Using a 3D ResNet to Determine Quality of Skull-Stripped Brain Images

Carl Brandenburg
School of Computing
Clemson University
Clemson, SC USA
brande3@clemson.edu

Keller Brandenburg
School of Computing
Clemson University
Clemson, SC USA
wbrande@clemson.edu

1. ABSTRACT

Scientific integrity initiatives and data sharing laws have increased the open access sharing of neuroimaging datasets, many of which include T1-weighted structural MRI scans. These images often retain identifiable facial features that create privacy risks. Researchers use skull-stripping tools to remove these identifiable features, but these tools may also unintentionally remove brain tissue. To ensure that identifiable facial features have been adequately removed while maintaining the brain tissue, tedious visual inspection of the images must be conducted to evaluate the data quality. Here, we describe an automated method that identifies facial features that may be left over and detects the loss of brain tissue to support the visual inspection of the data. Our facial feature model achieved 89.09% accuracy and our brain tissue loss detection model achieved 97.37% accuracy. These results demonstrate the feasibility of using residual neural network architectures to support automated quality control.

2. INTRODUCTION

Neuroimaging is an important part of brain research and clinical diagnostics. As the field grows, it gets more and more data that can be shared between research groups on which to create better tools for research. T1-weighted structural MRI scans are especially valuable due to their anatomical detail, and they are frequently included in shared databases.

The problem is that these MRI scans often contain identifiable facial features such as the eye sockets, forehead and brow ridge, jaw, and teeth. Laws such as HIPAA (Health Insurance Portability and Accountability Act of 1996) require the removal of identifiable data. In order to protect peoples' privacy and comply with the law, the MRI scans must be skull-stripped, which is the removal of all non-brain tissue. An unfortunate side effect of this process is the accidental removal of brain tissue. As part of the skull-stripping process, the scans must go through a quality assurance process to ensure that identifiable features have been removed. It is a tedious process to inspect the processed images for recognizable features and brain tissue loss. In this paper, we describe the development of a 3D residual neural network that can determine whether a skull-stripped MRI scan retains any identifiable facial features and whether or not the skull-stripping process has removed any brain tissue from the scan.

3. DESIGN AND METHODOLOGY

Our model is inspired by the model proposed by Luo et al. In their paper, they used a 3D convolutional neural network, and we will use a 3D residual neural network, which is basically an extension of the CNN. We will work with the same dataset, in which they "intentionally adjusted the parameters of these skull-stripping tools so that not all skull-stripped images are perfectly processed to mimic a real-world data de-identification process where there is varied extent of skull-stripping effectiveness."

3.1 Dataset Generation

In their study, Luo et al. created a dataset using raw 3D t1-weighted brain MRI scans. To preprocess the scans, they used two skull-stripping methods: the Brain Extraction Tool (BET) and the Brain Surface Extractor (BSE). They systematically adjusted the parameters on these tools to create four categories of images, corresponding to different levels of non-brain tissue removal. These categories ranged from images where nearly all facial features were

preserved to images where some brain tissue was starting to be removed. Altogether, the final dataset consisted of 2,656 labeled brain images.

For our work, we assigned each scan two binary labels based on manual classification done by Luo et al. These two labels were if there were any remaining identifiable features, and if there was significant brain tissue that was inadvertently removed. We normalized the scans by voxel intensity to be between 0 and 1, and converted them into PyTorch tensors. We then used a stratified split so that the proportion of positive and negative levels would be maintained to split the dataset into three subsets, 1912 training, 478 validation, and 266 testing. The training set is used by the model to learn patterns. The validation set is used to tune the model's hyperparameters and avoid overfitting. The testing set is used to assess how the final model performs on completely unseen data.

3.2 3D Residual Neural Network

The original study upon which this study is based uses a 3D CNN. A 3D ResNet is a specific type of CNN that uses skip connections to allow information to bypass layers. The advantages of a ResNet over a regular CNN are that it can be much deeper to capture more complex features and that, due to the layer skipping, it can be more efficient to train. ResNets also help deal with the problem of vanishing gradients, in which the gradients that are used to guide the learning process become extremely small, slowing or stopping the learning process.

Our network takes input voxel data of size 256x256x150. Voxel intensity values are normalized to between 0 and 1 to ensure stability. Our architecture, shown in Figure 1, consists of an initial 3D convolutional layer with a 7x7x7 kernel, followed by a max pooling layer and then four residual stages. Each residual block is made up of two 3x3x3 convolutional layers followed by batch normalization and ReLU activation. In the blocks where input and output dimensions are different, a 1x1x1 convolution is implemented to ensure that they can be properly added together. Each layer doubles the number of channels from 64 to 128 to 256 to 512 and decreases the spatial size of the image using a stride of 2. After the fourth residual layer, a global average pooling layer collapses the spatial dimensions into a 512 dimensional vector, which then goes through a dense layer which uses a sigmoid function to produce the final binary classification.

The network consists of 18 learnable layers, the 3D convolution layers plus the final linear layer. This architecture is deep enough to learn patterns, but still simple enough that one can analyze it and easily understand how it works. It is also not so big that it requires more GPU power or memory to which we reasonably have access.

4. IMPLEMENTATION

The deep learning methods in this study were implemented using PyTorch. The models were trained on the Clemson University Palmetto high performance computing cluster,

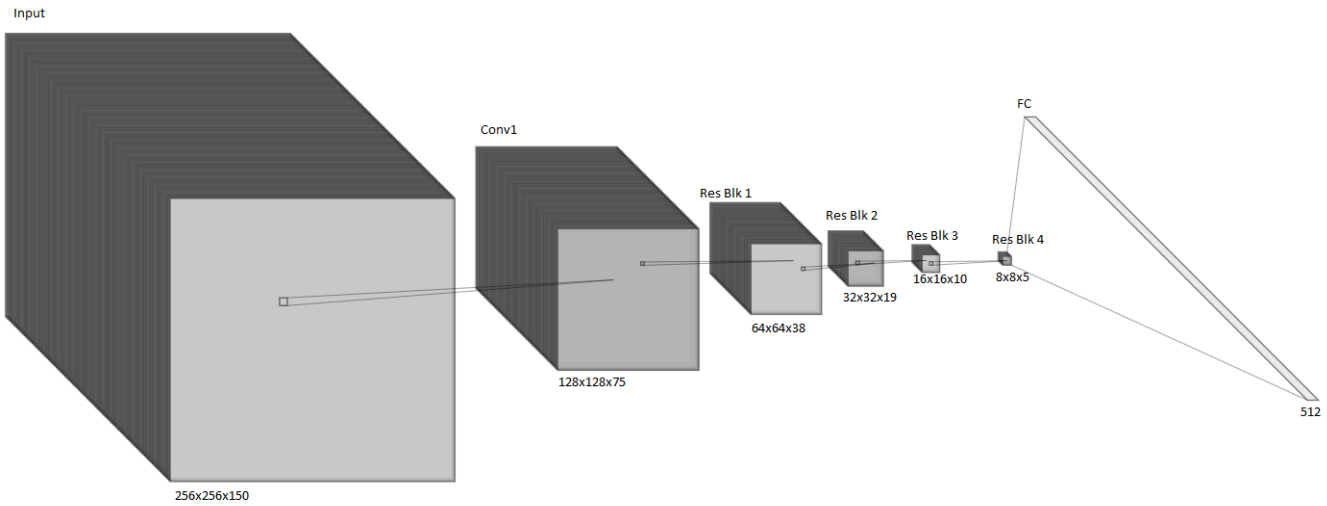

Fig. 1. Architecture of 3D Residual Neural Network

which allowed us use of a 16-core Intel processor, 64 GB of RAM, and an NVIDIA Tesla A100 GPU.

We applied a stratified split to the dataset, allocating 72% to training, 18% to validation, and 10% to testing, ensuring to maintain the proportion of classes among all subsets. Two separate 3D residual neural network models were trained using this data: one to classify the presence of identifiable facial features, and another to detect brain tissue loss due to over-aggressive skull-stripping.

Each model was trained using binary cross-entropy loss and used an Adam optimizer at a learning rate of 0.0001, and used a batch size of 1. Model performance was evaluated using standard classification metrics, accuracy, precision, recall, F1-Score, and confusion matrices.

5. EVALUATION

We evaluated the trained models using accuracy, sensitivity, and specificity on an unseen test set of MRI scans. As shown in Tables 1 and 2, our models achieved strong classification performance across both tasks. For facial feature detection, our 3D ResNet model achieved 89.09% accuracy, and outperformed the 3D CNN model with 97.92% Specificity. For brain tissue loss detection, we achieved 97.37 % accuracy, and were able to very slightly outperform the 3D CNN with 97.82% sensitivity on an unseen dataset. Overall, our model was not able to outperform the 3D CNN, falling short in key areas.

Table 1. Defacing classification results for the 3D CNN and 3D ResNet

Methods	Accuracy	Specificity	Sensitivity
3D CNN with Inception	95.49%	95.52%	94.42%
3D ResNet	89.09%	97.92%	84.12%

Table 2. Brain tissue loss classification results for 3D CNN and 3D ResNet

Methods	Accuracy	Specificity	Sensitivity
3D CNN with Inception	97.63%	97.62%	97.68%
3D ResNet	97.37%	96.88%	97.82%

The brain tissue model performed very well, with specificity and sensitivity indicating a good degree of confidence. The facial feature detection model, while less accurate, still performed decently well.

In figure 2A and 2B, we have histograms demonstrating the models' confidence by distributing the predicted probability for detecting facial features and identifying the loss of brain tissue.

Confusion matrices for the facial feature detection classifier and the brain tissue loss classifier are presented in Figures 3A and 3B respectively. The facial feature detection model is good at finding faces, but is a little bit weaker when it comes to rejecting non-face cases. The brain tissue loss detection model performs very well. It made very few errors, which makes it more reliable for use in real-world datasets.

For greater interpretability, Grad-CAM (Gradient-weighted Class Activation Mapping) visualizations were produced,

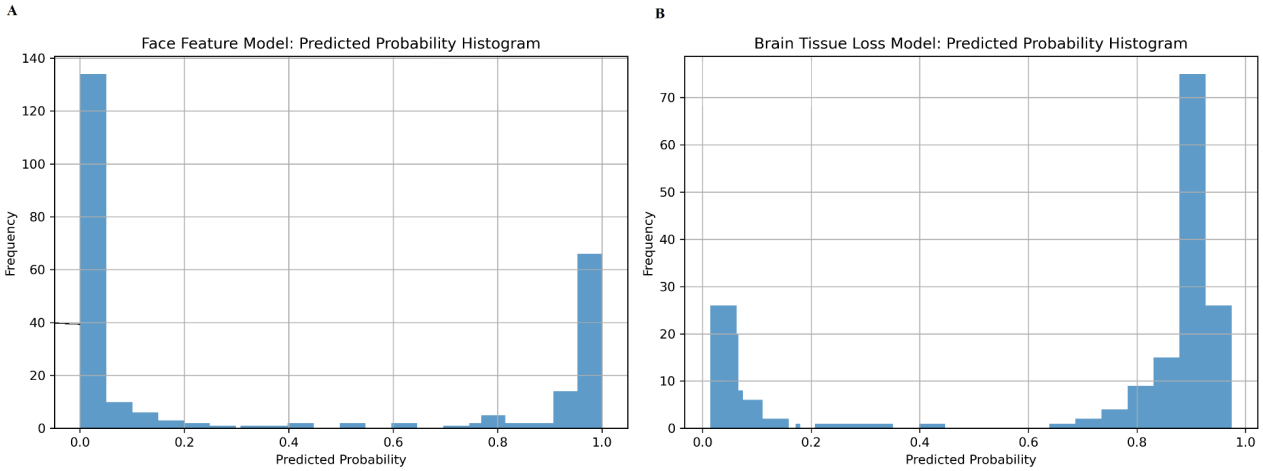

Fig. 2. The histogram distribution of predicted probability for facial feature classification (A) and brain tissue loss classification (B)

as shown in Figures 4 and 5. These visualizations highlight regions of an image on which the model is focusing. In Figure 4, the facial feature detection model, highlights were observed around sides of the head and the lower neck region. These areas may represent soft tissue or neck remnants that correlate with scans that are not de-identified. The model did not focus on what one would expect - the eye, nose, or jaw. The model may be learning to detect less obvious features that correlate better with incomplete defacing. In Figure 5, the brain tissue loss detection model, highlights were observed around the frontal lobe and the brain stem. The model could be activating around known areas where it may be difficult to remove non-brain tissue, like around the eyes, where there is a lot of tissue, and around the neck, similarly where there is a lot of tissue that may be difficult to distinguish.

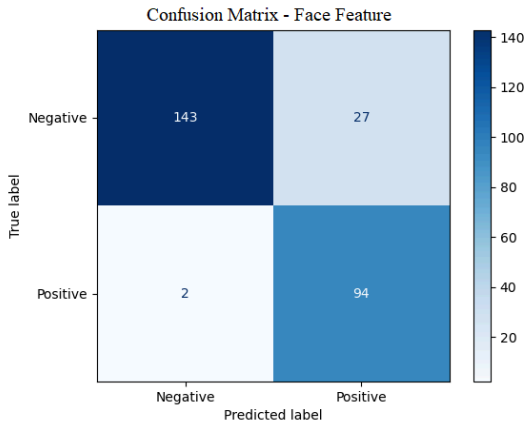

Figure 3A. Confusion matrix of the facial feature model used on an unseen dataset

6. CONCLUSION

In this paper, we developed two separate 3D Residual Neural Network (ResNet) models to identify facial features and to detect the loss of brain tissue in skull-stripped MRI brain scans. The facial feature detection model demonstrated only 89.09% accuracy in identifying recognizable facial features, but the brain tissue loss detection model demonstrated an impressive 97.37% accuracy, which while not better than the CNN model, was very close. This high classification accuracy can suggest strong generalization and reliability for real world quality control tasks.

To support our understanding of how the model works, we used Grad-CAM to visualize the regions each model focused on when making predictions. These showed that the model looked at unexpected places to make decisions, but it may have been able to associate some peripheral feature with an incomplete defacing.

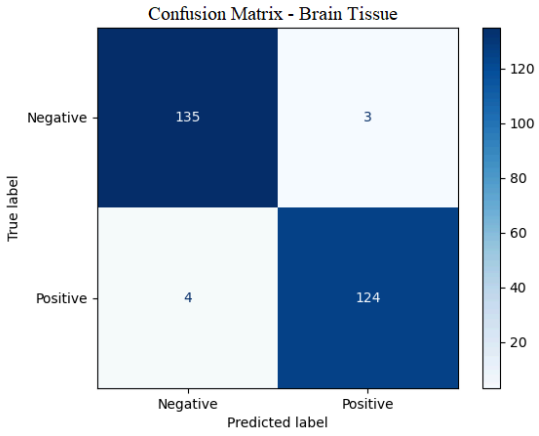

Figure 3B. Confusion matrix of the brain tissue model used on an unseen dataset

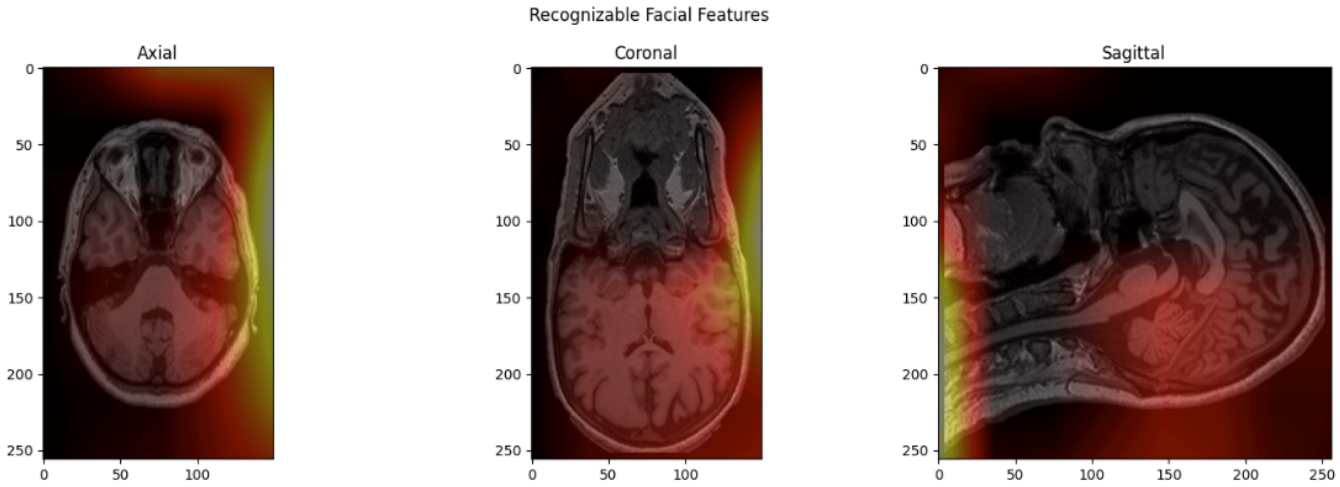

Figure 4. Grad-CAM visualization of the facial feature model

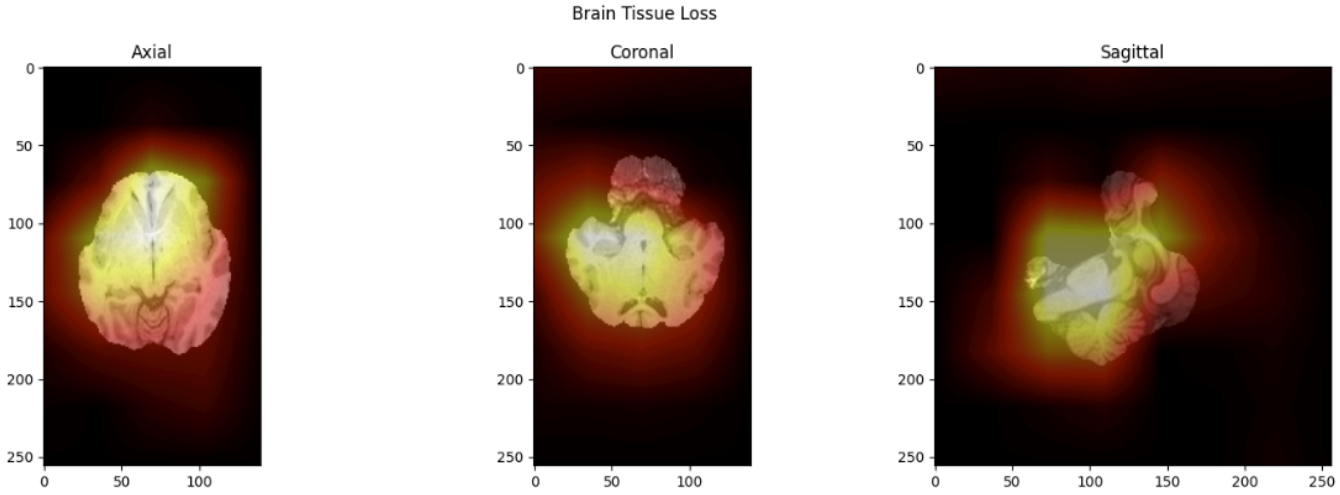

Figure 5. Grad-CAM visualization of the facial feature model

8. REFERENCES

[1]Li Luo, Rishikesh Phatangare, James Wang, Kenneth Vaden, Mark Eckert, *Deep Learning Methods to Evaluate Privacy and Quality of Skull-Stripped Brain Images*, The 17th International Conference on Brain Informatics (BI 2024), Bangkok, Thailand, 13-15 December 2024.

[2] He, K., Zhang, X., Ren, S., & Sun, J. 2015. Deep Residual Learning for Image Recognition. arXiv preprint arXiv:1512.03385.

[3] Vanshika Sharma. 2021. *ResNets: Why do they perform better than Classic ConvNets? (Conceptual Analysis)*. Towards Data Science. Retrieved April 17, 2025, from https://towardsdatascience.com/resnets-why-do-they-perform-better-than-classic-convnets-conceptual-analysis-6a9c82e06e53/

[4] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv preprint arXiv:1912.01703.

[5] Maier, A., Syben, C., Lasser, T., & Riess, C. (2019). A gentle introduction to deep learning in medical image processing. *Zeitschrift für Medizinische Physik*, 29(2), 86–101. https://doi.org/10.1016/j.zemedi.2018.12.003