Summer 2021

Predicting Severity of Traumatic Brain Injury: A Residual Learning Model from Magnetic Resonance Images

Dacosta Yeboah
Missouri State University, Dacosta123@live.missouristate.edu

PREDICTING SEVERITY OF TRAUMATIC BRAIN INJURY: A RESIDUAL LEARNING MODEL FROM MAGNETIC RESONANCE IMAGES

A Master's Thesis

Presented to

The Graduate College of

Missouri State University

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science, Computer Science

By

Dacosta Yeboah

July 2021

PREDICTING SEVERITY OF TRAUMATIC BRAIN INJURY: A RESIDUAL LEARNING MODEL FROM MAGNETIC RESONANCE IMAGES

Computer Science

Missouri State University, July 2021

Master of Science

Dacosta Yeboah

ABSTRACT

One of the most significant frontiers for computational scientists is the engineering of human healthcare delivery based on intelligent analysis of health data. In a variety of neurological disorders such as Traumatic Brain Injury (TBI), neuro-imaging information plays a crucial role in the decision-making regarding patient care and as a potential prognostic marker for outcome. TBI is a heterogeneous neurological disorder. Due to the economic burdens of the disorder, sorting out this heterogeneity could provide more insights and better understanding of TBI recovery trajectories, thus improving overall diagnosis and treatment options. Magnetic Resonance Imaging (MRI) is a non-invasive technique that for examining the anatomy and pathology of the brain. This work examines a residual convolutional neural network to build a predictive model for TBI severity using varied MR images and three different combinations of TBI assessment scores obtained from Federal Interagency Traumatic Brain Injury Research. To address the challenges of insufficient data and increase efficiency. The framework consists of five components which include data curation, data augmentation, residual learning model, model validation and clinical relevance assessment. The data curation phase pre-processes the images into a format reliable for use by the model. To address the problem of insufficient, unbalanced and highly skewed data, the data augmentation generates different forms of the images to improve the generalization capability of the model. The residual learning model integrates transfer learning by utilizing a network that has been pre-trained on general data and then fine-tuned for MR images to improve the model performance and reduce training time. Model validation consists of both quantitative and qualitative means. The clinical relevance assessment phase includes the identification of meaningful subgroups to better understand the how the results correlate with the MRI data. A mixed effects Analysis of Variance (ANOVA) model is performed using varied TBI outcome measures to assess the clinical significance of the results. The experimental results showed that our model achieve a high precision on the test sample.

KEYWORDS: deep learning, traumatic brain injury, residual learning, magnetic resonance imaging, transfer learning

PREDICTING SEVERITY OF TRAUMATIC BRAIN INJURY: A RESIDUAL LEARNING MODEL FROM MAGNETIC RESONANCE IMAGES

By

Dacosta Yeboah

A Master's Thesis
Submitted to the Graduate College
Of Missouri State University
In Partial Fulfillment of the Requirements
For the Degree of Master of Science, Computer Science

July 2021

Approved:

Tayo Obafemi-Ajayi, Ph.D., Thesis Committee Chair

Lloyd Smith, Ph.D., Committee Member

Siming Liu, Ph.D., Committee Member

Julie Masterson, Ph.D., Dean of the Graduate College

ACKNOWLEDGEMENTS

I would like to acknowledge everyone who played a role in my academic accomplishments. First of all, my parents, who supported me with love and understanding. Without you, I could never have reached this current level of success.

Secondly, my committee members, each of whom has provided patient advice and guidance throughout the research process. Thank you all for your unwavering support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

In recent years, deep learning has gained significant stride in machine learning translational research for a wide range of applications is utilized in healthcare for a variety of tasks ranging from computer-aided detection, prediction of medical events and supporting clinical decision making and survival analysis. Image classification has had series of breakthroughs as a result of deep learning [1]. Machine learning translational research is the application of machine learning on data set extracted from a specific domain, with the objective of obtaining useful insights or patterns that further helps to understand a problem related to that domain. With machine learning techniques and the availability of data, patterns that were not previously known could be easily discovered and aid domain experts in achieving desired goals. Medical images have been widely used in clinics, providing visual representations of under skin tissues in the human body [2]. A variety of imaging modalities including magnetic resonance imaging (MRI), computed tomography (CT), X-rays, and ultrasound are used for disease diagnosis and prognosis [3]. This work focuses on machine learning translational research based on deep convolutional neural network (CNN) in the medical imaging domain.

Deep CNNs have shown promising results for applications related to classification of medical images using MRI for a variety of disorder including neurological disorders such as Alzheimer's disease [4] [5]. However, there exist challenges of insufficient data, vanishing and exploding gradient problems, thus resulting in poor performance. Residual neural networks (ResNets) have been proposed [6] have demonstrated great potential in addressing the vanishing and exploding gradient problems [21] by introducing skip connections that short circuit shallow layers to deep layers. These connections between layers add the outputs from previous layers to the outputs of stacked layers. The accuracy of deep learning architectures can be further

improved by transfer learning to alleviate the problem of limited data and generalization. Transfer learning is the process of improving the learning of a new task by the transferring knowledge from a previously learned but related task [7]. Transfer learning is flexible and allows the use of weights from pre-trained models developed from standard computer vision benchmark data into new models. Integrating transfer learning into the deep CNN architecture could result in increased efficiency and a more robust performance [8].

In this work, we explore the application of a deep CNN method to Traumatic Brain Injury (TBI). TBI is a neurological disorder caused by a blow in the head that results in the disruption of brain function [9]. It is heterogeneous in cause, severity, pathology and prognosis. Sorting out the heterogeneity in TBI, though challenging, could reveal useful insights to clinical experts and aid in making more informed decisions. With the availability of enough patient data, sorting out the heterogeneity of TBI could be described as a typical machine learning problem. When the specific groups are known and associated with the available data, this problem translates to a supervised learning task in which a model could be trained to learn patterns in the data to accomplish the task of categorizing patients into known groups. Currently, the Glasgow Coma Scale (GCS) is an assessment score used to severalize patients with TBI into three categories based on TBI – severe (GCS 3-8), moderate (GCS 9-12), and mild (GCS 13-15) [10]. Abnormalities found in the CT scan of TBI patients are also used in quantifying the severity of the disorder. The Marshall and Rotterdam scores (Figure 1 and Figure 2) (both ranging from 1 to 6) are CT metrics used to predict TBI outcome [11] [12]. The Marshall score is based on abnormalities defined by visible presentation of increasing evidence of mass effect. The Rotterdam scores are based on the sum of specific CT scan elements that correlate with poor outcomes.

This work investigates a deep convolutional neural network based on residual learning to

predict the severity of TBI (as quantified by GCS score) from MRI brain scans. We extend the framework to include a joint predictor severity model based on both GCS and a CT derived metric (either Marshall or Rotterdam). The proposed model, as illustrated in Figure 1, utilizes the ResNet-50 architecture as a base model on top of which other fine-tuning layers are added. Our model integrates the concept of transfer learning by using information gained while learning from general image data set on MR images. Data curation is needed to pre-process the images using inhomogeneity correction and skull stripping into a format that is acceptable for the network. Data augmentation is utilized to expand the limited images available to ensure a better model performance as the model is generalized by various forms of the data**.** We evaluate the sensitivity of a model to detect anatomical changes in brain MRI scans that might correlate with outcome after TBI. The evaluation of the results is performed using both quantitative and qualitative analysis. The novelty of this work is that by transferring information that has been learned from general data set, we are able to fine-tune a residual neural network to perform joint classification tasks sufficiently well. We also incorporate CT scans information with MRI data, to examine correlations, if any, between both data modalities. To evaluate the effectiveness of our model, we utilize a commonly used plain CNN architecture, visual geometry group (VGG-16), without transfer learning for baseline comparisons.

Given the translational research focus of this work, we are interested in the interpretability and clinical relevance of the results obtained. This is dependent on the domain area, TBI, in this case. Varied outcome measures have been recommended to determine the baseline function of an individual at the beginning of treatment and to determine the progress and treatment efficacy. Thus, using a varied set of commonly acceptable TBI outcome measures [13], we conduct statistical analysis of the experimental results to validate the clinical relevance of the model for routine evaluation of TBI at the individual patient level. We analyze varied

severity subgroups correctly learned in comparison to the groups that the model failed to learn to further understand the correlations between TBI imaging modalities, clinical data, and outcome measures. Results from this predictive model would enhance personalized medicine for TBI patients by aiding decisions about key MRI features and the connections with patient prognosis and recovery outcomes.

The outline of the remainder of the paper is as follows. Firstly, we present an overview of CNNs, residual networks, transfer learning and data augmentation, then a review of current state of the art methods with respect to deep CNNs and neurological disorders classification. Secondly, we present preliminary works performed which include the application of residual network model on CIFAR-10 data set. Next, we discuss the detail of the residual learning model for sorting out the heterogeneity of TBI using MR images. Fourthly, we present and analyze the experimental results obtained. Lastly, we discuss the clinical interpretation and conclude with some suggestions for future work.

BACKGROUND

In this section, we briefly described key concepts including convolutional neural networks, residual neural networks, transfer learning and data augmentation utilized in this work to provide some context.

Convolutional Neural Networks (CNN)

Convolutional neural networks are a type of deep neural networks that derive their name from mathematical linear operation between matrices known as convolution [14]. The notable difference between CNNs and traditional networks and is that CNNs are mainly used in the field of pattern recognition within images [15]. They can be compared to the visual cortex (part of the brain that processes visual information). They also have multiple layers which include convolutional, pooling, and fully connected layers.

Figure 1 illustrates a general CNN architecture. The input layer of a CNN holds the pixel values of the input image. The convolutional layers contain kernels or filters whose parameters need to be learned. These are matrices that slide across the image to detect certain features. When the filter detects its feature on a sub-part of the image, it fires. At each layer, we end up with a feature map that can be passed to the next layer. CNNs obtain abstract features when input propagates toward the deeper layers [14] . The first CNN layers only extract basic features whilst deeper layers extract more concrete features. Pooling layers perform a down sampling operation in order to reduce the complexity for further layers. Max-pooling, partitions the image into sub-regions rectangles and only returns the maximum value each sub-region. The fully connected layers are similar to the neurons in traditional neural networks. Each neuron in a fully connected layer is directly connected to every node in the next and previous layers.

CNN architectures vary across different applications. A popular network GoogleNet [16], also commonly known as 'Inception' network, uses a unique type of layer called inception layer/block from which it drives its main strength. AlexNet is a deeper and wider CNN model that won the ImageNet challenge. Whereas still useful, AlexNet is no longer considered a state-of-the art network [17]. A network still applied frequently is VGG16, which is a plain CNN architecture with 16 layers. The ResNet architecture [6], based on residual learning has also gained high popularity in its ability to ease the training of networks that are substantially deeper. The DenseNet architecture [18] exploits the insights of residual learning to achieve the representation power similar to ResNet, but with a more compact network. However, this work utilizes the ResNet architecture due to its robustness and ability to solve a major problem faced by plain deep networks, which will be discussed in the next section.

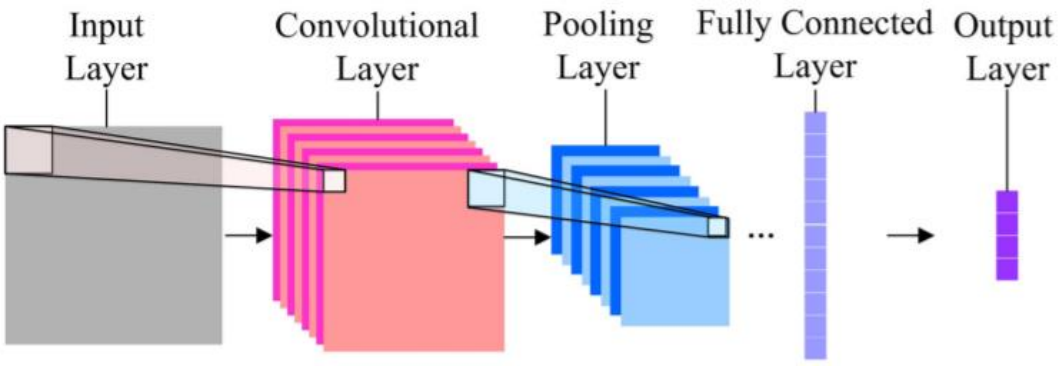


Figure 1. A general CNN architecture. This is adapted from Figure 1 in [19].

Residual Network

The ResNet architecture, consisting of residual blocks, utilizes residual learning to train deeper neural networks. This architecture solves the vanishing gradient problem found in plain deep CNNs by introducing skip connections that short circuit shallow layers to deep layers [21]. A residual block is shown in Figure 2. These connections between layers add the outputs from

previous layers to the outputs of stacked layers. The skip connections enable the network to learn residuals, performing a kind of boosting [3]. In residual learning [21], a building block can be defined as $y = F(x, W_i) + x$ where x and y are input and output vectors of the layers considered and F represents the residual mapping to be learned. The dimensions of x and F must be equal. To match them, if needed, a linear projection W_s is performed by the shortcut connection: $y = F(x, W_i) + W_s x$. The ResNet-50 architecture is utilized in this study.

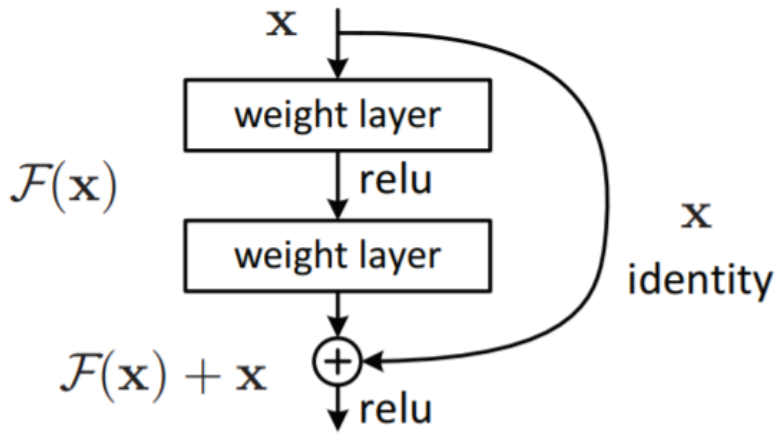


Figure 2. Residual learning: a building block. This is adapted from Figure 2 in [6].

Transfer Learning

Transfer learning is the improvement of learning in a new task through the transfer of knowledge from a related task that has already been learned [9]. Unlike traditional learning systems which are built specifically for separate tasks, transfer learning induces knowledge that has been learned from a previous task to a new task. Given a source domain D_s and learning task T_s, a target domain D_t and learning task T_t, transfer learning aims to help improve the learning of a target predictive function $F_t(.)$ in D_t using the knowledge in D_s and T_s, where $D_s \neq D_t$, or $T_s \neq T_t$. A domain consists of a feature space X and a probability distribution $P(X)$, where X $=$ {x1…, xn} € X. For a given domain, a task consists of a label space and an objective predictive

function which is learnt from the training data. In this work, transfer learning is utilized by transferring knowledge of parameters. For a neural network, the weights used to learn from D_s are transferred to D_t. Hence, information that has already been learned from D_s is applied to D_t. In Figure 3 illustrate the difference between traditional machine learning systems and transfer learning system. This work integrates the concept of transfer learning in the ResNet architecture to ensure a more robust system, given the limited clinical images available as well as speed up the learning process by reduce training time with pretrained models. Network weights that have been pre-trained on ImageNet data set are used to train the MR images.

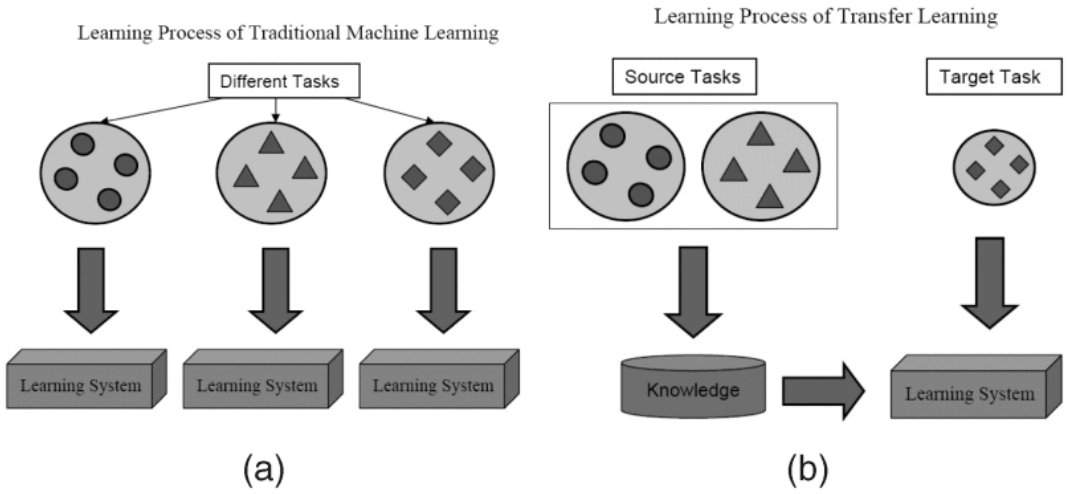


Figure 3. Traditional machine learning methods vs. Transfer learning. This is adapted from Figure 2 in [20].

Data Augmentation

A commonly encountered problem is the lack of sufficient amount of training data or even class imbalance within data sets. This can be alleviated by applying data augmentation technique [21]. Data augmentation generates multiple slightly different versions of images from each image in the original training set. If data set in a machine learning model is rich and

sufficient, the model performs better and more accurate [22]. Furthermore, gathering and labeling of data can be exhausting and expensive, hence transformations in datasets by using data augmentation techniques reduces operational costs [22]. Data augmentation techniques include flipping the image in various directions, translating the image within a range of distance, cropping the image in different ways, rotating the image within a range of angles, scaling the image over a range of factors, generating shape and intensity-transformed images by linear or non-linear methods [23]. This implies that variations of the training set data are likely to be seen by the model. However, one main challenge of data augmentation is that if the data set contains biases, the augmented data will contain biases too and lead to over-fitting. Hence, it is essential to identify the optimal data augmentation strategy.

RELATED WORK

A variety of CNN architectures have been used in the healthcare domain to accomplish a wide range of tasks. In [24], Altaf et al reviews current state of art methods of varied CNN architectures for varied medical applications such as Alzheimer's disease, breast cancer, and age-related macular degeneration**.** Applications of deep learning in medical image analysis including detection/localization, segmentation, registration and classification were discussed. These methods have been applied to various medical images which capture the brain, breast, eye, chest, abdomen, and others.

Deep CNN models have also been used to perform joint task whereby the model performs two or more related machine learning tasks for a given data input. In [25] , Liu et al proposed a framework for joint classification and regression of brain status using MRI and personal information from Alzheimer's disease data. The first step of their framework involves MRI image processing. This includes anterior commissure – posterior commissure correction, intensity correction, skull stripping and cerebellum removing. Then, using the landmark discovery algorithm, they extract patches based on landmarks. This is fed into a multi-channel CNN. The output from the CNN is combined with demographic information including age, gender and education and into a fully convolutional neural network which has two output branches for multi-class classification and clinical score regression respectively. The multi-class classification uses a SoftMax activation function to predict a patient into one of four classes. The branch for clinical score regression predicts four clinical scores pertaining to Alzheimer's disease. The framework yielded better results in both classification and regression when compared to other methods.

Despite the success of deep learning architectures, there is difficulty in training as the

number of layers in the network increases. Different deep learning architectures have been proposed to overcome this problem. Srivastava et al [26] proposed a deep learning architecture to overcome the difficulty of training neural networks as depth increases. The architecture, called highway networks, is inspired by Long Short-Term Memory recurrent neural networks and allows unimpeded information flow across many. Convolutional highway layers are constructed similar to fully connected layers The authors compared their architecture to plain networks with various depths. The test set performance obtained was competitive to state-of-the-art methods with much fewer parameters. One advantage of the highway architecture is that it can learn to dynamically adjust the routing of information based on the current input. However, solution to the problem of training deep networks is not only limited to the highway networks.

In [27], He et al also provided empirical evidence showing that residual networks, which are easier to optimize can gain accuracy from increased depth. Using the ImageNet data set, residual networks far deeper than VGG networks were evaluated. Analysis was also presented for CIFAR-10 and CIFAR-100 datasets. The images were resized and data augmentation was applied. Both residual and plain networks were compared. The experimental results obtained confirmed the superior performance of the residual networks over the plain networks.

Furthermore, the residual network architecture has been used to uncover hidden patterns in MRI scans for disease detection. In [28], Ebrahimi et al utilized ResNet-18 to detect Alzheimer's disease using MRI scans, where the classification classes were Alzheimer and Normal Control. The network was trained by transferring knowledge obtained from 2-dimentional data set to the 3D MR images. Image processing steps include intensity normalization, registration and augmentation. Their model achieved a good accuracy, sensitivity and specificity. Unlike our proposed model which performs joint tasks, their model performs a single classification task.

Our proposed framework is motivated by the work by Liu et al in [29]. The authors propose a multi-task deep model based on for automated lung nodule analysis using CT scans. Their method, based on ResNet architecture performed lung nodule malignancy classification task and attribute score regression task which are characteristics for malignancy assessment. Unlike our model which includes two classification components, the output of their framework is made of a classification module and regression module. The authors included a siamese network to address the problem of misclassification on ambiguous lung nodules, which was not included in our work. Image processing steps performed include random cropping, horizontal and vertical flipping. One limitation of their work is that the authors did not evaluate the effectiveness of the model in performing a single task independently. Our proposed model, inspired by this work extends it to MR images.

Overall Framework

The learning framework, as illustrated in Figure 4, consists of five phases (data curation, data augmentation, training of residual learning model, model validation, and assessment of clinical relevance).

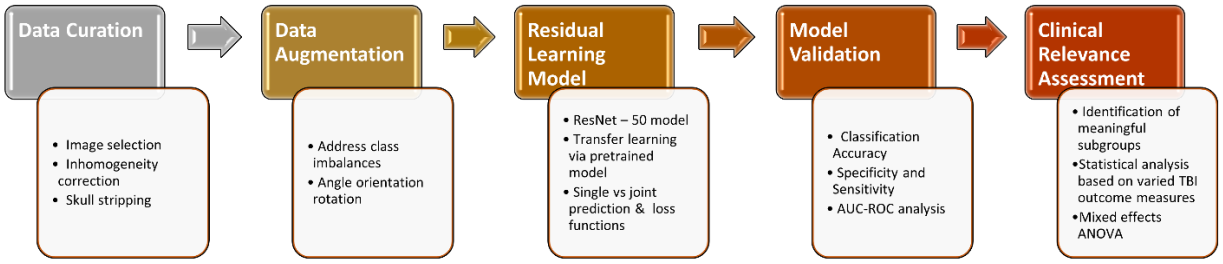


Figure 4. Deep learning framework for TBI severity prediction from MR images.

The data curation step, an essential process in any data driven learning model includes data extraction, cleaning, filtering, and pre-processing of the raw data to ensure that reliable data is available for modeling. Data augmentation, as discussed previously generates multiple different versions of the data. The model validation step evaluates the model performance using commonly used validation metrics. Clinical relevance assessment component allows us to evaluate whether identified groups from the model results have clinical significance. Statistical testing can aid in determining if the subgroups groups obtained from the modeling have predictive power for prognosis. A mixed effects analysis of variance (ANOVA) is performed to compare differences in the dependent variable (outcome measures) between two independent variables (predicted severity groups and time points).

Each of the key phases are described in detail in the next section. In order to first evaluate

the effectiveness of the residual learning model component of the overall model as well as the data augmentation component, we conduct some preliminary work using the commonly used CIFAR-10 data set.

Preliminary Work on CIFAR-10 Data Set

The ResNet framework for the CIFAR-10 was implemented using the Keras library with TensorFlow backend. The CIFAR-10 dataset consists of 60000 32x32 color images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images. The 10 different classes represent airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks. Figure 5 shows a random sample of images drawn from the CIFAR-10 data set.

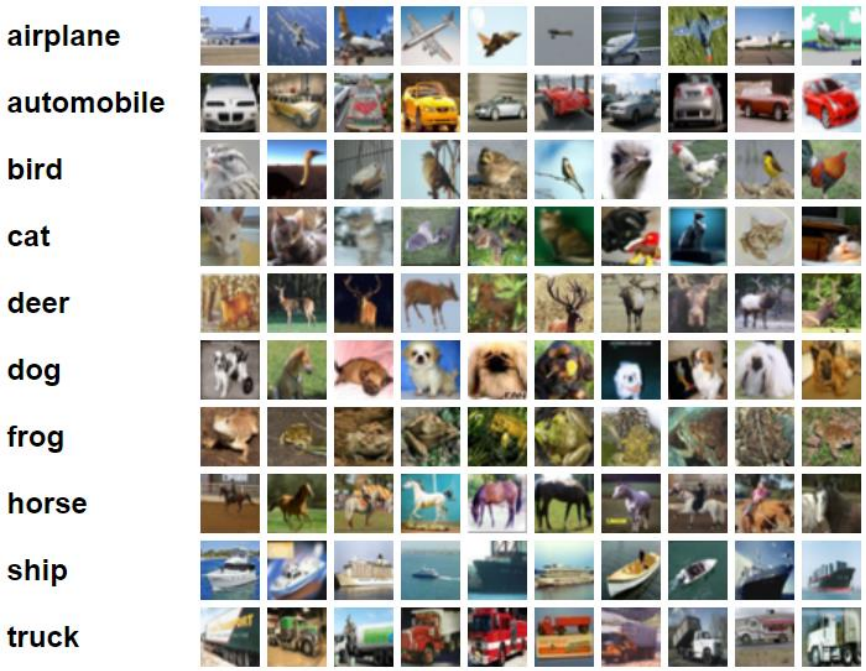
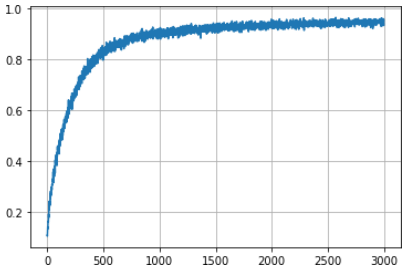


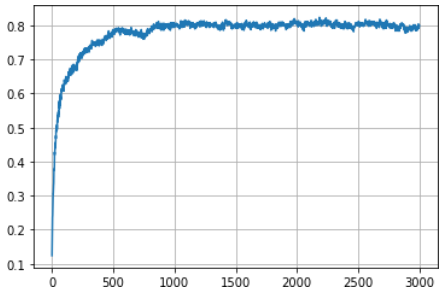
Figure 5. Classes from the CIFAR-10 dataset, with 10 samples from each class. Figure is adapted from [30].

For the purposes of this experiment, only a subset of this data samples was used. 500 images were sampled for training and 500 were sampled for testing. More training data was

acquired by applying data augmentation to the training data. 50 samples were augmented (each image i s varied 10 times). This produces 500 additional training samples of data. Hence, the total training samples were 1000. The data augmentation methods used were image rotation, width shifting, shearing, zooming, channel shifting and horizontal flipping. By integrating transfer learning, the network pre-trained on ImageNet dataset and the resulting weights were used to train the CIFAR-10 dataset. Since the dataset has 10 classification classes, a SoftMax activation layer with 10 neurons is used at the end of the Resnet model. Categorical cross-entropy was used as the loss function. The 32 × 32 images are placed into large 224 × 224-pixel images and can hence be heavily scaled, rotated and color augmented. The optimizer used was RMSprop with a learning rate of 1e-5. The test set consisting of 500 samples is used as the validation set. The test batch contained exactly 10 randomly-selected images from each class. A batch size of 10 was used in the training process. The graphs in Figure 5 indicate the trajectory training and validation accuracies of the model. A confusion table, used to describe the performance of the classification model is shown in Table 1.

From the experimental results, the model achieved a good accuracy in classifying images into their corresponding groups. This can be clearly drawn from the high accuracies achieved from Figure 5. and the confusion matrix in Table 1. The diagonal of the confusion matrix indicates the number of correctly classified samples for each class. The high performance obtained from the ResNet-50 model on the CIFAR-10 dataset motivates this work to utilize residual learning on MR images.





(a) Training accuracy (b) Validation accuracy

Figure 6. Training and validation accuracies of ResNet model on CIFAR-10 dataset.

Table 1. A confusion matrix between the ground truth and predicted labels of the CIFAR-10 testing data set.

n = 500	class 1	class 2	class 3	class 4	class 5	class 6	class 7	class 8	class 9	class 10
class 1 (57)	45	0	1	1	2	0	1	0	7	0
class 2 (41)	1	36	0	0	0	0	1	0	0	3
class 3 (51)	3	0	37	4	3	3	1	0	0	0
class 4 (49)	1	0	4	38	0	3	2	0	0	1
class 5 (40)	1	0	1	3	30	1	1	3	0	0
class 6 (48)	0	0	1	6	2	38	0	1	0	0
class 7 (54)	0	0	0	3	3	0	48	0	0	0
class 8 (47)	1	0	1	9	2	4	0	30	0	0
class 9 (57)	2	4	1	1	0	0	0	0	48	1
class10 (56)	0	2	0	0	0	0	0	0	4	50

METHODOLOGY

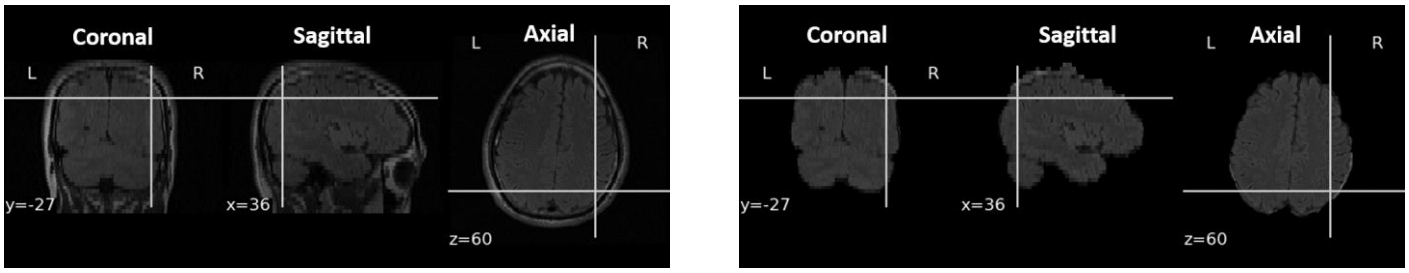
In this section, we explain the various components of our framework applied on MR images - data curation, data augmentation, residual learning model, model evaluation and the clinical relevance of the result.

Data Curation

Data curation is essential for any data driven learning model. Data curation includes data extraction, cleaning, filtering, and pre-processing of the raw data to ensure that reliable data is available for modeling. The TBI image data analyzed in this work is drawn from the Transforming Research and Clinical Knowledge in Traumatic Brain Injury (TRACK-TBI) pilot data set available via Federal Interagency Traumatic Brain Injury Research (FITBIR) [31] data repository to approved researchers. The data set comprised of 203 patients who underwent MRI brain scans. A variety of MRI sequences were available. Based on domain expert guidance, we focused on analysis of the fluid attenuated inversion recovery (FLAIR) images using all three planes (axial, coronal, and sagittal). This sequence has been shown to be sensitive to brain pathology and distinct between areas of brain injury.

Automated analysis of MR images is challenging due to intensity inhomogeneity, variability of the intensity ranges and contrast, and noise. Thus, preprocessing steps unique to image data are essential prior to the learning model phase. The brain MR scans were available in the Digital Imaging and Communications in Medicine (DICOM) open software format. Each DICOM image represents an individual slice of the brain. In order to utilize the spatial information, we converted the DICOM images into neuroimaging informatics technology (Nifti) volumes. Skull stripping was performed to remove the skull from images and focus on

intracranial tissues [32] . Inhomogeneity correction mitigates image contrast variations due to intensity inhomogeneity. We performed skull stripping and inhomogeneity correction on all the images using the R fslr package [33], as illustrated in Figure 7. The patient image data distribution based on the GCS, Marshall and Rotterdam scores are summarized in Table 2.



(a) Raw image scan

(b) Processed image after inhomogeneity correction and skull stripping

Figure 7. Image pre-processing of an MRI scan.

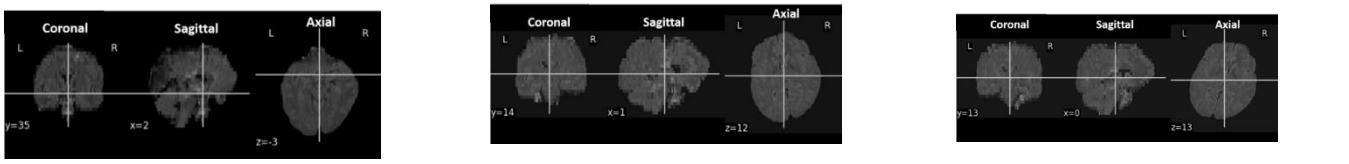
Table 2. Data distribution of 203 patients based on GCS, Marshall and Rotterdam scores.

Characteristic	Value
GCS	Mild: 77.8%, Moderate: 6.9%, Severe: 15.3
Marshall scores	1: 55.2%, 2: 31.0%, 3: 5.9%, 4: 1.5%, 5: 4.4%, 6: 2.0%
Rotterdam scores	1: 1.5%, 2: 70.4%, 3: 17.7%, 4: 7.4%, 5: 3.0%

Data Augmentation

The available data (Table 2) is of limited size and is class imbalanced (skewed towards the mild GCS group). Due to the imbalance in the GCS severity groups, data augmentation was

applied on the data set. The data augmentation process was applied by rotating images. Figure 8 shows an example of an image that has been augmented twice. After the data augmentation step, 144 moderate and 127 severe group images were added to the data set. This resulted in a total of 474 images with each GCS severity group having 158 images each. Figure 9 shows the data distribution on the Marshall and Rotterdam scores after the data augmentation process. We created an augmented data set of n = 474 subjects such that all three classes of GCS severity (mild, moderate, and severe) were balanced with 158 subjects each.



(a) Original image (b) First augmentation (c) Second augmentation

Figure 8. Data augmentation utilizing rotation to generate two additional images.

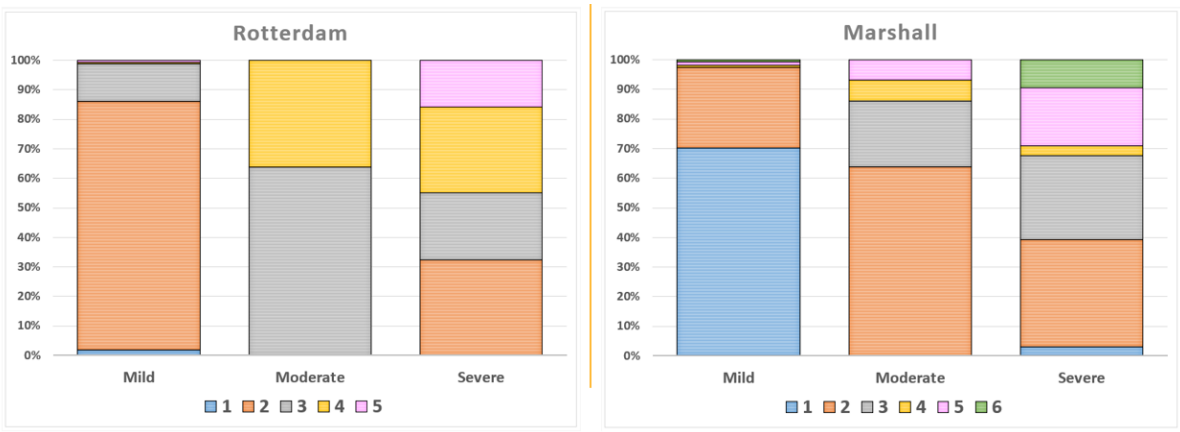


Figure 9. Skewed distribution of CT metric groups across augmented data.

Residual Learning Model

The residual learning model is illustrated in Figure 10. The ResNet-50 model consists of five stages, each having a convolutional (made up of three stacked layers) and identity block.

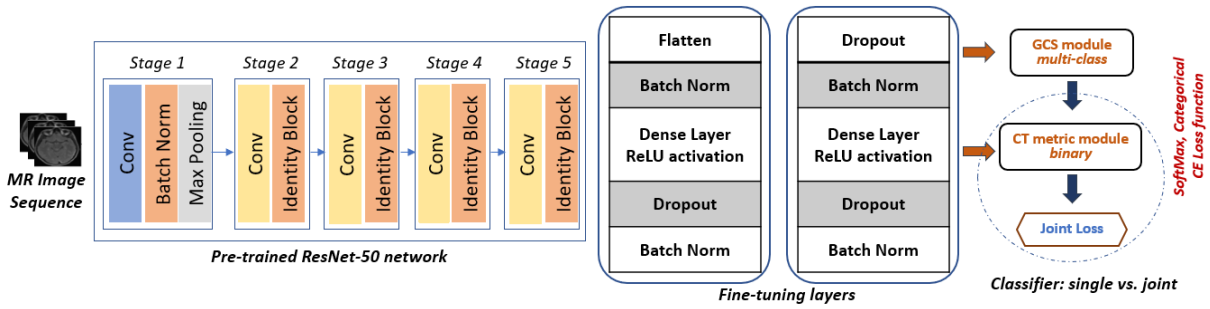


Figure 10. Residual learning model architecture.

The stage 1 block also includes a max-pooling operation that performs down-sampling. At the end of the last layer (stage 5), the data is passed in sequential order to the fine-tuning layers that flatten, batch normalize, and perform dropout regularization. It includes fully connected dense layers with ReLu activation function. The learning framework integrates transfer learning by adapting a well performing deep learning network (ResNet-50) trained on a large data set (ImageNet) and then subsequently fine-tuned on our smaller TBI MRI data. It has been shown that transferring the weights (and network parameters) from a pre-trained generic network to train on a specific dataset is better than random weight initialization of the network [32]. The weights of the layers in each stage of the ResNet-50 are fixed during the training process. In the subsequent fine-tuning layers, the network is trained with random weight initialization based on the transferred weights and parameters from the pre-trained model. Thus, the information learned from pre-trained model is used to aid the fine-tuning layers during the learning process.

The dropout layer sets the output of each hidden neuron to zero with a probability of 0.5. Dropout [8] is a technique that randomly removes neurons during training that creates slightly

different networks for each iteration of training. Hence, weights of the network are tuned based on optimization of multiple variations of the network. This allows the network to learn more robust features that are useful in conjunction with different random subsets of the other neurons. We also employ batch normalization which serves as a regularizer for the network [8]. This speeds up training and makes it less dependent on careful initialization of network parameters. It yields normalized activation maps by subtracting the mean divided by the standard deviation for each training batch.

Classification Tasks

The learning model is designed to perform three different prediction tasks using the RMSprop optimization function. The base configuration (single prediction model) is to determine the GCS severity group (mild, moderate, or severe) from a given MR image. The single prediction model relies only on the multiclass GCS module. It has a SoftMax activation layer with three neurons. Each neuron outputs the prediction probability of one GCS severity category. The neuron with the highest probability is selected as the predicted class. The model uses the categorical cross entropy (CCE) loss function defined in Equation 1.

$$\text{CCE} = -\frac{1}{N}\sum_{i=0}^{N}\ \sum_{j=0}^{J} y_j \,.\log\left(\hat{y}_j\right) + (1 - y_j)\,.\log(1 - \hat{y}_j) \quad (1)$$

where N denotes the number of samples and J, the number of classes. The actual probability that the input belongs to class j is given by y_j, while the estimated probability is $\hat{y}_j$.

To explore the model's ability to incorporate information that has been derived from the CT scan, we also train another model to jointly predict both the GCS severity category and a CT derived metric severity group. The remainder classification tasks are both joint prediction models: given an MR image, determine both the GCS severity group as well as the Marshall

score or both the GCS group and the Rotterdam score. Since the Marshall and Rotterdam are both CT derived metrics, the models have similar configurations which we denote as the GCS+CT metric classifier and utilize the same module, CT metric module (Figure 10). Due to the skewed distribution of the CT metric groups across the augmented data (Figure 9), we limit the prediction tasks as binary; either (Marshall: 1 vs. 2 or Rotterdam: 2 vs. 3). The joint prediction classifier utilizes the CT metric module along with the GCS module to compute the joint loss (Figure 10). Similar to the GCS module, the CT metric module also uses the SoftMax activation function and the CCE loss function. Since it is a binary classification, only 2 output neurons are used. The joint loss function is a summation of the CCE loss from both GCS and CT metric modules.

Model Evaluation

To evaluate the model performance, we utilize classification accuracy, sensitivity, specificity, and area under curve-receiver operating characteristics (AUC-ROC) metrics. Sensitivity and specificity measure the ability of a model to determine if a clinical condition is present or absent. A positive indicates the presence of the clinical condition while a negative implies absence of the condition. For a given sample, patients with the clinical condition that are correctly classified are known as true positives while false positives are patients without the condition incorrectly classified as having the condition. In contrast, true negatives are subjects correctly classified as not having the condition while false negatives denote subjects with the condition incorrectly classified as not having the condition. Sensitivity (also known as the true positive rate (TPR) or recall) is the ratio of the number of true positives to the total number of positives present in the data. Specificity (also known as the true negative rate (TNR)) is the ratio of the number of true negatives to the total number of negatives present in the data.

Mathematically, sensitivity is given as $\frac{TP}{P} = \frac{TP}{TP + FN}$ whiles specificity is given as $\frac{TN}{N} = \frac{TN}{TN + FP}$. The ROC curve is a graphical display of the relationship of sensitivity (y-axis) to the complement of specificity (x-axis). AUC is a measure of the overall performance as quantified by the average value of sensitivity for all possible values of specificity. Increasing AUC values imply better overall diagnostic performance of a model in predicting the severity group of each image.

Clinical Relevance

A set of outcome measures, selected by domain experts, are utilized to evaluate whether identified groups have clinical significance. We selected four TBI outcome measures Glasgow Outcome Scale-Extended (GOS-E), Brief Symptom Inventory 18 (BSI-18), Satisfaction with Life Scale (SWLS), Post Traumatic Stress Disorder (PTSD) Check List-Civilian (PCL-C). These evaluate functional and cognitive recovery levels to determine if the groups generated by MR image analysis have predictive value for clinical outcome. We briefly describe these measures, to provide a context for the statistical analysis.

The GOS-E is a global outcome measure that assesses the overall impact of TBI on the patient incorporating functional status, independence and role participation. It is an ordinal scale that ranges from 1 to 8: dead (1), vegetative state (2), lower severe disability (3), upper severe disability (4), lower moderate disability (5), upper moderate disability (6), lower good recovery (7), and upper good recovery (8). BSI-18 quantifies subject psychological health based on a brief self-report measure of psychological distress with three subscales (depression, anxiety, and somatization) and a global severity index. SWLS is used as a measure of the life satisfaction component of subjective well-being. Scores on the SWLS have been linked to measures of mental health and predictive future behaviors. It is a 7-point Likert style response scale, with the

scores ranging from 5-35 and a neutral point of 20. Scores from 5-9 indicate extreme dissatisfaction with life, and 31-35 indicate extreme satisfaction with life. PCL-C is a 17-item self-report measure composed of the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition symptoms of PTSD that quantifies a patient's psychological status. All four outcome measures were available at the 6-month and 12-month time points post injury.

EXPERIMENTAL RESULTS AND ANALYSIS

Experimental Setup

We constructed three different prediction models based on the imaging data: a single prediction model (GCS) and two joint prediction models (GCS + Marshall and GCS + Rotterdam). For the GCS single prediction model, a total of 474 samples (158 per GCS group) were used. For the joint prediction models (Table 3), the sample sizes were 317 (M1 - 116, M2 - 201), and 341 (R2 -184, R3 - 157) in the GCS + Marshall and GCS + Rotterdam models, respectively. The single prediction model was trained over 5000 epochs. The data set was split into training (75%) and testing (25%) subsets. For the joint prediction models, a stratified 4-fold cross validation with 1000 epochs per fold was used. All three models used a batch size of 12 and a learning rate of 0.001.

Table 3. Data distribution for joint prediction models.

	Marshall (317)		Rotterdam (341)	
	M1	M2	R2	R3
Mild	111	43	133	20
Moderate	0	101	0	101
Severe	5	57	51	36

Model Performance of ResNet on MR Images

Figure 11 shows the training performance of the GCS single prediction model. The model achieved a 90.08% training accuracy. The AUC-ROC analysis (Figure 12) reveals that the model performs well in classifying images into GCS categories (AUC > 0.94). Table 4 shows the specificity and sensitivity values when each GCS severity category is considered as a condition of interest independently. The mild group has the highest specificity (96.5%) and lowest sensitivity

(77.22%). The severe group achieved a sensitivity of 100% indicating the model accurately predicted all its images (no misses).

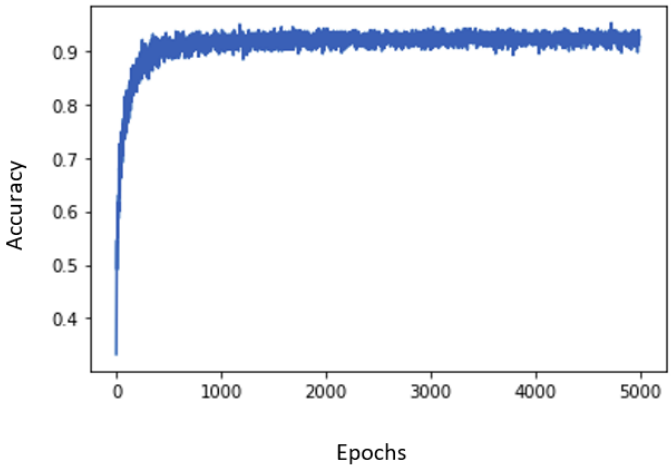


Figure 11. Training performance of the GCS severity prediction model.

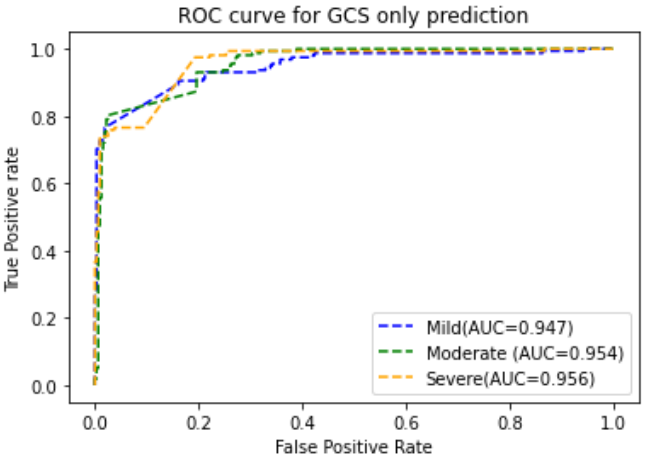


Figure 12. AUC-ROC performance of the GCS severity prediction model.

Table 4. Model performance to predict GCS based on MR images.

Severity Group	True Positive	False Negative	Sensitivity TPR (%)	Specificity TNR (%)
Mild (158)	122	36	77.22	96.5
Moderate (158)	147	11	93.04	88.61
Severe (158)	158	0	100	85.12

Table 5 shows the performance of the GCS + Marshall joint prediction model. The model achieved a classification accuracy of 100% for the M1 group and 92% for M2. The sensitivity for the M1-mild group is perfect but the model is unable to identify any of the M1-severe groups by the MR images. The sensitivities for the M2-mild and M2-moderate group are perfect though the specificities are not. The model is unable to identify any of the M2-severe groups.

Table 5. Joint prediction model performance (GCS + Marshall) using classification accuracy, sensitivity (TPR) & specificity (TNR).

	Marshall	M1 (116)		M2 (201)		
Ground Truth	GCS	mild (111)	severe (5)	mild (43)	Moderate (101)	Severe (57)
Predicted	Marshall	CA: 100%		CA: 92%		
		TPR/mild: 100%	TPR/severe: 0%	TPR/mild: 100%	TPR/moderate: 100%	TPR/severe: 0%
	GCS	TNR/mild: 0%	TNR/severe: 100%	TNR/mild: 64.0%	TNR/moderate: 43.0%	TNR/severe: 100%

CA: Classification Accuracy; TPR: True positive rate (Sensitivity); TNR: True negative rate (Specificity). Model uses MR image data to jointly predict the GCS and the Marshall score. Sensitivity and specificity are computed for each GCS severity group individually by considering each as the condition of interest for each M1 and M2 Marshall groups.

The outcome of the GCS + Rotterdam joint prediction model is shown in Table 6. The results indicate that the model performed better at classifying the images in group R2 (100%) than those in group R3 (85%). Similar to the GCS + Marshall model, the sensitivities for the R2-

mild, R3-mild, and R3-moderate groups are perfect. However, the model is unable to identify any of the severe groups as well. The specificity value for the R3-mild group is relatively high (74%).

Table 6. Joint prediction model performance (GCS + Rotterdam) using classification accuracy, Sensitivity (TPR) & Specificity (TNR).

	Rotterdam	R2 (184)		R3(157)		
Ground Truth	GCS	mild (133)	severe (51)	mild (20)	moderate (101)	severe (36)
Predicted	Rotterdam	CA: 100%		CA: 92%		
	GCS	TPR/mild: 100% TNR/mild: 0%	TPR/severe: 0% TNR/severe: 100%	TPR/mild: 100% TNR/mild: 74.0%	TPR/moderate: 100% TNR/moderate: 36.0%	TPR/severe: 0% TNR/severe: 100%

CA: Classification Accuracy; TPR: True positive rate (Sensitivity); TNR: True negative rate (Specificity). Model uses MR image data to jointly predict the GCS and the Rotterdam score. Sensitivity and specificity are computed for each GCS severity group individually by considering each as the condition of interest for each M1 and M2 Rotterdam groups.

Model Comparison using Plain CNN VGG-16

To evaluate the improvement of our proposed ResNet model, we compare to a plain CNN network model that has not been pre-trained. Figure 13 shows the basic VGG-16 architecture utilized. Using the MR images as the data input, the VGG-16 model is trained to predict the GCS severity group, the Marshall score and the Rotterdam score just as in our proposed model.

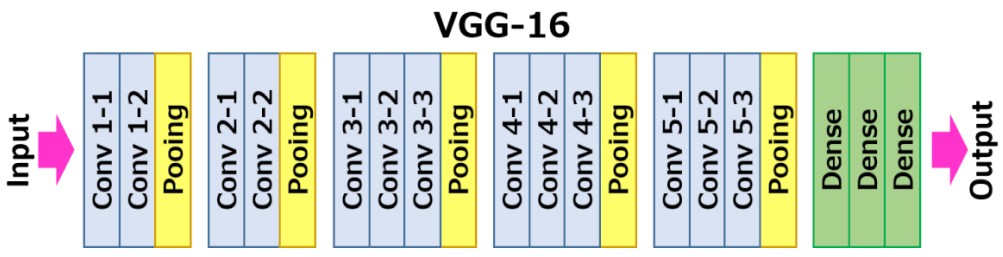


Figure 13. The VGG-16 architecture. Figure is adapted from [34].

Figure 14 shows the training curve for the GCS single prediction model based on the

VGG-16 plain and non-pretrained CNN. The final model prediction on the data set resulted in the assignment of moderate GCS class to all the images. This indicates that the VGG-16 model could not learn to distinguish between GCS severity groups from the MR image data. The joint prediction models, like the single prediction could not also distinguish between the classification groups for the images used. Both the GCS + Rotterdam and GCS + Marshall joint prediction models assigned the mild GCS severity group and group 2 (for both Rotterdam and Marshall) to all images, indicating an extremely poor performance.

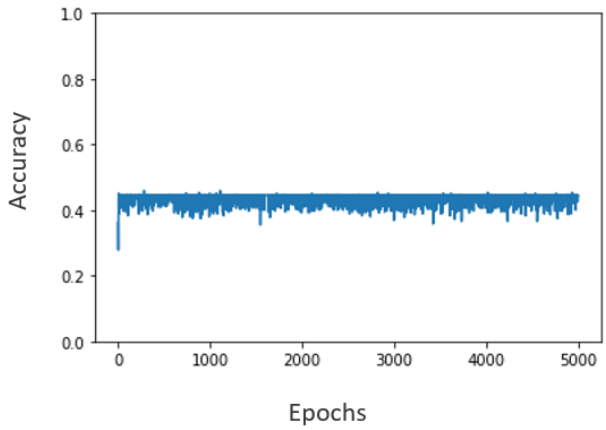
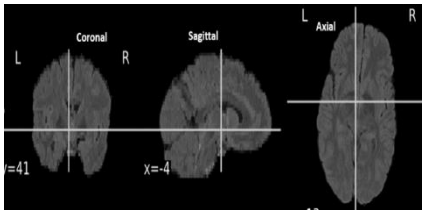


Figure 14. Training accuracy for GCS only prediction for VGG-16 model.

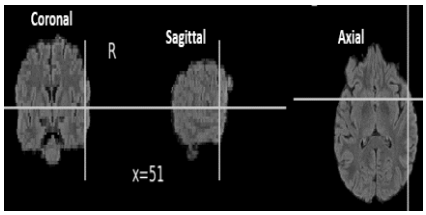
Qualitative Analysis of ResNet Based Model

In order to better understand the correlation between the model prediction and image information, we perform qualitative analysis via visual inspection with guidance from a domain expert. In a FLAIR image, the cerebrospinal fluid is inverted to black and any brain abnormality appears white. Hence, patients with less TBI severity are expected to have less white on their FLAIR images while the more severe patients should have more noticeable white. The domain expert visually inspected some images of patients within the GCS mild group that were accurately predicted as mild by the single prediction model. As shown in

Figure 15, there is little white in these images. Images from the mild GCS group that were classified as severe based on the MR images were also visually inspected. The white areas on the FLAIR scans (indicated by the red circles in Figure 16) may have led to the inconsistency between the actual mild GCS classification and the predicted severe group from the MR images.

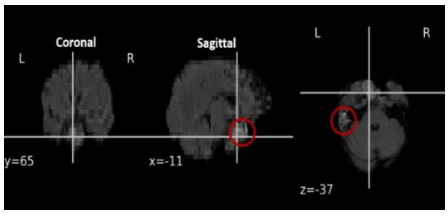


(a)

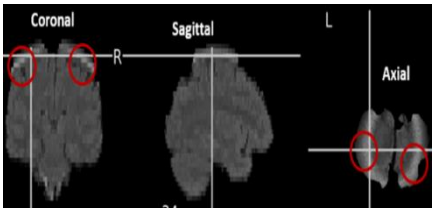


(b)

Figure 15. GCS mild cases classified as mild by model for two subjects.



(a)



(b)

Figure 16. GCS mild cases classified as severe by model for two subjects. Red circles highlight areas of artifact that may have misled classifier.

Clinical Relevance

From the prediction model results, we identified eight pairs of possible meaningful groups of interest for further analysis. Two pairs in the GCS model: (1) the mild correctly

classified (mild-CC) compared to the mild incorrectly classified as severe (mild-IC-sev) group, and (2) the moderate correctly classified (mod-CC) compared to the moderate incorrectly classified as severe (mod-IC- sev) group. Using similar notation for the joint prediction models (CC=correctly classified, IC=incorrectly classified), there are three comparisons of interest for the GCS + Marshall model: (1) M2-CC vs. M2-IC-M1, (2) M2-mod-CC vs. M2-sev-IC-mod, and (3) M1-mild-CC vs. M1-sev-IC-mild. There are also three comparisons of interest for the GCS + Rotterdam joint prediction model: (1) R3-CC vs. R3-IC-R2, (2) R3-mod-CC vs. R3-sev-IC-mod, and (3) R2-mild-CC vs. R2-sev-IC-mild.

Table 7 shows the mixed effects ANOVA results for the comparisons for BSI-18 at 6 and 12 months. For BSI-18, higher values of the measure indicate higher severity. The mean, standard deviation and percentage of not reported data are indicated. The time effect (◊) is significant in almost all of the comparisons of interest, indicating that this metric exhibit changes over time. The interaction effect (*) and/or mean differences in the identified subgroups (†) are significant in 5 of the 8 comparisons, suggesting that the differences in this outcome measure are explained by the predicted groups.

The mixed effects ANOVA results for the comparisons for SWLS at 6 and 12 months is shown in Table 8. Lower values indicate higher severity for SWLS. The mean, standard deviation and percentage of not reported data are indicated. In almost all of the comparisons of interest, the time effect (◊) is significant indicating that SWLS exhibit changes over time. The interaction effect (*) and/or mean differences in the identified subgroups (†) are significant in 5 of the 8 comparisons, suggesting that the groups predicted by the MR imaging models are important in explaining the difference in this outcome measure.

Table 9 shows the mixed effects ANOVA results for the comparisons for PCL-C at 6 and 12 months. Higher values of the measure indicate higher severity. The mean, standard deviation

and percentage of not reported data are indicated. The time effect (◊) is significant in 3 of the 8 comparisons. The results suggest that the groups predicted by the MR imaging models are important in explaining the difference in PCL-C since the interaction effect (*) and/or mean differences in the identified subgroups (†) are significant in 5 of the 8 comparisons.

Table 10 shows the mixed effects ANOVA results for the comparisons for GOS-E at 6 and 12 months. For GOS-E, lower values of the measure indicate higher severity. Due to the ordinal nature of GOS-E, the median values are reported. It is noticed that the interaction effect (*) and/or mean differences in the identified subgroups (†) are significant in 5 of the 8 comparisons suggesting that the predicted groups help explain the differences in this measure. Further analysis is required for comparisons where the interaction is significant.

Table 7. Statistical analysis of BSI-18 for identified subgroups of interest from single and joint prediction results.

	Significance	Brief Symptom Inventory -18			
		6 months		12 months	
		Mean (SD)	%NR	Mean (SD)	%NR
GCS Single Prediction					
Mild-CC (122) vs.		56.1 (11.8)	26.2	52.3 (10.9)	39.3
Mild-IC-Sev (36)	◊, *	55.4 (11.4)	8.3	50.0 (10.5)	41.7
Mod-CC (147) vs.		53.1 (11.3)	31.9	45.0 (7.0)	46.9
Mod-IC-Sev (11)	◊	54.0 (0)	0	49.0 (0)	0
GCS + Marshall Joint Prediction					
M2-CC (185) vs.		52.1 (8.0)	35.7	48.5 (9.1)	43.8
M2-IC-M1 (16)	◊	59.2 (8.4)	43.8	54.0 (10.2)	43.8
M1-Mild-CC (111) vs.		56.2 (11.9)	19.8	52.0 (10.5)	45
M1-Sev-IC-Mild (5)	◊	59.0 (0)	0	60.0 (0)	0
M2-Mod-CC (101) vs.		49.3 (5.5)	34.7	45.5 (8.1)	44.6
M2-Sev-IC-Mod (52)	◊, *	56.1 (7.1)	40.4	51.8 (6.2)	50
GCS + Rotterdam Joint Prediction					
R2-Mild-CC (133) vs.		55.3 (11.9)	19.5	51.3 (11.0)	40.6
R2-sev-IC-Mild (51)	◊, *, †	52.1 (10.7)	31.4	47.7 (8.4)	29.4
R3-Mod-CC (101) vs.		49.3 (5.5)	34.7	45.5 (8.1)	44.6
R3-Sev-IC-Mod (28)	◊, *	49.2 (6.9)	35.7	51.4 (6.1)	35.7
R3-CC (133) vs.		49.3 (5.9)	34.6	47.0 (7.9)	42.9
R3-IC-R2 (24)	◊, †	60.4 (8.5)	45.8	56.0 (9.3)	45.8

NR: Not Reported Data; SD: Standard Deviation.

Table 8. Statistical analysis of SWLS for identified subgroups of interest from single and joint prediction results.

	Significance	Satisfaction With Life Scale			
		6 months		12 months	
		Mean (SD)	%NR	Mean (SD)	%NR
GCS Single Prediction					
Mild-CC (122) vs.		20.7 (7.9)	73	22.5 (7.4)	62.3
Mild-IC-Sev (36)	◊	21.6 (8.5)	91.7	21.5 (8.9)	61.1
Mod-CC (147) vs.		25.0 (6.7)	32	24.4 (9.1)	54.4
Mod-IC-Sev (11)	◊, *, †	20.0 (0)	0	12.0 (0)	0
GCS + Marshall Joint Prediction					
M2-CC (185) vs.		23.7 (6.8)	0	22.0(7.8)	0
M2-IC-M1 (16)	◊	20.4 (9.2)	0	21.9 (7.7)	0
M1-Mild-CC (111) vs.		20.5 (8.0)	0	22.0 (8.1)	0
M1-Sev-IC-Mild (5)	◊, †	10.0 (0)	0	10.0 (0)	0
M2-Mod-CC (101) vs.		23.5 (7.4)	34.7	20.1(8.3)	55.4
M2-Sev-IC-Mod (52)	◊, *, †	24.8 (4.1)	40.4	24.7(6.2)	50
GCS + Rotterdam Joint Prediction					
R2-Mild-CC (133) vs.		20.9 (8.2)	79.7	22.1(7.9)	61.7
R2-sev-IC-Mild (51)	◊, *, †	24.9 (7.2)	78.4	25.3 (7.4)	70.6
R3-Mod-CC (101) vs.		23.5 (7.4)	34.7	20.1 (8.3)	55.4
R3-Sev-IC-Mod (28)	◊, *, †	22.3 (7.1)	35.7	22.7 (7.1)	35.7
R3-CC (133) vs. R3-IC-R2 (24)	Not Significant				

NR: Not Reported Data; SD: Standard Deviation.

Table 9. Statistical analysis of PCL-C for identified subgroups of interest from single and joint prediction results.

	Post-traumatic Stress Disorder Checklist – Civilian				
	Significance	6 months		12 months	
		Mean (SD)	%NR	Mean (SD)	%NR
GCS Single Prediction					
Mild-CC (122) vs.		34.8 (16.0)	26.2	29.5 (12.9)	37.7
Mild-IC-Sev (36)	◊	35.0 (14.6)	8.3	29.2 (11.3)	38.9
Mod-CC (147) vs.		26.9 (11.1)	39.5	20.3 (2.3)	61.9
Mod-IC-Sev (11)	†, *	21.0 (0)	0	25.0 (0)	0
GCS + Marshall Joint Prediction					
M2-CC (185) vs. M2-IC-M1 (16)		Not Significant			
M1-Mild-CC (111) vs.		35.8 (16.0)	19.8	29.8 (13.0)	43.2
M1-Sev-IC-Mild (5)	†, *	29.0 (8.4)	0	22.3 (7.0)	0
M2-Mod-CC (101) vs.		23.0 (2.7)	45.5	22.0 (0.8)	66.3
M2-Sev-IC-Mod (52)	◊, *, †	29.4 (8.9)	40.4	23.0 (4.5)	59.6
GCS + Rotterdam Joint Prediction					
R2-Mild-CC (133) vs.		34.7 (15.7)	19.5	29.2 (12.5)	39.1
R2-sev-IC-Mild (51)	◊, *, †	29.0 (8.4)	21.6	22.3 (7.0)	39.2
R3-Mod-CC (101) vs. R3-Sev-IC-Mod (28)		Not Significant			
R3-CC (133) vs.		23.5 (4.0)	42.9	22.2 (1.2)	63.2
R3-IC-R2 (24)	†, *	37.9 (14.7)	45.8	32.1 (13.2)	41.7

NR: Not Reported Data; SD: Standard Deviation.

Table 10. Statistical analysis of GOS-E for identified subgroups of interest from single and joint prediction results.

	Significance	Glasgow Outcome Scale - Extended			
		6 months		12 months	
		Median	%NR	Median	%NR
GCS Single Prediction					
Mild-CC (122) vs. Mild-IC-Sev (36)		Not Significant			
Mod-CC (147) vs.		7	31.97	7	46.9
Mod-IC-Sev (11)	◊, *	6	0	5	0
GCS + Marshall Joint Prediction					
M2-CC (185) vs.		7.0	25.4	7.0	36.2
M2-IC-M1 (16)	†	6.0	12.5	6.0	6.3
M1-Mild-CC (111) vs.		7.0	18	7.0	42.3
M1-Sev-IC-Mild (5)	*	8.0	0	7.0	0
M2-Mod-CC (101) vs.		7.0	34.7	7.0	44.6
M2-Sev-IC-Mod (52)	*	5.0	11.5	7.0	28.9
GCS + Rotterdam Joint Prediction					
R2-Mild-CC (133) vs. R2-sev-IC-Mild (51)		Not Significant			
R3-Mod-CC (101) vs.		7.0	34.7	7.0	44.6
R3-Sev-IC-Mod (28)	*	5.0	0	6.0	17.9
R3-CC (133) vs.		7.0	26.3	7.0	38.4
R3-IC-R2 (24)	◊	6.0	20.8	7.0	20.8

NR: Not Reported Data; SD: Standard Deviation.

MODEL INTERPRETATION

This work investigates a residual learning model using MR images to perform two main tasks: (1) classify TBI subjects according level of GCS severity; (2) jointly predict GCS and CT scan severity score (either Rotterdam or Marshall score). The model performed well on the first task to predict GCS severity level from MRI brain images. Both AUC-ROC and specificity was excellent for mild, moderate, and severe TBI patients. Sensitivity was excellent for both moderate and severe TBI. However, due to a large number of false negatives in the mild TBI group, sensitivity was lower in this group. Manual visual inspection of the misclassified images from the mild TBI group suggested that the model may have interpreted MRI artifacts on the images as brain abnormalities and erroneously assigned these images to a high level of TBI severity. This problem could possibly have been remediated if we had available a larger image set that would have allowed better training to recognize the artifacts.

On the second task to jointly predict the GCS and the CT score (either Rotterdam or Marshall), the prediction of the CT derived metric was reduced to a binary task (either M1 or M2 on the Marshall score or R2 or R3 on the Rotterdam score). The model showed a high classification accuracy in predicting both the Marshall score and the Rotterdam score. The model still displayed a high sensitivity (TPR) for mild TBI but a degraded sensitivity (TPR) for severe TBI. The model's inability to accurately classify severe TBI subjects on the joint prediction task and is puzzling since it accurately classified severe TBI subjects on the single prediction task. This could be also due to the inconsistency with GCS severe class being associated with CT derived metrics of relatively lower severity. Future work will be focused on improving the joint prediction learning tasks.

We also examined whether output from the single prediction model or either of the two

joint prediction models had predictive value beyond that already found in the GCS or two CT scores (Marshall and Rotterdam). Outcome measures were available at 6 and 12 months. Several measures showed a strong time effect with better scores at 12 months than 6 months consistent with improvement in most subjects over time. To evaluate whether there is latent predictive information in the predictive models based on the MR image data, we performed specific comparisons between groups in which the predictive model agreed with the ground truth assignments (GCS, Rotterdam score, or Marshall score) and groups in which the predictive model disagreed with the ground truth assignments. There were numerous instances across the six outcome measures, in which the groups with consistent classification (agreement between model and ground truth) differed in outcome from groups with a disagreement between model and ground truth. This suggests that the model may be able to detect predictive information that is not in the ground truth labels, but more investigation is needed to reveal the magnitude and direction of this latent predictive information.

CONCLUSION

In this work, we present a residual deep learning model for sorting out the heterogeneity of TBI based on patients' MR images. We utilized the ResNet-50 architecture on top of which we developed fine tuning layers. The model integrated the concept of transfer learning from pre-trained weights on general non-medical data set for an improved performance and a more robust system. Data extraction and cleaning were incorporated to present the raw data in a reliable format. Image processing steps including inhomogeneity correction and skull stripping were applied to the data set remove noise from the images. Data augmentation by rotation of images was applied to expand the insufficient images available. Our framework includes both a single and joint prediction models. The single prediction model predicts the severity of TBI based on GCS. The joint prediction model extends the single predictor model by incorporating CT derived metrics (Marshall and Rotterdam scores). By comparing our model to the plain non-pretrained VGG-16 CNN architecture to evaluate its effectiveness, the experimental results indicated that our proposed model's performance surpassed that of VGG-16. Our framework also includes a mixed ANOVA statistical analysis for identifying significant differences between relevant subgroups identified from the results to understand the correlation between data modalities and outcome measures. This framework has the potential to aid clinical experts in making decisions about key MRI features and the connections with patient prognosis and recovery outcomes.

FUTURE WORK

Some limitations of this study should be mentioned. First, the image data set was relatively small though we used data augmentation to partially address this issue. A larger data set would likely have resulted in more accurate predictions as well as enable the model to better discriminate between brain MRI artifacts and brain abnormalities. The model's accuracy could also be improved by extending training to other sequences (T1, T2, diffusion weighted, among others). Nevertheless, deep CNNs show promise for the interpretation of MR images to predict severity and outcome from TBI. More investigation is needed to determine whether deep learning models can uncover latent predictive information for outcome from TBI not already encapsulated in traditional measures such as the GCS, Marshall score, and Rotterdam score.

REFERENCES

[1] He, Kaiming;, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[2] Y. B, W. Y, S. D and Z. L, "Medical image synthesis via deep learning," *Deep Learning in Medical Image Analysis*, pp. 23-44, 2020.

[3] S. Serte, S. Ali and A. Fadi, "Deep learning in medical imaging: A brief review," *Transactions on Emerging Telecommunications Technologies*, p. e4080, 2020.

[4] S. S. S. R. Basheera, "Convolution neural network–based Alzheimer's disease classification using hybrid enhanced independent component analysis based segmented gray matter of T2 weighted magnetic resonance imaging with clinical valuation.," *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, vol. 5, pp. 974-986, 2019.

[5] D. Pan, "Early Detection of Alzheimer's Disease Using Magnetic Resonance Imaging: A Novel Approach Combining Convolutional Neural Networks and Ensemble Learning," *Frontiers in neuroscience*, vol. 14, 2020.

[6] K. He, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[7] L. Torrey and S. Jude, "Transfer learning," *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pp. 242-264, 2010.

[8] L. A. Selvikvåg and L. Arvid, "An overview of deep learning in medical imaging focusing on MRI," *arXiv e-prints*, 2018.

[9] Y. Dacosta, "An Explainable and Statistically Validated Ensemble Clustering Model Applied to the Identification of Traumatic Brain Injury Subgroups," *IEEE Access 8*, pp. 180690-180705, 2020.

[10] J. Yue, "Transforming research and clinical knowledge in traumatic brain injury pilot: multicenter implementation of the common data elements for traumatic brain injury," *Journal of neurotrauma*, vol. 30.22, pp. 1831-1844, 2013.

[11] A. I. Maas, "Prediction of outcome in traumatic brain injury with computed tomographic characteristics: a comparison between the computed tomographic classification and combinations of computed tomographic predictors," *Neurosurgery*, vol. 57.6, pp. 1173-1182, 2005.

[12] A. Deepika, "Comparison of predictability of Marshall and Rotterdam CT scan scoring system in determining early mortality after traumatic brain injury," *Acta neurochirurgica*, vol. 157.11, pp. 2033-2038, 2015.

[13] E. Wilde, "Recommendations for the use of common outcome measures in traumatic brain injury research," *Archives of physical medicine and rehabilitation*, vol. 91.11, pp. 1650-1660, 2010.

[14] S. Albawi, A. M. Tareq and A.-Z. Saad , "Understanding of a convolutional neural network," in *International Conference on Engineering and Technology (ICET). Ieee*, 2017.

[15] K. O'Shea and N. Ryan, "An introduction to convolutional neural networks.," *arXiv preprint*, vol. arXiv:1511.08458, 2015.

[16] C. Szegedy, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.

[17] F. Altaf, "Going deep in medical image analysis: concepts, methods, challenges, and future directions," *IEEE Access 7*, pp. 99540-99572, 2019.

[18] G. Huang, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[19] M. Peng, "Dual temporal scale convolutional neural network for micro-expression recognition," *Frontiers in psychology*, vol. 8, p. 1745, 2017.

[20] S. J. Pan and Y. Qiang , "A survey on transfer learning.," *IEEE Transactions on knowledge and data engineering*, vol. 22.10, pp. 1345-1359, 2009.

[21] A. Mikołajczyk and G. Michał, "Data augmentation for improving deep learning in image classification problem," *international interdisciplinary PhD workshop (IIPhDW). IEEE, 2018*, 2018.

[22] A. Takimoglu, "What is Data Augmentation? Techniques, Benefit and Examples," 2021. [Online]. Available: https://research.aimultiple.com/data-augmentation/.

[23] E. Goceri and G. Numan, "Deep learning in medical image analysis: recent advances and future trends.".

[24] F. Altaf, "Going deep in medical image analysis: concepts, methods, challenges, and future directions," *IEEE Access* , vol. 7, pp. 99540-99572, 2019.

[25] L. M, Z. J, A. E and S. D, "Deep multi-task multi-channel learning for joint classification and regression of brain status," in *International conference on medical image computing and computer-assisted intervention*, 2017.

[26] R. K. Srivastava, G. Klaus and S. Jürgen, "Highway networks," in *arXiv preprint arXiv:1505.00387*, 2015.

[27] K. He, "Deep residual learning for image recognition," in *IEEE conference on computer*

vision and pattern recognition, 2016.

[28] A. Ebrahimi, L. Suhuai and C. Raymond, "Introducing Transfer Leaming to 3D ResNet-18 for Alzheimer's Disease Detection on MRI Images," in *International Conference on Image and Vision Computing New Zealand (IVCNZ) IEEE*, 2020 .

[29] . L. Liu, "Mtmr-net: Multi-task deep learning with margin ranking loss for lung nodule analysis," *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support,* pp. 74-82, 2018.

[30] [Online]. Available: from https://www.cs.toronto.edu/~kriz/cifar.html.

[31] Y. K. J, J. V. M., F. L. H, R. C. S, D. O. Okonkwo, A. B. Valadka, W. A. Gordon, A. I. Maas, P. Mukherjee and E. L. Yuh, "Transforming research and clinical knowledge in traumatic brain," *Journal of neurotrauma,* vol. 30, p. 1831–1844, 2013.

[32] Z. Akkus, "Deep learning for brain MRI segmentation: state of the art and future directions.," *Journal of digital imaging,* vol. 30.4, pp. 449-459, 2017.

[33] J. Muschelli, "fslr: Connecting the FSL Software with R," *The R journal ,* vol. 7.1, p. 163, 2015.

[34] [Online]. Available: https://neurohive.io/en/popular-networks/vgg16/.

[35] P. Kalavathi and S. P. VB, "Methods on skull stripping of MRI head scan images—a review.," *Journal of digital imaging 29.3,* vol. 29.3, pp. 365-37, 2016.

[36] K. Noguchi, "MRI of acute cerebral infarction: a comparison of FLAIR and T2-weighted fast spin-echo imaging," *Neuroradiology,* vol. 39.6, pp. 406-410, 1997.

[37] J. N. Giedd and R. L. Judith , "Structural MRI of pediatric brain development: what have we learned and where are we going?," *Neuron,* vol. 67.5, pp. 728-734, 2010.

[38] T. G. B. Mahadewa, "Modified Revised Trauma–Marshall score as a proposed tool in predicting the outcome of moderate and severe traumatic brain injury.," *Open access emergency medicine,* vol. 10, p. 135, 2018.

[39] M. Mohammadifard, "Marshall and Rotterdam Computed Tomography scores in predicting early deaths after brain trauma," *European journal of translational myology,* vol. 28.3, 2018.

[40] Y. Fujikoshi, V. V. Ulyanov and R. Shimizu, "Multivariate statistics: High-dimensional and large-sample approximations," *John Wiley & Sons,* vol. 760, 2011.

[41] K. Liesemer, "Use of Rotterdam CT scores for mortality risk stratification in children with traumatic brain injury," *Pediatric critical care medicine: a journal of the Society of Critical Care Medicine and the World Federation of Pediatric Intensive and Critical Care Societies,* vol. 15.6, p. 554, 2014.

[42] M. S. L. Goh, "The Impact of Traumatic Brain Injury on Neurocognitive Outcomes in Children: a Systematic Review and Meta-Analysis," *ournal of Neurology, Neurosurgery & Psychiatry,* 2021.

[43] . X. S. Zhang, "Metapred: Meta-learning for clinical risk prediction with limited patient electronic health records," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019.

[44] J.-G. Lee, "Deep learning in medical imaging: general overview," *Korean journal of radiology,* vol. 18.4, 2017.

[45] M. A. Mazurowski, "Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI," *Journal of magnetic resonance imaging*, pp. 939-954, 2019.

[46] L. Torrey and S. Jude , "Transfer learning," *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques. IGI global*, pp. 242-264., 2010.

[47] . M. I. Razzak, N. Saeeda and Z. Ahmad , "Deep learning for medical image processing: Overview, challenges and the future," *Classification in BioApps,* pp. 323-350, 2018.

[48] F. Altaf, "Going deep in medical image analysis: concepts, methods, challenges, and future directions," *IEEE Access,* Vols. 99540-99572, 2019.

[49] A. Ebrahimi, L. Suhuai and C. Raymond , "Introducing Transfer Leaming to 3D ResNet-18 for Alzheimer's Disease Detection on MRI Images," *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ). IEEE,* 2020.

[50] B. A. Jónsson, "Brain age prediction using deep learning uncovers associated sequence variants," *Nature communications*, vol. 10.1 , pp. 1-10, 2019.

[51] M. Schönberger, "The Relationship between age, injury severity, and MRI findings after traumatic brain injury.," *Journal of neurotrauma,* vol. 26.12, pp. 2157-2167, 2009.

[52] H. Sajedi and P. Nastaran, "Age prediction based on brain MRI image: a survey," *Journal of medical systems,* vol. 43.8, p. 279, 2019.

[53] M. Liu, "Deep multi-task multi-channel learning for joint classification and regression of brain status," *International conference on medical image computing and computer-assisted intervention,* 2017.

[54] J. Springenberg, "Striving for simplicity: The all convolutional net.," *arXiv preprint arXiv*, vol. 1412.6806, 2014.

[55] R. K. Srivastava, G. Klaus and S. Jürgen , "Training very deep networks.," *arXiv preprint arXiv*, vol. 1507.06228, 2015.

[56] M. Pak and K. Sanghoon, "A review of deep learning in image recognition," in *2017 4th international conference on computer applications and information processing technology (CAIPT)*, 2017.

[57] L. Liu, "Multi-task deep model with margin ranking loss for lung nodule analysis," *IEEE transactions on medical imaging ,* vol. 39.3, pp. 718-728, 2019.

[58] P. Hridayami, G. D. P. Ketut and S. W. Kadek , "Fish species recognition using VGG16 deep convolutional neural network," *Journal of Computing Science and Engineering*, vol. 13.3, pp. 124-130, 2019.

[59] J. T. Springenberg, "Striving for simplicity: The all convolutional net," in *arXiv preprint arXiv:1412.6806*, 2014.