Final Report  
ECS 289L - AI for Health

**Multimodal Prediction of Alzheimer’s Disease Progression**

Members: Tejas Patil, Arvind Sudarshan, David Chu

Advisor: Prof. Vladimir Filkov

**Introduction**

Alzheimer's Disease (AD), one of the most common forms of dementia, is a progressive neurodegenerative disorder that significantly impacts memory, thinking, and behavior. Early diagnosis and accurate prediction of disease progression can enable better clinical interventions and improve patient outcomes. Traditional approaches often rely on a single modality of data, such as neuroimaging or cognitive testing. However, recent research highlights the benefits of multimodal analysis for capturing the complex nature of AD progression. The code for this project is available [here](https://github.com/tejassvpatil/multimodal-alzheimers-prediction.git).

**Goals**

1. To develop a multimodal deep learning framework to predict the progression of AD by integrating MRI brain scans, longitudinal cognitive assessment scores, and genetic information (specifically the APOE genotype and additional high-risk SNPs such as CLU and BIN1).
2. We leverage the advantage of longitudinal modeling with GRUs and integrate multi-domain modalities to improve MCI to AD conversion prediction.

**Research Questions**

1. Can multimodal data integration improve the accuracy of Alzheimer’s disease progression prediction when compared to single modality models?
2. Does incorporating additional genetic risk factors (e.g., CLU, BIN1) alongside APOE improve predictive performance?
3. Can longitudinal cognitive sequences modeled using RNNs capture temporal trends in cognitive decline?

**Literature Background**

Deep learning models trained on unimodal data have shown limited success in predicting Alzheimer’s disease progression, thus motivating the need for a new approach to improve predictive performance. Recent existing literature on the prediction of Alzheimer’s disease progression revolves heavily around applying deep learning algorithms to multimodal patient data, and shows promising results through the integration of neuroimaging, clinical, and genetic data.

As one of the earlier works on applying deep learning to Alzheimer’s disease progression, Lee et al. (2019) applied a multimodal recurrent neural network to predict conversion from mild cognitive impairment to probable Alzheimer’s disease. Their method contains multiple modality-specific gated recurrent unit (GRU) components that each accept a different modality of longitudinal patient data at baseline, such as longitudinal CSF, cognitive, and MRI data, which ultimately highlights the value of irregular sequence handling and non-overlapping sample learning. In the first training step, each of these GRU components takes the data and produces fixed-size feature vectors. In the second training step, these feature vectors are then concatenated together to form the input for the final prediction. The benefits of using the GRU component over any other alternatives is the ability to use any irregular length of data as an input without requiring preprocessing. This design enables the model to effectively use non-overlapping data to help with better representation learning, as each data contributes to the GRU component that it belongs to. The use of GRU components also allows for the ease of integrating other data modalities, as it uses a concatenation-based integration method when combining all of the feature representations together. Similar to their work, we follow the approach of concatenating all feature representations together, but we use different components to do so. For images, we plan to apply a 3D convolutional neural network (CNN), GRU for the longitudinal cognitive scores, and multi-layer perceptron (MLP) for the genetic data. Thus, we extend their work by enabling fusion informed representation learning.

To better capture the representations of each data modality, Venugopalan et al. (2021) use 3D CNNs for imaging data and stacked denoising autoencoders for clinical and genetic data. Specifically, Venugopalan et al. (2021) employ stacked denoising auto-encoders combined with post hoc shallow classifiers to model temporal patterns in cognitive decline. This approach features end-to-end trainable multimodal fusion and supports multitask learning, allowing simultaneous prediction of both the current disease stage and future cognitive trajectories. To justify the performance of using deep learning over shallow models to capture the representations of imaging, clinical, and genetic data, Venugopalan et al. (2021) compare the accuracy, precision, recall, and mean F1 scores of the respective deep model against k-nearest neighbor, support vector machine, decision trees, and random forest algorithms. They found that for single-modality data, such as imaging and clinical data, the performance of deep learning models, specifically of autoencoders and CNNs in this case, was always better than that of shallow models. They also found that the best three fusion setups for the task of predicting Alzheimer’s disease progression are electronic health records (EHR) and single nucleotide polymorphism (SNP), EHR and imaging and SNP, and EHR and imaging. Similar to their work, we plan on using the same deep learning algorithms for representation learning. However, we intend to implement a cross-modal attention fusion module that aims to improve upon the basic fusion model that they used. Rather than using static concatenation, our proposed fusion module will weigh the importance of each modality dynamically.

To improve upon current multimodal deep learning architectures, Wang et al. (2024) incorporate interaction effects to capture relationships between and within modalities to aid in predicting the progression of Alzheimer’s disease. Their model, Dual Interaction Stepwise Fusion Classifier, or DISFC, first learns an accurate representation for each data modality (image, clinical, and genetic data). However, before aggregating all the learned embeddings together, DISFC first aggregates the image and clinical representations together, because they found that empirically, this yielded the best results. By combining the representations for image and clinical representations, the model effectively learns a relationship between clinical data and brain structure. Next, this aggregated data is combined with the genetic representation, to create a more complete and comprehensive representation of the patient’s condition. This is fed into the stepwise-fusion classifier, which outputs a probability of whether the corresponding patient would convert to Alzheimer’s disease from mild cognitive impairment. Similar to DISFC, we approach Alzheimer’s disease prediction from a multimodal perspective, but we plan on using a cross-modal attention fusion module on all three representations at once, rather than split the representation aggregation into steps.

**Data and Methods**

A] Primary Dataset:

* ADNI ([Alzheimer’s Disease Neuroimaging Initiative](https://adni.loni.usc.edu/)): Includes T1 weighted MRI scans, time-series cognitive scores (e.g., MMSE, CDR), and APOE genotyping for patients.
* Patient Data Summary:

| **Category** | **Count** |
| --- | --- |
| Total MCI at baseline patients | 1214 |
| Stable MCI (sMCI) | 816 |
| Progressive MCI (pMCI) | 398 |
| Patients with complete cognitive sequences | 1099 |
| Patients with genotyping | 482 |
| Patients with T1 weighted MRI scans | 210 |

Table 1: Total number of MCI patients and class distribution between pMCI and sMCI

B] Methods:

1. Modality-Specific Representation Learning
   1. MRI Analysis: We use 3D Convolutional Neural Networks (CNNs) to process volumetric T1 weighted MRI scans. These models extract high-level spatial features from brain regions associated with AD pathology (e.g., hippocampus, amygdala), following preprocessing steps like skull stripping, normalization, and alignment to standard space.
   2. Cognitive Score Modeling: Longitudinal cognitive assessments (e.g., MMSE, CDR) are modeled using recurrent architectures, specifically Gated Recurrent Units (GRUs).
   3. Genetic Data Processing: Genetic features include APOE genotype and high risk SNPs such as CLU and BIN1. After VCF preprocessing (e.g., filtering low-quality variants and high-missing-rate loci), we use feature selection (e.g., mRMR) to retain informative SNPs. These features are encoded using a Multi Layer Perceptron (MLP) and scaled appropriately.
2. Multimodal Fusion:
   1. Each modality is encoded independently (CNN for imaging, GRU/Transformer for cognition, MLP for genetics), and their intermediate embeddings are merged via a fusion module that uses cross modal attention. This allows the model to weigh the importance of each modality dynamically, unlike static concatenation used in prior work.
3. Multi-Task Learning:
   1. The network will produce a classification head for diagnosis (MCI, AD). This encourages shared representation learning, potentially improving generalization. Evaluation metrics include accuracy, F1-score, AUC for classification, sensitivity and specificity.

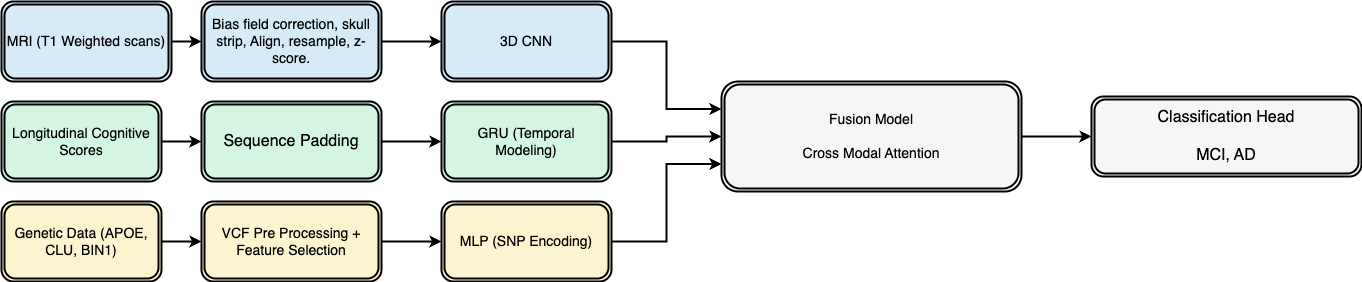
****

Figure 1: Pipeline of Multi-Modal Alzheimer’s Disease Progression Prediction

**How Our Methods Address the Research Questions:**

1. **RQ1:** The cross modal attention fusion module, trained end-to-end on CNN (MRI), GRU (cognition), and MLP (genetics), directly tests if integrating modalities improves prediction over single streams.
2. **RQ2:** The genetic data pipeline comprising VCF preprocessing + mRMR feature selection + MLP encoding evaluates the predictive gain from adding CLU and BIN1 to APOE.
3. **RQ3:** The GRU models applied to longitudinal cognitive scores (e.g., MMSE, CDR) test whether temporal modeling captures progression trends effectively.

**Results**

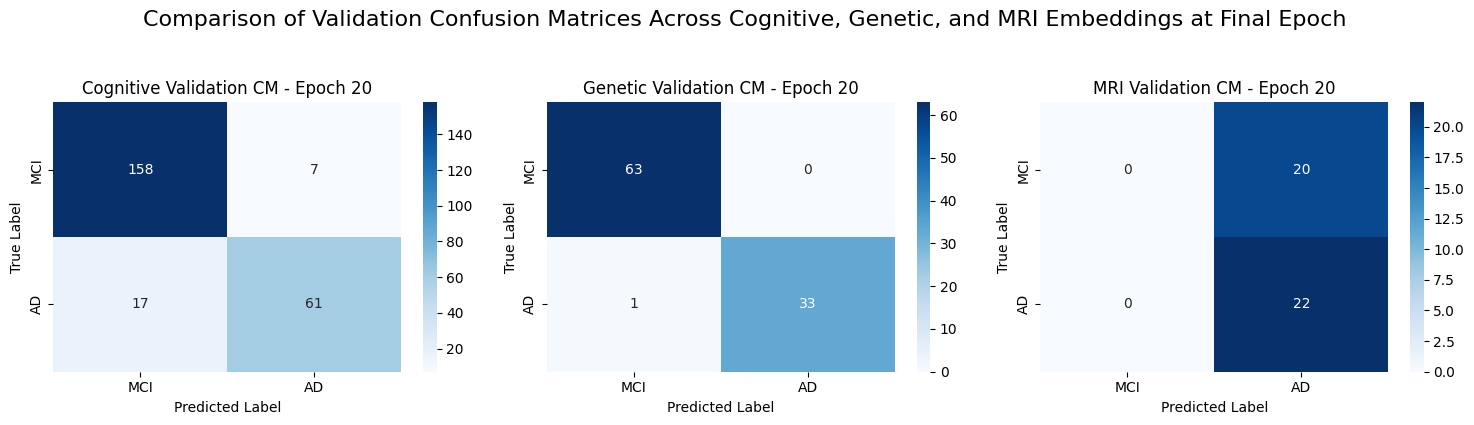


Figure 2: Confusion matrices for each modality (cognitive, genetic, and MRI) at the final epoch (20), where MCI is mild cognitive impairment and AD is Alzheimer’s disease

| Modality | Accuracy | Sensitivity | Specificity |
| --- | --- | --- | --- |
| Cognitive | 90.1% | 78.2% | 95.8% |
| Genetic | 98.9% | 97.1% | 100% |
| MRI | 52.4% | 100% | 0% |

Table 2: Sensitivity and specificity graphs for each modality (cognitive, genetic, and MRI data) at final epoch

| Modality | Accuracy | AUC | Precision | Sensitivity |
| --- | --- | --- | --- | --- |
| cog-mri | 47.6% | 0.9 | 0% | 0% |
| cog-gen | 96.7% | 0.991 | 97.9% | 94.1% |
| gen-mri | 80.9% | 0.797 | 75% | 80.9% |

Table 3: Accuracy, AUC, precision, and recall scores for combinations of the following two modalities: cognitive and MRI (cog-mri), cognitive and genetic (cog-gen), genetic and MRI (gen-mri)

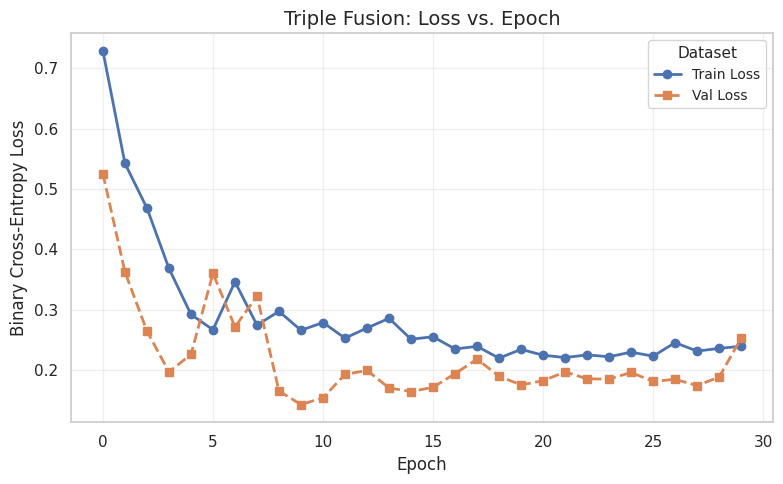


Figure 3. Loss curve illustration of the triple-modality fusion model across epochs

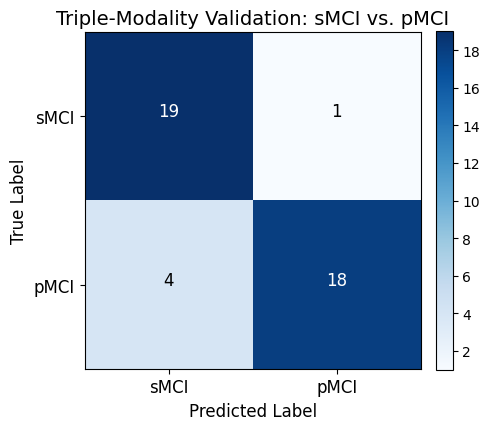


Figure 4. Confusion matrix of the triple-modal fusion approach (cog-mri-gen)

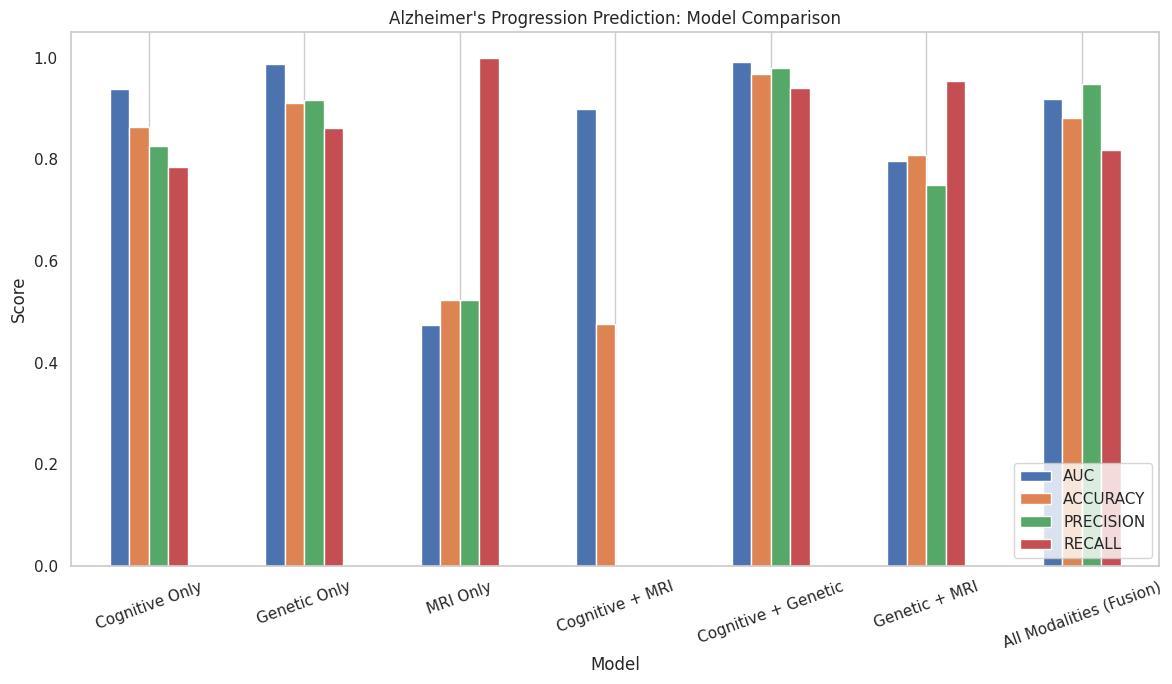


Figure 5. Comparison of all the modality combinations illustrating better performance of multi-modal approaches

**Discussion**

To first evaluate the performance of each single-modality encoder, the corresponding confusion matrices were created at the final epoch for validation for each modality, as shown in Figure 2. The cognitive encoder achieved a precision of 89.7% and an accuracy of 90.1%, the genetic encoder achieved a precision of 100% and an accuracy of 98.9%, and finally the MRI encoder achieved a precision of 52.4% and an accuracy of 52.4%. In conjunction with Table 2’s results, which contain the sensitivity and specificity for each modality, it can be concluded that the genetic modality is the best performing modality, as it achieved perfect specificity, indicating that it did not misclassify any false positives, had the highest accuracy out of all the modalities, and had a very high sensitivity of 97.1%, indicating that it almost classified all true positives correctly. The cognitive modality was the second best performing, with a strong precision score of 89.7%, indicating that it was typically correct when predicting a true positive, and an accuracy of 90.1%. However, the sensitivity of the cognitive modality was the lowest at 78.2% of all the modalities, indicating that it misclassified false negatives more than the other modalities. The MRI modality was the worst performing, achieving an accuracy of 52.4%, which is slightly better than random chance and a low precision of 52.4%. Because the accuracy and precision score are the same, the sensitivity score was 100%, and the specificity was 0%, this indicates that the model simply guessed the positive class, or AD, every time. It can be concluded that the MRI modality lacked discriminative ability and was unable to distinguish between classes effectively.

Our inclusion criteria for data required that each patient diagnosed with MCI had baseline visit data available across all three modalities, MRI, cognitive and genetic, which resulted in a much smaller cohort size than expected. We had a sample size of 210 which is far less than what is typically needed for training a deep 3d convolution neural network, which requires more samples to avoid overfitting and capture the diversity present in the brain imaging data. Although table 1 shows that the dataset includes over 1000 MCI patients with longitudinal records and complete cognitive sequences, only 210 subjects met all the criteria necessary for our multimodal analysis (having data across all three modalities). This stricter filtering ensured consistency across all the three modalities but reduced the cohort size. As a result of this combined constraints, the, discriminative power of the MRI modality in our study is significantly limited

In the large ADNI cohort, static classifiers trained on only baseline cognitive measures, such as MMSE, CDR-SB, FAQ and ADAS-Cog, generally achieve moderate discrimination. The AUCs were reported between roughly 0.59 and 0.77 in a previous sparse-learning study of 319 MCI subjects (Ye et al., 2012). Similarly, earlier work (Querbes et al., 2009) on cortical thickness combined with baseline MMSE and Trail Making Test B yields AUCs of 0.64 and 0.72, respectively. Our bidirectional GRU on longitudinal cognitive sequence contrastingly reaches an AUC of 0.946, accuracy of 0.900 and F1 score of 0.8608. This directly answers the **third research question** by demonstrating that explicitly modeling temporal trends in cognition substantially outperforms these non-temporal baselines.

To evaluate the performance of each bimodal fused representation, the accuracy, AUC score, precision, and recall metrics at the final epoch were analyzed. As shown in Table 3, the cognitive-genetic (cog-gen) modality pairing achieved the highest overall performance, with an accuracy of 96.7%, an AUC score of 0.991, a precision of 97.9%, and a recall of 94.1%. These results indicate that this modality pairing was highly precise in its positive predictions and able to correctly identify a large proportion of true positives, suggesting strong discriminative ability. The genetic-MRI (gen-mri) pairing was the second-best performing combination, achieving an accuracy of 80.9% and a recall of 80.9%, with slightly lower precision at 75.0% and an AUC score of 0.797. While this combination showed reasonable balance between sensitivity and precision, the performance was still substantially lower than cog-gen. In contrast, the cognitive-MRI (cog-mri) pairing demonstrated the worst performance across all metrics, with an accuracy of 47.6%, along with both precision and recall values being 0. This implies that the model was completely unable to identify any true positives and defaulted to only predicting the negative class, or MCI, despite a relatively high AUC score of 0.9. Thus, it can be concluded that cognitive and genetic features provide highly complementary and informative signals for Alzheimer’s disease progression prediction, while the MRI and cognitive features pairing contributes little to no predictive value. Thus, from these results, we can answer our **second research question**, that the predictive power of selected SNPs such as CLU and BIN1 alongside APOE do indeed improve predictive performance, as including the genetic data with either the MRI or cognitive modality significantly improved the performance when compared to only using either MRI or cognitive data.

**Tri-Modal Attention Fusion**

Our tri-modal fusion architecture, which integrates cognitive assessments, genetic factors, and MRI data, employs independent linear projections for each modality, followed by a multi-head self-attention layer to capture cross-modal interactions. The resulting fused representation is processed through layer normalization, dropout, and a final linear classifier for prediction. This approach achieved robust validation performance, with a best ROC-AUC of 0.9932 during training (Figure 3, loss curve). On the held-out test set, the tri-modal model reached an AUC of 0.918 and an accuracy of 0.881, alongside a high precision of 0.947 and recall of 0.818 (Figure 4, confusion matrix).

These findings directly address our primary research question (RQ1), demonstrating that multimodal integration can deliver strong and balanced predictive performance for distinguishing between stable (sMCI) and progressive (pMCI) MCI cases. Although the fusion model did not exceed the best-performing bimodal (cognitive+genetic, AUC 0.991) or single genetic-only (AUC 0.988) models in terms of AUC or accuracy, it achieved the highest precision. This indicates a more conservative classifier that reduces false positives. The slightly lower recall, compared to the MRI-only branch (which achieved perfect recall but very low specificity), reflects a better overall trade-off between sensitivity and specificity in the fusion model. It is also important to note that the tri-modal training was conducted on a relatively modest sample size (n = 210), which likely constrained performance; expanding the dataset in future work could further enhance the benefits of multimodal fusion.

Thus, from the results of the single modal, bimodal, and trimodal fused representations, **our first research question**, that incorporating multimodal data improves the accuracy of Alzheimer’s disease prediction when compared to single modal models does improve the predictive performance of the model(figure 5, comparative study). Specifically, from the data in Table 2 and Table 3, we can note that the choice of modality is correlated with the expected improvement when pairing together single modal representations. Most notably, it seems that genetic data carries the most information for accurately predicting Alzheimer’s disease progression, and MRI data carries the least information, while cognitive data is between the two. Pairing cognitive data and MRI data together yields a performance that is less than either the MRI or cognitive modality by itself, indicating that the two modalities do not strongly contribute to the prediction of Alzheimer’s disease progression. However, when pairing either modality with the genetic modality, the performance increases, further supporting the claim that genetic data carries the most predictive power for Alzheimer’s disease progression out of these three modalities.

**Conclusion**

In this work, we develop a multimodal deep learning framework to predict the progression of Alzheimer’s disease based on MRI brain scans, longitudinal cognitive scores, and genetic information. The core of our framework is the cross-modal attention fusion module, which fuses the features from the 3D CNN (MRI data), GRU (cognitive scores), and MLP (genetic data) that weighs the importance of each modality dynamically, unlike the static concatenations of previous work in multimodal Alzheimer’s disease progression prediction. Our results demonstrate that the genetic modality achieved the highest performance among the single-modality models, the cognitive-genetic modality pairing achieved the best performance among bimodal pairings, and that the multimodal fusion approach was able to achieve high scores across all evaluated metrics.

A fundamental limitation of this work is that it was trained solely on the ADNI dataset, in which the participants are predominantly white and from North America. This limits the generalizability of our model, which may perform poorly on people from other countries or regions. Furthermore, not all modalities of data were available for all patients, significantly lowering the sample size for training. This may impact the robustness and stability of our proposed multimodal fusion model, as the missing data may lead to biased representations based on the complete data. Finally, the ADNI dataset is based on a controlled research cohort with high levels of subject compliance and consistent data collection protocols, which may not reflect the common conditions of noise, variability, and missing data in real-world clinical settings.

Future work that can be conducted include incorporating more relevant sources of data that may improve accurate prediction of Alzheimer’s disease progression, such as PET scans or information collected from wearables (speech, behavioral, or motor information), that can further supplement the data modalities that we have explored. These additional modalities offer complementary insights into different aspects of the disease that can be measured in a non-invasive way, as PET scans can capture protein information in the brain directly and wearable-derived data can reflect subtle functional and cognitive changes in real-world settings that may not otherwise be collected or detected. Integrating these signals into a multimodal framework may enhance early detection, improve model generalization, and better enable personalized prediction across diverse patient populations.

**Team Member and Attestation of Work**

Team members **David Chu, Tejas Patil, Arvind Sudarshan** participated sufficiently.

**References**

[1] Taeho Jo, Kwangsik Nho, and Andrew J. Saykin. Deep learning in Alzheimer’s disease: Diagnostic classification and prognostic prediction using neuroimaging data. *Frontiers in Aging Neuroscience*, 11, August 2019.<https://doi.org/10.48550/arXiv.1905.00931>

[2] Querbes O., Aubry F., Pariente J., Lotterie J.-A., Demonet J.-F., et al. Early diagnosis of Alzheimer’s disease using cortical thickness: impact of cognitive reserve. Brain. 2009;132(8):2036–2047. <https://doi.org/10.1093/brain/awp088>

[3] Venugopalan, J., Tong, L., Hassanzadeh, H. R., & Wang, M. D. (2021). Multimodal deep learning models for early detection of Alzheimer’s disease stage. *Scientific Reports*, 11, 3254.<https://doi.org/10.1038/s41598-020-74399-w>

[4] Qiankun Zuo, Baiying Lei, Yanyan Shen, Yong Liu, Zhiguang Feng, and Shuqiang Wang. Multimodal representations learning and adversarial hypergraph fusion for early Alzheimer’s disease prediction, 2021.<https://doi.org/10.48550/arXiv.2107.09928>

[5] Lee, G., Nho, K., Kang, B., Sohn, K. A., Kim, D., & Alzheimer’s Disease Neuroimaging Initiative. (2019). Predicting Alzheimer’s disease progression using multi-modal deep learning approach. *Scientific Reports*, 9, 1952.<https://doi.org/10.1038/s41598-018-37769-z>

[6]Wang, Y., Gao, R., Wei, T. et al. Predicting long-term progression of Alzheimer’s disease using a multimodal deep learning model incorporating interaction effects. J Transl Med 22, 265 (2024). <https://doi.org/10.1186/s12967-024-05025-w>

[7] Mirabnahrazam, Ghazal, Ma, Da, Beaulac, Cédric, Lee, Sieun, Popuri, Karteek, Lee, Hyunwoo, Cao, Jiguo, Galvin, James E, Wang, Lei, & Beg, Mirza Faisal. (2023). Predicting time-to-conversion for dementia of Alzheimer’s type using multi-modal deep survival analysis. *Neurobiology of Aging*, 121, 139–156.<https://doi.org/10.1016/j.neurobiolaging.2022.10.005>

[8] Ye J., Farnum M., Yang E., Verbeeck R., Lobanov V., Raghavan N., et al. Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data. BMC Neurology. 2012;12:46. https://doi.org/10.1186/1471-2377-12-46