and three recently published convolutional neural networks. Model performance was quantified in terms of the Dice coefficient, a measure of spatial overlap between the model-identified lesion and the human expert-identified lesion, where 0 is no overlap and 1 is complete overlap. Training on the KF and MCW images and testing on the ATLAS images yielded a mean Dice coefficient of 0.54. This was reliably better than the next best previous model, UNet, at 0.47. Reversing the train and test datasets yields a mean Dice of 0.47 on KF and MCW images, whereas the next best UNet reaches 0.45. With all three datasets combined, the current system compared to previous methods also attained a reliably higher cross-validation accuracy. It also achieved high Dice values for many smaller lesions that existing methods have difficulty identifying. Overall, our system is a clear improvement over previous methods for automating stroke lesion segmentation, bringing us an important step closer to the inter-rater accuracy level of human experts.

Keywords: MRI, convolutional, neural network, deep learning, stroke, neuropsychology

*Corresponding author
Email address: usman@njit.edu (Usman W. Roshan)

1. Introduction

Neuropsychological studies of brain lesion-deficit relationships are an indispensable means of determining what brain areas are critical for carrying out particular functions. This contrasts with functional brain imaging techniques such as functional magnetic resonance imaging (fMRI), which while extremely popular and useful, cannot make strong claims about what brain areas are necessary for the functions being investigated. A major impediment to progress in brain lesion-deficit studies, however, is the labor-intensive and ultimately subjective step of having an expert manually segment brain lesions from MRI scans.

This has been highlighted in previous studies comparing inter-rater variability and speed of human compared to automatic lesion identification. Fiez et al. [1] report a 67% ($\pm$ 7%) agreement in overlapping voxels between two expert raters across ten subjects. More recently, other groups have reported an inter-rater overlap of 0.73 $\pm$ 0.2 between experts performing manual lesion segmentation for the ATLAS database [2]. When brain lesion segmentation is performed exclusively by experienced neuroradiologists, median inter-rater agreement has been shown to be as high as 0.78 [3]. However, the involvement of only a small number of patients (N = 14) and the use of lower-resolution scans (6.5 mm slices rather than the typical 1 mm slices used in research) suggests that an inter-rater agreement of 0.78 may be inflated relative to the 0.67 to 0.73 range that seems typical for research studies.

Aside from concerns with inter-rater reliability, manually segmenting lesions is also time consuming, often taking between 4.8 to 9.6 hours. Methods developed for automating this process, however, can segment lesions in roughly a minute [4]. However, manual lesion segmentation remains the method of choice, presumably due to the relatively poor accuracy of available automated methods [4, 5]. Clearly what is needed is a fast, automated method for brain lesion segmentation with a better accuracy than currently available methods.

Indeed, identifying lesions in brain MRI images is a key problem in medical imaging [6, 7]. Previous studies have examined the use of standard machine learning classifiers [8, 9, 10] and convolutional neural networks (CNN) [11, 12, 13, 14] for solving the problem of automating lesion segmentation. Machine learning methods like random forests tend to perform competitively [8] but fare below convolutional neural networks [15].

The first convolutional UNet [11] and subsequent models such as UResNet [14] take as input 2D slices of the MRI image in a single orientation. They predict the lesion for each slice separately and then combine the predictions into a volume. This approach has limited accuracy because it does not consider the other two planes in the image volume. Without some method, such as a post-processing mechanism, for considering views from other orientations, models such as this will be inherently limited by how well a lesion can be detected in a single orientation view. For example, a wide and flat lesion might be readily distinguishable from healthy tissue in an axial but not coronal view. Indeed, a lesion that is more visible in sagittal and coronal views than in the axial view is shown in Figure 9.

To address this limitation, CNN systems have been introduced that can accommodate multiple 2D slice orientations. The dual-path CNN, DeepMedic [14], while not considering multiple 2D orientations, does have two pathways, one for high and one for low resolution slices. Lyksborg et al. [16] use a three path network, one for each of the canonical axial, sagittal and coronal views. Indeed, multi-path systems with up to eight different network paths have been explored previously [17]. Adding paths, however, comes with a cost of having to fit many additional parameters for each path. Fitting these additional parameters leads to an increased risk of over-fitting, as has

been reported for multi-path systems [7].

Multi-path systems must also combine the predictions from each path into a final output. One approach to combining path predictions is a simple majority vote. This was the approach used by Lyksborg et al. [16]. However, this approach risks ignoring important but less frequently represented information, as the outputs from different paths are combined into a final voxel prediction by a simple majority vote. Also, the goal of their network was to segment tumors, where the pathology may present a somewhat different problem than stroke. Indeed, in the current work we show that majority vote performs less well on stroke lesion segmentation than a more inclusive 3D convolutional approach to combining outputs across paths.

We address shortfalls in previous approaches by proposing a novel nine-path system, where each path contains a custom U-Net to accommodate multiple MRI modalities or views, depending on the use case. For example, having both T1 and FLAIR modalities could be useful for segmenting sub-acute strokes that have occurred within, say, the last 5 weeks. For more chronic strokes having occurred more than 6 months previous, multiple T1 views might be more useful than combining with FLAIR. This possibility is tested in Table 1 below. Our system considers three different normalizations of the images along each of the three axial, sagittal and coronal views. Our custom U-Net is weak on its own but powerful as a component of our multi-path system. This makes sense in the context of ensemble learning where weak learners can perform better in an ensemble [18]. We also use a 3D convolutional kernel to merge 2D outputs from each path and show that it gives a better accuracy than majority vote. It is because of this combination of 2D and 3D approaches that we refer to our system as 2.5D.

Critically, we address the challenging issue of model over-fitting by performing a rigorous cross-study validation to evaluate accuracy of lesion identification across sites that differ in numerous ways such as scanner model, patient sample, and expert tracers. This is done by training a model on one set of patient MRIs and then testing the ability of those trained parameters to identify lesions in a separate validation (test) set. Cross-study validation gives a better estimate of the model's true accuracy compared to cross-validation, where train and test samples are simply re-shuffled from the same dataset [19].

Details of our model are provided below, followed by experimental results across three different datasets. We show that our system has significantly higher agreement with ground-truth segmentations by human experts compared to the recent CNN-based methods DeepMedic [14], the original UNet [11], a residual UNet [13], and two non-CNN based machine learning methods based on random forests [10] and naive Bayes [9].

2. Methods

2.1. Convolutional neural networks

Convolutional neural networks are the current state of the art in machine learning for image recognition [20, 21], including for MRI [7]. They are typically composed of alternating layers for convolution and pooling, followed by a final flattened layer. A convolution layer is specified by a filter size and the number of filters in the layer. Briefly, the convolution layer performs a moving dot product against pixels given by a fixed filter of size $k \times k$ (usually 3×3 or 5×5). The dot product is made non-linear by passing the output to an activation function such as a sigmoid or rectified linear unit (also called relu or hinge) function. Both are differentiable and thus fit into the standard gradient descent framework for optimizing neural networks during training. The output

of applying a $k \times k$ convolution against a $p \times p$ image is an image of size $(p-k+1) \times (p-k+1)$. In a CNN, the convolution layers just described are typically alternated with pooling layers. The pooling layers serve to reduce dimensionality, making it easier to train the network.

2.2. Convolutional U-network

After applying a series of convolutional filters, the final layer dimension is usually much smaller than that of the input images. For the current problem of determining whether a given pixel in the input image is part of a lesion, the output must be of the same dimension as the input. This dimensionality problem was initially solved by taking each pixel in the input image and a localized region around it as input to a convolutional neural network instead of the entire image [22].

A more powerful recent solution is the Convolutional U-Net (U-Net) [11]. This has two main features that separate it from traditional CNNs: (a) deconvolution (upsampling) layers to increase image dimensionality, and (b) connections between convolution and deconvolution layers. Another popular U-Net method is the residual U-Net (also known as UResNet [13]) that has residual connections to prevent the gradient from becoming zero (also called the vanishing gradient problem [23]).

2.3. U-Net systems

Since the introduction of the original U-net, several systems have been proposed for analyzing MRI images. DeepMedic is a popular multi-path 3D CNN model that combines high and low resolutions of input images. Previous systems like Lyksborg et. al. [16] consider the three axial, sagittal, and coronal planes in a multi-path ensemble, but use a potentially limiting majority vote approach to combine outputs from each path. Multi-path systems can be challenging to train, as can be seen in the work of Brebisson and Montana [17]. There they train eight networks in parallel to capture various aspects of the input image but report overfitting due to large number of parameters.

Post processing is another important component of U-Net systems to reduce false positives. The post processing methods range from simple ones like connected components and clustering [24, 25] to using 3D CNNs and conditional random fields [14]. The latter methods also end up accounting for temporal dependence between slices, resulting in a higher accuracy.

2.4. Our CNN system

2.4.1. Overview

We developed a modified U-network in a multi-path multi-modal system with a 3D convolutional kernel for post-processing shown in Figure 1. A 3D kernel is like a 2D one except that it has a third dimension that it convolves into as well, and thus it expects a 3D input. For example, in a 2D system kernels are typically 3×3 whereas in a 3D kernel it would be $3 \times 3 \times 3$. Details of our system are provided below, highlighting differences in our approach compared to previous ones.

2.4.2. Multiple paths

Our primary motivation for taking a multi-path approach is to optimize the ability of the model to identify brain lesions by capturing image information from all three angles as well as their normalizations. To return to the overview of our system shown in Figure 1(a), consider the three different normalizations of each of the three axial, sagittal, and coronal planes. For each plane we normalize (1) in the same plane, (2) across the third plane, and (3) both in the same plane first

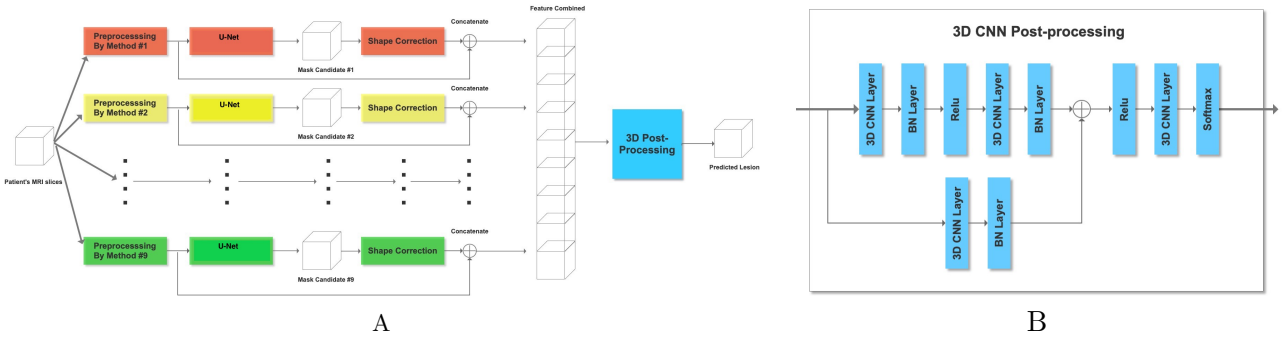


Figure 1: Overview of our entire nine-path system (A) and a zoomed in view of our 3D CNN post processor (B) for combining outputs from each path.

and then across the third, thus giving nine paths. These choices were motivated by our preliminary results not shown here and previous studies showing that different planes work best for different lesion locations [16], and that the best method of normalization may differ depending on image view [7].

2.4.3. Basic U-net

Encoder. First we look at details of our basic U-net that makes up the system. Our U-net used in each path is inspired by the original U-net [11] and a more recent one [26] that attains state of the art accuracies on the BRATS brain tumor MRI benchmark [27]. The encoder portion of our U-net is shown in Figure 2(c). After each convolution we perform a 2×2 average pooling with stride 2 to halve the image dimension. Features from the encoder are passed to the decoder. However, since there are two encoders (one for the original T1-weighted image and the other for its flipped version), corresponding features are combined using the block shown in Figure 2(e). Alternatively, the current network can be used with two different MRI modalities by substituting the T1 image and its flipped version with separate left hemisphere T1-weighted and Fluid-Attenuated Inversion Recovery (FLAIR) images.

Feature fusion. From each encoder we obtain a prediction of a lesion (in the respective normalization and plane) that we merge with a $2 \times 1 \times 1$ 3D convolutional kernel [26, 24]. We take the two feature maps each of dimension $32 \times x \times y$ where 32 is the number of convolutional filters from the encoder layer and $x \times y$ is the input size depending upon the encoder layer (see Figure 2(a)). Stacking refers to adding an extra dimension to make the input $32 \times 2 \times x \times y$ for the 3D kernel. The $2 \times 1 \times 1$ 3D kernel gives an output of $32 \times 1 \times x \times y$ which is "squeezed" to remove the unnecessary dimension to give an output of $32 \times x \times y$ to the decoder.

Decoder. The fused features are then given to the decoder, which we add to the output of deconvolutional layers (briefly explained below), a process shown as a $\oplus$ sign in Figure 2(c). The image dimensions are preserved because of the addition. The previous U-net that served as a starting point for our current effort [26] performed element-wise multiplication of fused features with deconvolved ones. However, this is unlikely to be useful for the current system. Our fused features and upsampled features have small values, so their product would even be smaller. This in turn would give a gradient with zero or near-zero values that would affect the training. Thus we prevent this by adding instead of multiplying fused and upsampled feature values.

Convolutional blocks. Shown in blue in Figure 2(d) are the convolutional blocks used in our encoder and decoder. We use 3×3 convolutional blocks with a stride of 1 and padding of one extra layer in the input to make the output dimensions same as the input. The previous U-net that inspired our design [26] performed Relu activation before adding fused features. Here we perform Relu activation twice. In the context of the decoder, this means Relu activation is performed after adding fused features to upsampled ones. Performing Relu activation after addition rather than before has been shown to be more accurate for image classification [28].

Deconvolutional blocks. Deconvolutional blocks (also known as transposed or fractionally strided convolutions) are meant to increase the dimensionality of images [29]. The term transpose arises from the fact that a deconvolution is simply the product of the transpose of the convolution weight matrix with the output when the stride is 1. If the stride is more than one we insert zeros in between the input to obtain the correct transpose result (as well-explained in Dumoulin and Visin [29]) We use 2×2 deconvolutions with a stride of 2 that doubles the image dimensions in both axes.

2.4.4. Post-processing

The output of each of the nine paths in our system is a 2D mask showing the predicted location of the lesion in the same view as the input image, as in Figure 2(a). The lesion prediction mask is binarized by rounding to 0 if the values in the mask are below 0.5, otherwise values are rounded up to 1. We stack each predicted lesion with the original input image and combine all slices to form a $2 \times 192 \times 224 \times 192$ volume. Since we have nine paths this becomes of size $18 \times 192 \times 224 \times 192$. This is passed to our 3D CNN post-processor as described below.

In the post-processor shown in Figure 1(b), we have a main path containing 36 3D $3 \times 3 \times 3$ kernels each with 18 channels, or equivalently 36 3D kernels each of size $18 \times 3 \times 3 \times 3$. Following that, the second 3D CNN in the main path has 9 3D $3 \times 3 \times 3$ kernels each with 36 channels, and two final 3D CNNs each of dimensions $3 \times 3 \times 3$ with 9 channels.

2.4.5. Loss function

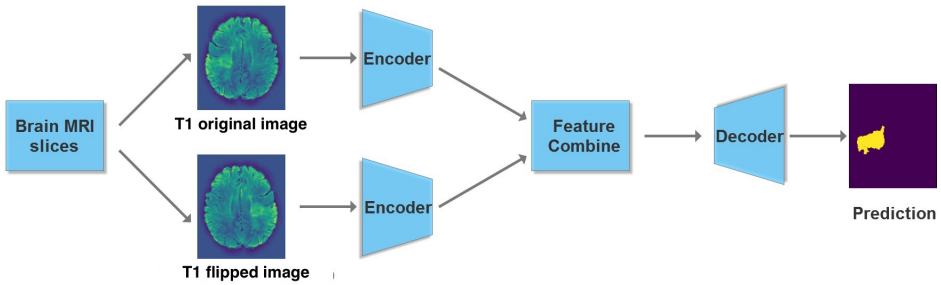
The final output from the post-processor has two channels each of dimensions $192 \times 224 \times 192$. The target lesion has the same dimensions but just one channel. The first channel in our output predicts the lesion and the second one predicts the complement of it. We convert the outputs of each channel into probabilities with softmax [30] and combined them into a modified Dice loss function [31, 32]. For a single channel output the Dice loss is defined to be $1 - D$ where

$$D(p) = \frac{2\sum_i p_i r_i}{\sum_i p_i^2 + \sum_i r_i^2}$$

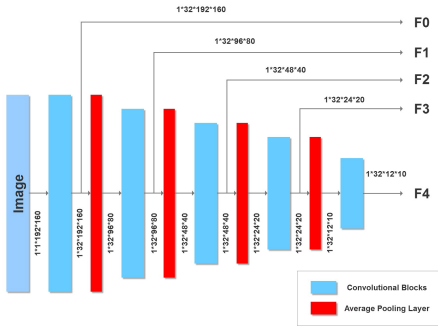
p_i are the predicted softmax outputs of the channel, and r_i is 1 if the voxel has a lesion and 0 otherwise. If we are predicting the complement of the lesion then the values of r_i are flipped from 0 to 1 and 1 to 0. With our two channel output p and q our loss becomes $2 - (D(p) + D(q))$ where the latter $D(q)$ is for the complement.

2.5. Imaging Data

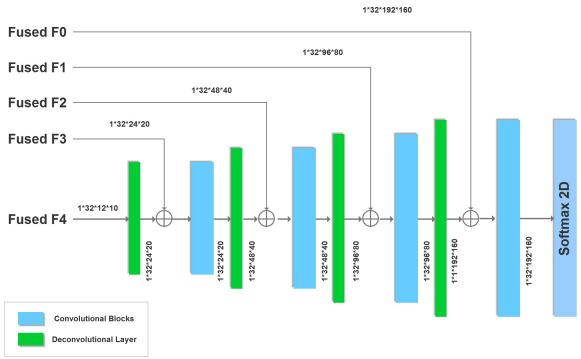
We obtained high-resolution (1 mm^3) whole-brain MRI scans from 25 patients from the Kessler Foundation (KES), a neuro-rehabilitation facility in West Orange, New Jersey. We also obtained 20 high-resolution scans from the Medical College of Wisconsin (MCW). Data heterogeneity is



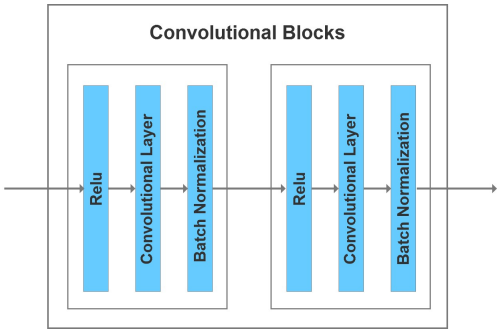
(a) Overview of our dual-path U-network. We have a separate encoder for the original T1 image of the brain scan and one for its flipped version. Alternatively, two different image modalities may also be used instead of two different hemispheres.



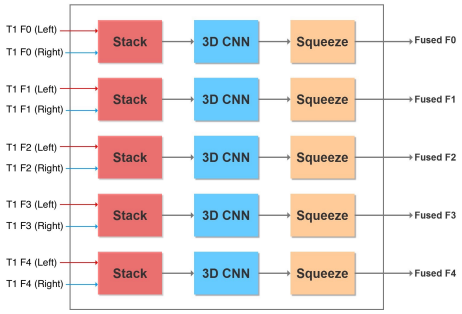
(b) U-Net Encoder with five convolutional blocks
Also shown are image dimensions after each convolution.



(c) U-Net Decoder with four convolutional and deconvolutional blocks
Also shown are image dimensions after each deconvolution.



(d) Convolutional blocks used in encoder above



(e) Fuse features from encoding the original and flipped images (or alternatively encoding from two different image formats)

Figure 2: Our U-network models with encoder and decoder details.

important for widespread applicability of the model. To that end, we included data from a variety of time points: subacute (< 5 weeks post stroke) and chronic (> 3 months post stroke). Strokes of both hemorrhagic and ischemic etiology were included. The lesions visualized on the scans were hand-segmented by a trained human expert, as described for the KES scans in [33] and the MCW scans [34, 35]. To move these scans into standard Montreal Neurological Institute (MNI) reference

space [36], we used the non-linear warping tool, 3dQwarp, from the AFNI software suite [37]. The segmented lesion was used as an exclusion mask so that the lesioned territory would be excluded from the warping procedure. This prevents non-lesioned brain tissue from being distorted to fill in the lesioned area. This transformation was performed on the T1 images, resulting in skull-stripped output in MNI space. This calculated transformation for each participant was then applied to the FLAIR image (KES only) and hand-traced lesion mask.

We also obtained scans and stroke lesion masks from the public ATLAS database [2] and processed them as just described for the KES and MCW data. We selected images according to the following criteria to focus on cases with single lesions in the left hemisphere:

```
Session = t01 [T1 Scans only]
LH_Cort + LH_SubCort = 1 [Cortical OR Sub-Cortical Lesion only]
RH_Cort = 0 [No Cortical Right Hemisphere Lesion]
RH_SubCort = 0 [No Sub-Cortical Right Hemisphere Lesion]
Other_Location = 0 [No Lesion elsewhere]
Hemisphere = Left [Left Hemisphere only]
```

This resulted in 54 images being selected from the ATLAS set. Thus we included a total of 99 images altogether across the three datasets. We divided these into two groups, ATLAS or Kessler+MCW, for cross-study comparisons. We then combined them to perform a five-fold cross-validation across all 99 images.

2.6. Comparison of CNN Methods

We compared our CNN to three state of the art recently published CNNs shown below. Our system was implemented using Pytorch [38], the source code for which is available on our GitHub site `available.upon.acceptance`. In each of our experiments we train our model, UNet, and UResNet with stochastic gradient descent and Nesterov momentum [39] of 0.9 and weight decay of .0001. We use a batch size of 32, starting from an initial learning rate of 0.01 with a 3% weight decay after each epoch for a total of 50 epochs. In DeepMedic we use the default settings of learning rate of 0.001, the RMSProp optimizer [39] with a weight decay of .0001, batch size of 10, and a total of 20 epochs.

- DeepMedic [14]: This is a popular dual-path 3D convolutional neural network with a conditional random field to account for temporal order of slices. DeepMedic contains a path for low- and a separate path for high-resolution of images. Its success was demonstrated by winning the ISLES 2015 competition to identify brain injuries, tumors, and stroke lesions. The code for implementing DeepMedic is freely available on GitHub, `https://github.com/Kamnitsask/deepmedic`.

- UResNet [13]: This is a convolutional neural network with residual connections [12]. The code for implementing UResNet is also freely available on GitHub, `https://github.com/DeepLearnPhysics/pytorch-uresnet`.

- UNet [11]: The was the original convolutional U-network proposed for biomedical image processing. Its code is also available on GitHub, `https://github.com/thonycc/PFE/tree/af9e804f71684b73cf7f3b25557edcf6a1b307b3`.

Two other non-CNN-based machine learning packages were also included because they have been made freely available to the brain imaging community and have been developed for ease of use. Both take a patch-based approach to automating lesion segmentation. That is, these methods convert the input image into multiple patches that are used to train the model. They are LINDA [10], based on a random forests algorithm, and a second method based on Gaussian naive Bayes [9].

2.7. Data analysis

2.7.1. Measure of accuracy: Dice coefficient

The Dice coefficient is typically used to measure the accuracy of predicted lesions in MRI images [40]. The output of our system and that of other methods is a binary mask of the dimensions as the input image, but with a 1 for each voxel calculated to contain a lesion, and a 0 otherwise. Comparison of the human expert-segmented lesion mask with that from automated methods is quantified with the Dice coefficient. Starting with the human binary mask as ground truth, each predicted voxel is determined to be either a true positive (TP, also one in true mask), false positive (FP, predicted as one but zero in the true mask), or false negative (FN, predicted as zero but one in the true mask). The Dice coefficient is formally defined as

$$DICE = \frac{2TP}{2TP + FP + FN} \tag{1}$$

2.7.2. Measure of statistical significance: Wilcoxon rank sum test

The Wilcoxon rank sum test [41] (also known as the Mann-Whitney U test) can be used to determine whether the difference between two sets of measurements is significant. More formally, it tests for the null hypothesis that a randomly selected point from a sample is equally likely to be lower or higher than a randomly selected one from a second sample. It is a non-parametric test for whether two sets of observations are likely to be from different distributions, without assuming a particular shape for those distributions.

3. Results

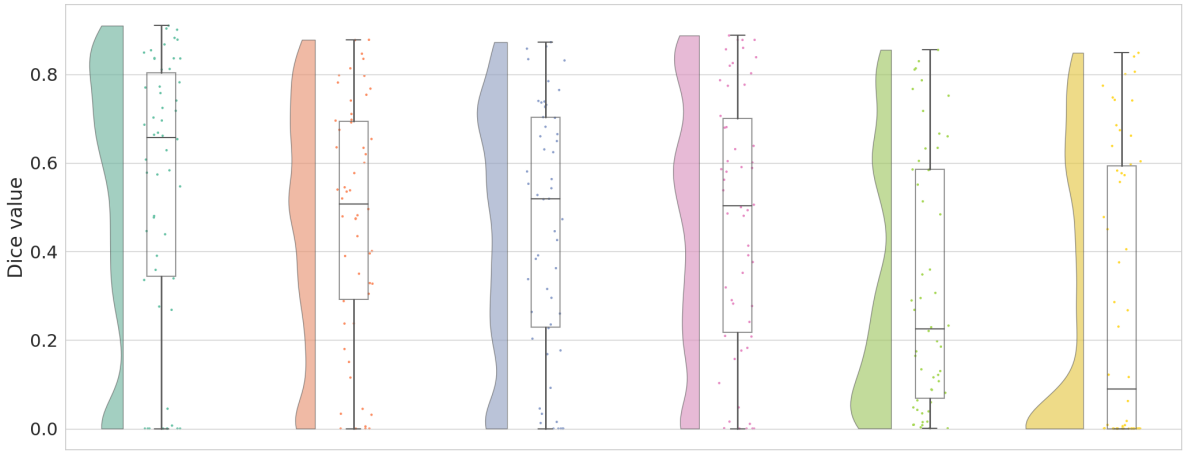
In the results presented below, we take the rare and rigorous step of performing cross-study validations across independent datasets [19]. We also examine results from cross-validation in the combined dataset from the three different sources (KES, MCW, and ATLAS).

3.1. Cross-study validation results

To create relatively balanced sets in terms of number of scans, we combine the KES and MCW datasets into one. This yielded 45 samples in KES+MCW and 54 in ATLAS. We first train all convolutional neural networks (CNNs) on the KES+MCW data and test their ability to predict lesion locations in the ATLAS set. We then repeat the same procedure but with the train and test datasets reversed. Since LINDA and GNB come pre-trained and were intended for out-of-the-box use rather than re-training, we ran them as-is. Both programs have skull-removal built into their pipelines. Because the ATLAS images were the largest dataset with the skull still intact, we restricted our test of the LINDA and GNB methods to the ATLAS dataset.

3.1.1. Train on KES+MCW, predict on ATLAS

Figure 3 shows the Dice coefficient values on the ATLAS test dataset with training performed on KES and MCW images. Results show that the current system, with a median Dice value of 0.66, yielded the best performance. This was not just due to a few high values, as its Dice values generally clustered toward the higher end. The Dice values of UNet, UResNet, and DeepMedic Dice have a more even distribution than our system and lower median values. Both LINDA and GNB have Dice values clustered toward the lower end. Figure 3 also shows that our system has the highest mean Dice value. This value is reliably higher than all other methods under the Wilcoxon rank test [41] ($p < 0.001$). All the convolutional networks achieve better median values than LINDA and GNB.



Method	Our system	UNet	UResNet	DeepMedic	LINDA	GNB
Mean Dice	0.54	0.47	0.45	0.47	0.32	0.29

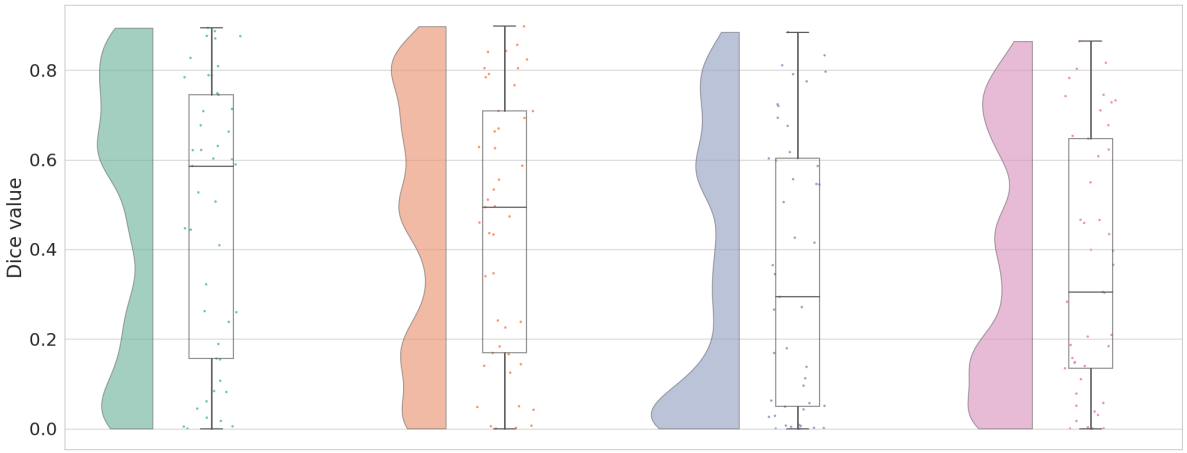
Figure 3: Raincloud plots of Dice coefficient values of all models trained on KES+MCW and tested on ATLAS. For each method we show the distribution of Dice coefficients across all test images as well as the five summary values: median (middle horizontal line), third quartile (upper horizontal line), first quartile (lower horizontal line), min (lowermost bar), and max (uppermost bar). All models except for LINDA and GNB are trained on KES+MCW. The Table below the graph contains the mean Dice coefficients of all models on the ATLAS test data.

3.1.2. Train on ATLAS, predict on KES+MCW

Figure 4 shows results from the other direction of the cross-study analysis: training on ATLAS and testing on KES+MCW. In this case, although our system has the highest median, its distribution of Dice values is no longer clustered toward the high end as it was previously. The mean Dice value of our system is marginally above that of UNet alone and not statistically distinguishable from it. Compared to UResNet and DeepMedic, however, our method performs better, as shown from its reliably higher Dice values ($p < 0.001$).

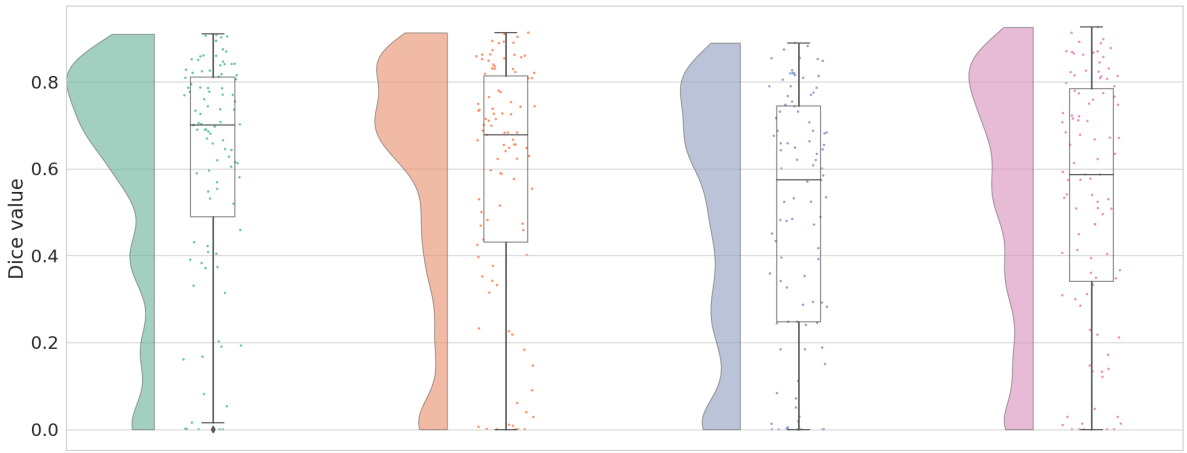
3.2. Cross-validation results on all datasets ATLAS, KES, and MCW combined

To take full advantage or our relatively large dataset, we combined images from all three sources to produce an overall dataset of 99 samples. We then performed a five-fold cross-validation on this combined dataset to evaluate the accuracy of each method. Figure 5 shows that our system again



Method	Our system	UNet	UResNet	DeepMedic
Mean Dice	0.47	0.45	0.35	0.37

Figure 4: Raincloud plots of Dice coefficient values for all models trained on ATLAS and tested on KES+MCW. Also shown in the table are mean Dice coefficients of each method, as tested on the KES+MCW set.



Method	Our system	UNet	UResNet	DeepMedic
Mean Dice	0.62	0.58	0.49	0.54

Figure 5: Raincloud plot of Dice coefficient values obtained by five-fold cross validation on all our data combined: ATLAS+Kessler+MCW. In the Table are the mean Dice coefficients given by cross-validation.

has the highest median Dice value. Our system also has the highest mean Dice value at 0.62, performing reliably better than the next best system, UNet, at 0.58. Indeed, our system performed better ($p < 0.001$) than all three of the other CNN-based systems.

In addition to reporting this advantageous numeric performance of our system, an overall illustration of how the lesion masks produced by the current model compared to those from the other CNN-based models is in Figure 6. The expert-traced lesions (A) are shown alongside those produced by our system (B) and the models (C-E).

One point to note is that while our system performed significantly better in terms of overlap with human expert tracings as measured by the Dice coefficient, visually all the automatic methods appear grossly similar to the human expert segmentations.

3.3. Distribution of Dice coefficients across lesion size

Lesions with $x \times y \times z$ dimensions less than $20 \times 20 \times 25$ mm were classified as small, and any lesions with dimensions greater than those were considered large. In Figure 7 we show a raincloud plot of Dice values obtained by our system in the cross-validation and cross-study settings.

Smaller lesions are generally harder to identify than larger ones [9, 10, 5]. To compare performance between lesion sizes, we split the lesions into small and large categories based on the distribution of lesion sizes in the overall set.

In all three cases, our method does very well on large lesions. In fact, when we train on KES+MCW and predict on ATLAS, the median Dice is above 0.8 for large lesions. In the cross-validation on all data combined, our model is significantly better than all methods except for DeepMedic, with p-values below 0.05. An example of a larger lesion is shown in Figure 8. The output lesion masks in red show our method and the other three to be qualitatively similar. An apparent exception is DeepMedic, which misidentifies tissue in the right-hemisphere as being lesioned. This mis-identification would seem to be an exception, however, given the similar numeric performance between our method and DeepMedic.

Smaller lesions, on the other hand, are associated with lower median Dice values overall, as generally expected. DeepMedic has particular difficulty with smaller lesions, whereas our system shows significantly greater accuracy than DeepMedic and UResNet. Interestingly, the distribution of Dice values for small lesions clusters towards the high end in the cross-validation

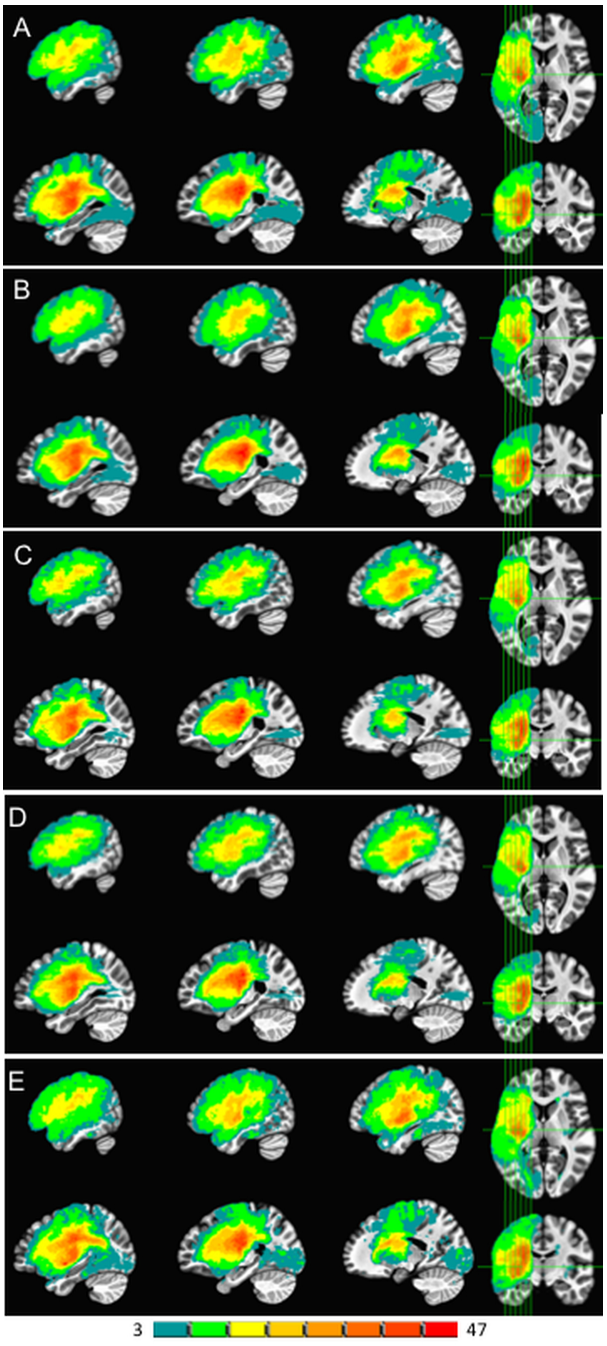
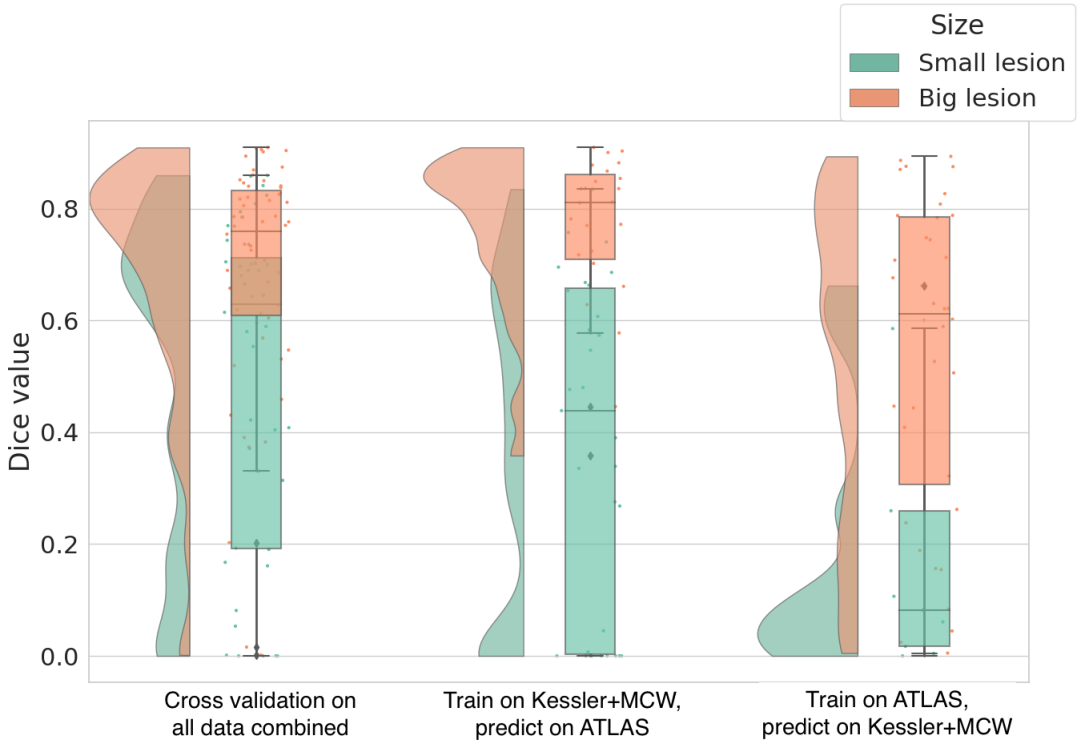


Figure 6: Lesion overlap map results from 5-fold cross-validation on the entire 99 scan dataset. The leftmost side of the color scale in teal shows locations with 3 spatially overlapping lesions, while the rightmost side in red shows a maximum of 47 overlapping lesions. Hand-segmented lesions are in panel A. Our 2.5D CNN model is in panel B. The UNet model is in panel C. The URestNet model is in panel D. And the DeepMedic model output is in E.



Method	Our system	UNet	UResNet	DeepMedic
Small lesions	0.5	0.48	0.36	0.33
Large lesions	0.69	0.66	0.59	0.68

Figure 7: Raincloud plot showing the distribution and five summary statistics of Dice coefficients in three different scenarios. The left panel shows Dice values given by cross-validation on all the data combined. The middle panel shows a cross-study scenario where the current model is trained on KES+MCW and tested on ATLAS. The right panel shows results from training on ATLAS and testing on KES+MCW. In the Table below the plots we show the mean Dice values of our system and the other CNNs on small and large lesions separately.

setup with the most training data (all three datasets combined). This suggests that still more data would enable the model to achieve better accuracy at identifying small lesions. An example of a smaller lesion classification for the combined data cross-validation scenario is shown in Figure 9. This figure shows how the similarity of the overall contours of the model-based lesion masks (C-F) match up with the hand-segmented lesion mask (B). It also illustrates the face validity of the Dice coefficient, where higher Dice values also qualitatively correspond better to the hand-segmented lesion mask.

CNNs are a type of neural network, and what neural networks learn depends on what information is in the training data [42]. In the cross-study scenario where we train on KES+MCW and test on ATLAS, the distribution of Dice values for smaller lesions is spread somewhat uniformly. However, when the network is trained on the ATLAS data and tested on the KES+MCW set, performance is worse. Thus the general rule that the information in the training dataset largely determines what the model can learn is also shown here for detecting small lesions.

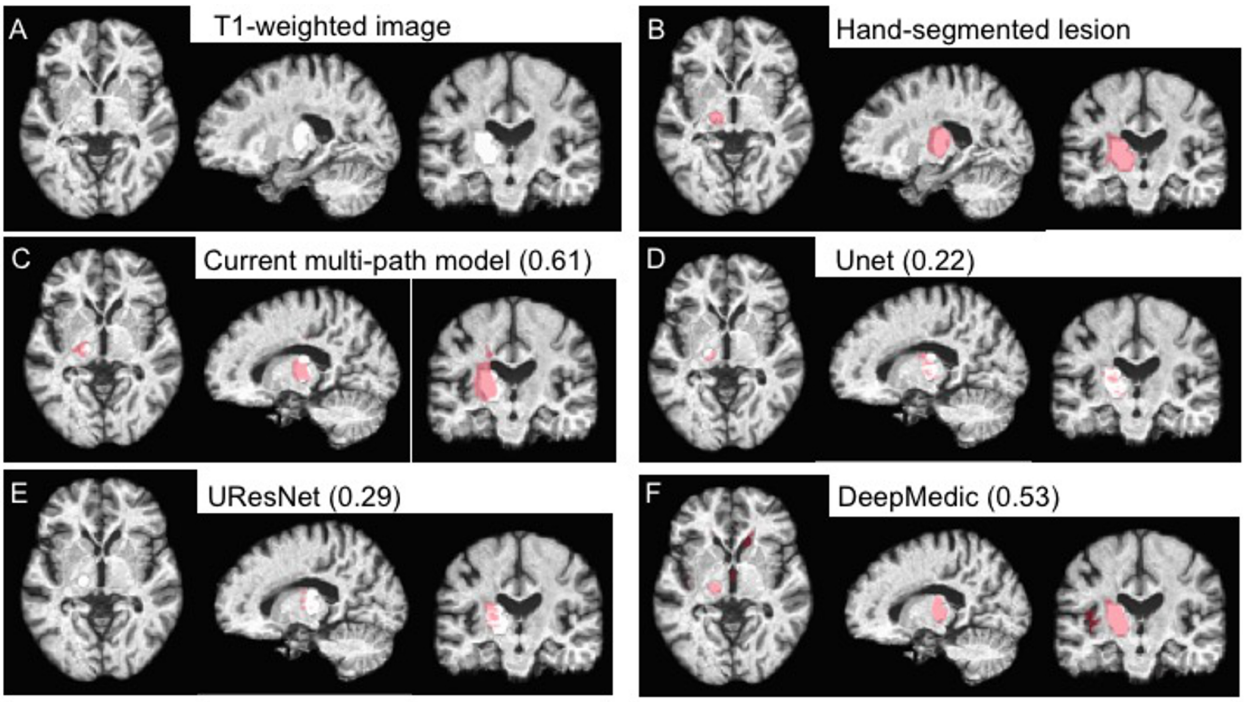


Figure 8: Example of a relatively large (10,739 mm^3) lesion (A) along with its hand-segmented mask (B). The remaining panels show the lesion masks derived from the 5-fold cross-validation with all 99 scans for our 2.5D model (C) and the other CNN-based approaches (D-F). The label for each model is followed by the corresponding Dice value for the lesion mask it produced in parentheses. Lesion masks overlaid in red are rendered semi-transparent to visualize the overlap between the lesion and the mask.

3.4. Consolidating multi-path outputs

Previous multi-path approaches use a majority vote to combine outputs from different paths [16]. We compare our 3D CNN for combining multi-path outputs to using the majority vote and a simple union. In the union method, if at least one pixel has a one across the paths then the aggregated output also has a one in that pixel. Figure 10 shows that the union clearly performs more poorly than majority vote and our 3D CNN. Between the two better performing methods, the 3D CNN is reliably better than majority vote by a 4% margin with a p-value of 0.004. Also compared to post-processing with majority vote, the Dice values of the 3D CNN are concentrated more towards the high end.

3.5. Mulit-modal T1 vs. T1+FLAIR

Our basic U-Net model is multimodal (specifically, bimodal) in that it allows for different image formats. Since the current project is focused exclusively on left hemisphere lesions, we present the model with T1 and FLAIR image formats of the lesioned left hemisphere. Below in Table 1 we show the cross-validation accuracy of our model on the KES and MCW images. When presented together, there is no significant difference between the two. However, if we look at just KES images that contain smaller lesions (and more recent, in the less than 5-week post-stroke range), then adding FLAIR confers a significant advantage. In the case of MCW images only that have lesions exclusively in the chronic epoch (at least 6 months post-stroke), the T1 images alone actually result in better performance than when the corresponding FLAIR images are added. This pattern

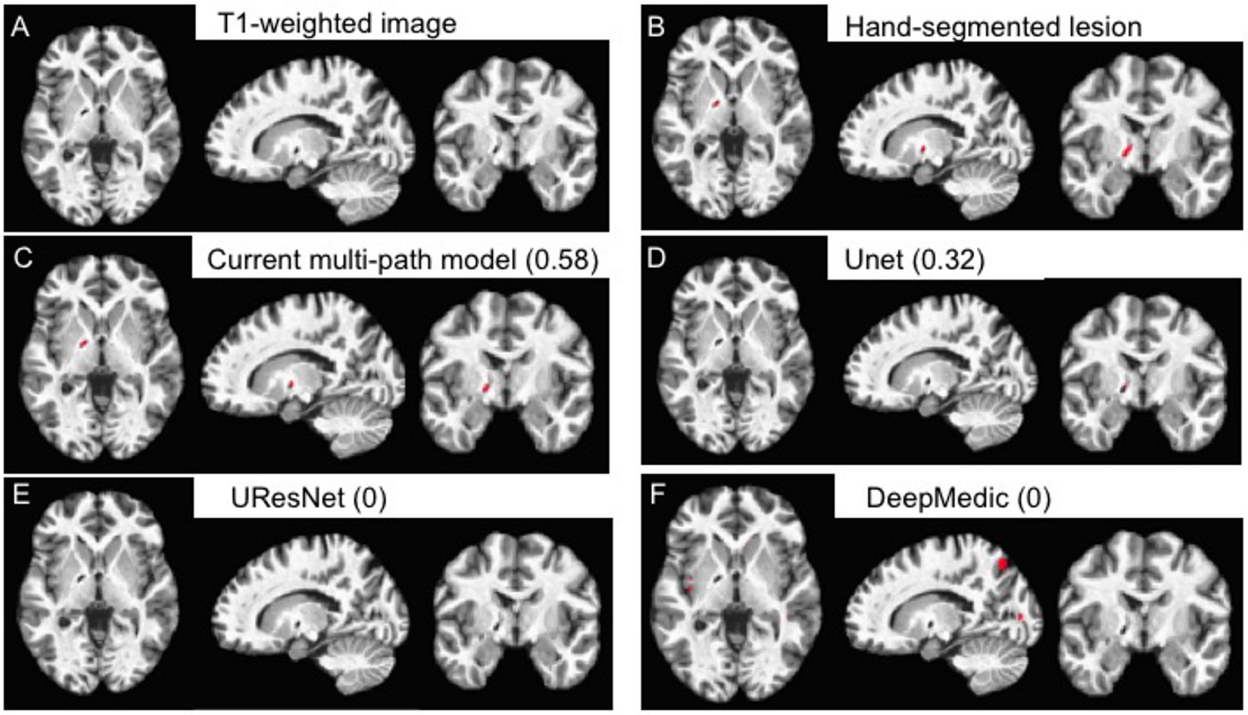


Figure 9: Example of a relatively small (85 mm^3) lesion (A) along with its hand-segmented mask (B). The remaining panels show the lesion masks derived from the 5-fold cross-validation with all 99 scans for our 2.5D model (C) and the other CNN-based approaches (D-F). The label for each model is followed by the corresponding Dice value for the lesion mask it produced in parentheses. Lesion masks are overlaid in red. Note that the lesion masks derived from the DeepMedic model (F) are valse positives rather than actual lesions.

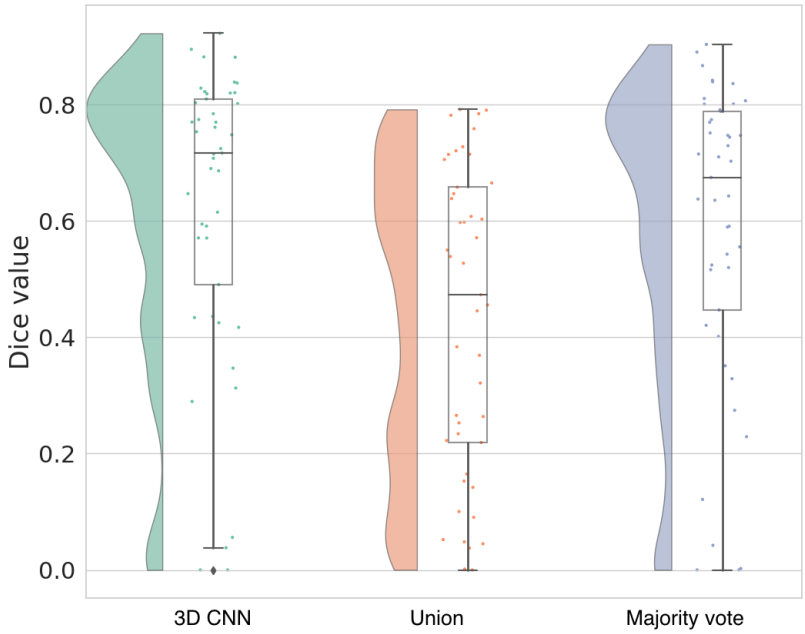
corresponds with the standard clinical observation that FLAIR scans are useful for more recent stroke lesions but less so for those in the chronic phase [43]. Such correspondence lends additional face validity to our model.

Data	T1	T1+FLAIR	Wilcoxon rank test p-value	Average lesion size (in pixels)
KES+MCW	0.59	0.63	0.2	58388
KES	0.47	0.58	0.004*	34054
MCW	0.74	0.68	0.002*	88804

Table 1: Mean Dice coefficients of our method on T1 vs. T1+FLAIR images on Kessler+MCW. Also shown are Wilcoxon rank test p-values and average lesion size of images in the combined and individual datasets

4. Discussion

Here we have created, trained, and tested a new multi-path 2.5D convolutional neural network. The fractional designation on the dimension comes from its use of nine different 2D paths, followed by concatenation of the learned features across the paths, which are then passed to a 3D CNN for post-processing. This 2.5D design combines flexible and efficient 2D paths that process the data in different canonical orientations and normalizations with a 3D CNN that combines the 2D features in a way that informs the final 3D image output. Comparison of our system to previous efforts



3D CNN	Union	Majority vote
0.62	0.43	0.58

Figure 10: Raincloud plot of Dice coefficient values of three different post-processing approaches in our system as given by five-fold cross validation on KES+MCW images combined. Mean Dice values for each approach are presented in the accompanying table.

shows that CNN-based systems outperform more traditional machine-learning approaches based on random forests or Gaussian naive Bayes algorithms. Compared to other CNN systems, our system shows reliably superior performance in its ability to automatically segment stroke lesions from healthy tissue in the left hemisphere.

As methods such as this continue to improve the automated segmentation of brain lesions, a question arises. How good is good enough? An intuitive answer to this question comes from human expert raters. As mentioned in the Introduction, human expert raters have been shown to produce lesion segmentations with overlapping volumes between raters in the 67% to 78% range [3, 1, 2], though 73% may be a more realistic upper value given the highly expert raters and limited scope of the data used by Neumann et al. [3] to obtain the 78% value. The Dice coefficient used here is a formal measure of degree of spatial overlap that ranges between 0 and 1. Therefore a Dice coefficient in the 0.67 range can be considered to be at the edge of the human expert gold standard. When combining the datasets and performing iterative training and testing using standard 5-fold cross-validation, the lesion traces from our model overlap with human experts with a mean Dice coefficient of 0.62. While the 0.67 to 0.73 human benchmark range should be interpreted with caution because those numbers are based on data that are not identical to the data considered here, the accuracy of our system relative to previous efforts does suggest that deep learning-based CNN methods are beginning to approach human expert level accuracy for stroke lesion segmentation.

4.1. Future directions

An alternative to our system is to have a multi-modal 3D U-Net instead of the current 2D ones. While promising, it may be difficult to implement in practice. Training a 3D CNN involves adjusting many more parameters than for a 2D CNN, and would therefore require more data to train. A second future direction is to extend our current left hemisphere-focused system to include lesions to the right hemisphere. This extension should be relatively straightforward, as nothing is preventing our current system from being trained and tested on images with lesions to either hemisphere.

4.2. Conclusion

We have presented a multi-path, multi-modal convolutional neural network system for identifying lesions in brain MRI images. While the data with which our model is trained and tested includes exclusively left hemisphere lesions, our model can be trained and tested on lesions present anywhere in the brain. In cross-study and cross-validation tests, our model shows superior performance compared to existing CNN and non-CNN based machine learning methods for lesion identification. Our method extends previous efforts showing relatively high segmentation accuracy for large lesions. Given sufficient data, it markedly improves on previous efforts by being able to segment smaller lesions as well. We provide freely available open source code to train and test our model.

This advance in performance is critically significant, as it brings the field closer to removing the bottleneck of having human experts spend numerous hours hand-segmenting brain lesions on MRI scans. Once automated methods are sufficiently accurate and widely available, they will free up researchers to focus their time on other critical aspects of neuropsychological data acquisition and analysis. The hope is this re-allocation of expert resources will help advance the pace at which we can further our understanding of the critical neural basis of thinking and behavior.

5. References

References

[1] Julie A. Fiez, Hanna Damasio, and Thomas J. Grabowski. Lesion segmentation and manual warping to a reference brain: Intra- and interobserver reliability. *Human Brain Mapping*, 9(4):192–211, 2000.

[2] Sook-Lei Liew, Julia M Anglin, Nick W Banks, Matt Sondag, Kaori L Ito, Hosung Kim, Jennifer Chan, Joyce Ito, Connie Jung, Nima Khoshab, et al. A large, open source dataset of stroke anatomical brain images and manual lesion segmentations. *Scientific data*, 5:180011, 2018.

[3] Anders B Neumann, Kristjana Y Jonsdottir, Kim Mouridsen, Niels Hjort, Carsten Gyldensted, Alberto Bizzi, Jens Fiehler, Roberto Gasparotti, Jonathan H Gillard, Marc Hermier, et al. Interrater agreement for final infarct mri lesion delineation. *Stroke*, 40(12):3768–3771, 2009.

[4] Marko Wilke, Bianca de Haan, Hendrik Juenger, and Hans-Otto Karnath. Manual, semi-automated, and automated delineation of chronic brain lesions: A comparison of methods. *NeuroImage*, 56(4):2038 – 2046, 2011.

[5] Kaori L Ito, Hosung Kim, and Sook-Lei Liew. A comparison of automated lesion segmentation approaches for chronic stroke t1-weighted mri data. *bioRxiv*, page 441451, 2018.

[6] Zeynettin Akkus, Alfiia Galimzianova, Assaf Hoogi, Daniel L Rubin, and Bradley J Erickson. Deep learning for brain mri segmentation: state of the art and future directions. *Journal of digital imaging*, 30(4):449–459, 2017.

[7] Jose Bernal, Kaisar Kushibar, Daniel S Asfaw, Sergi Valverde, Arnau Oliver, Robert Martí, and Xavier Lladó. Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. *Artificial intelligence in medicine*, 2018.

[8] Oskar Maier, Christoph Schröder, Nils Daniel Forkert, Thomas Martinetz, and Heinz Handels. Classifiers for ischemic stroke lesion segmentation: a comparison study. *PloS one*, 10(12):e0145118, 2015.

[9] Joseph C Griffis, Jane B Allendorfer, and Jerzy P Szaflarski. Voxel-based gaussian naïve bayes classification of ischemic stroke lesions in individual t1-weighted MRI scans. *Journal of neuroscience methods*, 257:97–108, 2016.

[10] Dorian Pustina, H Branch Coslett, Peter E Turkeltaub, Nicholas Tustison, Myrna F Schwartz, and Brian Avants. Automated segmentation of chronic stroke lesions using linda: Lesion identification with neighborhood data analysis. *Human brain mapping*, 37(4):1405–1421, 2016.

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] R Guerrero, C Qin, O Oktay, C Bowles, L Chen, R Joules, R Wolz, MC Valdés-Hernández, DA Dickie, J Wardlaw, et al. White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks. *NeuroImage: Clinical*, 17:918–934, 2018.

[14] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.

[15] Muhammad Rachmadi, Maria Valdés-Hernández, Maria Agan, and Taku Komura. Deep learning vs. conventional machine learning: Pilot study of wmh segmentation in brain mri with absence or mild vascular pathology. *Journal of Imaging*, 3(4):66, 2017.

[16] Mark Lyksborg, Oula Puonti, Mikael Agn, and Rasmus Larsen. An ensemble of 2d convolutional neural networks for tumor segmentation. In *Scandinavian Conference on Image Analysis*, pages 201–211. Springer, 2015.

[17] Alexander de Brebisson and Giovanni Montana. Deep neural networks for anatomical brain segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–28, 2015.

[18] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[19] Christoph Bernau, Markus Riester, Anne-Laure Boulesteix, Giovanni Parmigiani, Curtis Huttenhower, Levi Waldron, and Lorenzo Trippa. Cross-study validation for the assessment of prediction algorithms. *Bioinformatics*, 30(12):i105–i112, 2014.

[20] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[22] Dan Ciresan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in neural information processing systems*, pages 2843–2851, 2012.

[23] Sepp Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116, 1998.

[24] Matthew Lai. Deep learning for medical image segmentation. *arXiv preprint arXiv:1505.02000*, 2015.

[25] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017.

[26] Kuan-Lun Tseng, Yen-Liang Lin, Winston Hsu, and Chung-Yang Huang. Joint sequence learning and cross-modality convolution for 3d biomedical segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 3739–3746. IEEE, 2017.

[27] Bjoern Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elisabeth Gerstner, Marc-Andre Weber, Tal Arbel, Brian Avants, Nicholas Ayache, Patricia Buendia, Louis Collins, Nicolas Cordier, Jason Corso, Antonio Criminisi, Tilak Das, Hervé Delingette, Cagatay Demiralp, Christopher Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan Iftekharuddin, Raj Jena, Nigel John, Ender Konukoglu, Danial Lashkari, Jose Antonio Mariz, Raphael Meier, Sergio Pereira, Doina Precup, S. J. Price, Tammy Riklin-Raviv, Syed Reza, Michael Ryan, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos Silva, Nuno Sousa, Nagesh Subbanna, Gabor Szekely, Thomas Taylor, Owen Thomas, Nicholas Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging*, page 33, 2014.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[29] Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*, 2016.

[30] Ethem Alpaydin. *Machine Learning*. MIT Press, 2004.

[31] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016.

[32] Ken CL Wong, Mehdi Moradi, Hui Tang, and Tanveer Syeda-Mahmood. 3d segmentation with exponential logarithmic loss for highly unbalanced object sizes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 612–619. Springer, 2018.

[33] Olga Boukrina, AM Barrett, Edward J Alexander, Bing Yao, and William W Graves. Neurally dissociable cognitive components of reading deficits in subacute stroke. *Frontiers in human neuroscience*, 9:298, 2015.

[34] Sara B Pillay, Benjamin C Stengel, Colin Humphries, Diane S Book, and Jeffrey R Binder. Cerebral localization of impaired phonological retrieval during rhyme judgment. *Annals of neurology*, 76(5):738–746, 2014.

[35] Jeffrey R Binder, Sara B Pillay, Colin J Humphries, William L Gross, William W Graves, and Diane S Book. Surface errors without semantic impairment in acquired dyslexia: a voxel-based lesion–symptom mapping study. *Brain*, 139(5):1517–1526, 2016.

[36] Vladimir Fonov, Alan C Evans, Kelly Botteron, C Robert Almli, Robert C McKinstry, D Louis Collins, Brain Development Cooperative Group, et al. Unbiased average age-appropriate atlases for pediatric studies. *Neuroimage*, 54(1):313–327, 2011.

[37] Robert W Cox. Afni: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical research*, 29(3):162–173, 1996.

[38] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

[39] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[40] Alex P Zijdenbos, Benoit M Dawant, Richard A Margolin, and Andrew C Palmer. Morphometric analysis of white matter lesions in mr images: method and validation. *IEEE transactions on medical imaging*, 13(4):716–724, 1994.

[41] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

[42] David C Plaut, James L McClelland, Mark S Seidenberg, and Karalyn Patterson. Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological review*, 103(1):56, 1996.

[43] Peter E Ricci, Jonathan H Burdette, Allen D Elster, and David M Reboussin. A comparison of fast spin-echo, fluid-attenuated inversion-recovery, and diffusion-weighted mr imaging in the first 10 days after cerebral infarction. *American Journal of Neuroradiology*, 20(8):1535–1542, 1999.