

# Fusion-Based Multimodal Brain Tumor Detection Using Convolution Neural Network



Soumyabroto Banerjee, Sneha Roy, and Arpita Das

**Abstract** Early detection and proper treatment of brain tumor are essential to prevent permanent damage of brain even patient death. Present study proposes an automatic, effective approach to detect brain lesions in early stage including fusion of multimodal images to enrich the information content. As the convolution neural network (CNN) extracts the required features, fused images improve the quality of the feature bank which in turn enhances the classification accuracy. Present work also develops the modified architecture of CNN that contains only few parameters compared to the standard CNN model (VGG-16) available in Google Colab. Hence, the computation time is low, and this architecture is trainable on a local PC with standard RAM. Experimental results show the assessment of classification accuracy in terms of well-known receiver operating characteristic method, and the outcomes produce satisfactory results.

**Keywords** MRI · SPECT-Tc · Multimodal fusion · CNN architecture

## 1 Introduction

Brain tumor classification is one of the most important and difficult tasks in many medical-image applications because it usually involves a huge amount of data. Artifacts due to patient's motion, limited acquisition time, and soft tissue boundaries are usually not well defined. There are large class of tumor types which have variety of shapes and sizes. They may appear indifferent sizes and types with different image intensities. Some of them may also affect the surrounding structures that change the image intensities around the tumor. Before the treatment of chemotherapy, radiotherapy, or brain surgeries, there is a need for medical practitioners to confirm the boundaries and regions of the brain tumor and determine where exactly it is located and the exact affected area. Brain tumor classification acts as a pre-requisite stage

---

S. Banerjee (✉) · S. Roy · A. Das

Department of Radio Physic and Electronics, University of Calcutta, Kolkata, India

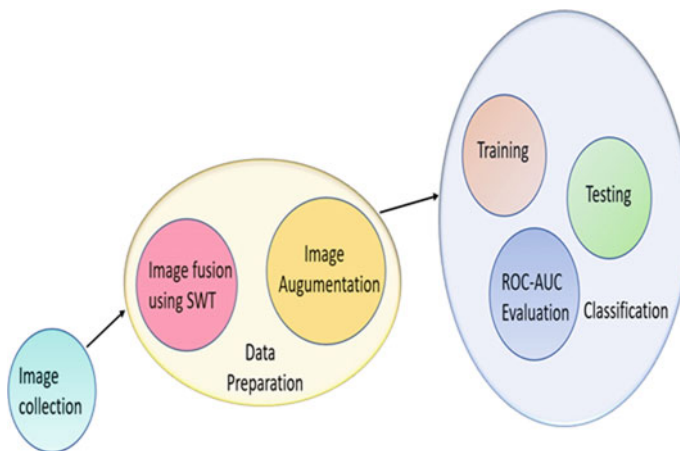
A. Das

e-mail: [adrpe@caluniv.ac.in](mailto:adrpe@caluniv.ac.in)

for doctors to identify the tumor before performing surgeries to identify the exact location of the tumor. A computer-aided diagnosis (CAD) system is designed to aid the radiologist in the diagnosis of such tumors.

However, a single imaging procedure alone cannot provide all the necessary information for medical diagnosis (Goyal and Wahla 2015). For example, in magnetic resonance imaging (MRI), T1 weighted scans, MRT1 imaging technique produces the detailed anatomical structure while, T2-weighted scans, MRT2 prominently highlights the differences between the normal and pathological structure of tissues. Hence, the anatomical features like shrinking of gray matter, enlargement of ventricles, etc., are visualized from MRI (Bhattacharya et al. 2012; Chang et al. 2002). On the other hand, positron emitted computed tomography (PET) and single photon emission computed tomography (SPECT) provide functional information like blood flow, food activity, and metabolism of affected organs. The goal of image fusion is to integrate complementary information from each images merged together to form a superior quality resultant image than any of the input images (Bhattacharya et al. 2012; Mukherjee and Das 2020; Horn et al. 2009). Hence, we have designed a simple CNN model which is trainable in general computer using the fused MRI and SPECT images. Proposed CNN model can extract the significant features of MRI and SPECT and classify the tumors more accurately than the single modality images. The schematic of the proposed work is described in Fig. 1.

The rest of the article is arranged as follows. Proposed methodology is described in Sect. 2. Section 3 gives the experimental results, and a comparative study between the proposed architecture and the standard VGG-16 architecture. Finally, a conclusion is drawn in Sect. 4.



**Fig. 1** Entire algorithmic overview

## 2 Methodology

### 2.1 Image Collection

The images were collected from the given website (<https://www.med.harvard.edu/aanlib/home>). Two types of images were mainly collected—**magnetic resonance imaging (MRI)** and **single photon emission computed tomography (SPECT) images**. The images were collected by changing the time axis and taking different slices along the axial plane and time axis.

### 2.2 Data Preparation

The final dataset was prepared using the following two steps:

1. **Image fusion using Shift Invariant Wavelet Transform (SWT):** The images were fused using SWT in order to extract multimodal features. It is a type of **Discrete Wavelet Transform** which omits both down sampling in the forward and up sampling in the inverse transforms (Sari-Sarraf and Brzakovic 1997). Primary advantages of SWT are it (a) produces less artifacts, (b) can better preserve the information of source images.

*Process:* Each channel of the two RGB images (MR-T2 and SPECT-TC) to be fused were first decomposed into *approximate matrix* and *details matrix* using SWT-based decomposition as shown in Fig. 2. The *approximation matrices* of both the images (MR-T2 and SPECT-TC) were *linearly blended* for each channel (RGB). The *details matrices* of each channel (RGB) were combined using *principal component analysis (PCA)* approach (Mishra et al. 2017) to integrate the information of three channels (RGB). Finally, inverse SWT was performed to produce the fused image containing information of all modality source images (MR-T2 and SPECT-TC). Figure 3 describes that PCA-based blending contains better clarity than the simple average blending.

2. **Image Augmentation:** Then, the number of collected images was not enough for efficient training of the CNN architecture. So, the number of images was increased with the help of image augmentation. Image data augmentation is a technique that can be used to artificially expand the size of a training dataset by creating modified versions of images in the dataset. Apart from creating more number of samples, it also helps in preventing over fitting. Some of the image augmentation techniques used for enhancing our dataset are mentioned as:
  - (a) **Flipping:** An image flip means reversing the rows or columns of pixels in the case of a vertical or horizontal flip, respectively.
  - (b) **Cropping:** Cropping can be used as a processing step for image data with mixed height and width dimensions of each image.

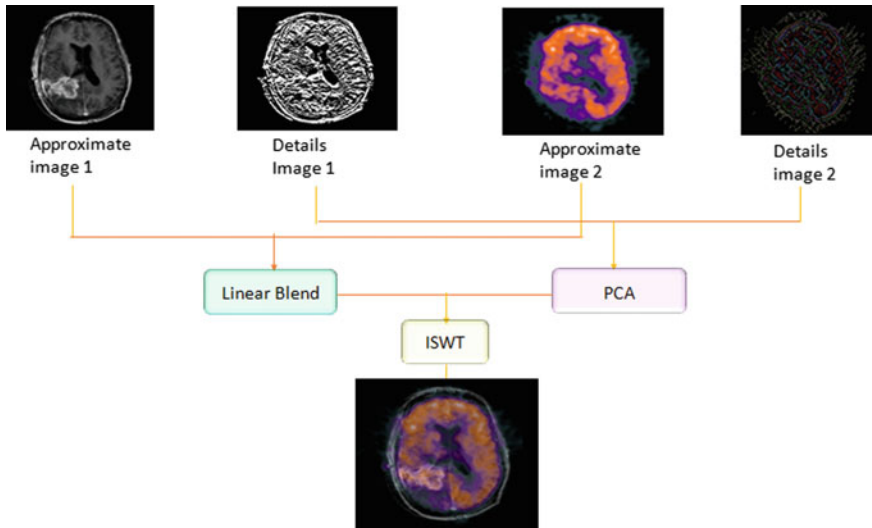


Fig. 2 Process flow of input images

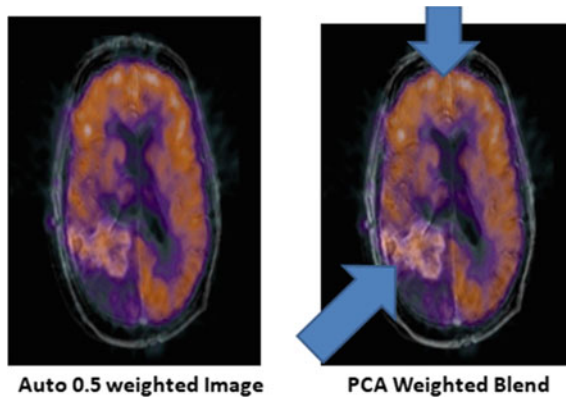


Fig. 3 (left) Auto-generated same weighted (0.5) blend; (right) PCA weighted blend using proposed algorithm

- (c) **Rotation:** Rotation augmentations are done by rotating the image right or left on an axis between  $1^\circ$  and  $359^\circ$ . After the final dataset was prepared, it was divided into **training, testing, and validation dataset—(80%–15%–5%, respectively)**.

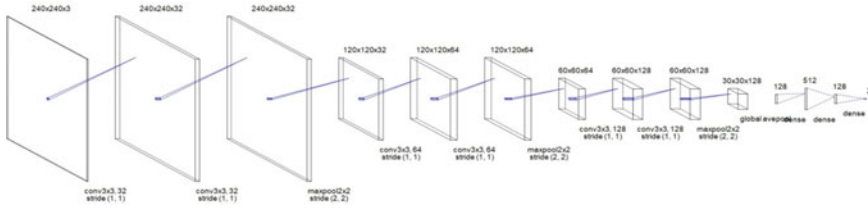


Fig. 4 Proposed architecture

### 2.3 Classifier Building

We build our CNN model which is trainable on a local PC having negligible GPU-CUDA support. We have built the proposed model as shown in Fig. 4 by keeping the reference VGG16 architecture as baseline model. The architecture of the proposed model is as follows:

- We use three CNN blocks of two CNN layers each.
- Each CNN block is followed by a MaxPool layer.
- Each MaxPool layer has a **Dropout of 0.2**.
- The CNN is translated into linear features using a global average pooling 2D layer to average out the intensities along channels.
- The CNN features are mapped into a dense network of two layers.
- We use a softmax to output the probabilities of two classes as follows:

- 1: Presence of Tumor
- 0: Absence of Tumor

### 2.4 Training and Testing of the Proposed Model

#### 2.4.1 Model Summarization

The model is written and compiled entirely using Tensor flow 2.2.0 and is compatible with versions >2.0+. Epochs trained over: **500**.

**Optimizer:** The model uses **Adam** optimizer (Kingma and Ba 2014). It uses a decay hyperparameter to optimize the learning rate,  $\beta_1 = 0.9$  and  $\beta_2 = 0.99$ . It computes the first and second order moments in order to estimate the decay rate of the steps. We use the Adam to back-propagate the gradients as well and optimize our loss. The main advantages of using **Adam** optimizer over other stochastic optimizers are listed as:

- (a) **Adaptive Gradient Algorithm** (AdaGrad) maintains per-parameter learning rate that improves performance on problems with sparse gradients (e.g., natural language and computer vision problems).

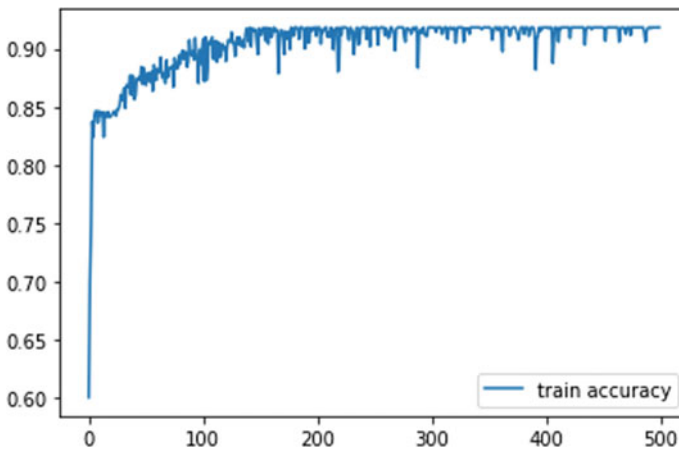
- (b) **Root Mean Square Propagation (RMSProp)** also maintains per-parameter learning rates that are adapted based on the average of recent magnitudes of the gradients for the weight (e.g., how quickly it is changing). This means the algorithm does well on online and non-stationary problems (e.g., noisy). It basically computes the learning rate not only based on the first moment—mean, but also based on the second moment—gradient. It uses an exponential moving average of gradients and also squared gradients over the loss plane to reach a much more global minima.
- (c) **Loss Function:** Categorical cross entropy (Ho and Wookey 2020) is used to compute the log loss.

$$\text{CE} = - \sum_{i=1}^n Y_i \log(\hat{Y}_i)$$

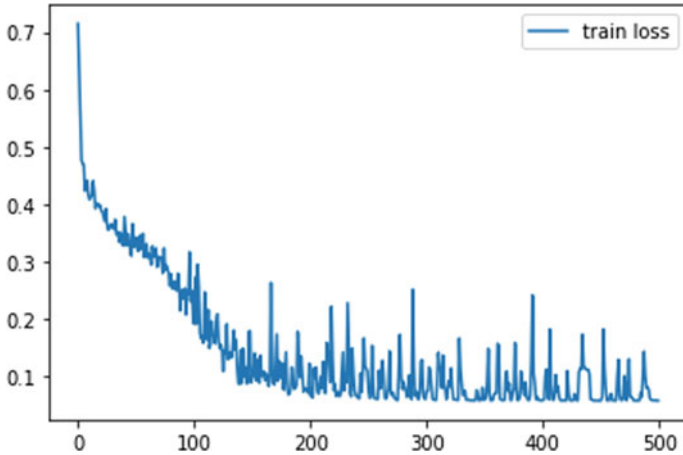
where  $n = 2$ ,  $\hat{Y}$  is the predicted label, and  $Y$  is the actual ground truth label.

### 2.4.2 Training Reports

The accuracy and loss of training process are shown in Figs. 5 and 6.



**Fig. 5** Accuracy plot of the training process



**Fig. 6** Loss plot of the training process

**Table 1** Comparison between the proposed architecture and the standard VGG-16

S. No.	Proposed architecture	VGG-16 (baseline architecture)
1	No. of trainable parameters: 418,978	No. of trainable parameters:138,000,000
2	No of layers: 9	No. of layers: 16
3	Training time: 5–9 h on CPU <sup>a</sup>	Training time: 1 day on CPU <sup>a</sup>
4	Requirement fit: Fits to our problem	Requirement fit: Over fits to our problem

<sup>a</sup>CPU in consideration is an Intel Core i3-6th Generation processor

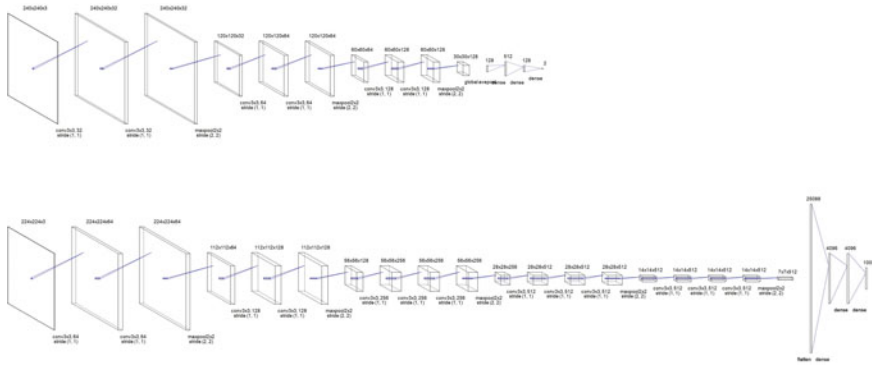
### 2.5 Comparison of the Proposed Architecture

A comparison has been described in Table 1 between the proposed architecture and the standard VGG-16 architecture. Figure 7 also shows that the proposed model consists of less number of parameters compared to VGG-16 architecture.

## 3 Results

### 3.1 Firing Patterns at the Different Layers of the Proposed Architecture

The main idea of using our CNN model is that we try to increase the number of channels in the later layers and to reduce the individual image dimensions as it progress through the network with less computation burden. As we do not intend to reduce the image dimensions in the progressing convolutional layers, we use padded

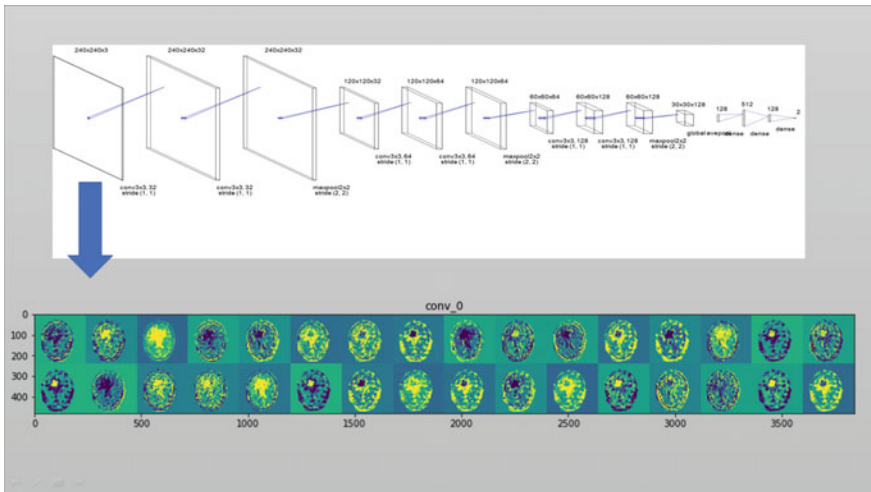


**Fig. 7** (top) Proposed architecture (with much lesser parameters than VGG-16 architecture); (bottom) Standard VGG-16 architecture

CNN blocks followed MaxPool block where the entire image dimension is reduced but expanded on the channels. We use *Relu Function* as shown in Eq. (1) which an activation function is used to map continuous values in positive range (Asadi and Jiang 2020).

$$\text{Relu Function: } f(x) = \max(0, x) \tag{1}$$

The firing pattern of each layer of the proposed architecture has been shown in detail in Fig. 8. The heat maps of the firings are clearly represented with alternating blue and yellow indicators.



**Fig. 8** Firing pattern of the first convolution layer of the first CNN block



The first layer of the CNN learns the basic visual level details. The overall structure of the image remains the same. The later layers use a fully connected dense network which is used to translate the 3D image channel structure into a linear structure. The dense network is used to feed into a *Softmax Layer* (Asadi and Hui 2020) as shown in Eq. (2) for probabilistic output of the classes: (Goyal and Wahla 2015).

$$\text{Softmax Layer: } f(x)_i = \frac{e^{x_i}}{\sum_{j=1}^K e^{x_j}} \tag{2}$$

The layers learn more and more complex features as we move deeper into the layers but as the layer increases, problem of vanishing gradients start to set in. Here, the initial layers may learn the various edge features and recognize those edges. All the convolutional layers use a  $3 \times 3$  filter with  $\text{stride} = 1$  and  $\text{padding} = \text{same}$ .

The following layers preserve/detect more sophisticated features and edges. These layers can understand features with more ‘inner’ meaning. Figure 9 describes the heat map firings of the second convolution layer of the second CNN block which learns more sophisticated features than the absolute initial layers but less sophisticated features than the third CNN Block (Hochreiter 1998). The dimensions of image are reduced, but the number of images is increased significantly.

As shown in Fig. 10, we use dropout (Srivastava et al. 2014) layer of 20% dropout after each CNN block to prevent over fitting of the images. These dropout layers are only used in training and do not contribute to model inference. The purpose of the dropout layer is to randomly drop 20% of the connections defined during a forward/backward pass through the network. Not only the nodes, but also the edges are dropped during the pass. Dropout is essential because of the inter-neuron co-dependency which exists during the training. It curbs the importance of the individual

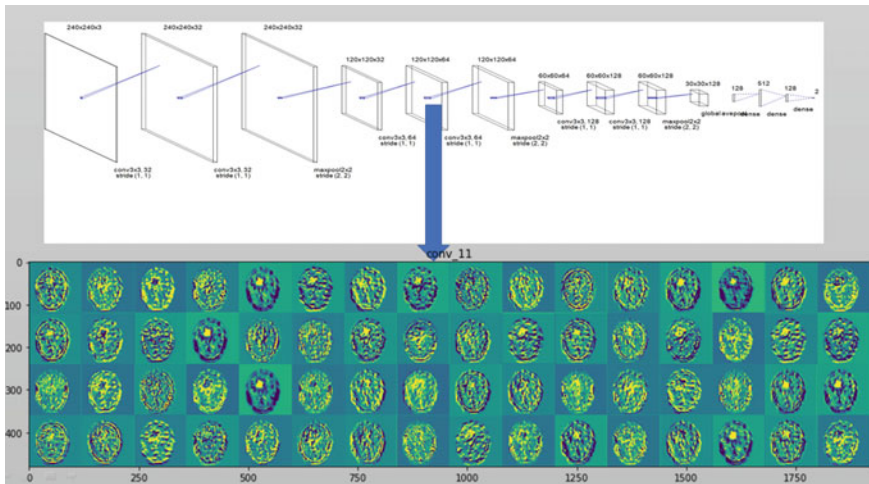


Fig. 9 Firing pattern of the second convolution layer of the second CNN block

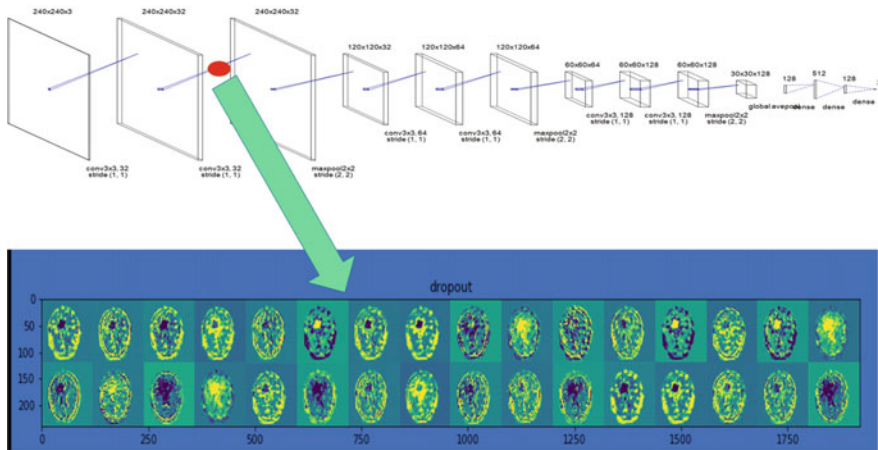


Fig. 10 Firing pattern of the dropout layer (0.2 dropout)

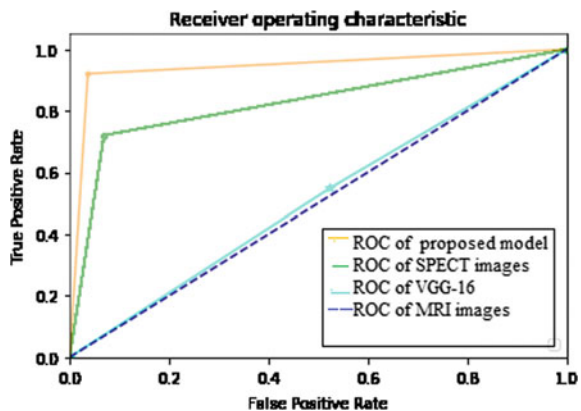
neurons and prevents over fitting. Dropout also forces the neurons to learn more robust features. It has been observed that dropout requires almost the double number of epochs to converge.

### 3.2 ROC Comparison

A comparative study of ROC of the proposed architecture using fused images, single modality MRI, SPECT images, and ROC of VGG-16 architecture is shown in Fig. 11. Table 2 also describes the values of the outcomes.

In the proposed architecture, as the fused images contain information of both modalities, number of significant features extracted from them is of better quality

Fig. 11 ROC analysis of the proposed architecture



**Table 2** Values of the outcomes

ROC analysis	TPR	FPR
Proposed architecture with fused image	0.980	0.040
Proposed architecture with MRI	0.500	0.502
Proposed architecture with SPECT	0.720	0.096
VGG-16 architecture with fused image	0.570	0.580

compared to the single modality. Hence, the classification accuracy of the fused images is superior to individual MR-T2 and SPECT-TC images.

## 4 Conclusion

As the proposed architecture of CNN has much lesser parameters (418,978) than the VGG-16 architecture (16,946,242), this model performs better as compared to the standard VGG-16. The main advantage of this model is that it is trainable on a local PC with standard RAM (about 8 GB) without any supporting GPUs (such as Google Colab) which are required in the VGG-16 architecture.

The areas of future work are:

- (1) Improving the existing dataset of images structure into 3D model view of the brain without sampling through the layers.
- (2) Developing U-Net like structures that can help build segmentation network to segment out critical locations.
- (3) Fine tune the parameters and hyperparameters even further to reduce training and inference time.
- (4) Extend the network as a generic network for various bio medical applications—liver, lung, prostate, etc.

## References

- Asadi B, Jiang H (2020) On approximation capabilities of ReLU activation and Softmax output layer in neural networks. [arXiv:2002.04060](https://arxiv.org/abs/2002.04060)
- Bhattacharya M, Das A, Chandana M (2012) GA based multiresolution fusion of segmented brain images using PD, T1 and T2 weighted MR modalities. *Neural Comput Appl* 21(6):1433–1447
- Chang DJ, Zubal IG, Gottschalk C, Necochea A, Stokking R, Studholme C, Corsi M, Slawski J, Spencer SS, Blumenfeld H (2002) Comparison of statistical parametric mapping and SPECT difference imaging in patients with temporal lobe epilepsy. *Epilepsia* 43(1):68–74
- Goyal S, Wahla R (2015) A review on image fusion. *IJIRCCE* 3(8):7582–7588
- Ho Y, Wookey S (2020) The real-world-weight cross-entropy loss function: modeling the costs of mislabeling. *IEEE Access* 8:4806–4813

- Hochreiter S (1998) The vanishing gradient problem during learning recurrent neural nets and problem solutions. *Int J Unc Fuzz Knowl Based Syst* 6:107–116. <https://doi.org/10.1142/S0218488598000094>
- Horn JF, Habert MO, Kas A, Malek Z, Maksud P, Lacomblez L, Giron A, Fertel B (2009). Differential automatic diagnosis between Alzheimer's disease and frontotemporal dementiabased on perfusion SPECT images. *Artif Intell Med* 47(2):147–158
- Johnson KA, Becker JA, The Whole Brain Atlas. Available online at: <https://www.med.harvard.edu/aanlib/home>
- Kingma D, Ba J (2014) Adam: a method for stochastic optimization. In: International conference on learning representations
- Mishra S, Sarkar U, Taraphder S, Datta S, Swain D, Saikhom R, Panda S, Laishram M (2017) Principal component analysis. *Int J Livestock Res*. <https://doi.org/10.5455/ijlr.20170415115235>
- Mukherjee S, Das A (2020) Vague set theory based segmented image fusion technique for analysis of anatomical and functional images. *Expert Syst Appl* 159:113592
- Sari-Sarraf H, Brzakovic D (1997) A shift-invariant discrete wavelet transform. *IEEE Trans Signal Process* 45(10):2621–2630
- Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 15:1929–1958