Brain Age Prediction Using 3D Convolutional Neural Networks: Analysis of Cognitive Decline in OASIS Dataset

Nolan Betts
University of Texas at Austin
Texas, USA
nolanfbetts@utexas.edu

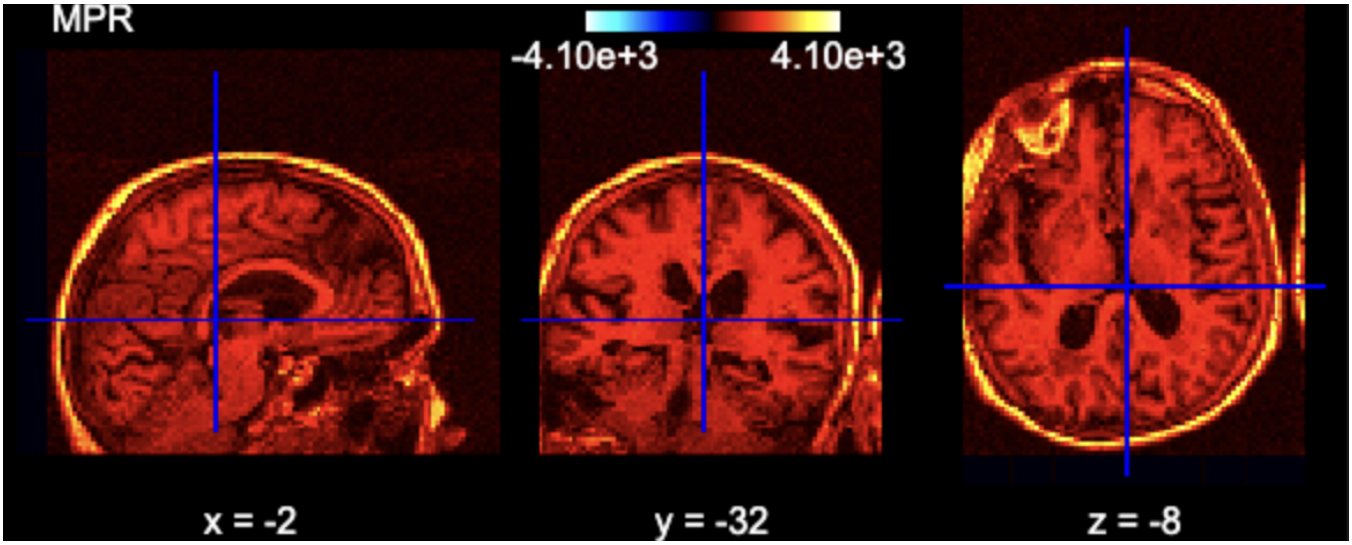

Figure 1: 3 axis view of an OASIS MRI brain scan.

Abstract

This study takes a deep learning approach for brain age prediction using MRI scans. We implement two distinct models, a baseline 3D Convolutional Neural Network (CNN) and an enhanced version incorporating explicit feature extraction from the MRIs. The models were trained on the OASIS-2 longitudinal dataset to predict age in Non-demented subjects and evaluated on subjects with dementia and those who converted to dementia throughout the study. Our enhanced model combines traditional convolutional features along with thresholding measurements of ventricle size, gray matter, and white matter, providing explicit features alongside age predictions. The study aims to detect accelerated brain aging in both dementia groups, offering potential early markers for neurodegenerative diseases.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
AI in Healthcare '25, Austin, TX
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/tbd

Keywords

Brain Age Prediction, Deep Learning, MRI Analysis, Neurodegeneration, Feature Engineering, Convolutional Neural Networks, Alzheimer's Disease, Biomarker Detection

ACM Reference Format:
Nolan Betts. 2025. Brain Age Prediction Using 3D Convolutional Neural Networks: Analysis of Cognitive Decline in OASIS Dataset. In *Proceedings of (AI in Healthcare '25).* ACM, New York, NY, USA, 8 pages. https://doi.org/tbd

1 Introduction

In this section we will introduce the background of neuro imaging and it's relation to Deep Learning as well as our research objectives.

1.1 Background

Brain age prediction has emerged as a crucial tool in neurology, serving as a potential biomarker for brain health and neurological disorders. When an individual's predicted brain age significantly exceeds their chronological age, it may indicate accelerated brain aging, which has been associated with increased risk of neurodegenerative conditions and cognitive decline. [1]

Neuroimaging, particularly MRIs, play a pivotal role in dementia diagnosis by enabling the assessment of brain structure and aging patterns. Traditionally, diagnostic approaches have relied on cognitive assessments and clinical symptoms, which often only become apparent after significant neurodegeneration has occurred.

MRI analysis, boosted by machine learning, can reveal structural changes in brain regions typically affected by dementia before clinical symptoms manifest. [2]

Early detection of dementia remains challenging due to the following:

- The early brain changes that may indicate disease are subtle and precede disease symptoms exhibited by the patient
- Individual nature of brain structure and aging
- Difficulty in differentiating normal brain aging from diseased brain aging
- Limited availability of longitudinal data tracking progression of dementia (converted patients)

This implementation uses Deep Learning approaches to attempt to address these challenges by learning relevant features from MRI data while incorporating engineered features related to brain structure and tissue.

1.2 Research Objectives

The objective of this study is to predict brain age given an MRI of a patient. We will train a model to predict age of patient given healthy brain scans. Using this model we will then predict the brain age of demented patients and converted patients. We will then study the brain age gap by taking the difference between predicted value and actual value in unhealthy patients. We hypothesize that if our model predicts unhealthy brain scans to be older than the actual value, the model could have clinical value in predicting neurodegenerative diseases.

2 Methods

Here we will explore the Dataset used in this study, our Model Architecture, and our Training Procedures.

2.1 Dataset

For this study we leveraged Longitudinal MRI Data in Non-demented and Demented Older Adults dataset from the OASIS-2 Longitudinal Dataset [3]. This dataset, provided by WashU Medicine, consists of 150 different patients aged 60-96. Each subject was scanned multiple times over separate visits for a total of 373 scannings sessions. The patients in this dataset present in different groups. Non-demented throughout the study, Converted from Non-demented to Demented throughout the study, and Demented.

Our healthy group consists of 190 imaging sessions totaling 691 images. Our demented group consists of 146 imaging sessions and our converted group consists of 37 imaging sessions totaling 676 images across the two groups. The MRI images used in this study are in a *.nifty.img* format. An example of this image can be seen in Figure 1.

2.2 Model Architecture

In this study we leverage a 3D CNN to predict brain age from MRI scans. Then network uses residual connections and progressive channel expansions. This network is designed to preserve features and achieve gradient stability while avoiding overfitting. The architecture includes residual blocks to facilitate better gradient flow and feature reuse by using skip connections. This design addresses the vanishing gradient problem common in deep neural networks. The network progressively increases channel depth and maintains spatial resolution until our final pooling stages.

The network structure consists of a feature extraction block, residual processing stages, feature aggregation, and regression head. The initial feature extraction using a 3D convolution layer with 8 output channels, batch normalization for training stability, ReLU activation, and a 0.2 dropout for regularization. There are two sequential residual blocks that have increasing channel dimensions. The first block goes from 8 to 16 channels with a 0.1 dropout. The second block goes from 16 to 32 channels with a 0.2 dropout. Each of these residual blocks have a 3D convolution path with batch normalization. We also leverage identity mappings for dimension matching in skip connections. Our feature aggregation uses strided convolution for downsampling, global averaging pooling for feature aggregation, and a progressive reduction in spatial dimensions while preserving feature channels. The regression head consists of two filly connected layers, ReLU activation with a dropout of 0.3 between layers, and a single output for age prediction.

This model is implemented using PyTorch. We use an AdamW optimizer with weight decay for regularization. Learning rate scheduling with a warm-up period. Gradient Clipping to ensure training stability. As well as early stopping based on validation performance. Our data processing pipeline has input normalization to zero mean and unit variance. We also implemented an augmentation strategy to better help out model generalize by introducing noise, random intensity scaling, and horizontal flipping. These augmentations are only applied during training. Our regularization strategy consists of progressive dropout rates, batch normalizations, weight decay optimization, and data augmentation during training.

In our enhanced model, we use the same network but we add extra layers for key brain related features. We extract ventricle features by attempting to extract total ventricle volume, number of ventricle regions, average ventricle size, ventricle asymmetry, and the distance of the ventricle center of mass from brain center totaling 5 features. We extract 13 gray matter features from the total gray matter volume, distribution statistics, and gray matter volume analysis. We extract 7 white matter features including the total white matter, distribution statistics, and connectivity features. To do this we employ intensity thresholding techniques similar to those found in the following projects. Below 0.2 for ventricle features, between 0.3 to 0.7 for gray matter features, and from 0.7 to 1.0 for white matter features. The potential limitations of this approach are explored in the conclusion [4] [5].

2.3 Training Procedure

The dataset is split into training and validation sets using a *GroupShuffleSplit* strategy. This ensures that all scans from the same patient are assigned to the same split. This prevents data leakage by guaranteeing that a patient won't have scans overlapping the test and training set. The patients are grouped by patient ID and the split is performed 80% for training and 20% for validation.

The model is trained using an AdamW optimizer with a learning rate of 0.001 and weight decay of 0.05. For the first three epochs there is a learning rate warmup that gradually increases to our learning rate. The training proceeds for up to 50 epochs. Early

stopping occurs if the validation loss does not improve for 10 epochs. A learning rate scheduler, *ReduceLROnPlateau*, reduces that learning rate by half after 5 epochs of validation loss plateau. Gradient clipping, *clip_grad_norm_*, is used with a max norm of 1.0 to help ensure straining stability.

Several metrics are tracked throughout the training process to analyze the model. The primary loss function is mean squared error between the predicted and the actual age. We use mean squared error, mean absolute error, and root mean squared error to track model performance. We track these metrics as denormalized ages on both the training and validation sets to help with interpretability. We track our max gradient norms over epochs to look for signs of exploding gradients which could be a sign of instability. We also track our learning rate at each epoch so that we can analyze training plateaus and analyze how our model training performs at different learning rates. These metrics are discussed in the Results section for both our standard CNN and enhanced CNN.

All training was performed using CPU supported methods on Apple Silicon. I attempted to leverage more advanced GPU specific functionality without success, the limitations and pain points are discussed in the conclusions. Given this limitation, I set a batch size of 8 to help with training/validation time and efficiency.

3 Results

In this section we will explore accuracy metrics across our different models. We will also make performance comparisons across the patient groups analyzing brain age gaps.

3.1 Model Performance CNN

In this section we will review the results of the CNN without any Feature Engineering. We will reference a figure and examine the results of that figure until the next figure is referenced.

(Figure 2) The training loss gradually decreases and stabilizes around 1.04. The validation loss quickly drops off and stabilizes around .87. The Mean Average Error (MAE) is around 6.8 years on training data and 5.7 years on validation data. The Root Mean Squared Error (RMSE) stabilizes around 8.7 years and 7.4 years for training and validation sets respectively. The learning rate scheduler shows a warm up phase, a long plateau, and a learning rate reduction before reaching early stopping at epoch 26. The max gradient shows some volatility in early training but is generally decreasing from around 4.5 to 2.0. Overall our best validation loss (.8662) was achieve at epoch 16 with the following characteristics, Training MAE 6.83 years, Validation MAE 5.75 years, Training RMSE 8.31 years, and Validation RMSE 7.41 years.

(Figure 3) Running the trained model against the Demented patient scans reveals the following statistics. Across 542 images we have an MAE of 5.64 years and an RMSE of 6.94 years. We have a Brain Age Gap of +1.02 years with a standard deviation of 6.87 years. The predictions cluster around 77-78 years with little variation.

(Figure 4) Running the trained model against the Converted patient scans reveals the following statistics. Across 134 images we have an MAE of 6.78 years and an RMSE of 7.72 years. We have a Brain Age Gap of -2.67 years with a standard deviation of 7.24 years. The predictions cluster around 77-78 years with little variation.

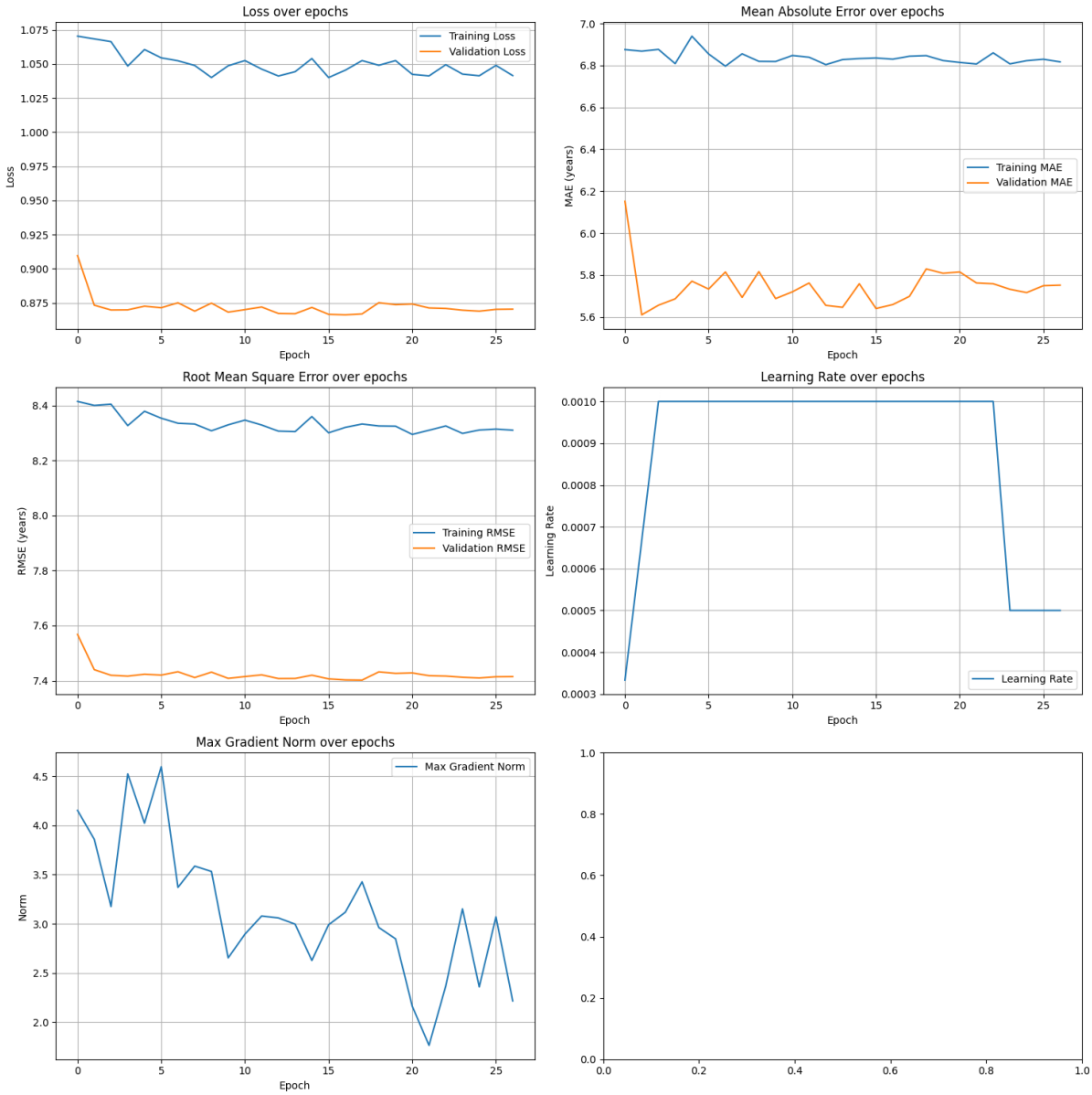

Figure 2: Training metrics for the CNN model, showing loss, MAE, RMSE, learning rate, and gradient norm over epochs.

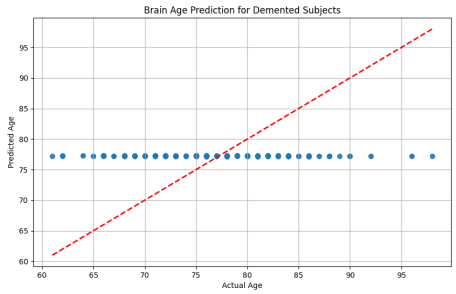

Figure 3: Brain age prediction results for subjects with dementia, showing the relationship between actual and predicted age.

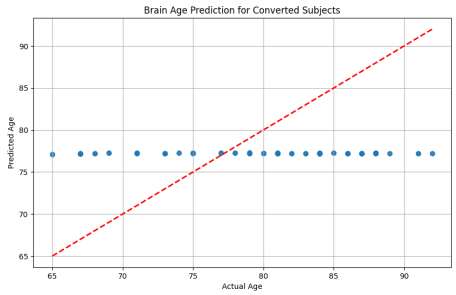

Figure 4: Brain age prediction results for converted subjects, showing the relationship between actual and predicted age.

3.2 Model Performance CNN with Feature Engineering

In this section we will review the results of the CNN with Feature Engineering. We will reference a figure and examine the results of that figure until the next figure is referenced.

(Figure 5) The training loss here decreases steadily from 1.07 to .72 while the validation loss is more volatile eventually reaching some semblance of convergence around .87. The training MAE steadily improves over time, decreasing from 6.84 years to 5.47 years. The validation MAE degrades overtime eventually reaching some stability in the final epochs around 6.45 years. The training RMSE reaches 6.92 years while the validation RMSE reaches 7.52 years. We observe a learning rate pattern as expected with a warm up phases, a long plateau, and a learning rate reduction before reaching early stopping at epoch 16. Gradient behavior is volatile peaking at 8.64 and 4.15 and increasing in later epochs.

(Figure 6) Running the trained model against the Demented patient scans reveals the following statistics. We observe an MAE 5.22 years and RMSE of 6.51 years. We have a Brain Age Gap of +1.23 years with a standard deviation of 6.40 years. The prediction variation ranges from 75 to 85 years.

(Figure 7) Running the trained model against the Converted patient scans reveals the following statistics. We observe an MAE 6.37 years and RMSE of 7.20 years. We have a Brain Age Gap of -2.89 years with a standard deviation of 6.60 years. The prediction variation ranges from 73 to 90 years.

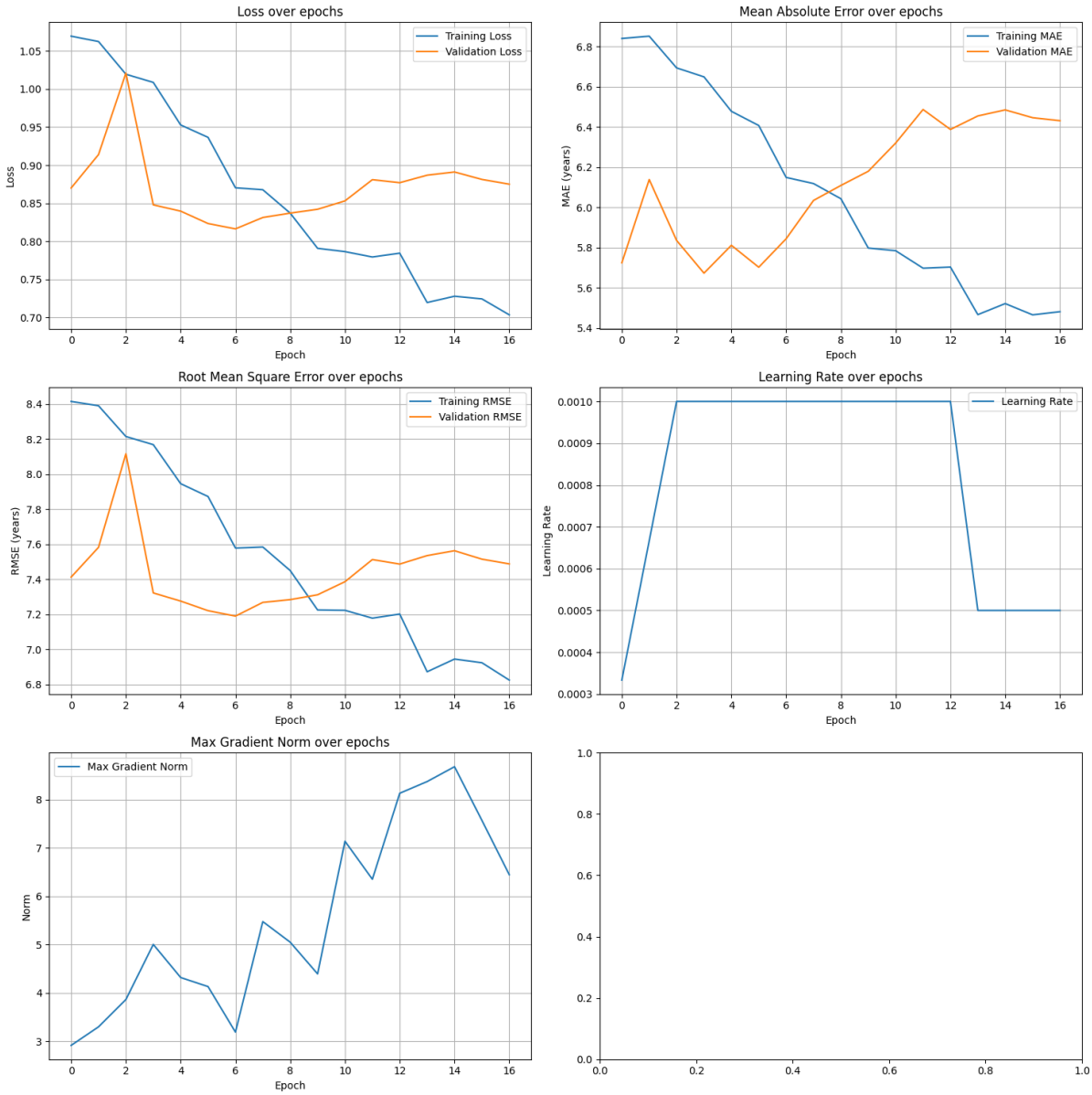

Figure 5: Training metrics for the CNN model, showing loss, MAE, RMSE, learning rate, and gradient norm over epochs (with Feature Engineering).

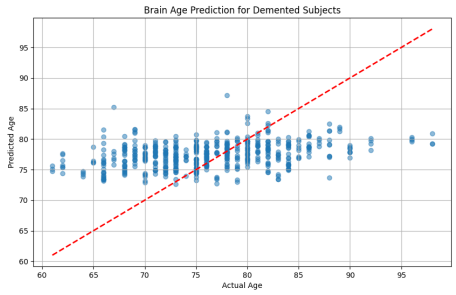

Figure 6: Brain age prediction results for subjects with dementia, showing the relationship between actual and predicted age (with Feature Engineering).

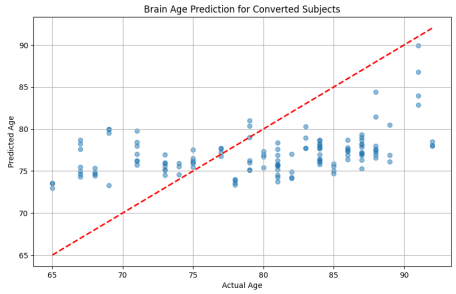

Figure 7: Brain age prediction results for converted subjects, showing the relationship between actual and predicted age (with Feature Engineering).

First, ensure the large figure appears properly

4 Discussion

In this section we will discuss the two different models presented in this study. We will start with the CNN model and follow up with the CNN model with Feature Engineering.

4.1 Interpretation of Results CNN

(Figure 2) The model shows consistent performance with stable error metrics. However, the validation set performs better than the training set which may be suggestive of underfitting. Otherwise, we see smooth loss curves and a generally healthy looking learning pattern. We see a Validation MAE 5.75 years which is fairly reasonable for our task. However, our age range is relatively small (60-97 years of age). Our inflated RSME of 7.41 years is suggestive of our model guessing being further off, as RSME heavily penalizes larger errors. Overall we do not see a progressive learning pattern here. As we will show next, the model is not performing well for the given task.

(Figure 3) In our Demented patient group it is clear that the model has cratered towards a narrow prediction of 77-78 years. The model has extremely limited to no predictive power. The model is not making any differentiated decisions.

(Figure 4) In our Converted group we experience the same model failures mentioned above.

This failure may suggest that the features learned for healthy brain aging don't transfer well to brains affected by dementia. However, it is more likely that our model does not have enough information or features to make appropriate decisions with respect to brain age. There is no clinical value in this model and it should not be used.

4.2 Interpretation of Results CNN with Feature Engineering

(Figure 5) This model shows much better overall performance. We experience ~20% improvement in training MAE and ~17% improvement in training RMSE. We also see faster convergence (16 epochs vs 26 epochs). This model is a much more aggressive learner showing steeper validation loss, higher gradient norms, and a more pronounced training/validation gap. While this is a markedly improved performance, we now see more volatile validation metrics and signs of overfitting. This could suggest that our model may be too specific with feature extraction or that our dataset is not large enough to generalize well. While our feature extraction may not be perfect, we will see in the next section that it has provided value in age prediction not seen in our previous model.

(Figure 6) Feature Engineering has lead to some meaningful improvements in our model. We see ~7% reduction in MAE for demented subjects. We observe much less severe cratering offering reasonable age guesses within the 75-85 year range. There is more variance in prediction suggesting better feature utilization.

(Figure 7) In our converted group we see ~6% reduction in MAE. We also see a larger range of prediction (73-90 years). Our age gap has grown slightly becoming more negative (-2.89 years).

The improved accuracy on training data suggests better capture of age/ disease relevant features. Our model is still cratering a bit which can be observed by the horizontal banding. It remains difficult to understand why exactly the model is having difficulty recognizing bio markers related to degradation. While we are seeing improvements, this model is still not performing well enough to be consider real world ready.

5 Conclusion

In this section we will explore key findings, potential improvements, and future direction.

5.1 Key Findings

Unfortunately, the results obtained in this study leave a lot to be desired. In our first attempt, with minimal feature engineering, our model cratered toward mean age. It showed little, if any, predictive power and is apparently useless at predicting neural age. With this in mind, we took a second approach to extend the network to include some brain features such as Grey Matter, White Matter, and other Ventricle features. Unfortunately, despite providing the model with additional information, we still failed to captured truly meaningful patterns in our model. There is an argument to be made that our thresholding attempt at capturing brain features did marginally improve results, with this in mind it is plausible that more improvements down this path could yield more fruitful

results. It is also difficult to determine whether the model is having difficulty learning normal age progression in comparison with disease progression. We will explore these thoughts in the next section.

5.2 Improvements

There are three vectors of improvement that should be considered to improve results. The first issue is that our dataset is relatively small. As mentioned before, this dataset has only 150 patients and ultimately provides a little under 1400 images. With the limited number of patients and only around half of them being Non-demented, training a model because a challenging task. With a larger size dataset, our model could better generalize to less pronounced patterns.

While the small dataset size contributed to the lack of results, I was also limited by my system requirements. I trained this neural network using a relatively new Apple Macbook M4. However, tensor still lacks the support to effectively use apple silicon (mps). While I experimented with using some known supported functionality, I ultimately opted for CPU supported functionality. This led to much slower training times and less access to the more robust functionality that tensor offers. One example of this was my inability to leverage the *maxpool3d* function, a function that is supported and functional using CUDA/NVIDIA hardware. [6]

Along with the system specific limitations noted above, we could also consider our potential options with more compute power. With a more powerful GPU we could increase our batch sizes and resolutions. This would allow us to form deeper networks on higher resolution images. Higher resolution could lead to more apparent feature on the image scans. With more memory for our system we could improve training times and accuracy by leveraging strategies such as mixed precision training. We could also further expand our feature set/ data set sizes. This particular dataset was 20GB worth of image data alone. It was cumbersome for my system alone to handle. Training could improve if provided the benefits of multiple CUDA enabled GPU's and improved memory allocation.

One other potential area for improvement would be to use a more robust brain feature extraction process. While we experimented with feature extraction for things like Grey Matter, White Matter, and other Ventricle features, our results leave much to be desired. It is likely that our feature extraction is not truly extracted the features required for the neural network to learn patterns that represent an aging brain. One potential candidate for this feature extraction would be to use FSL. FSL is a comprehensive library of analysis tools for FMRI, MRI and DTI brain imaging data. [7]

The specialized tools provided by FSL would provide much more sophisticated feature engineering opportunities than our basic thresholding approaches. This tooling contains advances methods included but not limited to tissue segmentation, subcortical structure segmentation, and white matter analysis. We could perform this more advanced feature extraction and engineering as part of our data loading phase. It is likely that the documented features would help our neural network better recognize patterns across the 3D image scans rather than cratering towards average age values. [8]

5.3 Future Direction and Final Thoughts

Overall, the task of predicting brain age proved to be extremely challenging. I believe that by expanding our dataset, increasing our computational power, and leveraging more advancing feature engineering or feature extracting tooling could push this effort further in the right direction. While the task remains challenging, the benefits of success would be life changing for patients, their families, and healthcare professionals.

Acknowledgments

I would like to thank all of the students, Teaching Assistants, and Dr. Ying Ding for a wonderful semester. I found that the Ed Discussion and Discord groups helped me tremendously. This course, especially this project, were extremely thought provoking and challenging. It is typical in academia that students are given a task with strict confines. While I wish I could have produced more encouraging results, I learned a tremendous amount as I journeyed through this open ended project. I would also like to acknowledge Cursor and ChatGPT for helping me formulate ideas and implement the code used in this study. This course and project provided an invaluable learning experience. Thank you!

References

[1] Iman Beheshti. Brain age: A promising biomarker for understanding aging in the context of cognitive reserve. *medRxiv*, 01 2025. Preprint.

[2] Lin Y Zhu, Lin Shi, Yiwei Luo, Jason Leung, and Timothy Kwok. Brain mri biomarkers to predict cognitive decline in older people with Alzheimer's disease. *Journal of Alzheimer's Disease*, 88(2):763–769, 2022.

[3] D.S. Marcus, R.L. Buckner, J. Csernansky, and J.C. Morris. Open access series of imaging studies (oasis-2): Longitudinal mri data in nondemented and demented older adults. Dataset, 2010. Principal Investigators: D. Marcus, R. Buckner, J. Csernansky, J. Morris; Funded by grants: P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382.

[4] X. Wang, Y. Yang, M. Nan, G. Bao, and G. Liang. Optimum multilevel thresholding for medical brain images based on tsallis entropy, incorporating bayesian estimation and the cauchy distribution. *Applied Sciences*, 15(5):2355, 2025.

[5] Deepa V., Benson C. C., and Lajish V. L. Gray matter and white matter segmentation from mri brain images using clustering methods. *International Research Journal of Engineering and Technology (IRJET)*, 2(8):913–919, 2015.

[6] PyTorch Contributors. Add support for maxpool3d on the mps backend. GitHub issue #100674, 2023. Open issue on GitHub repository.

[7] S.M. Smith, M. Jenkinson, M.W. Woolrich, C.F. Beckmann, T.E.J. Behrens, H. Johansen-Berg, P.R. Bannister, M. De Luca, I. Drobnjak, D.E. Flitney, R. Niazy, J. Saunders, J. Vickers, Y. Zhang, N. De Stefano, J.M. Brady, and P.M. Matthews. Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, 23(S1):S208–S219, 2004.

[8] M.W. Woolrich, S. Jbabdi, B. Patenaude, M. Chappell, S. Makni, T. Behrens, C. Beckmann, M. Jenkinson, and S.M. Smith. Bayesian analysis of neuroimaging data in FSL. *NeuroImage*, 45:S173–S186, 2009.