Deep Transfer Learning for Brain Magnetic Resonance Image Multi-class Classification

Yusuf Brima, *Member, IEEE,* Mossadek Hossain Kamal Tushar, *Member, IEEE,*
Upama Kabir, *Member, IEEE* and Tariqul Islam

Abstract—Magnetic Resonance Imaging (MRI) is a principal diagnostic approach used in the field of radiology to create images of the anatomical and physiological structure of patients. MRI is the prevalent medical imaging practice to find abnormalities in soft tissues. Traditionally they are analyzed by a radiologist to detect abnormalities in soft tissues, especially the brain. The process of interpreting a massive volume of patient's MRI is laborious. Hence, the use of Machine Learning methodologies can aid in detecting abnormalities in soft tissues with considerable accuracy. In this research, we have curated a novel dataset and developed a framework that uses Deep Transfer Learning to perform a multi-classification of tumors in the brain MRI images. In this paper, we adopted the Deep Residual Convolutional Neural Network (ResNet50) architecture for the experiments along with discriminative learning techniques to train the model. Using the novel dataset and two publicly available MRI brain datasets, this proposed approach attained a classification accuracy of 86.40% on the curated dataset, 93.80% on the Harvard Whole Brain Atlas dataset, and 97.05% accuracy on the School of Biomedical Engineering dataset. Results of our experiments significantly demonstrate our proposed framework for transfer learning is a potential and effective method for brain tumor multi-classification task.

Index Terms—Convolutional Neural Network, Magnetic Resonance Imaging (MRI), Transfer Learning, K-Nearest Neighbour, Brain Tumor,

1 Introduction

Recent advances in Artificial Intelligence (AI) and Machine Learning (ML) techniques are massively revolutionising healthcare as new capabilities of automation are being applied in electronic patient record analysis, radical personalization, medical image analysis, drug discovery, etc. [1]. The application AI and ML in different sectors of Healthcare is impacting its outcomes in new and profound ways. One of these outcomes is observed in Magnetic Resonance Imaging analysis [2].

Magnetic Resonance Imaging is a leading modality used in radiology to study anatomical and the physiological processes of the patients. It is the prevalent medical imaging method to identify tomors in brain scans of patients. MRI is frequently used to provide soft-tissue contrast because of its non-invasive approach towards medical imaging [3]. MRI images are traditionally analyzed by a radiologist to detect abnormalities of the brain. This process of interpreting huge volumes of patient MRI scans is painstakingly difficult and time-consuming [4]. In applications where distinguishing between abnormal and healthy tissue is delicate, precise interpretations become imperative [5]. Machine learning has shown considerable ability to classify, detect correctly, and segment images with precise accuracy and processing speed [2]. In this research, we propose a framework that uses Deep Transfer Learning to perform fine-grain classification of brain MRI images. This process entails specific categorization of brain tumor from MRI scans of patients. Our proposed framework uses a Convolutional Neural Network (CNN) based on ResNet50 architecture [6].

Brain MRI image classification is an area of extensive research at the intersection of Computer Vision (CV), Machine Learning, and Biomedical Imaging. Researchers have proposed novel methods that are geared toward tackling this research problem [7], [5], [8]. These methods range from learning the statistical characteristics of images to dimension reduction techniques to present-day Machine Learning approaches. However, these approaches thus far present inherent limitations of dataset size requirements and computational cost.Current Machine Learning approaches require massive training datasets and extended training time to generalize unseen samples better. Deep learning methodologies have been proposed in [9], [10], [11], [12]; in their studies, they found that a better performance is achieved, though the method required substantial data sets that require massive computational requirements and training procedures. We propose a Deep Transfer Learning framework in this paper to improve further the classification performance of brain tumor using MRI images.

Our proposed approach attained a classification accuracy of 86.40% on the National Institute of Neuroscience & Hospitals, Bangladesh dataset, 93.80% on the Harvard Whole Brain Atlas dataset and 97.05% accuracy on the Southern Medical University School of Biomedical Engineering dataset. Our experimental results indicate the pro-

- *Yusuf Brima is a research student at the University of Dhaka.*
- *Dr. Mossadek Hossain Kamal Tushar is a professor at the University of Dhaka.*
- *Dr. Upama Kabir is a professor at the University of Dhaka.*
- *Dr. Tariqul Islam is a research scientist at the National Institute of Neuroscience and Hospital, Dhaka, Bangladesh.*

posed framework for transfer learning is robust, feasible and efficient for brain tumor classification using MRI images.

The remaining portions of the paper are organized as follows: Section 2 presents the problem statement of MRI brain tumor classification and contribution of this research. Section 3 explains the preliminaries of Deep Transfer Learning. Section 4 the proposed methods are discussed while section 5 experimental results, and section 6 discussion and conclusion.

1.1 Related Work

The task of classifying brain MRI is an area that is actively undergoing research. There has been an extensive corpus of the literature of proposed approaches amongst researchers studying brain MRI classification. These approaches are at the intersection of Computer Vision, Image Processing and Artificial Intelligence, and Machine Learning, which have been evolving over the years. The maturation of big data, compute power, and robust Machine Learning approaches have led to predominant use cases. Supervised learning algorithms [13] like Support Vector Machine, Artificial Neural Network, Decision Tree and Random Forest, and unsupervised learning approaches like Fuzzy C-Means, Self-Organising Map, K-Mean Clustering etc., have been pivotal to these range of use cases [5].

In their paper, Zacharaki et al. [7] reported a multi-stage framework with a Support Vector Machine and K-Nearest Neighbour classification algorithms to first predict three glioma grades with an accuracy of 85% and to further the predict the glioma case into either high or low grade with an accuracy of 88%. Many pieces of literature exist on frameworks for solving brain tumor classification using MRI images [9], [10], [11], [12]. Paul et al., in their 2017 work, presented two deep learning methods to predict brain tumor using axial MRI data. They proposed a fully connected neural network as their first method and a deep convolutional neural network as their second approach with a 91.43% overall accuracy [14].

Chaplot et al., in their 2006 paper[11], proposed methods using Self-Organizing Map, Artificial Neural Network, and a Support Vector Machine for brain MRI classification. El-Dahshan et al.[12], in their work, presented a novel technique consisting of Discrete Wavelet Transform, Principal Component Analysis to reduce feature space of the dataset, and a Feed-Forward Neural Network for a binary classification task of 101 normal and abnormal brain MRI images. Prior research in this direction strongly underscores the relevance of automating brain MRI classification to aid neuroscientists improve healthcare outcomes. Contrary to conventional classification methods that extensively depend on feature engineering as a precursor to the classification task, a Convolutional Neural Networks learns feature representations hierarchically and directly using the data [5]. This approach is crucial to enabling the classifier to learn complex, rich representations of factors of variation from phenomena. Ertosun and Robin utilized a Convolutional Neural Network pipeline comprising two modules to detect Lower Grade Glioma (LGG) vs. Glioblastoma Multiforme (GBM) and determination of tumor grade using pathological data from The Cancer Genome Atlas [15]. Their experimental results achieved 96% accuracy of GBM v LGG classification, and 71% for the identification of the tumor grades. Paul et al., proposed two approaches using Fully Connected Neural Networks and Convolutional Neural Network for the classification of three tumor variants: meningioma, glioma, and pituitary [14]. Their results show that using axial brain images increases the classification accuracy of the neural network. Afshar et al., in [16] have proposed a Capsule Network that combines brain MRI images with coarse tumor boundaries to learn spatial relations to enable better classification performance. Their study obtained an accuracy of 90.89%. Anaraki et al., in their 2019 paper [17], presented an approach that combines a Convolutional Neural Network with a Genetic Algorithm for Optimization used to classify three Glioma sub-types achieving 90.0% accuracy and in another experiment Glioma, Meningioma, and Pituitary tumor variants obtaining a 94.2% accuracy. Zang et al. [18] have presented a hybrid approach that utilized Digital Wavelet Transform as feature extractor and Principal Component Analysis for dimensionality reduction, and a Kernel Support Vector Machine with Particle Swarm Optimization algorithm for the estimation of C and σ parameters. Saritha et al. [19] demonstrated Wavelet Entropy-based Spider Web Plots as features coupled with a Probabilistic Neural Network to classify brain tumor images achieving 100% accuracy. Wang et al., in [20] used a pipeline of Stationary Wavelet Transform (SWT) for feature extraction, Principal Component Analysis for dimensionality reduction combined with Particle Swarm Optimization, and Artificial Bee Colony for MRI brain images. Zhang et al. [21] in their 2015 paper, proposed a novel Computer-Aided Diagnostic system that comprises Discrete Wavelet Packet Transform (DWPT), which serves as a wavelet packet coefficient extractor, Shannon and Tsallis entropies which obtained entropy features from DWPT. Finally, they used a generalized Eigenvalue Proximate Support Vector Machine with a Radial Basis Function to classify MRI brain images into normal and abnormal classes.

Nayak et al., in their 2016 work, proposed a novel classification framework the utilized 2-D discrete Wavelet Transform to extract feature vectors, Probabilistic Principal Component Analysis for feature space reduction, and AdaBoost with Random Forests classifiers. Nayak et al. [22] further proposed a Computer-Aided Diagnosis framework the employed contrast limited adaptive histogram equalization as a mechanism to enhance tumor regions in brain MRI images, 2-D stationary wavelet transform for feature extraction, AdaBoost and Support Vector Machine algorithms for normal and abnormal brain MRI classification achieving 99.45% accuracy [23]. Gudigar et al., in a comparative study assessed three multi-resolution analysis methods - Discrete Wavelet Transform, Curvelet Transform and Shearlet Transform, and Particle Swarm Optimization for textual feature extraction from transformed images using Support Vector Machine for classification [24]. [25] have proposed a Deep Neural Network model using Convolutional Neural Network to classify three tumor variants (glioma, meningioma and pituitary tumor) and three glioma sub-types (Grade II, Grade III and Grade IV) achieving an overall performance accuracy of 96.13% and 98.7% respectively.

Deep learning (DL) is becoming a cornerstone to a great many applications across disciplines [26], [27], [28], [29]. Back-propagation algorithm has enabled DL models to construct hierarchies of concepts and representations through layered non-linear transforms with minimal feature engineering and domain knowledge [30]. Various DL architectures are applied increasingly to problems across disciplines, from the reconstruction of satellite imagery to drug discovery and protein structure/function prediction. The ability to learn complex hierarchies of factors of variability in signals is a focal reason for the surge in Deep Learning research and practice [31]. Deep Learning models are applied to a range of use cases across domains and disciplines [32], [33], [34], [35], [36], [37], [38].

2 Problem Statement and Contribution

2.1 Problem Statement

The field of neuroscience is traditionally tasked with analyzing MRI data for the detection of tumors. The procedure, however, grapples with a considerable challenge and requires a substantial level of domain expertise through extensive formal skills acquisition. Researchers have proposed various methods to address this classification task using feature selection and reduction techniques as a precursor to classification [10], [11], [12]. In our research, we propose a framework using Deep Transfer Learning as a mechanism to solve the multi-class classification task of brain MRI images. Our research uses a Deep Residual CNN based on the ResNet50 architecture.

2.2 Contributions

Our main contribution in research is a novel end-to-end Deep Transfer Learning framework using a ResNet50 CNN architecture that performs multi-classification of brain MRI images. We aim to investigate the transferability of source domain invariant knowledge to the target domain to reduce the algorithmic training time in conjunction with improving the generalization performance of the trained model using minimal target domain dataset. Our second contribution is a publication of a novel brain MRI dataset of 130 patients consisting of 5285 images belonging to 37 categories in collaboration with the National Insitute of Neuroscience & Hospitals (NINS) in Bangladesh. In our study, we conducted experiments with our proposed NINS dataset and two public benchmark datasets: the Biomedical School of Engineering brain tumor dataset [39] that consists of 3064 T1-weighted MRI images, and the Whole Brain Atlas that contains 1133 T2-weighted Harvard Medical School benchmark dataset [8].

3 Preliminaries

3.1 Transfer Learning

3.1.1 *Formal definition and notation*

The framework for transfer learning comprise **domain**, **task**, and **marginal probabilities** [40], [41] defined as follows: A domain, D, is a two-element tuple that consists of input space indicated by χ, and a marginal probability, $P(X)$, X is a sample data point. Thus, we can represent the domain formally as

$$D = \{\chi, P(X)\} \tag{1}$$

A Domain consist of two elements: $D = \{\chi, P(X)\}$ where:

- Input Space: χ
- Marginal distribution:

$$P(X), X = \{x_1, x_2, x_3, \cdots, x_n\}, \forall x_i \in \chi \tag{2}$$

Hence x_i represents a specific vector as represented in the above depiction. A task, T, on the other hand, is a two-element tuple of the label space, γ, and target function, $f(.)$. From a probabilistic standpoint, the target function can be stated thus: $P(\gamma|X)$ as a conditional probability. Given a domain D, the definition of a Task T is thus stated by two elements:

$$T = \{\gamma, P(Y/X)\} = \{\gamma, f(.)\}Y = \{y_1, y_2, y_3, ..., y_n\}, , \forall y_i \in \gamma \tag{3}$$

- A Label Space: γ
- A predictive function $f(.)$, learned from the feature vectors/label pairs $(x_i, y_i), x_i{\in}X, \forall y_i{\in}\gamma$
- For each feature vector in the domain, $f(.)$ predicts its corresponding label: $f(x_i) \equiv y_i$

The formulation of a Transfer Learning setting where D_S is the source domain, T_S is the source task, D_T is the target domain, and T_T is the target task is thus:

$$D_S = \{\chi_S, P(X_S)\}, X_S = \{x_{S_1}, x_{s_2}, x_{S_3}, ...x_{S_n}\}\forall x_{S_i} \in \chi_S \tag{4}$$

$$T_S = \{\gamma_S, P(Y_S|X_S)\}, Y_S = \{y_{S_1}, y_{s_2}, y_{S_3}, ...y_{S_n}\}\forall y_{S_i} \in \gamma_S \tag{5}$$

$$D_T = \{\chi_T, P(X_T)\}, X_T = \{x_{T_1}, x_{T_2}, x_{T_3}, ...x_{T_n}\}\forall x_{T_i} \in \chi_T \tag{6}$$

$$T_T = \{\gamma_T, P(Y_T|X_T)\}, Y_T = \{y_{T_1}, y_{T_2}, y_{T_3}, ...y_{T_n}\}\forall y_{T_i} \in \gamma_T \tag{7}$$

Considering a classification task T, with X as the input space and Y as the set of labels. Given two sets of samples drawn from both the source and target domains:

$$D_S = \{x_i, y_i\}_{i=1}^{n} \sim P(X_S) \tag{8}$$

$$D_T = \{x_i, y_i\}_{i=n+1}^{N} \sim P(X_T) \tag{9}$$

The goal of Transfer Learning is to build a functional mapping $f(.) : X \rightarrow Y$ with a low target classification error on the target domain assuming the D_S and D_T are **i.i.d.** $R_{D_T} = Pr(f(x) \neq y) \leftarrow$ probability of miss-classification $(x, y) \sim D_T$. Therefore, Transfer Learning as follows is defined as: given a source domain D_S with a source task T_S,and a target domain D_T with its target task T_T, the objective of Transfer Learning is to allow the transferability of target conditional probability distribution $P(T_T|D_T)$ in D_T using latent domain invariant knowledge from D_S and T_S given $D_S \neq D_T$ or $T_S \neq T_T$. Transfer Learning is well suited to cases where the number of target labels is disproportionately smaller than the number of source labels [42].

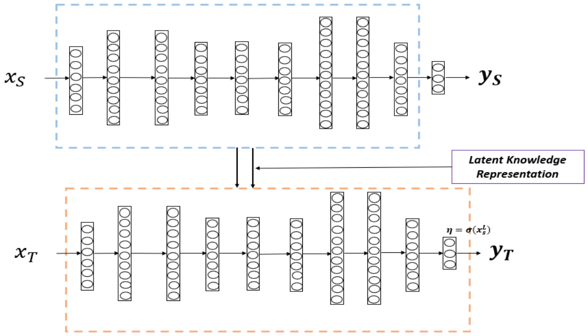

Fig. 1: High-level formal representation of Transfer Learning

4 Proposed Method

4.1 System Architecture

In this paper, we used ImageNet [43] as the source domain, a dataset with over one million images in one thousand categories of concepts from which the model learned and transferred invariant factors of variation to the target domain Brain MRI. Fig. 2 shows the architectural framework for this Representational Transfer Learning task. Transfer learning is a well-suited framework for healthcare computer vision tasks where target domain datasets for learning are significantly small, and model generalisation is a key consideration [40].

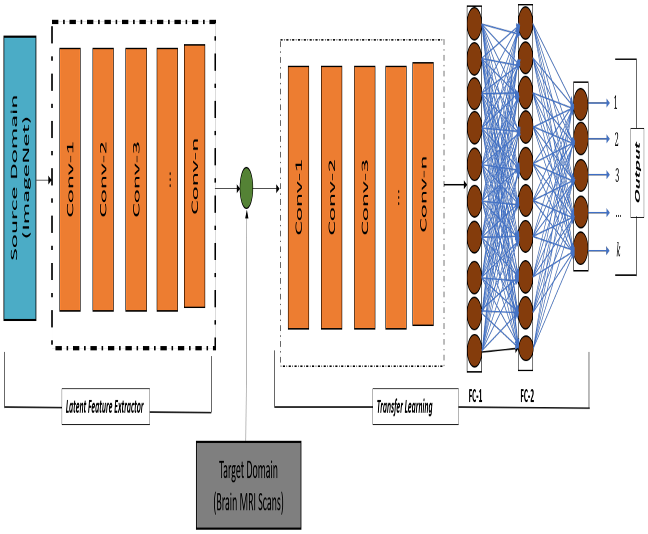

Fig. 2: Proposed System Architecture

4.2 System Model and Assumptions

We have shown a high-level overview of our proposed Transfer Learning framework in Fig. 3. The initial stage entails the setup of the dataset which includes loading both the brain MRI image data and associated class labels followed by batch normalization and cross-validation split into train and validation sets. We have used various data augmentation approaches to overcome overfitting by creating virtual copies of brain MRI images. This includes methods such as zooming, flipping, rotation, mirroring etc. to up-sample the datasets. Afterwards, we used ResNet50 CNN architecture, Optimal Learning Rate Finder, Gradient Descent with Restart (SGDR), and Adaptive Moment Estimation (ADAM). We trained the model through the three stages presented in Fig. 4 accordingly.

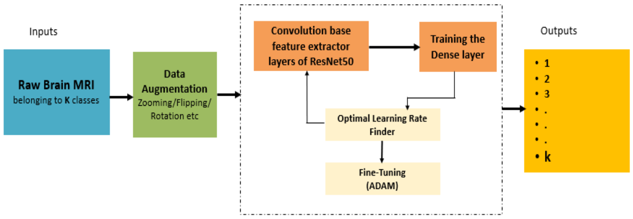

Fig. 3: A flow diagram of our proposed transfer learning framework

4.2.1 Mechanism of Deep Transfer Learning Model

This work proposes a Transfer Learning approach to the fine-grain brain MRI classification problem using state-of-the-art pre-trained ResNet50 as the CNN architecture due to its robustness to the exploding gradient problem [6]. We have used a pre-trained network on the ImagetNet, a dataset that comprises over a million images in a thousand classes of concepts. ResNet architecture has applications in a variety of Computer Vision tasks due to its faster convergence speed relative to Inception and Visual Geometric Group, AlexNet. The architecture is very robust, simple, and useful.

4.2.2 Proposed stages of Deep Transfer Learning

The framework of Deep Transfer Learning overcomes the problem of lack of labelled data, for example, in medical imaging applications by utilizing domain-invariant tacit knowledge to solve related problems. This approach towards learning can lead to better generalization of models with minimal target domain data. Fig.4 illustrates the three stages of our proposed transfer learning framework.

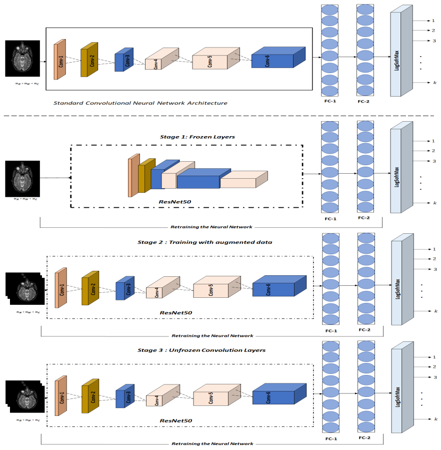
Conv-1
Conv-2
Conv-3
Conv-4
Conv-5
Conv-6
Standard Convolutional Neural Network Architecture
FC-1
FC-2
LogSofrMax
1
2
3
k
Stage 1: Frozen Layers
ResNet50
Retraining the Neural Network
Stage 2 : Training with augmented data
Stage 3 : Unfrozen Convolution Layers

Fig. 4: Deep Transfer Learning Stages 1: Retraining ResNet50; 2: Retraining the frozen layers of the network with augmented data; 3:Unfreezing and retraining the network with augmented data

4.3 Dataset

In this paper, we have published a dataset consisting of 5285 T1-weighted MRI images belonging to 37 categories in collaboration with the National Insitute of Neuroscience & Hospitals (NINS) of Bangladesh. Besides that, we have used two other public datasets, the School of Biomedical Engineering brain MRI dataset that comprises 3064 T1-weighted contrast-enhanced images of 233 study patients indicated in Table2 [39]. The second dataset is the Harvard Medical School Whole Brain Atlas brain MRI benchmark [8] that consists of 1133 T2-weighted brain MRI images of 38 patients shown in Table 3. The size of all images is 256x256 pixels in the axial plane. The School of Biomedical Engineering, Southern Medical University brain MRI dataset comprises of three classes of brain tumors, Meningioma, Glioma, and Pituitary tumor, as depicted in Table 2 [39]. The Harvard Medical School dataset entails five classes, Cerebrovascular class (stroke), Neoplastic, Degenerative, and Inflammatory disease, as shown in Table 3. We have used a 5-fold cross-validation technique to perform the network re-training, which ensures that we overcome the size limit of our dataset without over-fitting.

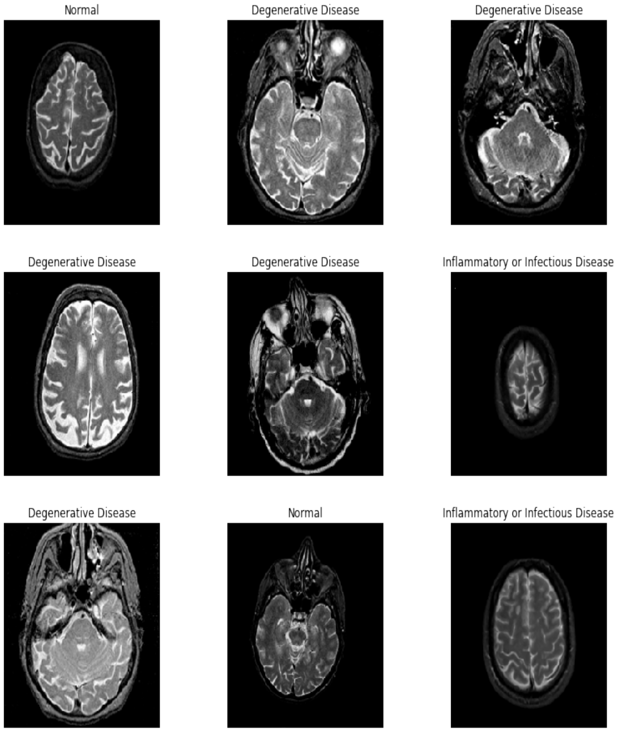

Fig. 5: Sample Harvard Whole Brain MRI Dataset

4.4 Essential deep learning techniques

In this research, we have applied state-of-the-art techniques in Deep Learning, like, data augmentation, Adam optimizer, Optimal Learning Rate Finder algorithm, and neural network hyper-parameter fine-tuning for the proposed framework.

4.4.1 Data augmentation:

To avoid overfitting the model, which happens because of small training set size, we have used data augmentation techniques such as flipping (vertical and horizontal), rotation, mirroring, and zooming to create virtual copies of the brain MRI images to improve model generalization performance.

Category	Total Patients	Total Slices
Brain Atrophy	8	264
Brain Infection	2	38
Brain Infection with abscess	2	76
Brain Tumor	3	76
Brain Tumor (Ependymoma)	1	36
Brain Tumor (Hemangioblastoma Pleomorphic xanthroastrocytoma metastasis)	2	74
Brain tumor (Astrocytoma Ganglioglioma)	1	38
Brain tumor (Dermoid cyst craniopharyngioma)	1	38
Brain tumor - Recurrenceremnant of previous lesion	3	114
Brain tumor operated with ventricular hemorrhage	1	76
Cerebral Hemorrhage	1	36
Cerebral venous sinus thrombosis	1	76
Cerebral abscess	1	36
demyelinating lesions	1	38
Encephalomalacia with gliotic change	2	76
focal pachymeningitis	1	36
Glioma	2	76
Hemorrhagic collection	1	38
Ischemic change demyelinating plaque	1	38
Left Retro-orbital Haemangioma	2	112
Leukoencephalopathy with subcortical cysts	1	38
Malformation (Chiari I)	1	38
Microvascular ischemic change	2	72
Mid triventricular hydrocephalus	1	38
NMOSD ADEM	1	36
Normal	44	1749
Obstructive Hydrocephalus	2	76
Post-operative Status with Small Hemorrhage	1	38
Postoperative encephalomalacia	1	38
Small Vessel Diease Demyelination	1	38
Stroke (Demyelination)	1	38
Stroke (Haemorrhage)	13	564
Stroke(infarct)	19	906
White Matter Disease	1	36
meningioma	2	76
pituitary tumor	1	76
small meningioma	1	36

TABLE 1: National Institute of Neurscience & Hospitals brain MRI dataset that contains 5285 T1-weighted contrast-enhanced images

Category	Total Patients	Total Slices
Meningioma	-	708
Glioma	-	1426
Pituitary Tumor	-	930

TABLE 2: Southern Medical University, School of Biomedical Engineering brain MRI dataset that contains 3064 T1-weighted contrast-enhanced images

Category	Total Patients	Total Slices
Normal	2	65
Degenerative Disease	8	223
Neoplastic Disease	8	277
Inflammatory Infectious Disease	5	189
Cerebrovascular Disease	15	376

TABLE 3: Harvard Whole Brain Atlas dataset contains 1133 T2-weighted contrast-enhanced images

4.4.2 Optimal Learning Rate Finder (OLRF

Hyper-parameter tuning is vital towards model performance improvement, but this process is often cumbersome [44]. Therefore, we used the Optimal Learning Rate Finder algorithm to identify a stable set of learning rates to boost model generalization performance. The learning rate indicates the step size used to update the model weights during training. This affects the rate of convergence; if it is too small, convergence to the optimum on the error surface takes much time, and tiny updates on the model weights are performed. Conversely, when the learning rate is too high, the optimization algorithm shoots over the minimum, leading to divergence, thereby affecting model performance. The strategy used to set the step size has a critical role in out of sample generalization performance.

4.4.3 Stochastic Gradient Descent with Restarts (SGDR)

SGDR is a form of learning rate annealing which uses cosine annealing to incrementally reduce the step size during the process of training a neural network. This approach tends to make minimal changes when the optimizer tends towards desired weight parameter updates of the model [45].

4.4.4 Classification Cost function: Softmax Classifier(Multinomial Logistic Cost)

The goal in a classification task is to maximize log-likelihood or associate higher probability mass for correct class and lower probability mass for incorrect classes. Therefore we define the loss and cost functions are defined as thus:

$S \ = f(x_i, W)$ scores is the unnormalized log probabilities of the classes.

$$P(Y = k|X = x_i) = \frac{e^{s_k}}{\sum_j e^{s_j}}$$

$$\mathcal{L}_i = -\log P(Y = y_i|X = x_i)$$

Loss Function $\mathcal{L}_i = -\log\left(\frac{e^{s_{y_i}}}{\sum_j e^{s_j}}\right)$

Cost Function $\mathcal{L}(W) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_i(f(x_i, W), y_i)\ \{(x_i, y_i)\}_{i=1}^{N}$

4.5 Performance Evaluation

This subsection presents the metrics we have applied to evaluate the performance of the proposed learning framework. We have used the standard classification performance metric to evaluate the model.

4.5.1 Performance metrics

$$Precision = \frac{A}{(A + \Theta)} \tag{10}$$

$$Recall = \frac{A}{(A + B)} \tag{11}$$

$$F1Score = 2 \cdot \frac{(Precision \cdot Recall)}{(Precision + Recall)} \tag{12}$$

$$Accuracy = \frac{A + \Psi}{(P + N)} \tag{13}$$

where,

False Negatives (B) are class labels that are predicted negative which in the ground truth are positive. The is also known as a type two error.

False Positives (Θ) are class examples that are predicted to be positive which in the ground truth are negative. The is also known as a type one error.

True Positive (A) are class labels that are predicted to be positive which in the ground truth are positive.

True Negative (Ψ) are class labels that are predicted to be negative which in the ground truth are negative.

4.6 Simulation Environment

In this research, we carried out all experiments in Python using fast.ai [46], a wrapper framework built on top of PyTorch library. PyTorch is used for Graphics Processing Unit-based accelerated computing [47]. We carried out all experiments on a Linux Server (Ubuntu 10.04.4 LTS) Tesla V100. To evaluate the learning performance, we applied five-fold cross-validation using 80% - 20% dataset split.

5 Experimental Results

Experimental results of our work throughout the three progressive stages of transfer learning are presented. We present the model performance for the three stages of learning using ResNet50, ResNet34, AlexNet, and VGG19 deep learning CNN architectures. Training the neural network comprises three consecutive steps:

1) Training the network with base layers frozen
2) Retraining the network with augmented data
3) Unfreezing the Network and fine-tuning with the augmented data

5.1 Experiment I: National Institute of Neuroscience and Hospitals dataset

5.1.1 Stage I: Retraining the Network

In the first stage, we re-trained the model as a benchmark by freezing the convolution base layers of the network and did not update its weights during this stage. The weight updates only occurred in the fully-connected layers of the network. We utilized the multinomial logistic cost function to measure the loss and a step size of 0.01096. We set the number of epochs to 8 to ensure the model does not overfit on the training set because small, and highly imbalanced dataset tends towards model overfitting, which affects out-of-sample performance. On completing the first stage of the training, the model showed an overall accuracy of 87.43% using a fivefold cross-validation strategy. Fig 6 shows the top miss-classified images during this phase of training.

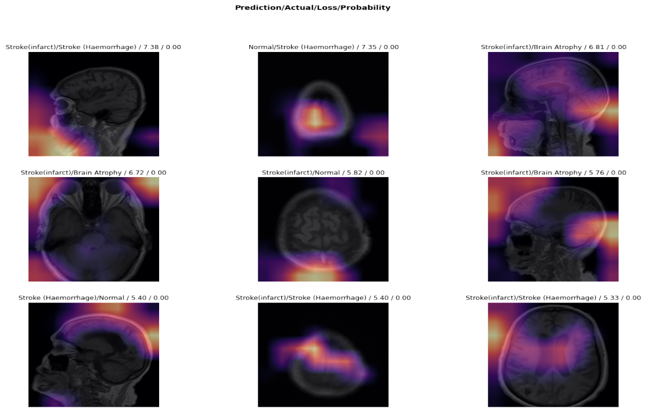
Prediction/Actual/Loss/Probability
Stroke(infarct)/Stroke (Haemorrhage) / 7.38 / 0.00
Normal/Stroke (Haemorrhage) / 7.35 / 0.00
Stroke(infarct)/Brain Atrophy / 6.81 / 0.00
Stroke(infarct)/Brain Atrophy / 6.72 / 0.00
Stroke(infarct)/Normal / 5.82 / 0.00
Stroke(infarct)/Brain Atrophy / 5.76 / 0.00
Stroke (Haemorrhage)/Normal / 5.40 / 0.00
Stroke(infarct)/Stroke (Haemorrhage) / 5.40 / 0.00
Stroke(infarct)/Stroke (Haemorrhage) / 5.33 / 0.00

Fig. 6: Illustration of top miss-classified images after stage I of training

We applied OLRF combined with SGDR algorithm to learn a step size in the weight space. This was achieved by boosting the learning rate with respect to the validation loss to discover an optimal learning rate to train the model. SDGR uses cosine annealing to reset the learning rate to traverse regions of the error surface to find the optimal minimum.

5.1.2 Stage II: Retraining the Network with Augmented Data

In this second stage, we retrained the model with the augmented data and set the learning rate hyper-parameter by utilizing OLRF and SGDR algorithms between $2e^{-4}$ to $2e^{-2}$. After this stage of retraining, we did not notice a decrease in the train and validation losses as well as the accuracy as shown in Fig 10.

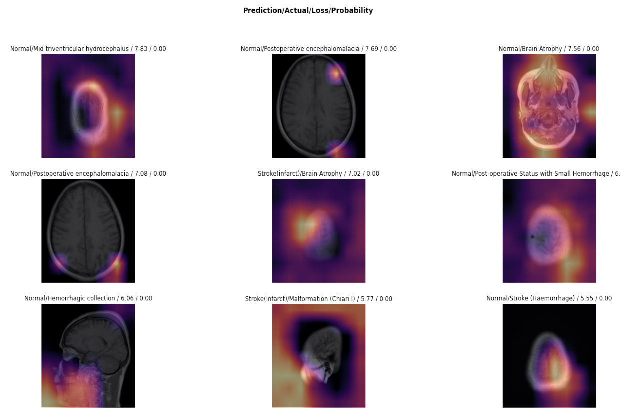
Prediction/Actual/Loss/Probability
Normal/Mid triventricular hydrocephalus / 7.83 / 0.00
Normal/Postoperative encephalomalacia / 7.69 / 0.00
Normal/Brain Atrophy / 7.56 / 0.00
Normal/Postoperative encephalomalacia / 7.08 / 0.00
Stroke(infarct)/Brain Atrophy / 7.02 / 0.00
Normal/Post-operative Status with Small Hemorrhage / 6.63 / 0.00
Normal/Hemorrhagic collection / 6.06 / 0.00
Stroke(infarct)/Malformation (Chiari I) / 5.77 / 0.00
Normal/Stroke (Haemorrhage) / 5.55 / 0.00

Fig. 7: Illustration of top miss-classified images after stage II of training with augmented training data

5.1.3 Stage III: Unfreezing the Network and Fine-tuning it with the Augmented Data

At this third and final stage, we unfroze all the base convolution layers, fine-tuned the network, and jointly trained them using the augmented data. This is Stage-3 in Fig 9 graph of the error rate across the model architectures. We trained the network for four epochs at this stage to adjust the trained weights to preserve the learned representation. To attain this goal, we set the step size of the last layers higher than the preceding ones during this fine-tuning process. We set the step size in the range of $3e^{-6}$ to $4e^{-3}$ across the network. After fine-tuning, we achieved a validation accuracy of 84.40% due to the imbalanced nature of the dataset and a degree of noise in it as well.

We have plotted the error rates of the models, as shown in Fig.9. ResNet50 converges faster during the last four epochs of the third training stage compared to VGG, Alexnet, and ResNet34. This convergence property of ResNet50 is demonstrated through the nearly smooth learning curve, which indicates the model was able to find a stable set of weights though the data is highly noisy, and classes are imbalanced. The VGG model is second to ResNet50 regarding convergence to the local minimum during the training phases consistently across the three stages of training, which indicates superior performance to Alexnet and ResNet34.

In Table 4, we have presented a comparison of metrics across the three stages of Transfer Learning in this experiment. The graph in Fig10 gives information on the training and validation accuracy for all training stages for the four CNN model architectures. We observed the network underfitted since the validation error was lower than the training error. This was solved by extending the number of epochs. At stage two, we used techniques such as vertical flipping, max zooming, and max lighting to augment the MRI images.

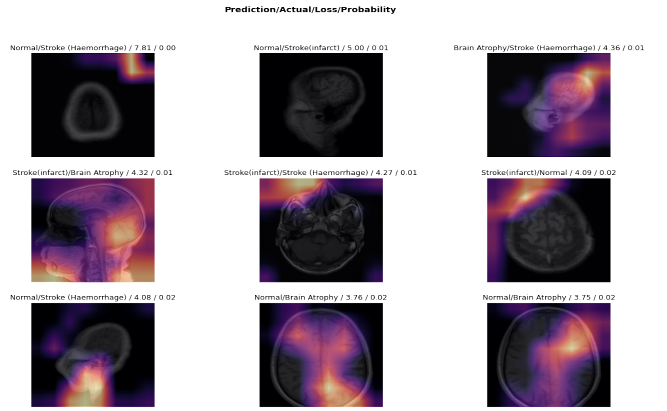
Prediction/Actual/Loss/Probability
Normal/Stroke (Haemorrhage) / 7.81 / 0.00
Normal/Stroke(infarct) / 5.00 / 0.01
Brain Atrophy/Stroke (Haemorrhage) / 4.36 / 0.01
Stroke(infarct)/Brain Atrophy / 4.32 / 0.01
Stroke(infarct)/Stroke (Haemorrhage) / 4.27 / 0.01
Stroke(infarct)/Normal / 4.09 / 0.02
Normal/Stroke (Haemorrhage) / 4.08 / 0.02
Normal/Brain Atrophy / 3.76 / 0.02
Normal/Brain Atrophy / 3.75 / 0.02

Fig. 8: Illustration of top miss-classified images after stage III of fine-tuning

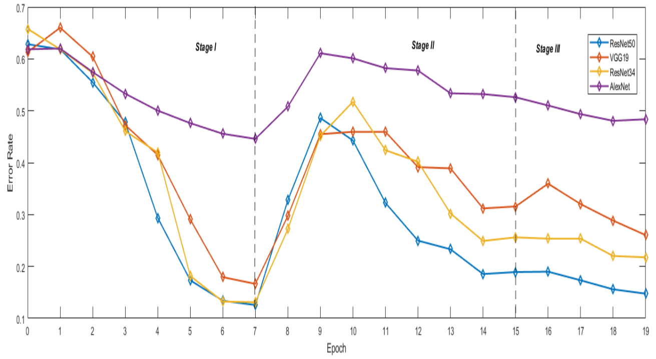

Fig. 9: Error Rates across the four model architectures

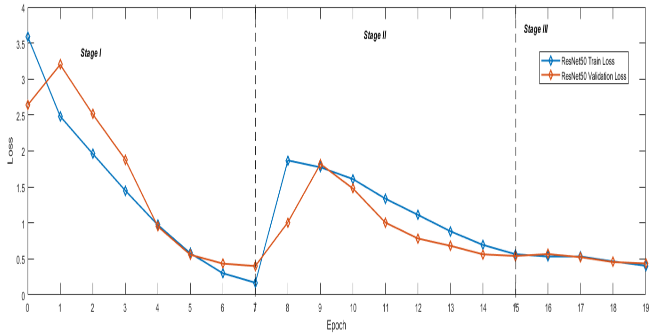

Fig. 10: Training vs. Validation Loss for all stages with ResNet50

We assess the classification performance on the validation data for the three stages. During the five-fold cross-validation training, we obtained the confusion matrices, which are shown in Fig.12 through Fig. 14. The matrices give a summary of the classification performance of the model, where the leading diagonals indicate correctly predicted

Stage	Train Loss	Validation Loss	Error Rate	Accuracy	Precision	Recall	F1-Score
I	0.169824	0.401473	0.125662	0.874338	0.874338	0.874338	0.874338
II	0.565008	0.541435	0.189251	0.810749	0.810749	0.810749	0.810749
III	0.434329	0.460459	0.155942	0.844058	0.844058	0.844058	0.844058

TABLE 4: Comparison of metrics across the 3 stages of training with NINS dataset

classes and while miss-classified samples are outside the diagonals. At the end of stage I, the model reached an accuracy of 87.43% as shown in Fig.11. However, the average five-fold cross-validation accuracy value did not improve at the end of stage 2, which is mostly due to class-imbalance problem with most classes in the dataset having a smaller number of examples.

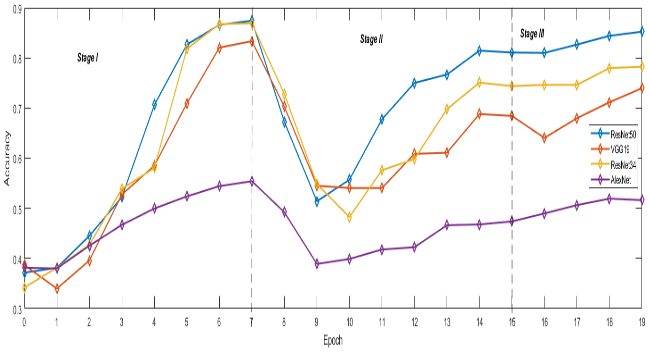

Fig. 11: Accuracy across four different model architectures

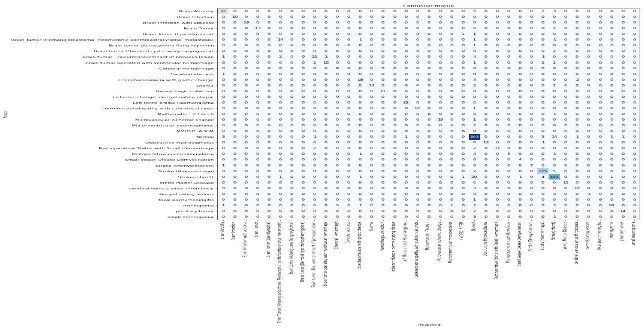

Fig. 12: Stage I confusion matrix after training the network with base layers frozen

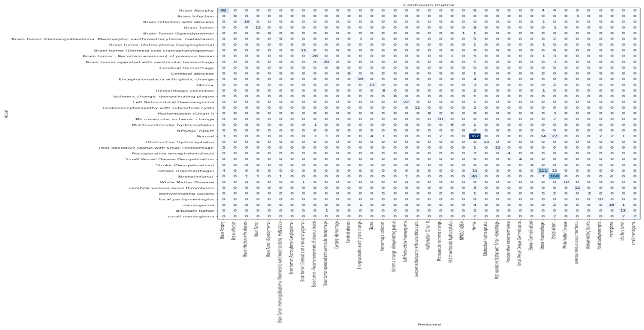

Fig. 13: Stage II confusion matrix after retraining the network with augmented data

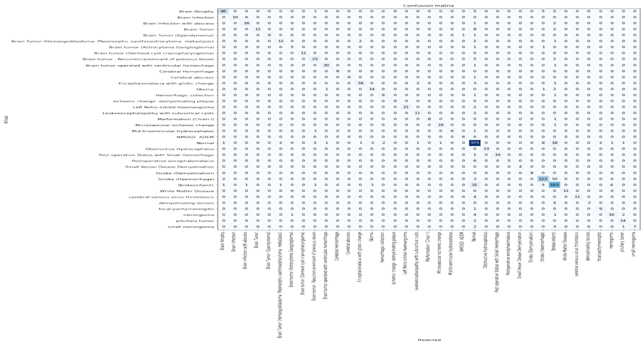

Fig. 14: Stage III confusion matrix after unfreezing the Network and fine-tuning with the augmented data

5.2 Experiment II: Harvard Whole Brain Atlas Dataset

5.2.1 Stage I: Retraining the Network

Following the same methodology used in experiment 5.1 stage 5.1.1, we froze the convolution layers of the network and did not update it during this stage. The learning rate hyperparameter was assigned $1e^{-3}$ and the number of epochs was set to 6. The epoch set to 6 to avoid overfitting the model on the training set. At this first stage, the training model achieved an overall accuracy of 92.47% in a fivefold cross-validation strategy. Fig 15 shows the top miss-classified images during this phase of training.

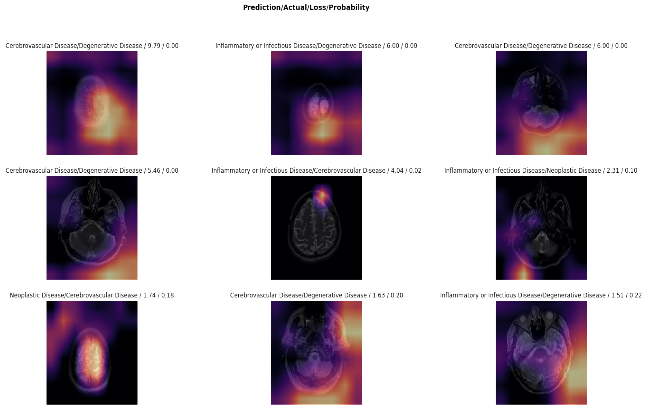
Prediction/Actual/Loss/Probability
Cerebrovascular Disease/Degenerative Disease / 9.79 / 0.00
Inflammatory or Infectious Disease/Degenerative Disease / 6.00 / 0.00
Cerebrovascular Disease/Degenerative Disease / 6.00 / 0.00
Cerebrovascular Disease/Degenerative Disease / 5.46 / 0.00
Inflammatory or Infectious Disease/Cerebrovascular Disease / 4.04 / 0.02
Inflammatory or Infectious Disease/Neoplastic Disease / 2.31 / 0.10
Neoplastic Disease/Cerebrovascular Disease / 1.74 / 0.18
Cerebrovascular Disease/Degenerative Disease / 1.63 / 0.20
Inflammatory or Infectious Disease/Degenerative Disease / 1.51 / 0.22

Fig. 15: Illustration of top miss-classified images after stage I of training

5.2.2 Stage II: Retraining the Network with Augmented Data

Having completed stage I, we used the OLRF algorithm combined with SDGR to find a stable step size in the weight space. SDGR uses cosine annealing to reset the learning rate to traverse regions of the error surface to find the minimum. We retrained the model with the augmented data and a learning rate set between the intervals $3e^{-3}$ to $2e^{-2}$. After this stage of retraining, we saw a decrease in the train and validation losses while the accuracy, however, did not increase.

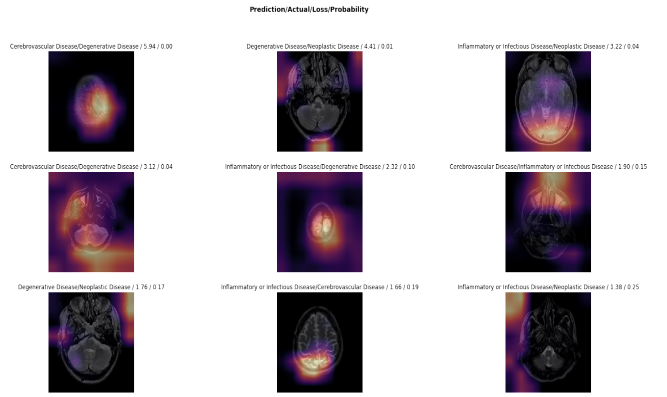
Prediction/Actual/Loss/Probability
Cerebrovascular Disease/Degenerative Disease / 5.94 / 0.00
Degenerative Disease/Neoplastic Disease / 4.41 / 0.01
Inflammatory or Infectious Disease/Neoplastic Disease / 3.22 / 0.04
Cerebrovascular Disease/Degenerative Disease / 3.12 / 0.04
Inflammatory or Infectious Disease/Degenerative Disease / 2.32 / 0.10
Cerebrovascular Disease/Inflammatory or Infectious Disease / 1.90 / 0.15
Degenerative Disease/Neoplastic Disease / 1.76 / 0.17
Inflammatory or Infectious Disease/Cerebrovascular Disease / 1.66 / 0.19
Inflammatory or Infectious Disease/Neoplastic Disease / 1.38 / 0.25

Fig. 16: Illustration of top miss-classified images after stage II of training with augmented training data

5.2.3 Stage III: Unfreezing the Network and Fine-tuning it with the Augmented Data

Similar to the experiment 5.1.3, we unfroze all the convolution base layers of the network, fine-tuned it, and trained them all using the augmented data. This is the Stage-3 in Fig.18 graph of the error rate across the four model architectures. We have set the step size in the range of $1e^{-6}$ to $4e^{-3}$ across the network. After fine-tuning, we re-trained the model and achieved a validation accuracy of 93.80%.

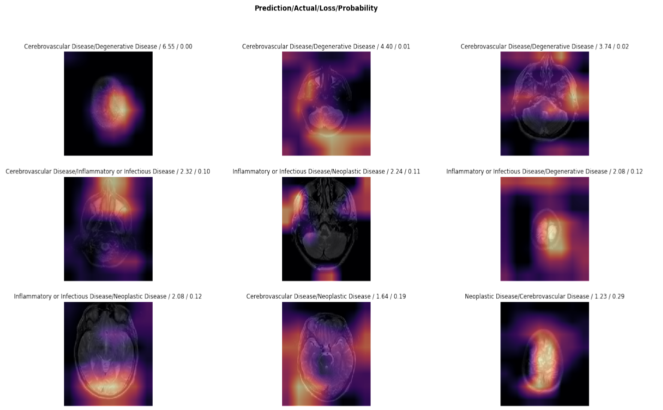
Prediction/Actual/Loss/Probability
Cerebrovascular Disease/Degenerative Disease / 6.55 / 0.00
Cerebrovascular Disease/Degenerative Disease / 4.40 / 0.01
Cerebrovascular Disease/Degenerative Disease / 3.74 / 0.02
Cerebrovascular Disease/Inflammatory or Infectious Disease / 2.32 / 0.10
Inflammatory or Infectious Disease/Neoplastic Disease / 2.24 / 0.11
Inflammatory or Infectious Disease/Degenerative Disease / 2.08 / 0.12
Inflammatory or Infectious Disease/Neoplastic Disease / 2.08 / 0.12
Cerebrovascular Disease/Neoplastic Disease / 1.64 / 0.19
Neoplastic Disease/Cerebrovascular Disease / 1.23 / 0.29

Fig. 17: Illustration of top miss-classified images after stage III of fine-tuning

We have plotted the error rates of the models as shown in Fig.18. ResNet50 converges faster during the last two epochs of the third training stage compared to VGG, Alexnet, and ResNet34. This convergence property of ResNet50 is demonstrated through the nearly smooth learning curve, which indicates the model was able to find a stable set of weights. The VGG model is second to ResNet50 regarding convergence to the local minimum during the training phases consistently across the three stages of training, which indicates superior performance to Alexnet and ResNet34.

In Table. 5, we have presented the train and validation accuracy graph for the experiment. The graph in Fig.19 gives information on the training and validation accuracy for all three stages of training. We observed the network under-fitted since the validation error was lower than the training error. This was solved by extending the number of epochs. At stage two, we used techniques such as vertical flipping, max zooming, and max lighting to augment the MRI images.

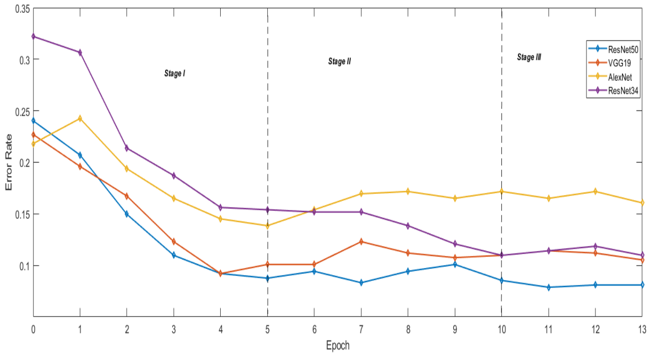

Fig. 18: Error Rates across the four model architectures

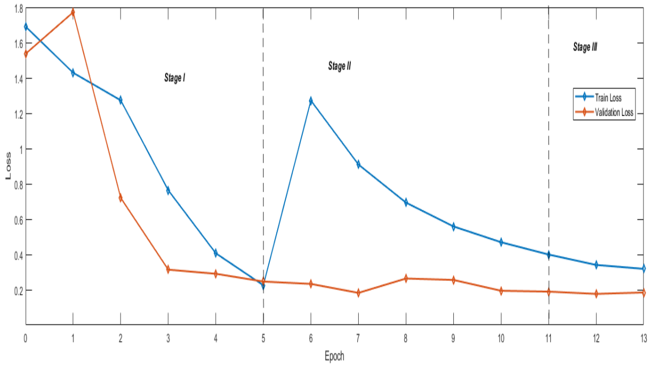

Fig. 19: Training vs. Validation Loss for all stages with ResNet50

We assess the classification performance on the validation data for the three stages. During the five-fold cross-validation training, we obtained the confusion matrices, which are presented in Fig.21 through Fig. 23. The matrices give a summary of the classification performance of the model were the leading diagonals indicate correctly predicted classes and while miss-classified samples are outside the diagonals. At the end of stage I, the model reached an accuracy of 92.47% with 17 incorrectly classified images, as shown in Fig.21. This model correctly classified all normal cases in stage II, as shown in Fig. 19.

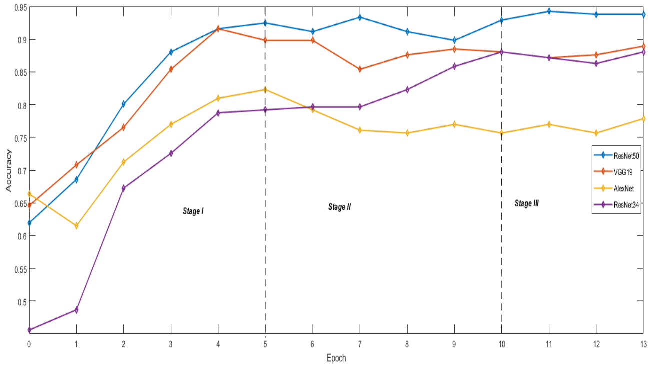

Fig. 20: Accuracy across four different model architectures

Stage	Train Loss	Validation Loss	Error Rate	Accuracy	Precision	Recall	F1-Score
I	0.225288	0.246468	0.075221	0.924779	0.924779	0.924779	0.924779
II	0.399370	0.189434	0.057522	0.942478	0.942478	0.942478	0.942478
III	0.318317	0.184569	0.061947	0.938053	0.938053	0.938053	0.938053

TABLE 5: Comparison of metrics across the 3 stages of training with Harvard dataset

However, the average five-fold cross-validation accuracy value improved from 92.24% to 94.42% at the end of stage 2. Finally, Fig.23 depicts stage III results. 14 images were incorrectly classified, and the model reached 93.80% average five-fold cross-validation accuracy. The whole training procedure takes 310 seconds.

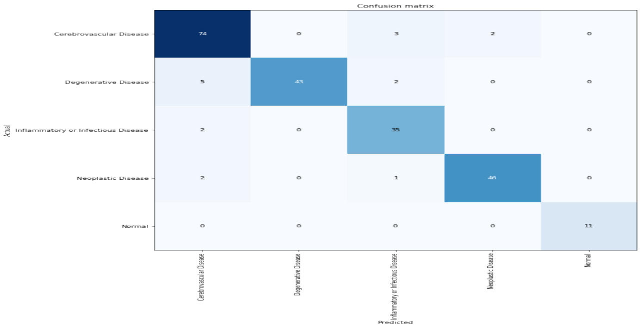

Fig. 21: Stage I confusion matrix after training the network with base layers frozen

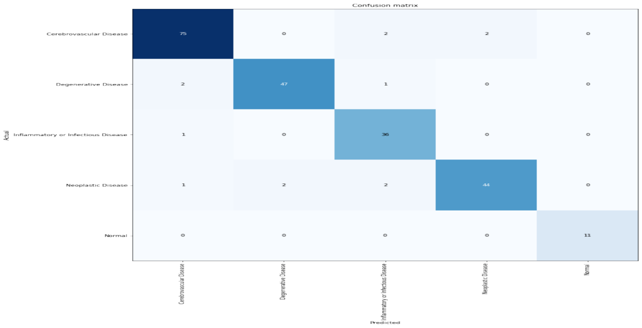

Fig. 22: Stage II confusion matrix after retraining the network with augmented data

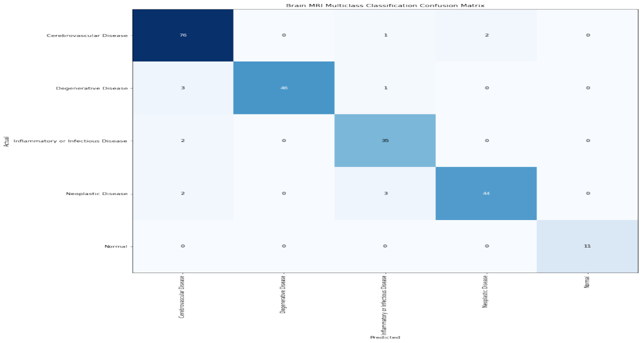

Fig. 23: Stage III confusion matrix after unfreezing the Network and fine-tuning with the augmented data

5.3 Experiment III: Biomedical School Brain MRI dataset

5.3.1 Stage I: Retraining the Network

In this third and final experiment, we used the brain MRI dataset from the School of Biomedical Engineering, which consists of T-1 weighted images. We re-trained the model as a benchmark. We froze the convolution layers of the network and did not update it during this stage. We only trained the weights of the fully-connected layers of the model. We used the multinomial logistic cost function to measure the loss and error rate with a step size of $1e^{-3}$ and trained the model for 4 epochs. We set the number of epochs to 4 to ensure the model does not overfit on the training set. The model memorises the given small dataset which affects out-of-sample performance. On completing the first stage of the training, the model showed an overall accuracy of 96.73% using a fivefold cross-validation strategy. Fig 24 shows the top miss-classified images during this phase of training.

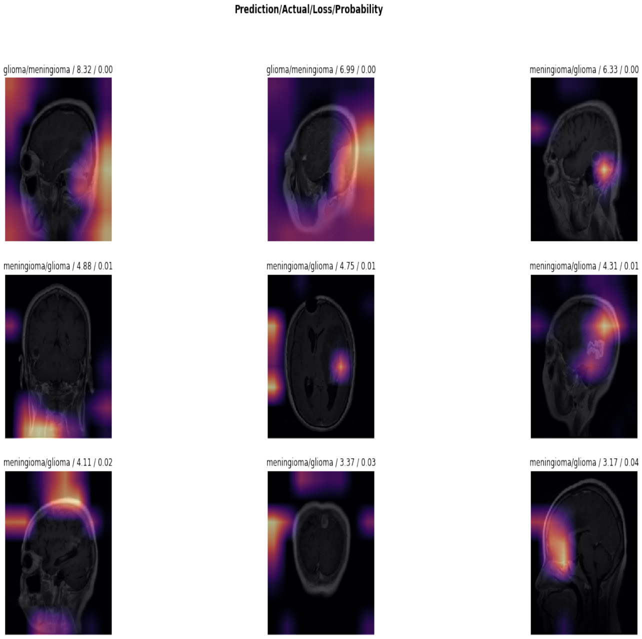
Prediction/Actual/Loss/Probability
glioma/meningioma / 8.32 / 0.00
glioma/meningioma / 6.99 / 0.00
meningioma/glioma / 6.33 / 0.00
meningioma/glioma / 4.88 / 0.01
meningioma/glioma / 4.75 / 0.01
meningioma/glioma / 4.31 / 0.01
meningioma/glioma / 4.11 / 0.02
meningioma/glioma / 3.37 / 0.03
meningioma/glioma / 3.17 / 0.04

Fig. 24: Illustration of top miss-classified images after stage I of training

5.3.2 Stage II: Retraining the Network with Augmented Data

Similar to the preceding experiments, we utilized the OLRF with SGDR algorithms to learn an optimal step size in the weight space. This is achieved by boosting the learning rate with respect to the validation loss to discover an optimal learning rate to train the model. We used an initial learning rate set between the intervals $3e^{-3}$ to $2e^{-2}$ in this stage of the experiment. SDGR uses cosine annealing to reset the learning rate to traverse regions of the error surface to find the minimum. After this stage of retraining, we saw a decrease in the train and validation losses, as shown in Fig.28. We used techniques such as vertical flipping, max zooming, and max lighting to augment the MRI images. The performance metrics are presented in Table. 6.

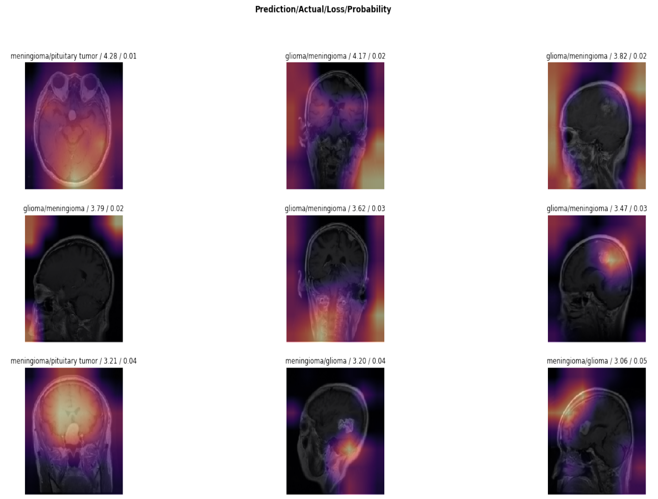
Prediction/Actual/Loss/Probability
meningioma/pituitary tumor / 4.28 / 0.01
glioma/meningioma / 4.17 / 0.02
glioma/meningioma / 3.82 / 0.02
glioma/meningioma / 3.79 / 0.02
glioma/meningioma / 3.62 / 0.03
glioma/meningioma / 3.47 / 0.03
meningioma/pituitary tumor / 3.21 / 0.04
meningioma/glioma / 3.20 / 0.04
meningioma/glioma / 3.06 / 0.05

Fig. 25: Illustration of top miss-classified images after stage II of training with augmented training data

5.3.3 Stage III: Unfreezing the Network a and fine-tuning it with the Augmented Data

Similar to 5.1.3 and 5.2.3, we unfroze all the convolution layers of the network, fine-tuned the it, and jointly trained them using the augmented data. This is the Stage-3 in Fig.27 graph of the error rate across the model architectures. At this stage, we trained the fully connected layers for 4 epochs, and we want the trained weights to adjust in line with training steps. To attain this goal, we set the step size of the last layers higher than the preceding layers during this process of fine-tuning. For this reason, we varied the learning rate across the layers of the network. At this stage, we set the learning rate to $1e^{-6}$ through $4e^{-3}$ across the network. After fine-tuning, we re-trained the model and achieved a validation accuracy of 97.05%.

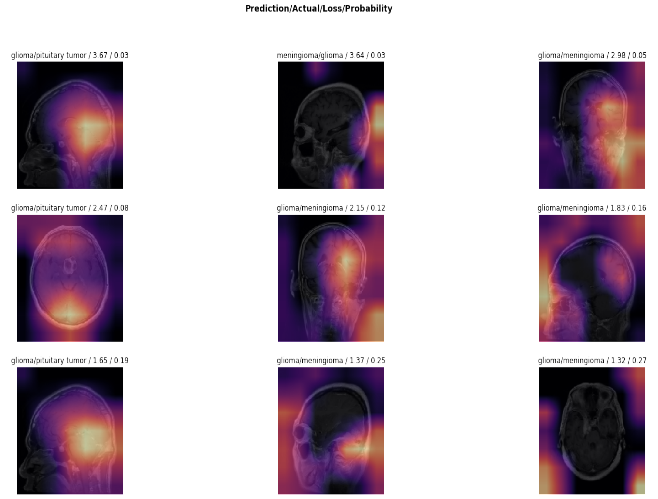
Prediction/Actual/Loss/Probability
glioma/pituitary tumor / 3.67 / 0.03
meningioma/glioma / 3.64 / 0.03
glioma/meningioma / 2.98 / 0.05
glioma/pituitary tumor / 2.47 / 0.08
glioma/meningioma / 2.15 / 0.12
glioma/meningioma / 1.83 / 0.16
glioma/pituitary tumor / 1.65 / 0.19
glioma/meningioma / 1.37 / 0.25
glioma/meningioma / 1.32 / 0.27

Fig. 26: Illustration of top miss-classified images after stage III of fine-tuning

We have plotted the error rates of the models, as presented in Fig.27. ResNet50 converges faster during the last two epochs of the third training stage compared to VGG, Alexnet, and ResNet34. This convergence property

Stage	Train Loss	Validation Loss	Error Rate	Accuracy	Precision	Recall	F1-Score
I	0.136737	0.121995	0.032680	0.967320	0.967320	0.967320	0.967320
II	0.114440	0.104330	0.035948	0.964052	0.964052	0.964052	0.964052
III	0.084613	0.067653	0.029412	0.970588	0.970588	0.970588	0.970588

TABLE 6: Comparison of metrics across the 3 stages of training with School of Biomedical Engineering dataset

of ResNet50 is demonstrated through the nearly smooth learning curve, which indicates the model was able to find a stable set of weights. The VGG model is second to ResNet50 regarding convergence to the local minimum during the training phases consistently across the three stages of training, which indicates superior performance to Alexnet and ResNet34. In Table. 6, we have presented the training and

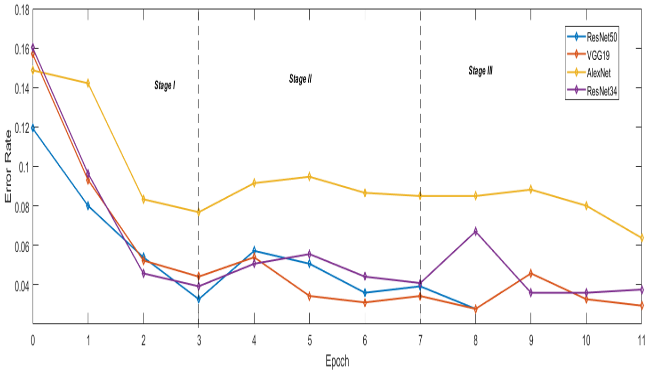

Fig. 27: Error Rates across four different model architectures

validation accuracy values for the experiment. The graph in Fig.28 gives information on the training and validation accuracy for all stages of training. We observed the network under-fitted since the validation error was lower than the training error. This was solved by extending the number of epochs. At stage two, we used techniques such as vertical flipping, max zooming, and max lighting to augment the MRI images.

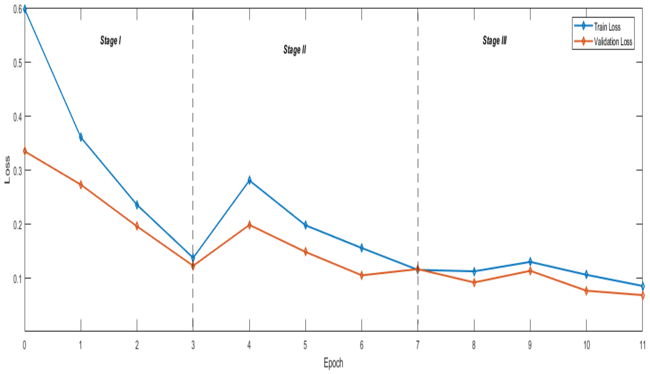

Fig. 28: Training vs. Validation Loss for all stages with ResNet50

We assess the classification performance on the validation data for the three stages. During the five-fold cross-validation training, we obtained the confusion matrices, which are shown in Fig.30 through Fig. 32. The matrices give a summary of the classification performance of the model were the leading diagonals indicate correctly predicted classes and while miss-classified samples are outside the diagonals. At the end of stage I, the model reached an accuracy of 96.73% with 20 incorrectly classified images across the three classes, as shown in Fig.30. The model in stage II, as shown in Fig. 19 had a steady number of incorrectly classified images. However, the average five-fold cross-validation accuracy value improved from 96.73% to 97.05% at the end of stage III. Finally, Fig.32 depicts stage III results. 17 images were incorrectly classified and the model reached 97.05% average five-fold cross-validation accuracy. The whole training procedure takes 430 seconds.

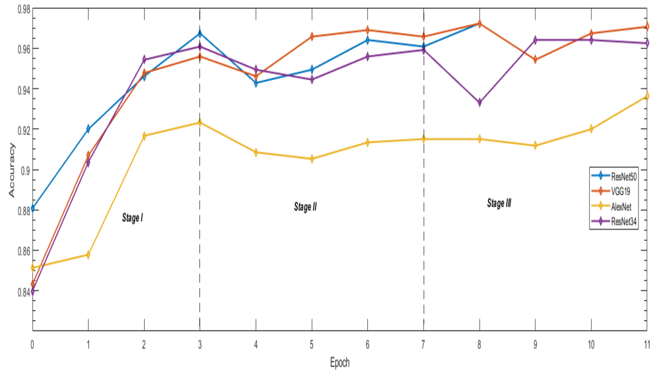

Fig. 29: Accuracy across four different model architectures

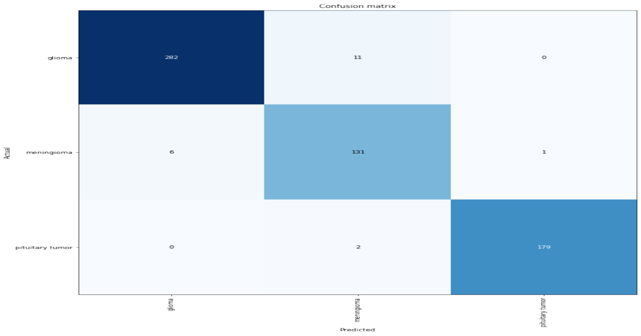

Fig. 30: Stage I confusion matrix after training the network with base layers frozen

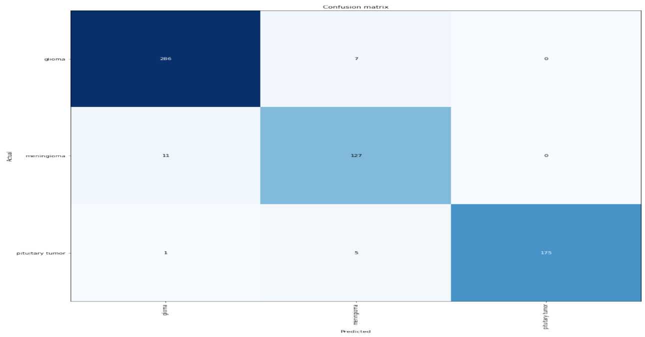

Fig. 31: Stage II confusion matrix after retraining the network with augmented data

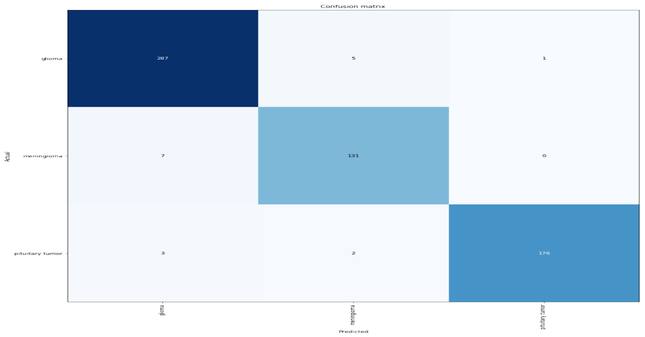

Fig. 32: Stage III confusion matrix after unfreezing the Network and fine-tuning with the augmented data

6 Discussion and Conclusion

6.1 Discussion

Many researchers have proposed methods and techniques in Machine Learning using standard publicly available datasets for brain MRI classification tasks. These methods vary in their approaches from feature engineering, dimensionality reduction to deep learning-based. Table 7 presents a comparison of research on the classification of MRI brain images. The table presents our proposed dataset Study I and three publicly available datasets: Study II (Harvard Whole Brain Atlas) [8], Study III (The Cancer Imaging Archive) [48], Study IV (School of Biomedical Engineering, Guangzhou University) [49]. In their paper, Chaplot et al. [11] proposed a 3-level Wavelet Transform method for feature extraction and a Support Vector Machine binary brain MRI classifier using 52 MRI datapoints obtaining an accuracy of 98%. Gudigar et al. [24] in a comparative study assessed three multi-resolution analysis methods - Discrete Wavelet, Curvelet, and Shearlet Transform, and Particle Swarm Optimization for textual feature extraction from transformed images using SVM classifier reported accuracy of 97.38% with a dataset of 612 brain MRI scans based on Shearlet Transform. In their 2015 paper, Zhang et al. [21], proposed a novel Computer-Aided Diagnostic system that comprises Discrete Wavelet Packet Transform (DWPT), which serves as a Wavelet Packet Coefficient Extractor, Shannon and Tsallis entropies which obtained entropy features from DWPT. Finally, they used a Generalized Eigenvalue Proximate Support Vector Machine with Radial Basis Function for brain MRI binary classification achieved 99.33% accuracy using a dataset of 255 brain MRI images. Saritha et al. [19] have used Wavelet Entropy-based Spider Web Plots' for dimensionality reduction combined with a Probabilistic Neural Network (PNN) for the classification reported accuracy of 100%. El-Dahshan et al. [12] in their work presented a novel technique consisting of Discrete Wavelet Transform, PCA to reduce the feature vectors dimensions, and a Feed-Forward Artificial Neural Network to classify 101 brain MRI images achieved 98.6% accuracy. Wang et al. [20] used a pipeline of Stationary Wavelet Transform for feature extraction, Principal Component Analysis to reduce the feature space combined with Particle Swarm Optimization and Artificial Vee Colony to classify MRI brain images reported accuracy of 99.45%. Nayak et al.[23] proposed a Computer-Aided Diagnosis framework the em-

Study	Accuracy Study I	Accuracy Study II	Accuracy Study III	Accuracy Study IV	Task	Method
Cheng et al [49]	-	-	-	91.28%	Mutli	SVM and KNN
Paul et al[14]	-	-	-	91.43%	Multi	CNN
Afshar et al [16]	-	-	-	90.89%	Multi	CNN
Anaraki et al[17]	-	-	90.9%	94.2%	Multi	GA-CNN
Zacharaki et al[7]	-	-	85%	-	Multi	SVM and KNN
Zacharaki et al[7]	-	-	88%	-	Binary	SVM and KNN
El-Dahshan et al[12]	-	-	98%	-	Binary	ANN and KNN
Ertosum et al[15]	-	-	71%	-	Multi	CNN
Ertosum et al[15]	-	-	96%	-	Binary	CNN
Sultan et al[25]	-	-	98.7%	96.13%	Multi	CNN
Gudigar et al.[24]	-	97.38%	-	-	Binary	PSO and SVM
Zhang et al.[21]	-	99.33%	-	-	Binary	GEPSVM
Talo et al.[4]	-	100%	-	-	Binary	Deep Transfer Learning
Proposed Study	**85.23%**	**93.80%**	-	**97.05%**	Multi	Deep Transfer Learning

TABLE 7: A comparative survey of brain MRI image classification methodologies

ployed contrast limited adaptive histogram equalisation as a mechanism to enhance tumor regions in brain MRI images, 2-D Stationary Wavelet Transform to extract features, AdaBoost with a Support Vector Machine algorithms for normal and abnormal brain MRI classification achieving 99.45% accuracy with 255 MRI images. In this paper, we have found in the existing literature that feature engineering, dimensionality reduction, and other hybrid approaches have been predominantly used to solve the brain MRI image classification problem. Various dimensionality reduction methods like Principal Component Analysis are employed to reduce the feature space of training datasets. Most recently, Talo et al. [4] used Transfer Learning to perform binary classification of the brain, but our study proposes an expressive, fine-grain classifier that was trained on three distinct datasets. We have obtained an accuracy of 84.40% on the National Institute of Neuroscience and Hospitals dataset which comprises 5,285 T1-weighted, 93.80% on the Harvard Whole Brain Atlas comprising 1,133 T2-weighted images and 97.05% on the Biomedical School of Engineering which comprise 3,064 T1-weighted brain images using 5-fold cross-validation with ResNet50 CNN model. In Fig. 33, we have presented a proposed system deployment diagram depicting the infrastructural setup and configuration of our proposed framework.

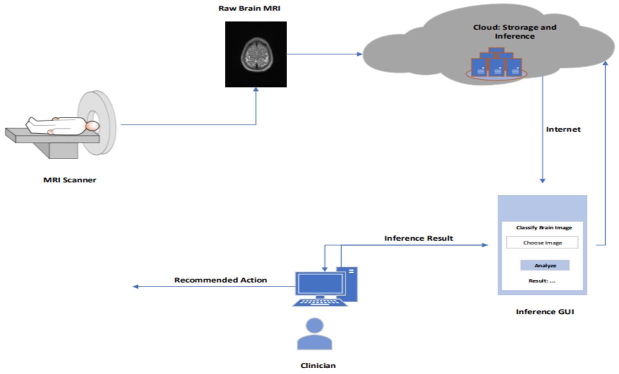

Fig. 33: Proposed system deployment diagram

6.2 Conclusion and Future Direction

identified the novelty of this research work that uses Deep Transfer Learning for multi-class classification of brain MRI images. Deep Transfer Learning demonstrates the potential of quickly adapting a model to solving a problem rather than building one from scratch and has shown to be suitable in terms of evaluation performance. Transfer Learning, as an active area of Machine Learning research, highlights the potential of accelerated adaptation of models across varied problem domains where underlying knowledge is invariant and transferable. We conducted three case studies using our novel dataset and two other publicly available datasets with separate contrast-enhancement techniques. Our research utilized data augmentation to address the limitations of the small dataset set size for the multi-class classification problem. We have shown the potential impact of Deep Transfer Learning as a potential research direction in solving the brain MRI multi-class classification task.

References

[1] M. Oguz and A. P. Cox, “Machine learning biopharma applications and overview of key steps for successful implementation.”

[2] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, “A survey on deep learning in medical image analysis,” *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[3] Z.-P. Liang and P. C. Lauterbur, *Principles of magnetic resonance imaging: a signal processing perspective*. SPIE Optical Engineering Press, 2000.

[4] M. Talo, U. B. Baloglu, Özal Yıldırım, and U. R. Acharya, “Application of deep transfer learning for automated brain abnormality classification using mr images,” *Cognitive Systems Research*, vol. 54, pp. 176 – 188, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389041718310933

[5] G. Mohan and M. M. Subashini, “Mri based medical image analysis: Survey on brain tumor grade classification,” *Biomedical Signal Processing and Control*, vol. 39, pp. 139–161, 2018.

[6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[7] E. I. Zacharaki, S. Wang, S. Chawla, D. Soo Yoo, R. Wolf, E. R. Melhem, and C. Davatzikos, “Classification of brain tumor type and grade using mri texture and shape in a machine learning scheme,” *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 62, no. 6, pp. 1609–1618, 2009.

[8] D. Summers, “Harvard whole brain atlas: www.med.harvard.edu/aanlib/home.html,” *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 74, no. 3, pp. 288–288, 2003. [Online]. Available: https://jnnp.bmj.com/content/74/3/288

[9] C. H. Moritz, V. M. Haughton, D. Cordes, M. Quigley, and M. E. Meyerand, “Whole-brain functional mr imaging activation from a finger-tapping task examined with independent component analysis,” *American Journal of Neuroradiology*, vol. 21, no. 9, pp. 1629–1635, 2000.

[10] S. G. Mallat, “A theory for multiresolution signal decomposition: the wavelet representation,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 7, pp. 674–693, 1989.

[11] S. Chaplot, L. Patnaik, and N. Jagannathan, “Classification of magnetic resonance brain images using wavelets as input to support vector machine and neural network,” *Biomedical signal processing and control*, vol. 1, no. 1, pp. 86–92, 2006.

[12] E.-S. A. El-Dahshan, H. M. Mohsen, K. Revett, and A.-B. M. Salem, “Computer-aided diagnosis of human brain tumor through mri: A survey and a new algorithm,” *Expert systems with Applications*, vol. 41, no. 11, pp. 5526–5545, 2014.

[13] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, “Comparing different supervised machine learning algorithms for disease prediction,” *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, pp. 1–16, 2019.

[14] J. S. Paul, A. J. Plassard, B. A. Landman, and D. Fabbri, "Deep learning for brain tumor classification," in *Medical Imaging 2017: Biomedical Applications in Molecular, Structural, and Functional Imaging*, vol. 10137. International Society for Optics and Photonics, 2017, p. 1013710.

[15] M. G. Ertosun and D. L. Rubin, "Automated grading of gliomas using deep learning in digital pathology images: A modular approach with ensemble of convolutional neural networks," in *AMIA Annual Symposium Proceedings*, vol. 2015. American Medical Informatics Association, 2015, p. 1899.

[16] P. Afshar, K. N. Plataniotis, and A. Mohammadi, "Capsule networks for brain tumor classification based on mri images and coarse tumor boundaries," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 1368–1372.

[17] A. K. Anaraki, M. Ayati, and F. Kazemi, "Magnetic resonance imaging-based brain tumor grades classification and grading via convolutional neural networks and genetic algorithms," *Biocybernetics and Biomedical Engineering*, vol. 39, no. 1, pp. 63–74, 2019.

[18] Y. Zhang, S. Wang, G. Ji, and Z. Dong, "An mr brain images classifier system via particle swarm optimization and kernel support vector machine," *The Scientific World Journal*, vol. 2013, 2013.

[19] M. Saritha, K. P. Joseph, and A. T. Mathew, "Classification of mri brain images using combined wavelet entropy based spider web plots and probabilistic neural network," *Pattern Recognition Letters*, vol. 34, no. 16, pp. 2151–2156, 2013.

[20] S. Wang, Y. Zhang, Z. Dong, S. Du, G. Ji, J. Yan, J. Yang, Q. Wang, C. Feng, and P. Phillips, "Feed-forward neural network optimized by hybridization of pso and abc for abnormal brain detection," *International Journal of Imaging Systems and Technology*, vol. 25, no. 2, pp. 153–164, 2015.

[21] Y. Zhang, Z. Dong, S. Wang, G. Ji, and J. Yang, "Preclinical diagnosis of magnetic resonance (mr) brain images via discrete wavelet packet transform with tsallis entropy and generalized eigenvalue proximal support vector machine (gepsvm)," *Entropy*, vol. 17, no. 4, pp. 1795–1813, 2015.

[22] D. Ranjan Nayak, R. Dash, and B. Majhi, "Stationary wavelet transform and adaboost with svm based pathological brain detection in mri scanning," *CNS & Neurological Disorders-Drug Targets (Formerly Current Drug Targets-CNS & Neurological Disorders)*, vol. 16, no. 2, pp. 137–149, 2017.

[23] D. R. Nayak, R. Dash, and B. Majhi, "Brain mr image classification using two-dimensional discrete wavelet transform and adaboost with random forests," *Neurocomputing*, vol. 177, pp. 188–197, 2016.

[24] A. Gudigar, U. Raghavendra, T. R. San, E. J. Ciaccio, and U. R. Acharya, "Application of multiresolution analysis for automated detection of brain abnormality using mr images: A comparative study," *Future Generation Computer Systems*, vol. 90, pp. 359–367, 2019.

[25] H. H. Sultan, N. M. Salem, and W. Al-Atabany, "Multi-classification of brain tumor images using deep neural network," *IEEE Access*, 2019.

[26] P. Pławiak, "Novel genetic ensembles of classifiers applied to myocardium dysfunction recognition based on ecg signals," *Swarm and evolutionary computation*, vol. 39, pp. 192–208, 2018.

[27] Ö. YILDIRIM and U. B. BALOĞLU, "Texture classification system based on 2d-dost feature extraction method and ls-svm classifier," *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, vol. 21, no. 2, pp. 350–356, 2017.

[28] P. Pławiak, "An estimation of the state of consumption of a positive displacement pump based on dynamic pressure or vibrations using neural networks," *Neurocomputing*, vol. 144, pp. 471–483, 2014.

[29] K. Rzecki, T. Sośnicki, M. Baran, M. Niedźwiecki, M. Król, T. Łojewski, U. Acharya, Ö. Yildirim, and P. Pławiak, "Application of computational intelligence methods for the automated identification of paper-ink samples based on libs," *Sensors*, vol. 18, no. 11, p. 3670, 2018.

[30] Y. Bengio and H. Lee, "Editorial introduction to the neural networks special issue on deep learning of representations," *Neural Networks*, vol. 64, no. C, pp. 1–3, 2015.

[31] C. Cao, F. Liu, H. Tan, D. Song, W. Shu, W. Li, Y. Zhou, X. Bo, and Z. Xie, "Deep learning and its applications in biomedicine," *Genomics, proteomics & bioinformatics*, vol. 16, no. 1, pp. 17–32, 2018.

[32] Ö. Yıldırım, P. Pławiak, R.-S. Tan, and U. R. Acharya, "Arrhythmia detection using deep convolutional neural network with long duration ecg signals," *Computers in biology and medicine*, vol. 102, pp. 411–420, 2018.

[33] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, M. Adam, A. Gertych, and R. San Tan, "A deep convolutional neural network model to classify heartbeats," *Computers in biology and medicine*, vol. 89, pp. 389–396, 2017.

[34] S. Kiranyaz, T. Ince, and M. Gabbouj, "Real-time patient-specific ecg classification by 1-d convolutional neural networks," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 3, pp. 664–675, 2015.

[35] Ö. Yildirim, "A novel wavelet sequence based on deep bidirectional lstm network model for ecg signal classification," *Computers in biology and medicine*, vol. 96, pp. 189–202, 2018.

[36] U. R. Acharya, H. Fujita, O. S. Lih, Y. Hagiwara, J. H. Tan, and M. Adam, "Automated detection of arrhythmias using different intervals of tachycardia ecg segments with convolutional neural network," *Information sciences*, vol. 405, pp. 81–90, 2017.

[37] O. Yildirim, R. San Tan, and U. R. Acharya, "An efficient compression of ecg signals using deep convolutional autoencoders," *Cognitive Systems Research*, vol. 52, pp. 198–211, 2018.

[38] Ö. Yıldırım, U. B. Baloglu, and U. R. Acharya, "A deep convolutional neural network model for automated identification of abnormal eeg signals," *Neural Computing and Applications*, pp. 1–12, 2018.

[39] J. Cheng, "brain tumor dataset," 4 2017. [Online]. Available: https://figshare.com/articles/brain˙tumor˙dataset/1512427

[40] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[41] M. C. Mabray and S. Cha, "Advanced mr imaging techniques in daily practice," *Neuroimaging clinics of North America*, vol. 26, no. 4, p. 647—666, November 2016. [Online]. Available: https://doi.org/10.1016/j.nic.2016.06.010

[42] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," *CoRR*, vol. abs/1808.01974, 2018. [Online]. Available: http://arxiv.org/abs/1808.01974

[43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[44] L. N. Smith, "Cyclical learning rates for training neural networks," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2017, pp. 464–472.

[45] G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger, "Snapshot ensembles: Train 1, get m for free," *arXiv preprint arXiv:1704.00109*, 2017.

[46] J. Howard *et al.*, "fastai," https://github.com/fastai/fastai, 2018.

[47] N. Ketkar, "Introduction to pytorch," in *Deep learning with python*. Springer, 2017, pp. 195–208.

[48] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle *et al.*, "The cancer imaging archive (tcia): maintaining and operating a public information repository," *Journal of digital imaging*, vol. 26, no. 6, pp. 1045–1057, 2013.

[49] J. Cheng, W. Huang, S. Cao, R. Yang, W. Yang, Z. Yun, Z. Wang, and Q. Feng, "Correction: Enhanced performance of brain tumor classification via tumor region augmentation and partition," *PloS one*, vol. 10, no. 12, p. e0144479, 2015.