
01 Jan 2021

A Deep Learning Model to Predict Traumatic Brain Injury Severity and Outcome from MR Images

Dacosta Yeboah

Hung Nguyen

Daniel B. Hier


Missouri University of Science and Technology, hierd@mst.edu

Gayla R. Olbricht

Missouri University of Science and Technology, olbrichtg@mst.edu

et. al. For a complete list of authors, see https://scholarsmine.mst.edu/chem_facwork/3237

Follow this and additional works at: https://scholarsmine.mst.edu/chem_facwork

 Part of the [Chemistry Commons](#), [Mathematics Commons](#), and the [Statistics and Probability Commons](#)

Recommended Citation

D. Yeboah et al., "A Deep Learning Model to Predict Traumatic Brain Injury Severity and Outcome from MR Images," *2021 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2021*, Institute of Electrical and Electronics Engineers, Jan 2021.

The definitive version is available at <https://doi.org/10.1109/CIBCB49929.2021.9562848>

This Article - Conference proceedings is brought to you for free and open access by Scholars' Mine. It has been accepted for inclusion in Chemistry Faculty Research & Creative Works by an authorized administrator of Scholars' Mine. This work is protected by U. S. Copyright Law. Unauthorized use including reproduction for redistribution requires the permission of the copyright holder. For more information, please contact scholarsmine@mst.edu.

A deep learning model to predict traumatic brain injury severity and outcome from MR images

Dacosta Yeboah
Computer Science Dept.
Missouri State University
Springfield, MO USA
dacosta123@live.missouristate.edu

Hung Nguyen
Computer Science Dept.
Missouri State University
Springfield, MO USA
hung249@live.missouristate.edu

Daniel B. Hier
Electrical & Computer Eng. Dept.
Missouri University of Science & Technology
Rolla, MO, USA
hierd@mst.edu

Gayla R. Olbricht
Mathematics and Statistics Dept.
Missouri University of Science & Technology
Rolla, MO USA
olbrichtg@mst.edu

Tayo Obafemi-Ajayi
Engineering Program
Missouri State University
Springfield, MO USA
tayoobafemijayi@missouristate.edu

Abstract—For many neurological disorders, including traumatic brain injury (TBI), neuroimaging information plays a crucial role determining diagnosis and prognosis. TBI is a heterogeneous disorder that can result in lasting physical, emotional and cognitive impairments. Magnetic Resonance Imaging (MRI) is a non-invasive technique that uses radio waves to reveal fine details of brain anatomy and pathology. Although MRIs are interpreted by radiologists, advances are being made in the use of deep learning for MRI interpretation. This work evaluates a deep learning model based on a residual learning convolutional neural network that predicts TBI severity from MR images. The model achieved a high sensitivity and specificity on the test sample of subjects with varying levels of TBI severity. Six outcome measures were available on TBI subjects at 6 and 12 months. Group comparisons of outcomes between subjects correctly classified by the model with subjects misclassified suggested that the neural network may be able to identify latent predictive information from the MR images not incorporated in the ground truth labels. The residual learning model shows promise in the classification of MR images from subjects with TBI.

Index Terms—Traumatic Brain Injury, MRI, Deep learning, Medical Imaging, Transfer learning

I. INTRODUCTION

There has been recent progress in the use of deep convolutional neural networks (CNN) [1] for analysis of medical images [2]–[4]. A variety of imaging modalities including magnetic resonance imaging (MRI), computed tomography (CT), plain X-rays, and ultrasound are used for disease diagnosis and prognosis [5]. Healthcare data sets are complex, context dependent, encompass many modalities, and typically heterogeneous with regards to both classes and features. Class imbalances or limited data set size makes the extraction of knowledge by machine learning challenging [6]. Nonetheless, deep learning models based on medical images have shown promise in identifying embedded patterns that have diagnostic and prognostic value [6].

Traumatic brain injury (TBI) is a disruption of brain function caused by a blow to the head [7]. It is heterogeneous

in cause, severity, pathology, and prognosis [8]. It is a major cause of death and disability, accounting for over 2.8 million emergency department (ED) visits in the United States and a negative economic impact of \$76.5 billion annually [9]. Long term effects of TBI include physical impairments (movement, vision, hearing), emotional changes (depression, personality changes), or cognitive impairments (memory loss, attention, linguistic, and others). The consequences of TBI can worsen without timely diagnosis and treatment [10].

The heterogeneity of TBI combined with the lack of precise outcome measures is a challenge [11]. Currently, the Glasgow Coma Scale (GCS) score is used to stratify TBI patients into three severity categories: severe (GCS 3-8), moderate (GCS 9-12), and mild (GCS 13-15) [11]. The GCS score is based on best response in three areas: eye-opening, motor, and verbal.

TBI severity can also be stratified by cranial CT abnormalities as CT scans are routinely obtained within the first 24 hours after brain injury [10], [12]. The Marshall and Rotterdam scores are CT derived metrics used to predict TBI outcome [13], [14]. The Marshall scoring system is based on morphological abnormalities on CT scans as defined by visible presentation of increasing evidence of mass effect (1 is best, 6 is worst). The Rotterdam scores are based on the sum of CT scan elements that could correlate with poor outcomes including cistern compression, midline shift, epidural mass lesion, and traumatic subarachnoid hemorrhage or intraventricular hemorrhage (1 is best, 6 is worst).

Neuroimaging plays a critical role in the evaluation of TBI patients. A CT of the head is first-line imaging in the ED [15]. For many brain disorders, including TBI, MRI is a powerful imaging technique for diagnosis and assessment [16]. It has a high spatial resolution and provides key information on the anatomical structure, often in greater detail than CT. [17]. Though CT and MRI are used in the TBI clinical setting, no single imaging modality has proven sufficient for all patients due to the heterogeneity of TBI presentation [18]. In particular

for mild TBI (mTBI), which constitutes about 75% of TBI [16], studies indicate that CT scanning may be of limited usefulness in clinical evaluation of patients presenting to the ED [18]. Though MRI is more sensitive to brain injury than CT scans, there remains concerns that MRI does not fully correlate with anticipated clinical outcomes [16], [18]. Further research is needed on predictive power of MRI in TBI.

Deep learning CNN models have shown promising results for the automated classification of disease severity in neurological disorders such as Alzheimer’s disease based on MRI images [19], [20]. Building on this success, in this work, we investigate a deep CNN model framework to predict the severity of TBI (as quantified by GCS score) from 3T MRI brain scans. Given that CT has demonstrated usefulness in the TBI clinical setting, we are also interested in exploring a more comprehensive data-driven predictor model for TBI. 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 deep CNN model framework utilizes a residual learning method implemented with the residual network (ResNet-50) architecture [21]. Residual learning does not allow error accumulation on the convolutional layers thus enabling a better representation of the content in these layers [5]. A drawback of deep CNN models is the long training times. Transfer learning improves learning of a new task by the transfer of knowledge from a previously learned but related task [22]. Models trained on one problem can be used as a starting point for training new models on a related problem. Transfer learning is flexible and allows the use of weights from pre-trained models developed from standard computer vision benchmark data into new models. Thus, we minimize the training time drawback by integrating transfer learning in our deep CNN model framework.

The overall objective is to evaluate the sensitivity of a CNN model to detect anatomical changes in a brain MRI scan that might correlate with outcome after TBI. The evaluation of the results is performed using both quantitative metrics and qualitative analysis (based on visual examination by domain expert - D.B.H). In addition, using a varied set of commonly acceptable TBI outcome measures [23], 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 are interested in the critical analysis of the 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.

II. METHODS

The learning framework, as illustrated in Fig. 1, consists of five phases (data curation, data augmentation, training of residual learning model, model validation, and assessment of clinical relevance).

TABLE I
CLINICAL AND DEMOGRAPHIC SUMMARY OF STUDY SUBJECTS (N=203)

Characteristic	Value
Gender	Male 70.94%; Female 29.06%
Age *	40.20 ± 15.8 [18, 79]
MRI days post injury*	13.54 ± 16.9 [0, 196]
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%
LOC	Yes: 71.9%, No: 18.7%, Unknown: 9.4%
Injury severity score > 15 (Yes/No)	Yes: 36.9%, No: 63.1%
Abbrev. injury score (Head or Neck)	Yes: 71.9%, No: 18.7%, Unknown: 9.4%

LOC: Loss of consciousness; *: Mean ± SD, Range

A. 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 [11] available via Federal Interagency Traumatic Brain Injury Research (FITBIR) [24] data repository to approved researchers. The TRACK-TBI study [11] is a multicenter observational pilot study aimed at validating the feasibility of implementing the TBI Common Data Elements that span demographics, clinical care, genetics and proteomic biomarkers, neuroimaging, and a battery of outcome measures. A subset of these subjects (252) 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). FLAIR, although lacking the spatial resolution of some other MRI sequences, is sensitive to brain pathology and facilitates the distinction between cerebrospinal fluid and areas of brain injury. It has high sensitivity to a wide range of central nervous system pathologies [25]. Since age could have an effect on the brain scans [26] or outcome measures, we focused on images from patients between 18 and 79 years old, reducing the available sample size for this study (n = 203). The clinical and demographics characteristics of the subjects are summarized in Table I.

Automated analysis of MR images is challenging due to intensity inhomogeneity, variability of the intensity ranges and contrast, and noise [27]. Thus, preprocessing steps unique to image data are essential prior to the learning model phase. The brain 3T MRI 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 [27]. Inhomogeneity correction mitigates image contrast variations

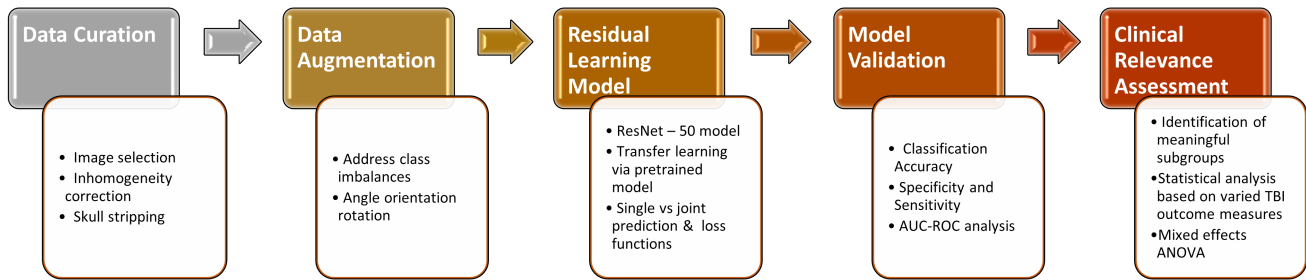
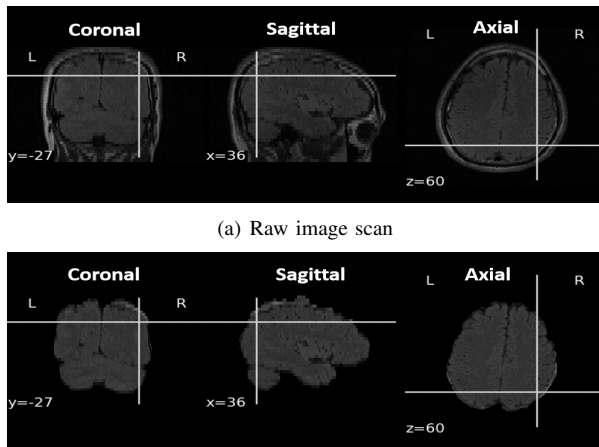


Fig. 1. Deep learning framework for TBI severity prediction from MRI FLAIR sequences.



(a) Raw image scan
(b) Processed image after inhomogeneity correction and skull stripping
Fig. 2. Image pre-processing of an MRI scan.

due to intensity inhomogeneity. We performed skull stripping and inhomogeneity correction on all the images using the R fsr package [28], as illustrated in Fig. 2.

B. Data Augmentation

The available data (Table I) is of limited size and is class imbalanced (skewed towards the mTBI group). Deep learning models generally yield more accurate results on large data sets. The problems of overfitting due to limited sample size or class imbalances can be mitigated by data augmentation [20], [29]. This involves enlarging the image data set using varied label preserving transformations [1]. This technique generates multiple different versions of the images from the original slices by rotation, translation, gamma correction, random noise addition, scaling, and random affine transformation [20]. We applied data augmentation to the images by rotating the volumes within a range of angles. Fig. 3 shows an example of an image that has been augmented twice. 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.

C. Residual Learning Model

The residual learning model, as illustrated in Fig. 4, utilizes a ResNet-50 [21] CNN architecture implemented in Keras with Tensorflow backend [30]. The ResNet architecture solves

the vanishing gradient problem found in plain deep CNNs by introducing skip connections that short circuit shallow layers to deep layers [21]. 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 model (Fig. 4) consists of 5 stages, each having a convolutional (made up of 3 stacked layers) and identity block. 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.

Our learning framework integrates transfer learning by adapting a well performing deep learning network (ResNet-50) trained on a large data set (ImageNet [31]) 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 data set is better than random weight initialization of the network [27]. 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 [3] 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

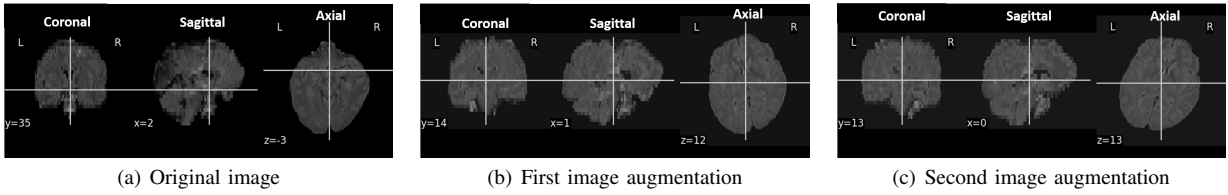


Fig. 3. Data augmentation utilizing rotation to generate two additional images

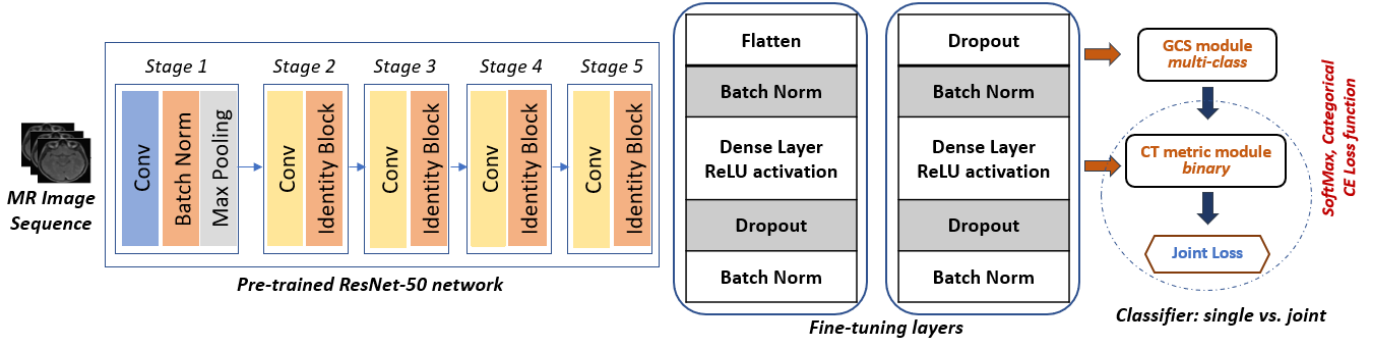


Fig. 4. Residual learning model architecture.

the network [3]. 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 classification/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 (Fig. 4). 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 as defined in Eq. 1.

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

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. Thus, 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

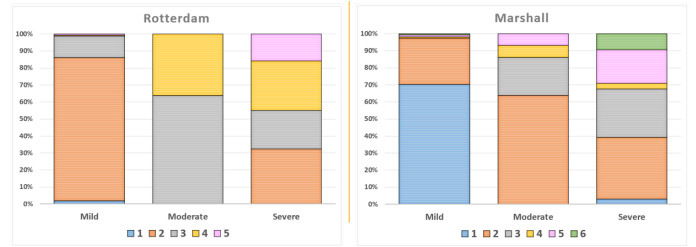


Fig. 5. Skewed distribution of CT metric groups across augmented data.

denote as the *GCS+CT metric* classifier and utilize the same module, *CT metric module* (Fig. 4). Due to the skewed distribution of the CT metric groups across the augmented data (Fig. 5), 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 (Fig. 4). 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.

D. 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* denotes 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.

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.

E. Clinical Relevance Assessment

A set of outcome measures, selected by domain experts, can be used to evaluate whether identified groups have clinical significance. We selected six TBI outcome measures [23] (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), California Verbal Learning Test-II (CVLT), and Wechsler Adult Intelligence Scale-III Processing Speed Index (PSI)) that 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 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. Increasing values indicate higher psychological stress. 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. CVLT is a neuropsychological impairment measure that assesses the patient's verbal learning and memory. It evaluates the recollection and recognition of two lists of words

over five learning trials. Subject free recall and cued recall are assessed after both short-term and long-term delays. Increasing values indicates lower impairment. PSI is a score relevant to a patient's ability to identify, discriminate, integrate, derive a choice about information, and to respond to both visual and verbal information. All six outcome measures were available at the 6-month and 12-month time points post injury.

Statistical testing can aid in determining if the severity 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 (PSGs) and time points). A separate mixed effects ANOVA is done for each outcome measure and PSG comparison of interest. The PSG factor is a fixed "between-subject" effect and the time factor is a random "within-subject" effect. The interaction between PSG and time is also included in the model to represent situations where the effect of one factor depends on the value of the other factor. If the interaction term is significant, it implies that both the PSG and time are important in explaining differences in the outcome measure, but further analysis is needed to understand the nature of these differences. If the interaction is not significant, the main effects of PSG and time can be interpreted individually. A significant PSG effect reveals a difference in the average outcome measure between the PSGs, suggesting the model provides some prognostic value. A significant time effect shows a difference in the average outcome measure between 6 and 12 months, indicating that the outcome measure is capturing an aspect of TBI recovery that is changing over time. A significance level of $\alpha=0.05$ is used for all statistical tests.

III. RESULTS

A. 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 II), 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.

B. Model Performance

Fig. 6 shows the training performance of the GCS single prediction model. The model achieved a 90.08% training accuracy. The AUC-ROC analysis (Fig. 7) reveals that the model performed well in classifying images into GCS categories (AUC > 0.94). Table III shows the specificity and sensitivity values when each GCS severity category is considered as a

TABLE II
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

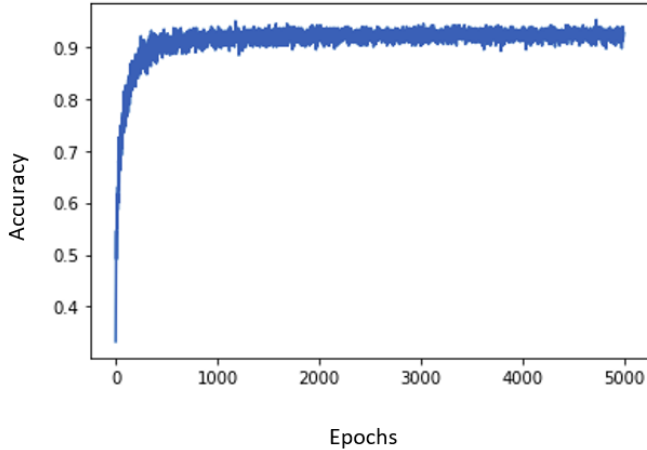


Fig. 6. Training performance of the GCS severity prediction model

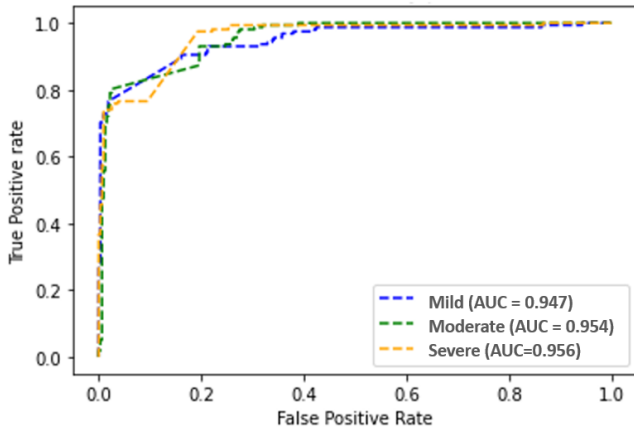


Fig. 7. AUC-ROC performance of the GCS severity prediction model

condition of interest independently. The mild group had 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).

Table IV 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 was perfect but the model was unable to identify any of the M1-severe groups by the MR images. The sensitivities for the M2-mild and M2-moderate group were perfect though the specificities were not. The model was unable to identify any of the M2-severe groups.

The outcome of the GCS+Rotterdam joint prediction model

TABLE III
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	96.50
Moderate (158)	147	11	93.04	88.61
Severe (158)	158	0	100.00	85.12

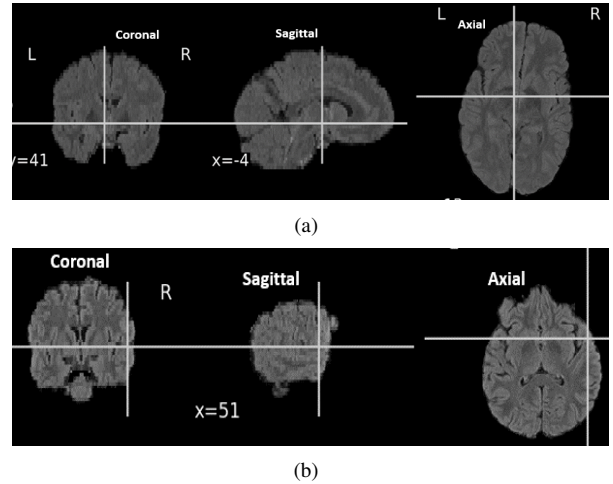


Fig. 8. GCS mild cases classified as mild by model for two subjects.

is shown in Table V. 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 were perfect. However, the model was unable to identify any of the severe groups as well. The specificity value for the R3-mild group was relatively high (74%).

Qualitative analysis: 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 Fig. 8, 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 Fig. 9) may have led to the inconsistency between the actual mild GCS classification and the predicted severe group from the MR images.

C. Clinical Relevance Assessment

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

TABLE IV
JOINT PREDICTION MODEL PERFORMANCE (GCS+MARSHALL) USING CLASSIFICATION ACCURACY, SENSITIVITY (TPR) & SPECIFICITY (TNR)

Ground Truth	Marshall GCS	M1 (116)		mild (43)	M2 (201)	
		mild (111)	severe (5)		moderate (101)	severe (57)
Predicted	Marshall	CA: 100%		TPR/mild: 100%	CA: 92.0%	
	GCS	TPR/mild: 100%	TPR/severe: 0%		TPR/moderate: 100%	TPR/severe: 0%
		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.

TABLE V
JOINT PREDICTION MODEL PERFORMANCE (GCS+ROTTERDAM) USING CLASSIFICATION ACCURACY, SENSITIVITY (TPR) & SPECIFICITY (TNR)

Ground Truth	Rotterdam GCS	R2 (184)		mild (20)	R3 (157)	
		mild (133)	severe (51)		moderate (101)	severe(36)
Predicted	Rotterdam	CA: 100%		TPR/mild: 100%	CA: 85.0%	
	GCS	TPR/mild: 100%	TPR/severe: 0%		TPR/moderate: 100%	TPR/severe: 0%
		TNR/mild: 0%	TNR/severe: 100%	TNR/mild: 74.0%	TNR/moderate: 36.0%	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 Rotterdam score. Sensitivity and specificity are computed for each GCS severity group individually by considering each as the condition of interest for each R2 and R3 Rotterdam groups.

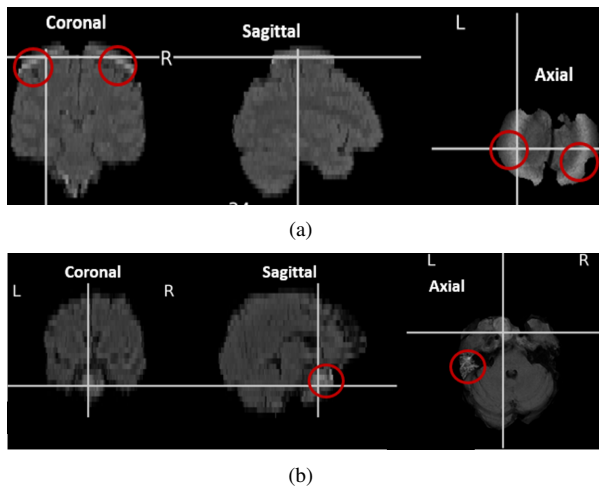


Fig. 9. GCS mild cases classified as severe by model for two subjects. Red circles highlight areas of artifact that may have misled classifier.

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.

The mixed effects ANOVA results for these comparisons in the six TBI outcome measures at 6 and 12 months are shown in Table VI. For GOS-E, the median values are reported due to its ordinal nature. For all other outcome measures, the mean and standard deviation are reported. The time effect was significant (\diamond) in almost all of the comparisons of interest for BSI-18 and SWLS, indicating that these metrics exhibit changes over time. Time was also significant in 3 of the 8 comparisons for PCL-C and PSI. The interaction effect ($*$) and/or mean

differences in the identified subgroups (\dagger) were significant in 5 of the 8 comparisons for BSI-18, SWLS, PCL-C, GOS-E, and CVLT (and 4 of the comparisons for PSI) suggesting that the groups predicted by the MR imaging models are important in explaining the difference in these outcome measures. Further analysis is required for comparisons where the interaction is significant.

IV. DISCUSSION AND CONCLUSION

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 (Fig. 6 and Table III). Both AUC-ROC and specificity was excellent for mild, moderate, and severe TBI patients (Fig. 7, Table III). 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 (Table III). Manual visual inspection of the misclassified images from the mild TBI group suggested that the model may have interpreted MRI artifacts (Fig. 9) 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 (Table IV) and the Rotterdam score (Table V). 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

TABLE VI
STATISTICAL ANALYSIS OF TBI OUTCOME MEASURES FOR IDENTIFIED SUBGROUPS OF INTEREST FROM SINGLE & JOINT PREDICTION RESULTS.

	↑ Brief Symptom Inventory -18					↓ Satisfaction With Life Scale					↑ Post-traumatic Stress Disorder Checklist - Civilian				
	Sig.	6 mths		12 mths		Sig.	6 mths		12 mths		Sig.	6 mths		12 mths	
		mean (std)	NR (%)	mean (std)	NR (%)		mean (std)	NR (%)	mean (std)	NR (%)		mean (std)	NR (%)	mean (std)	NR (%)
GCS Single Prediction															
Mild-CC (122) vs. Mild-IC-Sev (36)	◇, *	56.1 (11.8)	26.2	52.3 (10.9)	39.3	◇	20.7 (7.9)	73.0	22.5 (7.4)	62.3	◇	34.8 (16.0)	26.2	29.5 (12.9)	37.7
Mod-CC (147) vs. Mod-IC-Sev (11)	◇	55.4 (11.4)	8.3	50.0 (10.5)	41.7	◇, *, †	21.6 (8.5)	91.7	21.5 (8.9)	61.1	†, *	35.0 (14.6)	8.3	29.2 (11.3)	38.9
		53.1 (11.3)	31.9	45.0 (7.0)	46.9		25.0 (6.7)	32.0	24.4 (9.1)	54.4		26.9 (11.1)	39.5	20.3 (2.3)	61.9
		54.0 (0)	0	49.0 (0)	0		20.0 (0)	0	12.0 (0)	0		21.0 (0)	0	25.0 (0)	0
GCS+Marshall Joint Prediction															
M2-CC (185) vs. M2-IC-M1 (16)	◇	52.1 (8.0)	35.7	48.5 (9.1)	43.8	◇	23.7 (6.8)	0	22.0 (7.8)	0	Not Significant				
M1-Mild-CC (111) vs. M1-Sev-IC-Mild (5)	◇	59.2 (8.4)	43.8	54.0 (10.2)	43.8	◇, †	20.4 (9.2)	0	21.9 (7.7)	0	†, *	35.8 (16.0)	19.8	29.8 (13.0)	43.2
M2-Mod-CC (101) vs. M2-Sev-IC-Mod (52)	◇, *	56.2 (11.9)	19.8	52.0 (10.5)	45.0	◇, *, †	20.5 (8.0)	0	22.0 (8.1)	0	†, *	29.0 (8.4)	0	22.3 (7.0)	0
		59.0 (0)	0	60.0 (0)	0		10.0 (0)	0	10.0 (0)	0		23.0 (2.7)	45.5	22.0 (0.8)	66.3
		49.3 (5.5)	34.7	45.5 (8.1)	44.6		23.5 (7.4)	34.7	20.1 (8.3)	55.4	◇, *, †	29.4 (8.9)	40.4	23.0 (4.5)	59.6
		56.1 (7.1)	40.4	51.8 (6.2)	50.0		24.8 (4.1)	40.4	24.7(6.2)	50.0					
GCS+Rotterdam Joint Prediction															
R2-Mild-CC (133) vs. R2-sev-IC-Mild (51)	◇, *, †	55.3 (11.9)	19.5	51.3 (11.0)	40.6	◇, *, †	20.9 (8.2)	79.7	22.1 (7.9)	61.7	◇, *, †	34.7 (15.7)	19.5	29.2 (12.5)	39.1
R3-Mod-CC (101) vs. R3-Sev-IC-Mod (28)	◇, *	52.1 (10.7)	31.4	47.7 (8.4)	29.4	◇, *, †	24.9 (7.2)	78.4	25.3 (7.4)	70.6	†, *	29.0 (8.4)	21.6	22.3 (7.0)	39.2
R3-CC (133) vs. R3-IC-R2 (24)	◇, †	49.3 (5.5)	34.7	45.5 (8.1)	44.6	◇, *, †	23.5 (7.4)	34.7	20.1 (8.3)	55.4	Not Significant				
		49.2 (6.9)	35.7	51.4 (6.1)	35.7		22.3 (7.1)	35.7	22.7 (7.1)	35.7		23.5 (4.0)	42.9	22.2 (1.2)	63.2
		49.3 (5.9)	34.6	47.0 (7.9)	42.9	Not Significant					†, *	37.9 (14.7)	45.8	32.1 (13.2)	41.7
		60.4 (8.5)	45.8	56.0 (9.3)	45.8										
GCS Single Prediction															
Mild-CC (122) vs. Mild-IC-Sev (36)	Not Significant					Not Significant					Not Significant				
Mod-CC (147) vs. Mod-IC-Sev (11)	◇, *	7.0	31.97	7.0	46.9	†, *	54.3 (10.2)	54.4	60.2 (8.4)	61.9	◇, *	95.9 (17.6)	39.36	100.7 (14.7)	54.4
		6.0	0	5.0	0		67.0 (0.0)	0	62.0 (0.0)	0		108.0 (0.0)	0	94.0 (0.0)	0
GCS+Marshall Joint Prediction															
M2-CC (185) vs. M2-IC-M1 (16)	†	7.0	25.4	7.0	36.2	Not Significant					Not Significant				
M1-Mild-CC (111) vs. M1-Sev-IC-Mild (5)	*	6.0	12.5	6.0	6.3	Not Significant					Not Significant				
M2-Mod-CC (101) vs. M2-Sev-IC-Mod (52)	*	7.0	18.0	7.0	42.3	Not Significant					Not Significant				
		8.0	0	7.0	0										
		7.0	34.7	7.0	44.6	†	57.8 (9.8)	56.4	62.0 (8.5)	55.4	◇, *, †	96.4 (16.9)	45.5	100.4 (16.1)	44.6
		5.0	11.5	7.0	28.9		40.5 (3.3)	40.4	48.0 (12.7)	71.2		86.2 (14.8)	40.4	90.0 (13.4)	71.2
GCS+Rotterdam Joint Prediction															
R2-Mild-CC (133) vs. R2-Sev-IC-Mild (51)	Not Significant					†	54.4 (11.5)	70.7	55.0 (10.6)	56.4	Not Significant				
R3-Mod-CC (101) vs. R3-Sev-IC-Mod (28)	*	7.0	34.7	7.0	44.6	†	45.3 (7.0)	68.6	45.2 (11.1)	43.1	◇, *, †	96.4 (16.9)	45.5	100.4 (16.1)	44.6
R3-CC (133) vs. R3-IC-R2 (24)	◇	5.0	0	6.0	17.9	†	57.8(9.8)	56.4	62.0 (8.5)	55.4	†, *	80.1 (15.9)	35.7	90.6 (14.3)	53.6
		7.0	26.3	7.0	38.4	†	40.8 (2.0)	35.7	56.6 (8.2)	71.4		91.7 (18.1)	42.9	98.2(16.1)	46.6
		6.0	20.8	7.0	20.8		52.4 (11.3)	51.1	60.6 (9.0)	58.6	*	95.8 (8.1)	50.0	94.0 (11.3)	45.8
							46.0 (8.2)	54.2	47.3 (17.8)	45.8					

↓: lower values of the measure indicates higher severity; †: higher values of the measure indicates higher severity; NR: Not Reported Data; SD: Standard Deviation;

on the joint prediction task (Table IV and V) is puzzling since it accurately classified severe TBI subjects on the single prediction task (Table III). 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 (Table VI). 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.

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, etc.). 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.

ACKNOWLEDGMENT

We wish to thank the TRACK-TBI Pilot Investigators for access to the CT derived scores and TBI outcome measures.

REFERENCES

- [1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.
- [2] S. R. Swarna, S. Boyapati, V. Dutt, and K. Bajaj, "Deep learning in dynamic modeling of medical imaging: A review study," in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*. IEEE, 2020, pp. 745–749.
- [3] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on mri," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, 2019.
- [4] L. Liu, Q. Dou, H. Chen, I. E. Olatunji, J. Qin, and P.-A. Heng, "Mtnet: Multi-task deep learning with margin ranking loss for lung nodule analysis," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 74–82.
- [5] S. Serte, A. Serener, and F. Al-Turjman, "Deep learning in medical imaging: A brief review," *Transactions on Emerging Telecommunications Technologies*, p. e4080, 2020.
- [6] S. Dua, U. R. Acharya, and P. Dua, *Machine learning in healthcare informatics*. Springer, 2014, vol. 56.
- [7] D. Yeboah, L. Steinmeister, D. B. Hier, B. Hadi, D. C. Wunsch, G. R. Olbricht, and T. Obafemi-Ajayi, "An explainable and statistically validated ensemble clustering model applied to the identification of traumatic brain injury subgroups," *IEEE Access*, vol. 8, pp. 180690–180705, 2020.
- [8] A. J. Masino and K. A. Folweiler, "Unsupervised learning with glmr feature selection reveals novel traumatic brain injury phenotypes," *arXiv preprint arXiv:1812.00030*, 2018.
- [9] C. for Disease Control, "Traumatic brain injury & concussion," 2019.
- [10] H. Yao, C. Williamson, J. Gryak, and K. Najarian, "Automated hematoma segmentation and outcome prediction for patients with traumatic brain injury," *Artificial Intelligence in Medicine*, vol. 107, p. 101910, 2020.
- [11] J. K. Yue, M. J. Vassar, H. F. Lingsma, S. R. Cooper, D. O. Okonkwo, A. B. Valadka, W. A. Gordon, A. I. Maas, P. Mukherjee, E. L. Yuh *et al.*, "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, no. 22, pp. 1831–1844, 2013.
- [12] K. K. Wang, Z. Yang, T. Zhu, Y. Shi, R. Rubenstein, J. A. Tyndall, and G. T. Manley, "An update on diagnostic and prognostic biomarkers for traumatic brain injury," *Expert review of molecular diagnostics*, vol. 18, no. 2, pp. 165–180, 2018.
- [13] A. I. Maas, C. W. Hukkelhoven, L. F. Marshall, and E. W. Steyerberg, "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, no. 6, pp. 1173–1182, 2005.
- [14] A. Deepika, A. Prabhuraj, A. Saikia, and D. Shukla, "Comparison of predictability of marshall and rotterdam ct scan scoring system in determining early mortality after traumatic brain injury," *Acta neurochirurgica*, vol. 157, no. 11, pp. 2033–2038, 2015.
- [15] M. Wintermark, P. C. Sanelli, Y. Anzai, A. J. Tsiouris, C. T. Whitlow, T. J. Druzgal, A. D. Gean, Y. W. Lui, A. M. Norbash, C. Raji *et al.*, "Imaging evidence and recommendations for traumatic brain injury: conventional neuroimaging techniques," *Journal of the American College of Radiology*, vol. 12, no. 2, pp. e1–e14, 2015.
- [16] E. L. Yuh, P. Mukherjee, H. F. Lingsma, J. K. Yue, A. R. Ferguson, W. A. Gordon, A. B. Valadka, D. M. Schnyer, D. O. Okonkwo, A. I. Maas *et al.*, "Magnetic resonance imaging improves 3-month outcome prediction in mild traumatic brain injury," *Annals of neurology*, vol. 73, no. 2, pp. 224–235, 2013.
- [17] P. Kalavathi and V. S. Prasath, "Methods on skull stripping of mri head scan images—a review," *Journal of digital imaging*, vol. 29, no. 3, pp. 365–379, 2016.
- [18] F. Amyot, D. B. Arciniegas, M. P. Brazaitis, K. C. Curley, R. Diaz-Arrastia, A. Gandjbakhche, P. Herscovitch, S. R. Hinds, G. T. Manley, A. Pacifico *et al.*, "A review of the effectiveness of neuroimaging modalities for the detection of traumatic brain injury," *Journal of neurotrauma*, vol. 32, no. 22, pp. 1693–1721, 2015.
- [19] S. Basheera and M. S. S. Ram, "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.
- [20] D. Pan, A. Zeng, L. Jia, Y. Huang, T. Frizzell, and X. Song, "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.
- [21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [22] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI global, 2010, pp. 242–264.
- [23] E. A. Wilde, G. G. Whiteneck, J. Bogner, T. Bushnik, D. X. Cifu, S. Dikmen, L. French, J. T. Giacino, T. Hart, J. F. Malec *et al.*, "Recommendations for the use of common outcome measures in traumatic brain injury research," *Archives of physical medicine and rehabilitation*, vol. 91, no. 11, pp. 1650–1660, 2010.
- [24] "Federal interagency traumatic brain injury research (fitbir)," <https://fitbir.nih.gov/>, accessed: 2019-09-20.
- [25] K. Noguchi, T. Ogawa, A. Inugami, H. Fujita, J. Hatazawa, E. Shimosegawa, T. Okudera, K. Uemura, and H. Seto, "Mri of acute cerebral infarction: a comparison of flair and t2-weighted fast spin-echo imaging," *Neuroradiology*, vol. 39, no. 6, pp. 406–410, 1997.
- [26] B. A. Jónsson, G. Bjornsdottir, T. Thorgerisson, L. M. Ellingsen, G. B. Walters, D. Gudbjartsson, H. Stefansson, K. Stefansson, and M. Ulfarsson, "Brain age prediction using deep learning uncovers associated sequence variants," *Nature communications*, vol. 10, no. 1, pp. 1–10, 2019.
- [27] Z. Akkus, A. Galimzianova, A. Hoogi, D. L. Rubin, and B. J. Erickson, "Deep learning for brain mri segmentation: state of the art and future directions," *Journal of digital imaging*, vol. 30, no. 4, pp. 449–459, 2017.
- [28] J. Muschelli, E. Sweeney, M. Lindquist, and C. Crainiceanu, "fslr: Connecting the fsl software with r," *The R journal*, vol. 7, no. 1, p. 163, 2015.
- [29] H.-P. Chan, R. K. Samala, L. M. Hadjiiski, and C. Zhou, "Deep learning in medical image analysis," *Deep Learning in Medical Image Analysis*, pp. 3–21, 2020.
- [30] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.
- [31] 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.