Data Preprocessing for the Classification of Alzheimer’s Disease

Farhaan Pishori

Mentor: Dr. Juhao Wu

SLAC National Laboratory

ABSTRACT

This research focused on developing a deep learning model to improve early detection of Alzheimer’s Disease (AD), a challenging but crucial task for managing this progressive neurological disorder. The model classified patients into three categories: Alzheimer’s Disease (AD), Cognitively Normal (CN), and Mild Cognitive Impairment (MCI) using realistic MRI data from the Alzheimer’s Disease Neuroimaging Initiative (ADNI). This data, unlike overly simplified pre-curated datasets, provided essential insights into brain changes such as shrinkage and lateral ventricle enlargement, making the model more accurate and generalizable. Extensive data preprocessing, including converting MRI images, organizing them systematically, thorough image selection, and enhancing image clarity, ensured that the model could effectively identify key features linked to Alzheimer’s. The final model, combining two neural network architectures, achieved 82% accuracy, highlighting the importance of using high quality and realistic data with systematic preprocessing in developing a diagnostic tool for medical conditions like Alzheimer’s Disease.

INTRODUCTION

Alzheimer's Disease is a progressive neurodegenerative disorder that affects millions of people worldwide. Early detection and diagnosis are crucial for managing disease and improving the quality of life for patients. However, accurately diagnosing Alzheimer's, especially in its early stages, remains a significant challenge in the medical community. The need for reliable biomarkers and advanced diagnostic tools is paramount in addressing this challenge.

Biomarkers play a vital role in the detection and monitoring of Alzheimer's Disease. These biological measures provide critical information about the presence and progression of the disease, aiding in early diagnosis and informing treatment strategies. For this project, various types of biomarkers have been explored, including genetic, proteomic, and imaging biomarkers. Each type of biomarker offers unique insights into the disease, but they also come with their limitations as datasets.

In this project, the focus was on developing a deep learning model to classify Alzheimer's Disease into three categories: Alzheimer's Disease (AD), Cognitively Normal (CN), and Mild Cognitive Impairment (MCI). The project aimed to address the limitations of existing models that rely on unrealistic, overly curated datasets by using raw, realistic data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The use of realistic data is critical for developing a model that can be applied effectively in real-world clinical settings.

The decision to focus on imaging biomarkers, specifically MRI scans, was driven by the extensive historical data available and the ability of MRI images to provide detailed insights into brain changes associated with Alzheimer's. Unlike genetic and proteomic biomarkers, which often suffer from limited historical data, imaging biomarkers, like MRI scans, offer extensive information about the disease, including brain shrinkage and lateral ventricle enlargement, which are key indicators of Alzheimer's.

The project's scope also included addressing the challenge of data quality in Alzheimer's research. Early attempts to build a model using prebuilt datasets, such as those available on platforms like Kaggle, resulted in models with high accuracy but limited real-world applicability due to the unrealistic nature of the data. This highlighted the need for a model based on raw MRI images that could be genuinely useful in clinical diagnostics.

In summary, this project sought to develop a deep learning model for Alzheimer's classification using realistic, high-quality MRI data. By focusing on imaging biomarkers and leveraging the extensive historical data from the source ADNI, the project aimed to create a model that not only achieves high accuracy but also has practical applications in the early detection and diagnosis of Alzheimer's Disease.

BIOMARKER SELECTION

Selecting appropriate biomarkers was a critical step in developing an effective model for detecting Alzheimer's Disease. Biomarkers are indicators that provide insights into the presence or severity of a disease. For this project, we considered three main types of biomarkers: genetic, proteomic, and imaging.

Genetic and Proteomic Biomarkers

Genetic biomarkers, such as the Apolipoprotein E (APOE) ε4 allele, are well-known in Alzheimer's research for their role in indicating a genetic predisposition to the disease. The presence of the APOE ε4 allele increases the risk of developing Alzheimer's and often correlates with an earlier onset diagnosis[1]. Proteomic biomarkers, which involve the study of protein levels in the brain, were also considered. Tau and Beta-Amyloid proteins are key proteomic markers associated with Alzheimer's. Elevated levels of these proteins contribute to the formation of neurofibrillary tangles and amyloid plaques, hallmark features of Alzheimer's Disease[2]. However, the use of genetic and proteomic biomarkers is constrained by the relatively limited historical data available, which poses challenges for large-scale model training.

Imaging Biomarkers

Imaging biomarkers, especially those derived from Magnetic Resonance Imaging (MRI), offer valuable insights into structural changes in the brain associated with Alzheimer's Disease. MRI scans can detect critical signs of neurodegeneration, such as brain shrinkage or atrophy, which are commonly observed in Alzheimer's patients. The lateral ventricle, a specific region of interest, often shows enlargement in individuals with Alzheimer's, indicating the loss of surrounding brain tissue[3]. Additionally, unlike genetic and proteomic biomarkers, MRI imaging has been more widely used in clinical settings and has much more accessible data. This availability of historical data and the ability of MRI scans to clearly illustrate brain changes relevant to Alzheimer's were the reasons why MRI images were chosen as the primary biomarker for this study.

DATA COLLECTION AND PREPROCESSING

The dataset used for this project was sourced from the ADNI database. It consisted of 1,716 MRI scans, totaling 84,423 individual images. The acquisition plane for these images is the axial angle and the series description is Axial PD/T2 FSE.

Data Source

A key aspect of this project was the decision to use realistic data sourced from ADNI rather than relying on pre-curated datasets, such as those available on platforms like Kaggle. While curated datasets often allow for quick model development and high accuracy due to their carefully selected samples, they do not accurately represent the complexity and variability of real-world data. For instance, a preliminary model built using a Kaggle dataset achieved a high accuracy of 96%, but upon closer examination, it became clear that this data was overly simplified and not representative of actual clinical scenarios[4].

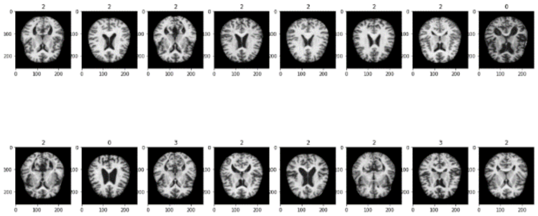


Figure 1: Kaggle Dataset [4]

In contrast, the raw MRI data from ADNI presented a more accurate reflection of the diversity found in clinical settings, including variations in image quality, patient demographics, and disease progression. Using such realistic data is crucial to be able to develop a model that could generalize effectively to new and unseen cases[5].

Acquisition Plane and Series Description

The MRI scans were captured using the Axial acquisition plane, selected for several reasons. The Axial plane provided the most data compared to other planes (the Coronal or Sagittal angles). Additionally, during a consultation with Dr. Steven Z. Chao, a neurologist from the Stanford Neurological Department, he emphasized that the Axial plane offers greater variation in brain sections relevant to Alzheimer’s, particularly enhancing the visibility of the lateral ventricle, a critical region for Alzheimer’s diagnosis.

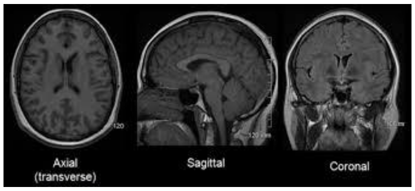


Figure 2: The Acquisition Plans of MRI scans [6]

Within the Axial plane, the series description chosen was Axial PD/T2 FSE. This series was selected for two primary reasons: first, it offered better color differentiation between different regions of the brain, making it easier to identify critical areas relevant to Alzheimer’s compared to other series like the 3D plane localizer. Second, it also had the most extensive data available compared to other series descriptions, providing a large enough dataset for model development.

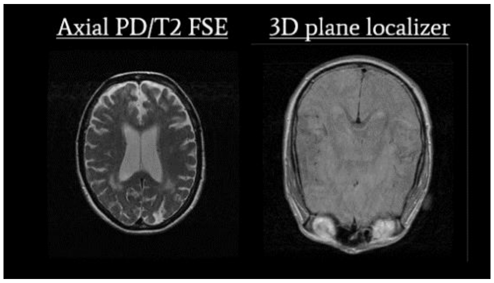


Figure 3: Series Description Comparison [5]

Image Conversion and Organization

The MRI scans obtained from ADNI were originally provided in DICOM format, a standard format for storing and transmitting medical images. However, for use in machine learning, these images needed to be converted into a more manageable format. To achieve this, a Python script was written to convert the DICOM files into JPEG format. This conversion was necessary to make the images easily accessible for further processing and model training.

Once converted, the images were organized into a machine-readable format. This involved renaming the files to ensure they were correctly ordered according to the progression of the brain slices and categorizing them based on their corresponding class (AD, CN, or MCI). A

CSV file provided by ADNI, containing classification information for each MRI scan was used to guide this reorganization process. This step was essential to maintain consistency and accuracy in the data used for training the model.

Image Selection

Initially, all 50 MRI slices per scan were used for model training. However, this approach introduced significant challenges due to the excessive noise present in the data, which led to a low model accuracy of 35%. This accuracy level suggested that the model was essentially guessing the classifications, highlighting the need for a more refined approach to data selection.

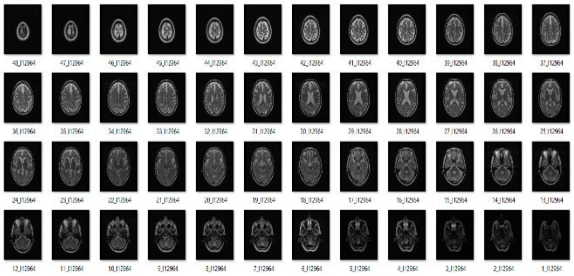
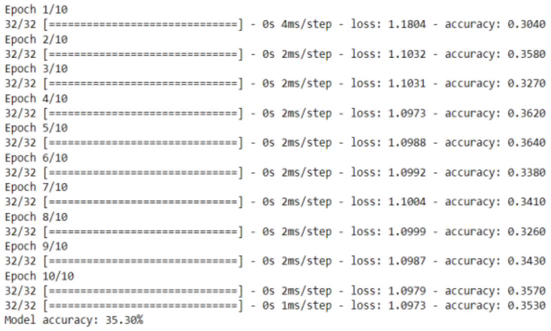


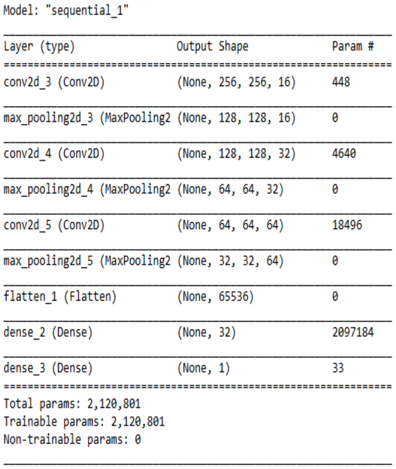
Figure 4: All 50 slices of 1 MRI scan



```
Epoch 1/10
32/32 [==============================] - 0s 4ms/step - loss: 1.1804 - accuracy: 0.3040
Epoch 2/10
32/32 [==============================] - 0s 2ms/step - loss: 1.1032 - accuracy: 0.3580
Epoch 3/10
32/32 [==============================] - 0s 2ms/step - loss: 1.1031 - accuracy: 0.3270
Epoch 4/10
32/32 [==============================] - 0s 2ms/step - loss: 1.0973 - accuracy: 0.3620
Epoch 5/10
32/32 [==============================] - 0s 2ms/step - loss: 1.0988 - accuracy: 0.3640
Epoch 6/10
32/32 [==============================] - 0s 2ms/step - loss: 1.0992 - accuracy: 0.3380
Epoch 7/10
32/32 [==============================] - 0s 2ms/step - loss: 1.1004 - accuracy: 0.3410
Epoch 8/10
32/32 [==============================] - 0s 2ms/step - loss: 1.0999 - accuracy: 0.3260
Epoch 9/10
32/32 [==============================] - 0s 2ms/step - loss: 1.0987 - accuracy: 0.3430
Epoch 10/10
32/32 [==============================] - 0s 2ms/step - loss: 1.0979 - accuracy: 0.3570
32/32 [==============================] - 0s 1ms/step - loss: 1.0973 - accuracy: 0.3530
Model accuracy: 35.30%
```

Figure 5: First iteration of the model with 35% accuracy

To improve the model’s accuracy, the focus was narrowed to the lateral ventricle, a region identified as critical in Alzheimer’s diagnosis. A custom Convolutional Neural Network (CNN) with 2 million parameters and an accuracy of 97% was developed to sift through all 50 slices of each MRI scan and identify the images where the lateral ventricle was most prominently displayed. From these, the middle four slices were selected for model training, as they provided the most consistent and relevant information. This selection process was essential to reduce noise and enhance the accuracy of the model.



```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_3 (Conv2D)            (None, 256, 256, 16)      448
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 128, 128, 16)      0
_________________________________________________________________
conv2d_4 (Conv2D)            (None, 128, 128, 32)      4640
_________________________________________________________________
max_pooling2d_4 (MaxPooling2 (None, 64, 64, 32)        0
_________________________________________________________________
conv2d_5 (Conv2D)            (None, 64, 64, 64)        18496
_________________________________________________________________
max_pooling2d_5 (MaxPooling2 (None, 32, 32, 64)        0
_________________________________________________________________
flatten_1 (Flatten)          (None, 65536)             0
_________________________________________________________________
dense_2 (Dense)              (None, 32)                2097184
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 33
=================================================================
Total params: 2,120,801
Trainable params: 2,120,801
Non-trainable params: 0
_________________________________________________________________
```

Figure 6: The CNN model to separate Lateral Ventricle specific Images

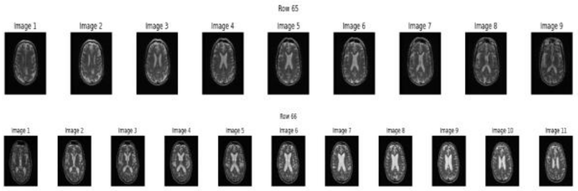


Figure 7: The Lateral Ventricle specific Images

IMAGE ENHANCEMENT

To further refine the selected MRI images and enhance the focus on critical brain regions, specific image enhancement techniques were applied. The goal was to improve the clarity and uniformity of the images, making them more suitable for the deep learning model.

Zooming into the Lateral Ventricle

Given the importance of the lateral ventricle in Alzheimer’s diagnosis, the first step in image enhancement was to zoom into this specific area of the brain. By zooming into this area, the model could better focus on the features most relevant to the disease, reducing the influence of less pertinent regions of the brain and minimizing noise.

Application of Filter Enhancements

After focusing on the lateral ventricle, the next step was to apply filters on the images to enhance the brain. The first applied filter was gamma correction. Gamma correction was used to adjust the brightness and contrast of the images, making the important features within the lateral ventricle more distinguishable. This adjustment helped ensure that variations in image exposure did not affect the model’s ability to accurately interpret the data.

Following gamma correction, adaptive thresholding was applied to sharpen the edges within the images. Adaptive thresholding was particularly effective in creating a uniform appearance across the dataset, especially given that MRI scans can vary in quality and clarity. By enhancing the edges, this technique allowed the model to more consistently identify and analyze the structural details of the lateral ventricle, which are critical for Alzheimer’s detection.

These enhancement techniques improved the quality and uniformity of the data, ensuring that the model was trained on images that were not only clear and focused but also consistent across the entire dataset.

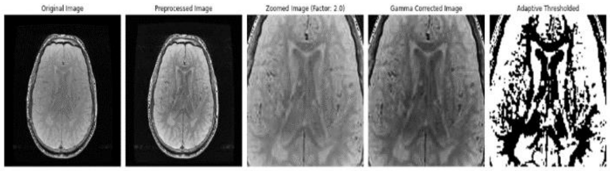
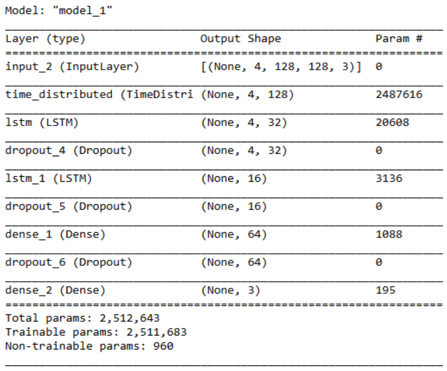


Figure 8: Image Enhancements

MODEL DEVELOPMENT AND RESULTS

CRNN Model

The first model developed during this project was a Convolutional Recurrent Neural Network (CRNN). This model architecture was chosen for its ability to handle sequential data, making it particularly well-suited for analyzing the progression of MRI slices. The CRNN model was designed with approximately 2.5 million parameters, which allowed it to capture a significant amount of information from the MRI images while maintaining computational efficiency.



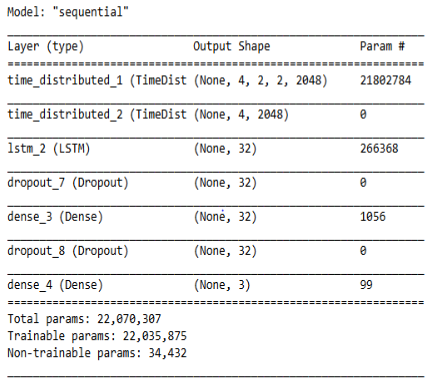
```
Model: "model_1"
_________________________________________________________________
Layer (type)                Output Shape              Param #
=================================================================
input_2 (InputLayer)        [(None, 4, 128, 128, 3)]  0
_________________________________________________________________
time_distributed (TimeDistri (None, 4, 128)           2487616
_________________________________________________________________
lstm (LSTM)                 (None, 4, 32)             20608
_________________________________________________________________
dropout_4 (Dropout)         (None, 4, 32)             0
_________________________________________________________________
lstm_1 (LSTM)               (None, 16)                3136
_________________________________________________________________
dropout_5 (Dropout)         (None, 16)                0
_________________________________________________________________
dense_1 (Dense)             (None, 64)                1088
_________________________________________________________________
dropout_6 (Dropout)         (None, 64)                0
_________________________________________________________________
dense_2 (Dense)             (None, 3)                 195
=================================================================
Total params: 2,512,643
Trainable params: 2,511,683
Non-trainable params: 960
_________________________________________________________________
```

Figure 9: CRNN architecture

In terms of performance, the CRNN achieved an accuracy of 75% in classifying the MRI images into the three categories. While this accuracy was promising, it also highlighted areas for improvement, particularly in refining the model’s ability to accurately distinguish between the subtle differences in brain structures associated with these conditions.

Inception Net Model

Following the development of the CRNN, the decision was made to explore a different model architecture to further enhance classification accuracy. The Inception Net model was chosen due to its proven effectiveness in handling complex image data through its unique architecture, which allows for the simultaneous application of multiple convolutional filters.



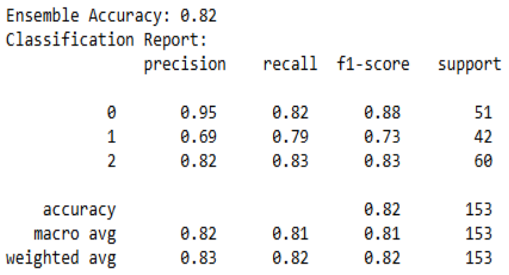
```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
time_distributed_1 (TimeDist (None, 4, 2, 2, 2048)     21802784
_________________________________________________________________
time_distributed_2 (TimeDist (None, 4, 2048)           0
_________________________________________________________________
lstm_2 (LSTM)                (None, 32)                266368
_________________________________________________________________
dropout_7 (Dropout)          (None, 32)                0
_________________________________________________________________
dense_3 (Dense)              (None, 32)                1056
_________________________________________________________________
dropout_8 (Dropout)          (None, 32)                0
_________________________________________________________________
dense_4 (Dense)              (None, 3)                 99
=================================================================
Total params: 22,070,307
Trainable params: 22,035,875
Non-trainable params: 34,432
_________________________________________________________________
```

Figure 10: Inception Net Architecture

The Inception Net model was significantly more complex, incorporating approximately 22 million parameters. This increase in model complexity allowed it to capture finer details within the MRI images, particularly in the critical regions such as the lateral ventricle. The Inception Net model achieved an accuracy of 81.7%, a substantial improvement over the CRNN model. This result demonstrated the effectiveness of the Inception Net architecture in accurately classifying MRI images across the three categories.

Model Ensemble

To leverage the strengths of both the CRNN and Inception Net models, an ensemble approach was implemented. By combining the predictions from both models, the ensemble aimed to improve overall accuracy by capitalizing on the complementary strengths of each architecture. The ensemble model effectively balanced the CRNN's ability to analyze sequential data with the Inception Net's capacity for detailed image processing. The ensemble approach resulted in an overall accuracy of 82%.



```
Ensemble Accuracy: 0.82
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.82      0.88        51
           1       0.69      0.79      0.73        42
           2       0.82      0.83      0.83        60

    accuracy                           0.82       153
   macro avg       0.82      0.81      0.81       153
weighted avg       0.83      0.82      0.82       153
```

Figure 11: Ensemble Accuracy

FUTURE DIRECTIONS

Exploring Additional Datasets

To further improve the model's accuracy, one of the next steps involves expanding the dataset by incorporating additional reputable sources such as the Oasis-1 and Oasis-2 datasets[7]. These datasets, like ADNI, provide rich collections of MRI scans and other relevant data. By integrating these datasets, we can enhance the diversity and volume of data available for training, which is likely to improve the model's ability to generalize to a wider range of cases, thereby increasing its accuracy and reliability.

FreeSurfer Image Enhancement

Another area of exploration involves the use of advanced image enhancement techniques. One promising tool is FreeSurfer, a widely used software suite for processing and analyzing brain MRI images. FreeSurfer specializes in segmenting and labeling different brain regions, which could significantly enhance the features extracted from MRI scans. By incorporating FreeSurfer into our preprocessing pipeline, we can further refine the MRI images, potentially leading to even better model performance by providing more detailed and accurate representations of brain structures.

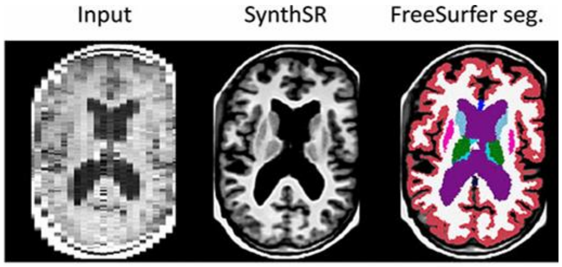


Figure 12: FreeSurfer Enhancement [8]

Incorporating Other Biomarkers

Finally, to create a more comprehensive diagnostic tool, there is significant potential in integrating additional biomarkers, such as genetic and proteomic markers. Specifically, incorporating the APOE ε4 allele and Tau protein levels could provide valuable information that complements the imaging data. By combining these biomarkers with MRI data, we can develop a multi-modal model that offers a more holistic assessment of Alzheimer's Disease, potentially leading to earlier and more accurate diagnoses. This approach would allow us to leverage the strengths of both imaging and non-imaging biomarkers, making the model a more powerful tool in the fight against Alzheimer’s.

CONCLUSION

This project highlighted the role of data preprocessing in the development of a deep learning model aimed at classifying Alzheimer’s Disease into three categories: Alzheimer’s Disease, Cognitively Normal, and Mild Cognitive Impairment. By leveraging realistic, raw MRI data from ADNI, the project ensured that the resulting model would be applicable in real world settings. The extensive preprocessing pipeline comprising of the conversion and organization of DICOM images, strategic selection of key MRI slices, and the application of advanced image enhancement techniques was significant in addressing the inherent variability and noise in raw data.

These preprocessing efforts laid a strong foundation for the model’s success. The focus on critical brain regions, particularly the lateral ventricle, and the enhancement of image clarity and uniformity significantly contributed to the model’s ability to generalize across diverse clinical data. The final model, developed using an ensemble of CRNN and Inception Net architectures, achieved an accuracy of 82%. This performance displays the effectiveness of the preprocessing techniques used, as well as the integration of multiple model architectures.

The project’s success shows us that investing in thorough data preprocessing is essential for developing high-performing deep learning models, particularly in complex fields like medical imaging. The quality of the data used in training is directly linked to the model’s ability to provide reliable and accurate classifications.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my mentor, Dr. Juhao Wu, for his invaluable guidance, support, and encouragement throughout this project. His expertise and insights were instrumental in shaping the direction of my research and in overcoming the challenges encountered along the way.

Additionally, my sincere thanks go to the staff at SLAC National Accelerator Laboratory, the SULI program, especially Ms. Hillary Freeman, and my fellow interns for providing a stimulating research environment and the resources necessary to complete this project. The opportunity to engage in meaningful research at this prestigious institution has been an invaluable experience in my academic and professional development.

Finally, I would like to thank the Department of Energy’s SULI program for making this internship possible. The program not only allowed me to contribute to important research but also facilitated my growth as a researcher, equipping me with skills and knowledge that will be essential in my future endeavors.

REFERENCES

[1] Scheltens, P., De Strooper, B., Kivipelto, M., Holstege, H., Chételat, G., Teunissen, C. E., Cummings, J., & van der Flier, W. M. (2021). Alzheimer's disease. *Lancet (London, England)*, *397*(10284), 1577–1590. https://doi.org/10.1016/S0140-6736(20)32205-4

[2] Bloom G. S. (2014). Amyloid-β and tau: the trigger and bullet in Alzheimer disease pathogenesis. *JAMA neurology*, *71*(4), 505–508. https://doi.org/10.1001/jamaneurol.2013.5847

[3] How biomarkers help diagnose dementia | National Institute on Aging, https://www.nia.nih.gov/health/alzheimers-symptoms-and-diagnosis/how-biomarkers-help-diagnose-dementia.

[4] Dubey, S. (2019). Alzheimer’s Dataset (4 class of images) [Dataset]. In *Kaggle Dataset*. Kaggle. https://www.kaggle.com/datasets/tourist55/alzheimers-dataset-4-class-of-images/suggestions?status=pending&yourSuggestions=true

[5] Alzheimer’s disease neuroimaging initiative. ADNI. (n.d.). https://adni.loni.usc.edu/

[6] *Positioning terminology*. Clark’s Radiography. (n.d.). https://raditechies.blogspot.com/p/positioning-terminology.html

[7] OASIS-1: Cross-Sectional: https://doi.org/10.1162/jocn.2007.19.9.1498

[8] Juan E. Iglesias *et al.* SynthSR: A public AI tool to turn heterogeneous clinical brain scans into high-resolution T1-weighted images for 3D morphometry. DOI:10.1126/sciadv.add3607