Exploring Alzheimer's Disease Patterns Using Multimodal Machine Learning

1. Problem

Alzheimer's is a progressive brain disorder that slowly worsens cognitive skills. It is the most common type of dementia and affects a significant portion of elderly above the age of 85. Early detection of Alzheimer's is very difficult, as many of its symptoms are easily mistaken for normal aging (forgetfulness, mild confusion, etc). There are also not enough diagnostic tests, as Alzheimer's affects a wide span of variables that aren't routinely taken (biomarkers, imaging, etc). However, this doesn't mean that early detection of Alzheimer's isn't important. Treatments have been proved to be the most effective when the least brain damage has been done, as it slows the disease's progression significantly. MRI, PET biomarkers, and demographic data are very important when detecting Alzheimer's disease because they provide insight on structural, molecular, and patient-specific levels, allowing for high diagnostic accuracy. MRI scans show brain atrophies and other structural damage. AD specifically causes the shrinking of the hippocampus and entorhinal cortex – two of the main areas for memory. PET biomarkers show amyloid and tau levels. Negative amyloid levels eliminate the possibility of AD, while high concentration areas of tau proteins may indicate AD. Finally, demographic and clinical data are essential because they provide context to the other more scientific data. They also include cognitive test scores that are crucial when determining the level of dementia and the possibility of AD. This is a real problem in the medical field because there are no definitive tests that truly detect Alzheimer's. The diseases' complexity and symptoms overlap with many others so it is hard to differentiate and detect early.

2. Research Question: Can machine learning models trained on MRI imaging and clinical biomarker data capture patterns associated with Alzheimer's disease?

2.2 Objective:

The objective of this project was to develop and evaluate ML models based on image and clinical data, and to interpret their strengths and limitations through an interactive interface. There is no one give-away symptom for Alzheimer's. Instead, there are multiple factors that need to be taken into account – MRI scans, PET biomarkers, cognitive scores, etc. A multimodal AI approach is appropriate in this scenario because it can take in different types of data and train accordingly. Instead of just focusing on one variable, it has the ability to focus on many.

3. System Overview

The MRI model takes in a singular MRI slice of the brain from the axial view and outputs the probabilities for 4 classes of dementia – Non Demented, Very Mild Demented, Mild Demented, and Moderate Demented. The clinical/biomarker model takes in multiple features (amyloid biomarker, tau biomarker, MMSE scores, education, gender) and classifies the subject into one of three classes – Cognitively Normal, Mildly Cognitive Impaired, and Alzheimer's Disease. The Streamlit interface that ties all of this together is a simple dashboard. When entering the dashboard, the user can choose one of the two models and then fill out the necessary inputs before getting a report of their predictions.

4. Data Sources

The dataset used for the MRI model was obtained through Hack4Health, who got it from Kaggle. This is a public dataset that anyone can access and use. It has 5120 patients, and for each patient there is a corresponding brain image and dementia classification (Non Demented, Very Mild Demented, Mild Demented, Moderate Demented). Unlike the MRI dataset, the dataset used for the clinical model is from Alzheimer's

Disease Neuroimaging Initiative (ADNI). In order to obtain access to this dataset, the team had to submit registration and gain approval. Due to this, the included repository does not include the dataset and instead has a preloaded model. This dataset has six features: AMY_BIOMARKER, TAU_BIOMARKER, MMSE_TOTAL, CDR_SB, SEX, and EDUC. There is also a diagnosis column that classifies the subject into three classes (Cognitively Normal, Mildly Cognitive Impaired, and Alzheimer's Disease). The MRI dataset came pre-split into a train file and a test file. During training, this train file was further split between train/val on a 80/20 split. The ADNI dataset required the merging of numerous datasets into one main one. The basis for merging was the date of the clinical visit and patient, as this kept data in the same timeframe and for the same patient linked together. After creating one singular dataset from the numerous initial ones in the training file, this was saved for future use in the evaluation file. As described above, the team had to apply to get access to the ADNI dataset, so it is not included in the repository due to data use and access policies. In order to gain access themselves, one must first register through ADNI and then download the necessary files before putting them into the correct slot in the directory.

5. Model Methodology

5.1 MRI Model

This model takes an MRI slice of the brain from the axial view. The bytes in this image are converted to pixels to aid the model when training and predicting. This model is a CNN model made with TensorFlow. It has 3 convolution layers with input shapes matching that of the inputted images, 128x128x1. The filters go from 32 to 64 to 128, doubling by each layer. After these layers, the model does MaxPooling and then there are two dense layers. The final output layer is a softmax layer with the objective of classifying into 4 classes (hence output shape of 4). All hidden layers have activation 'relu.' Finally, during fitting the model using the adam optimizer. First, the MRI images were converted from raw bytes to gray scale pixel arrays with the input shape of 128x128x1. Pixel values were also normalized to a range of 0-1. Before adding noise, the training dataframe was split 80/20 between training and validation. A separate test set was reserved and not used during training/validation. To prevent overfitting, Gaussian noise augmentation was applied to the training images. During training, the model used the adam optimizer with categorical cross-entropy loss and label smoothing, overall mitigating any overconfident predictions for noisy images. Finally, evaluation metrics were only computed on the separate test set, including weighted F1 score, precision, recall, and confusion matrices.

5.2 Clinical Model

The input for this model were tabular clinical and biomarker features. These included an amyloid PET biomarker, a tau PET biomarker, the total score on the cognitive MMSE test, the Clinical Dementia Rating score, sex, and years of education. The output was a classification into one of three classes: CN (Cognitively Normal), MCI (Mildly Cognitive Impaired), and AD (Alzheimer's Disease). The clinical model was implemented using XGBoost, a gradient-boosted decision tree ensemble that works well on biomedical data. The model was trained as a multi-class classifier (which is appropriate for classifying into three classes) by using the soft probability objective (multi:softprob). To prevent data leakage and ensure temporal consistency, all clinical and biomarker measurements were merged and aligned based on the date of the subject's diagnosis using the closest available observation in a fixed time window. From there, the dataset was split into training, validation, and a held-out test set using a stratified train/validation/test split. During training, gradient boosting with early stopping was implemented so that the model would stop once performance stopped increasing. This prevented overfitting. Finally, predictions were generated by using the best validation performance and

evaluation metrics were computed only on the held-out test set (weighted F1 score, precision, recall, and confusion matrix). These helped us evaluate our model's performance and make any necessary changes.

6. Evaluation and Results

When evaluating, both models used a held-out test set that the model never saw during training. This evaluation was done after training was complete and in a completely separate file. Next, I collected the following metrics: accuracy, weighted F1 score, precision, recall, and confusion matrices. The MRI CNN achieved a weighted F1 score of 0.971, with a very strong sensitivity for non-demented and very mild dementia classes. The clinical XGBoost model achieved a weighted F1 score of 0.906, with the best performance in CN and AD classifications. Finally, an analysis of the confusion matrix showed that adjacent disease stages proved to have the most misclassifications (MCI vs AD, Moderate vs Mild, etc). Detailed performance metrics and ethical considerations are in the accompanying model cards.

7. Application Interface

To tie everything together, the Streamlit dashboard was utilized to help the user understand the different models and how to use them. When the user first logs onto the page, they are presented with the title and a small note describing the website. Then, there are two buttons – one for each model – that describe the model in a short blurb. When the user clicks on the button for the MRI model, they are redirected to a page where they can upload their MRI image. After the model classifies the image into different classes, they are presented with a report. This report includes a primary and secondary classification, a bar chart, uncertainty/confidence scores, and a dataframe with the separate probabilities for each class. On the clinical model page, this similar report is reflected, but instead of uploading an MRI image slice the user fills out a quick form. It is important for the user to be able to interact with their results in multiple ways (charts/frames/numbers) so that 1) they can understand the results, 2) they can see that they aren't 100% into one class and there is variability, and 3) it can provide insight into how the model works for future research.

8. Limitations and Ethics

The clinical dataset from the ADNI cohort may not be demographically representative of the entire global population, and it also may have selection bias when considering research volunteers. Because of this, the model performance may vary depending on the demographic groups that use it. Additionally, the MRI dataset only has a specific image included, and it uses 2D slices instead of full 3D volumes. This means that the model only has one piece of the puzzle, and could lead to overgeneralizing. In no cases can this model be considered a diagnosis. Even though its accuracy is high, there is always a chance that it can identify a false positive/negative, which could be detrimental in a clinical setting. Therefore, this model must only be used as a research aide, and any predictions it makes must be verified by a human in a medical position before any actions are taken.

9. Conclusion and Future Work

This project shows the vast amount of variables when it comes to detecting Alzheimer's. It highlights the need for a multimodal approach, as everything else would be overgeneralized. This project could be extended by combining the two MRI and clinical models into one cohesive model. This would improve its accuracy and also prevent any generalization. Additionally, other factors such as sleep, speech, and movement would be important to add, as they all have a role in Alzheimer's. Overall, this was a meaningful research direction because it brought attention to the need for early detection and it created a useful research tool that utilizes not only MRIs but also clinical and biomarker data.