Early Detection of Alzheimer's Disease Using Machine Learning

Ben Searchinger

Problem Description

Alzheimer's Disease (AD) is a neurodegenerative disease marked by severe cognitive decline over a period of time, often resulting in death. Roughly 10% of adults over age 65 in the United States suffer from AD currently, and this proportion is only expected to increase in the near future (1). Despite its prevalence, the pathology and progression of AD are relatively poorly understood. Other than age, the most significant risk factor associated with developing AD is the presence of the apolipoprotein-e4 (APOE-e4) gene, however its presence cannot be used to predict an eventual AD diagnosis (1). The most prominent brain changes observed in AD patients are the accumulation of abnormal levels of beta-amyloid and tau proteins, as well as a gradual loss of overall brain volume, especially in the hippocampus (2). However, beta-amyloid and tau proteins also accumulate in healthy aging brains, so their presence alone cannot be used to diagnose AD. Cerebrospinal fluid and tau PET scans can be used to measure beta-amyloid and tau levels, and MRI scans can determine regional volumetric atrophy over time. These brain scan measurements coupled with abnormal performance on cognitive assessments can be used to diagnose AD.

The Alzheimer's Disease Neuroimaging Initiative (ADNI) was created primarily to find an early detection method using demographic and biochemical data from

AD patients and cognitively normal (CN) controls. This initiative has significantly progressed scientists' collective understanding of AD, but many AD patients and their doctors won't have access to a full suite of bioimaging and genotyping technology. Tau PET scans in particular are prohibitively expensive for many patients, and are unlikely to be reimbursed through insurance providers, especially Medicaid (3). MRI scans are typically more accessible than PET scans, although multiple scans over time are needed to assess regional volumetric atrophy (4).

Our goal is to create a model that 1) predicts amyloid positivity based on more readily available patient information and biochemical data, thereby selecting patients at a higher risk for developing AD as candidates for MRI and tau PET scans. Then 2) using those brain scans, we will apply deep learning models to predict which of these high-risk patients will eventually develop AD, which can be thought of as early detection of AD. This framework is illustrated in Figure 1. This two stage approach should reduce the use of unnecessary invasive and costly brain scans for low-risk patients, and improve predictive diagnostic power for patients who are deemed high-risk. This framework can be used to gather more meaningful data in clinical trials where funding may only cover a fixed number of scans, and in practice to predict an eventual AD diagnosis before the disease has fully manifested.

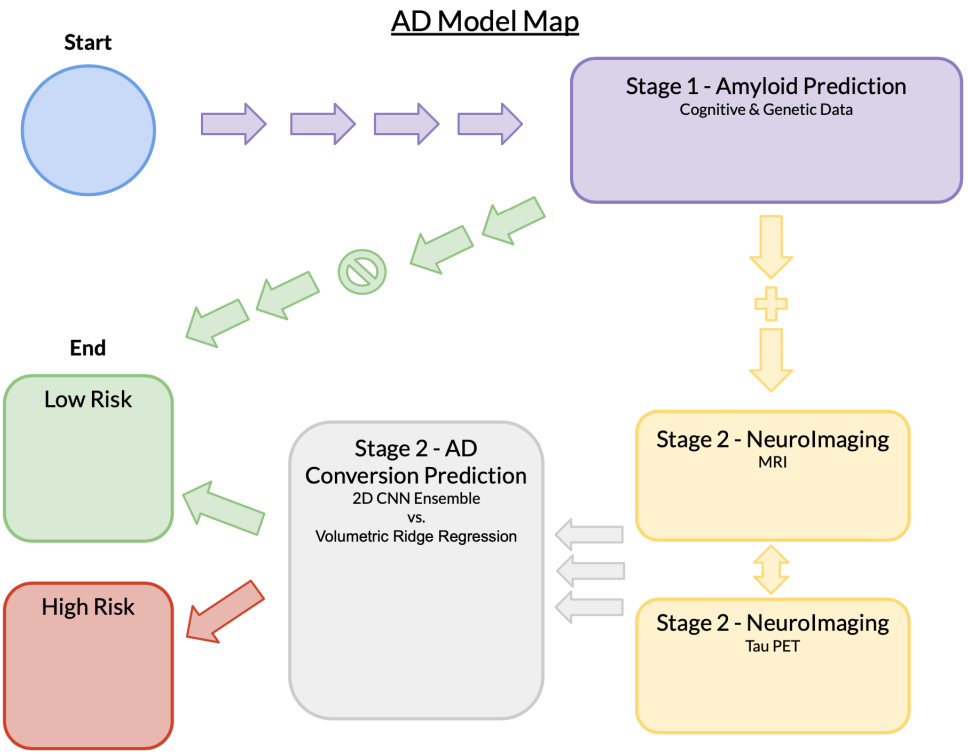

Figure 1. Two Stage model framework

Data Description

ADNI began in 2004, and has enrolled 1,821 patients across four phases. The study adds new patients in each phase and attempts to retain existing patients through each phase, but unfortunately the dropout rate of the study is high due to the age and overall health of many of these patients. As a result, most patients have demographic information, APOE4-e4 genotyping data, cognitive assessment scores, amyloid positivity status, and at least one brain scan recorded.

Stage 1

Figure 2 shows the number of patients who converted to AD by their baseline diagnosis and amyloid positivity status, which are the two target variables corresponding to

the two stages of our model. Note that there n = 1,641 patients here due to missing AD conversion or amyloid positivity observations.

AD.Conversion	BaselineDX	Total	Amyloid.Pos	Amyloid.Neg	Percent.Pos
0	AD	275	253	22	0.920
0	CN	364	181	183	0.497
0	EMCI	285	155	130	0.544
0	LMCI	219	139	80	0.635
0	SMC	182	83	99	0.456
1	CN	28	24	4	0.857
1	EMCI	54	47	7	0.870
1	LMCI	232	203	29	0.875
1	SMC	2	1	1	0.500

Figure 2. AD conversion counts by amyloid positivity status given baseline diagnosis

We observe that overall, amyloid positivity is strongly associated with both baseline AD diagnosis and eventual AD conversion, although it is not predictive of either category. That is, amyloid positivity cannot predict a current or future AD diagnosis, but current or future AD patients are much more likely than not to be amyloid positive. This relationship makes amyloid positivity a natural filtering criteria for low vs. high risk of AD conversion, which is the goal of stage 1 in our model.

The new acronyms in Figure 2 represent Early Mild Cognitive Impairment (EMCI), Late Mild Cognitive Impairment (LMCI), and Significant Memory Concern (SMC). Typically, the progression of AD is represented as $CN \rightarrow EMCI \rightarrow$

$LMCI \rightarrow AD$, although within ADNI, patients are typically only rediagnosed if they convert to AD. The diagnostic criteria for these categories is primarily based on cognitive assessments, and SMC represents patients who scored normally on cognitive assessments but showed cognitive impairment in everyday life (5). For the purposes of our research, we ignore the SMC category due to the lack of observations, and consider only CN, AD, and combined EMCI and LMCI as MCI.

Stage 2

Brain scan data adds a degree of complexity, in terms of both file formatting and study design. Some patients have multiple brain scans at different time points recorded, although few patients have both MRI and tau PET scans, since tau PET scans were not included in the study until the most recent phase. Figure 3 shows the total number of patients and scans for each modality, accounting for the overlap between the two groups. Only CN and AD patients are considered for the non-overlapping groups (i.e. MRI, no tau PET and tau PET, no MRI) since these two groups are used for training and validation in a classification model. The overlapping groups (MRI, has tau PET and tau PET, has MRI) are comprised of CN and MCI patients, but here the target variable becomes AD conversion, so the baseline diagnoses are largely ignored. In Figure 3, the 28 'CN' patients actually represent patients who did not convert to AD, regardless of baseline diagnosis. This is the holdout set, and the objective is to predict future AD conversion given baseline scans.

	Unique Patients	Total Scans	CN Patients
MRI, no tau PET	404	1,478	876
tau PET, no MRI	440	641	542
MRI, has tau PET	36	176	28
tau PET, has MRI	36	57	28

Figure 3. Scan and patient counts by image modality

Given the limited number of scans available, we treat each scan independently, even though some scans are from the same patient at different points in time. This approach is not uncommon when working with ADNI data (6), and we carefully ensure that the training/validation and holdout datasets are constructed based on patient ID to avoid any overlapping patients between the two datasets.

MRI scans and tau PET scans inherently encode different information about the brain. MRI technology measures energy given off by realigning molecules within the body using a strong magnet. This creates a clear structural image. In contrast, PET scans use radioactive tracers injected in the bloodstream that bind to specific cell types to show regional depositions. For tau PET scans, the tracer 18F-flortaucipir binds to tau proteins in the brain, showing the location and concentration of tau buildups. Both of these imaging technologies create a series of 2 dimensional slices that can be stitched together to infer the 3 dimensional structure of the brain. They are both extremely sensitive to noise created by differences in scanner machine models, head position, and other externalities, so sophisticated image processing techniques must be applied to the raw scans to allow for comparisons between patients. ADNI provides image scans with different degrees of processing, and we use FreeSurfer Cross-Sectional Processing brainmask images for MRI scans and AV1451 Coreg, Avg, Std Img and Vox Siz, Uniform Resolution images for PET scans. These levels of processing, respectively, provide the highest degree of noise reduction and consistency between patients. Figure 4 shows some central slice examples of these images. Note that the cortical decay in the AD MRI images is clearly visible, and the AD tau PET image shows signs of tau protein deposits on the hippocampus in slice 30.

Regional brain volume measurements and tau accumulation can be computed from MRI and tau PET scans directly, but this method requires additional sophisticated and time-consuming image processing. As a result, this information is available for only a small subset of patients within the ADNI study.

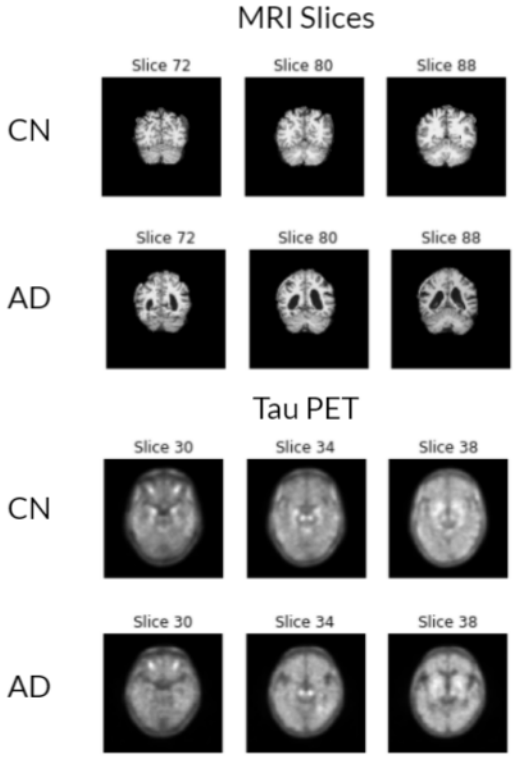

Figure 4. Central brain slices for MRI and tau PET scans

Methods

Stage 1

The goal of the Stage 1 model is to predict amyloid positivity for individuals to then rule out or recommend additional costly and invasive medical procedures that more accurately predict their risk of eventually developing AD. With this in mind, we only use patient data that can be gathered through questioning, medical history, and blood draws, while intentionally omitting data from lumbar punctures and brain scans. This approach is very similar to other research in the field (7).

We first select n=1,565 patients with data on age, education, gender, cognitive assessments, APOE-e4 genetic markers, and amyloid positivity status. This becomes

the training dataset for a series of logistic regressions, where the target variable is amyloid positivity status. Note that the baseline diagnosis (i.e. CN, MCI, AD) is not considered, as the goal of Stage 1 is to create a generalized model that can be applied to future patients before they receive these diagnoses. We use 5 fold cross-validation to select the best of 7 logistic regression models that account for different combinations of cognitive assessment scores, age, education and APOE-e4 genetic markers. Discussion of these results follows in the next section, and the equation of the model with the highest cross-validated training accuracy score is as follows:

$$Amyloid = \beta_0 + \beta_1 * Age + \beta_2 * Male + \beta_3 * CogScore + \beta_4 * APOE + \beta_5 * Age * CogScore * Educ$$

This is somewhat of a simplification, since multiple cognitive assessment scores are considered independently, and APOE-e4 genetic markers are measured based on two alleles, but the equation above summarizes the information captured by our logistic regression. We then run our n=36 patient holdout set through the trained logistic regression to predict their amyloid positivity statuses, and patients who were predicted to be amyloid positive comprise the filtered holdout set for Stage 2.

Stage 2

The goal of Stage 2 is to predict AD conversion for patients who were flagged as high-risk due to their predicted amyloid positivity status in Stage 1. We use ridge regression on data generated from brain scans to create a baseline/benchmark model to compare to a multimodal ensemble of Convolutional Neural Networks.

Baseline: Volumetric and SUVR Penalized Regression

The information contained in MRI and tau PET scans can be further distilled from an image array into a series of measurements given time and labor intensive post-processing. Brain scan images must be mapped to an atlas brain to infer consistency,

then volumetric measurements by brain region can be taken. Atrophy over time within patients can be shown using this method, as well as the difference between individual volumetric measurements and those of the atlas brain. This method can be applied to bot MRI scans and tau PET scans. Tau PET scans also allow for additional standard uptake value ratio (SUVR) measurements. SUVR measurements can more granularly show regional tau depositions by effectively renormalizing the radioactive tracer's pixel intensity in a given region against a given patient's baseline (8). Basically, the radioactive tracers work differently in different individuals, and it's not uncommon for CN individuals to have elevated levels of tau protein throughout the brain. The primary concern in AD patients is a disproportionate buildup of tau proteins in certain brain regions, which is best assessed using SUVR measurements.

Given volumetric and SUVR measurements, we use cross-validation to select the best hyperparameters for an elastic-net penalized logistic regression on 158 measurements generated from n=778 scans of CN and AD patients. The model with the highest validation accuracy on classifying CN vs. AD is selected. This model is then used to generate predictions for AD conversion for the holdout set. The holdout set here is a n=32 subset of the holdout set specified above, since not all of the 36 patients in the main holdout set have the required post-processed volumetric and SUVR measurements required.

Convolutional Neural Networks

Once the baseline model has been established, to follow the promising current research trend we use 2 Dimensional Convolutional Neural Networks (2D CNNs) trained to classify CN vs. AD (6) to then predict AD conversion for CN and MCI patients in the holdout set. During the training phase, 1,478 MRI scans from n=404 patients at different time points are used to train 10 CNNs corresponding to 10 central 2D brain slices along the coronal axis. Independently, 641 tau PET scans from n=440 patients are used to train 10 CNNs for 10 central 2D brain slices along the axial axis. The

best model for each slice is chosen based on 5 fold cross-validated validation accuracy after 30 epochs per fold. The architecture for each of these CNNs is shown in Figure 5. Note that input dimensions for tau PET scans are (160, 160) as shown in Figure 5, while the input dimensions for MRI scans are (196, 196), so the dimensions of each of the 9 layers are slightly different, although the overall architecture is the same.

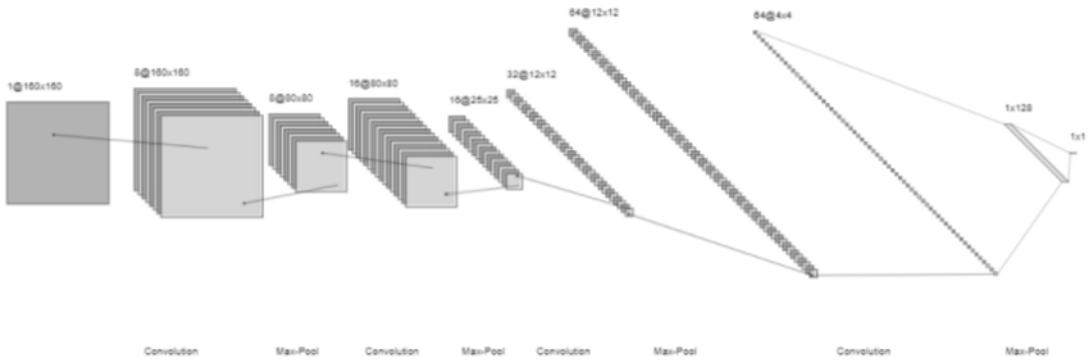

Figure 5. CNN architecture for each slice for each image modality

At the end of the training phase, we have 10 softmax probabilities of AD corresponding to 10 slices for both MRI scans and PET scans. To assess training accuracy, we need to convert these 10 probabilities into a single classification for each scan. We use the threshold AD $= 1(\sum_1^{10} \frac{P_i(AD)}{10} > .4)$, and CN otherwise. That is, if the average softmax probability of AD across all 10 slices is greater than 40%, then we classify that scan as AD. Otherwise we classify as CN This threshold is designed to be more sensitive to AD classification because a false CN classification for a true AD patient could significantly hamper their medical care and quality of life.

Once the models are trained, we freeze the weights and run 176 MRI scans through the MRI CNNs and 57 tau PET scans through the tau PET CNN. These are all of the scans for the n=36 holdout patients. Each scan generates 10 softmax probabilities of AD as above, though these probabilities now represent the likelihood of eventual AD conversion. We use the same rule as above to generate a single label for each scan, either 'will not convert to AD' (simplified to 'CN') or 'will convert to

AD' (simplified to 'AD Conv.'). We now have multiple labels for each patient that must be unified to generate a meaningful result and assess performance. We define a rule AD Conv. $= 1(\sum_i^n \frac{1_{MRI}(label_i = ADConv.)}{2} + \sum_j^m \frac{1_{tauPET}(label_j = ADConv.)}{2} \geq .5)$. In other words, if all slices from one modality predict AD conversion, or if there's a strong consensus for AD conversion from one modality and a weak consensus for the other, then the joint prediction will be AD conversion. Otherwise the prediction will be no AD conversion.

Results

Stage 1

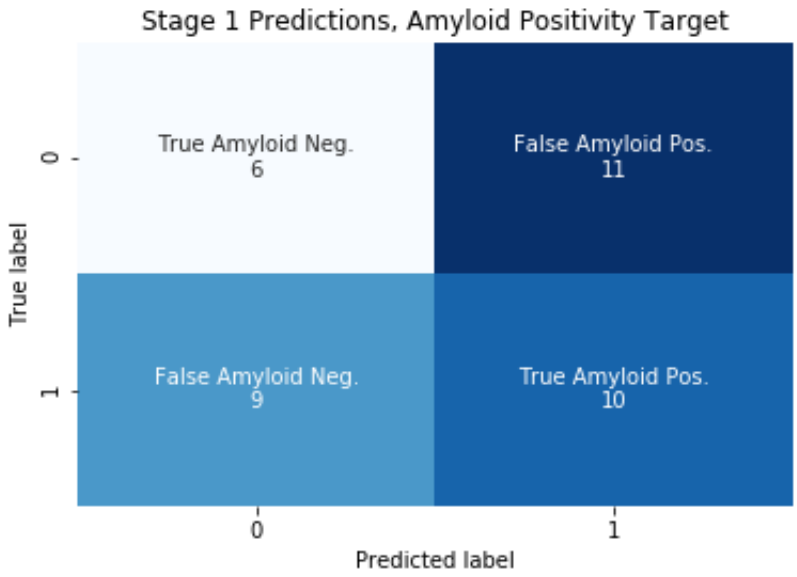

Figure 6. Amyloid positivity predictions vs. true amyloid positivity

The results of the amyloid positivity prediction logistic regression applied to the 36 patient holdout set are shown in the confusion matrix in Figure 6. The overall accuracy of this model is 44.4%, and the false negative rate is 25%. Given the potential ramifications of our framework, the false negative rate is critically important

as patients who are falsely classified as CN at any stage could be overlooked and fail to receive additional monitoring or early intervention before their cognitive condition degrades to full AD. However, amyloid positivity itself is not necessarily indicative of eventual AD conversion. Knowing this, we compare amyloid positivity predictions to eventual AD conversion in Figure 7. We now see the overall accuracy improve to 52.8%, and more importantly the false negative rate drops to 5.6%. This means that the filtering approach in Stage 1 reduces the holdout set by 41.7%, but at the cost of leaving 5.6% of eventual AD patients out of Stage 2. See the Discussion section for further analysis.

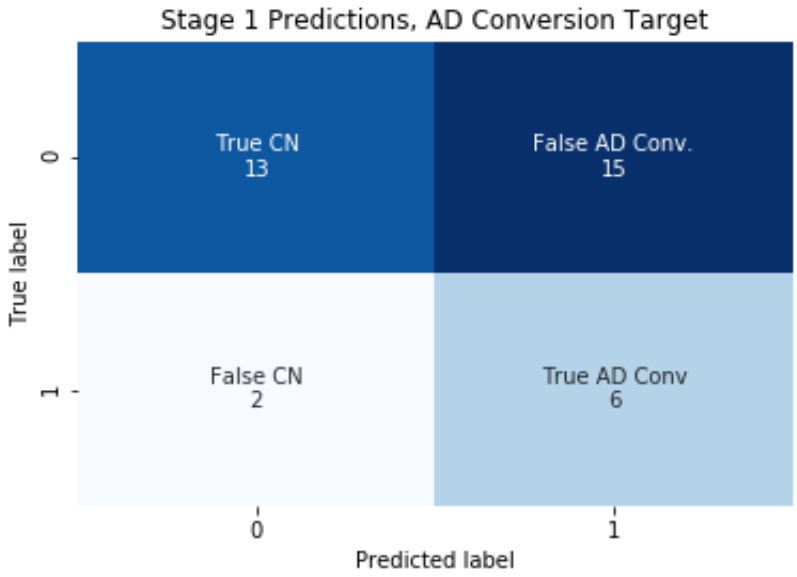

Figure 7. Amyloid positivity predictions vs. true AD conversion

Stage 2

Figure 8 shows the validation accuracy scores for each Stage 2 models, using the AD classification threshold mentioned above. We see that the baseline model and the MRI CNN perform similarly on the CN vs. AD classification task, and the PET CNN is the best classifier. This suggests that there may be some regional interaction

in tau deposition that contributes to AD beyond just the regional buildup amounts considered independently, as that's the main source of unique information contained in a PET scan image compared to the SUVR readings used in the baseline model.

Model	Validation Set Size	Validation Accuracy
Baseline Volumetric & SUVR	239	80.5%
MRI CNN	148	78.4%
tau PET CNN	65	95.4%

Figure 8. Validation accuracies for Stage 2 models

Figure 9 shows a confusion matrix of holdout subset predictions made from the baseline model. Note that Stage 1 filtering is not considered here, as the two holdout sets are not entirely consistent. We see an overall accuracy of 75% and an overall false negative rate of 9.4%, however of the true AD conversion patients in this holdout set, 75% are predicted incorrectly.

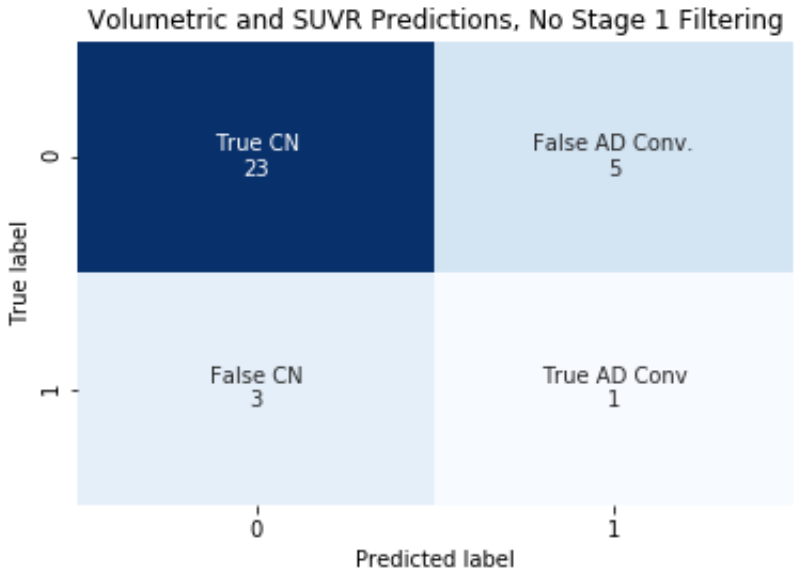

Figure 9. Baseline Stage 2 predictions vs. true AD conversion

The results of the multimodal CNN ensemble applied to the main 36 patient holdout set are shown in Figure 10. The overall accuracy here is 80.6% and the overall false negative rate is 13.9%. Of the true AD conversion patients in the holdout set, 62.5% are predicted incorrectly. These results are not perfectly comparable to the baseline since the holdout sets are different, although overall the results are similar. Any improvements in overall accuracy are dampened by increases in the false negative rate. Figure 10 does not account for Stage 1 filtering for comparison sake, but these results are included in Figure 11.

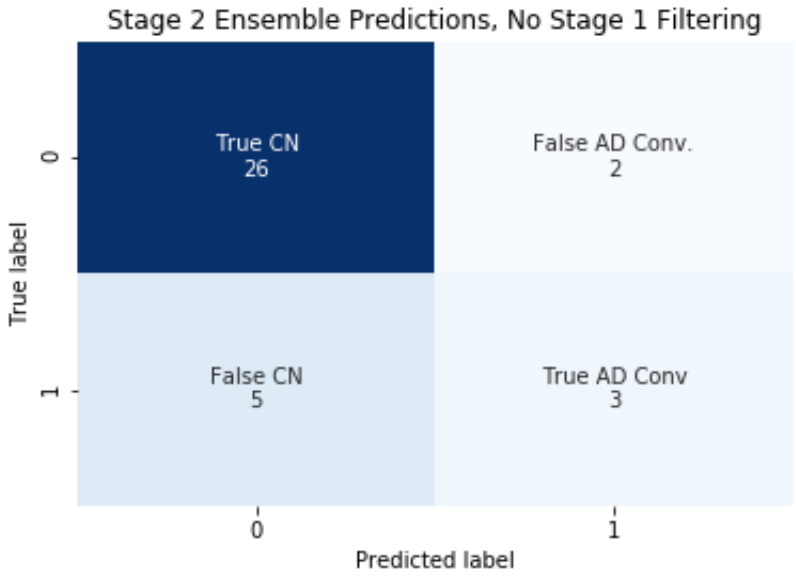

Figure 10. 2D CNN ensemble predictions vs. true AD conversion, without Stage 1 filtering

Figure 11 shows the combined results of the Stage 1 and Stage 2 models. That is, the 'high risk' patients identified in Stage 1 are then run through the Stage 2 model. Here we see the overall accuracy improve to 81%, the false negative rate becomes 14.3%, and of the true AD conversion patients, 50% are predicted incorrectly.

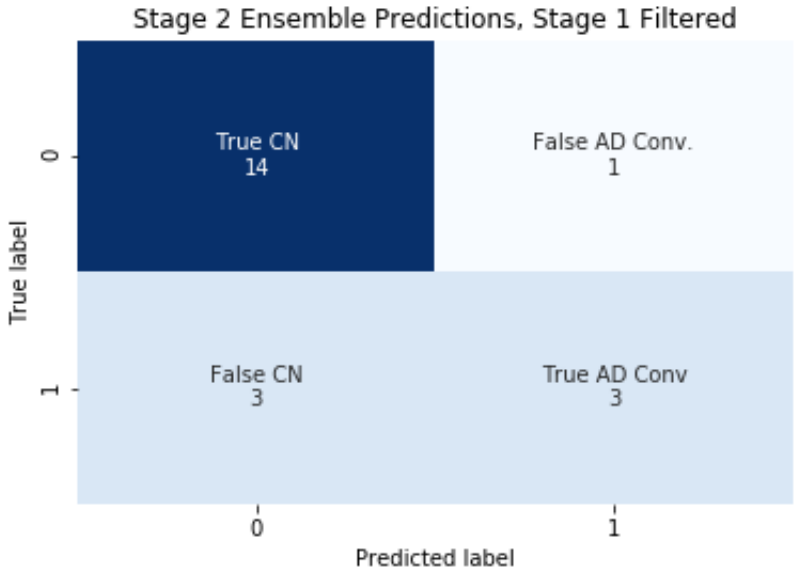

Figure 11. 2D CNN ensemble predictions vs. true AD conversion, with Stage 1 filtering

Discussion & Follow Up

Overall, the task of predicting AD before the disease manifests is an extremely challenging task that has yet to be solved. We've created a series of highly accurate classification models, but applying these models to the prediction task yields less useful results. The combined Stage 1 and Stage 2 results in Figure 11 appear promising, but 14.3% of eventual AD patients would be overlooked. This combined with the 5.6% of eventual AD patients who were filtered out through Stage 1 yields the same 62.5% of patients incorrectly predicted not to convert to AD as the unfiltered CNN ensemble. However, the timeline for patients was not accounted for. It's possible that the scans examined for individuals predate their respective AD diagnoses by a decade or more, which could lead to incorrect predictions if the disease had not progressed enough to be detected at the time of the scan. Given the limited amount of data available, this is the most promising area to examine for further analysis.

The Stage 1 model could have immediate utility in selecting patients for clinical trials. If the goal of a clinical trial is to select patients who are most likely to convert to AD, then they could screen candidates in a single office visit using minimally invasive procedures and use the Stage 1 model to select patients who are predicted to be amyloid positive. This is a preferable alternative to measuring true amyloid positivity, which requires a costly and extremely painful lumbar puncture. The clinical utility of the Stage 2 model is more difficult to qualify. It would certainly be helpful for MCI patients and their clinicians to receive a prediction for their future prognosis, and given an amyloid positive prediction from Stage 1, a Stage 2 prediction of AD conversion is 75% accurate. However, this is based on a small sample size, so we would need more data to have more confidence in these results. In addition, some of the true and predicted AD conversion patients are diagnosed as CN at the time of their scans. It's highly unlikely someone with no signs of cognitive impairment would be subject to this battery of tests and scans, so further analysis could be done to segment the results based on the original diagnosis as well as the prediction.

References

[1] "2020 alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 16, no. 3, pp. 391–460, 2020.

[2] L. G. Apostolova, R. A. Dutton, I. D. Dinov, K. M. Hayashi, A. W. Toga, J. L. Cummings, and P. M. Thompson, "Conversion of Mild Cognitive Impairment to Alzheimer Disease Predicted by Hippocampal Atrophy Maps," *Archives of Neurology*, vol. 63, no. 5, pp. 693–699, 2006.

[3] "National coverage determination (ncd) for pet scans (220.6),"

[4] K. Johnson, N. Fox, R. Sperling, and W. Klunk, "Brain imaging in alzheimer disease," 2012.

[5] "Adni study design," *ADNI Website*.

[6] M. Böhle, F. Eitel, M. Weygandt, and K. Ritter, "Layer-wise relevance propagation for explaining deep neural network decisions in mri-based alzheimer's disease classification," *Frontiers in Aging Neuroscience*, vol. 11, Jul 2019.

[7] N. Maserejian, S. Bian, W. Wang, and J. Jaeger, "Practical algorithms for amyloid probability in subjective or mild cognitive impairment," October 2019.

[8] K. Knesaurek, G. Warnock, L. Kostakoglu, C. Burger, and f. A. D. , "Comparison of standardized uptake value ratio calculations in amyloid positron emission tomography brain imaging," *World Journal of Nuclear Medicine*, vol. 17, no. 1, pp. 21–26, 2018.