CS 437 Project - Literature Review

Project: VLM-Based Solutions for Early Alzheimer's Disease Detection using MRI Data

Umer Raja - 26100063
Muhammad Ahmad Ashraf - 26100169

Recent advances in vision-language models (VLMs) are transforming automated Alzheimer's disease (AD) diagnosis from brain scans by integrating imaging with textual information. Traditional deep learning approaches – e.g., CNN or autoencoder-based models on MRI/PET data – achieved high classification accuracies (around 85–90% for AD vs. controls) [1, 2]. For instance, Janghel and Rathore's CNN model (VGG-16) combined fMRI and PET features and slightly outperformed earlier methods [3]. However, such models typically treat imaging in isolation and often lack interpretability [4]. To address these gaps, state-of-the-art VLM-based methods leverage multimodal learning (images + text) for AD detection, aiming to improve accuracy and provide human-interpretable outputs.

One notable approach represents the *joint structure* of images and descriptive language as a graph. Lee et al. (2025) construct a bipartite graph linking salient regions of the standardized "Cookie Theft" picture (a cognitive test stimulus) with the participant's spoken description of that image [5]. A pre-trained VLM (BLIP) computes image-text similarity, which forms the edge weights in the graph connecting image nodes to text nodes [5]. A graph convolutional network (GCN) then classifies each subject as AD or healthy based on the graph structure. This GNN+VLM model achieved 88.7% classification accuracy on the ADReSSo challenge dataset, surpassing prior state-of-the-art results [5]. Notably, ablation studies showed that removing the image-text edges degraded performance, highlighting the value of multimodal graphs [5]. The graph framework also enabled explainability: by analyzing the learned node embeddings, the authors could identify which descriptive sentences and keywords were most predictive of AD [5], offering insights into linguistic markers of cognitive decline.

Another line of research uses contrastive learning in VLMs to align MRI features with textual descriptions of neuroanatomical abnormalities. *VisTA* (Vision-Text Alignment) is a multimodal model proposed by Can et al. (2025) that fine-tunes a large pre-trained VLM (BiomedCLIP) on a small expert-curated dataset of MRI scans [6]. Each MRI in the training set is paired with a verified description of observed brain abnormalities (e.g., hippocampal atrophy), and VisTA employs a contrastive objective to align image embeddings with the corresponding text [6]. Uniquely, the VisTA pipeline produces four outputs for a given patient: the predicted type of brain abnormality, retrieval of similar cases from reference data, an evidence-based textual explanation, and the final AD vs. healthy diagnosis [6]. Despite fine-tuning on only 170 image–text samples, VisTA markedly improved both the retrieval of relevant cases and dementia prediction accuracy [6]. It achieved 88% accuracy (AUC 0.82) for AD diagnosis – a substantial gain over baseline CNN models [6]. Perhaps most importantly, VisTA's generated explanations closely matched clinicians' reasoning [6], demonstrating how VLMs can enhance interpretability in neuroimaging by justifying model decisions with human-readable descriptions.

To fully exploit VLMs in the AD domain, researchers are also tackling the shortage of textual annotations for MRI data. Chiumento et al. (2024) address the fact that, unlike X-ray or CT datasets, brain MRI repositories often lack detailed radiology reports [7]. They generate synthetic reports for MRI scans to serve as training data for VLMs [7]. In their framework, tabular clinical data from the OASIS-4 MRI dataset (663 subjects) were fed to a GPT-based model to produce pseudo-radiology reports describing each patient's brain condition [7]. Using these GPT-generated texts as ground truth, they then trained a vision-language model (combining BiomedCLIP for vision and a T5 text decoder) to generate diagnostic summaries directly from MRI images [7]. The results showed encouraging text-generation quality (e.g., BLEU-4 ≈0.18 and ROUGE-L ≈0.37) [7], indicating the feasibility of MRI report synthesis [7]. This approach opens the door to VLM-based diagnosis in neuroimaging by creating the needed image-text pairs artificially – an approach that could be further improved as large language models become more accurate medical scribes. In a related vein, other work has explored using large language models to integrate multimodal patient data. For example, Feng et al. (2023) augmented an AD MRI classifier with an LLM that ingests non-imaging data (demographics, cognitive tests, genomics, etc.), enabling the model to utilize rich clinical context [8]. This multimodal LLM-enhanced method achieved state-of-the-art performance on the ADNI MRI dataset [8], underscoring the benefit of combining neuroimaging with textual patient information. Other SOTA VLM-based approaches to tackling multimodal medical data whilst incorporating principles from Explainable AI include MedCLIP, BioMedGPT and Med-Flamingo [9, 10, 11]. These models leverage the power of capabilities such as Visual Questioning and Answering (VQA), Semantic Segmentation, Classification and Zero Shot Predictions within the context of medical image/text data to demonstrable success. However, their efficacy within the domain of neurodegenerative diseases remains to be demonstrably proven, and bridging the gap between general medical data and MRI-based analysis of neurodegenerative diseases using these technologies therefore forms a fundamental goal for this project.

In summary, vision-language techniques are emerging as a powerful paradigm for early AD detection. By marrying MRI features with textual descriptions or clinical data, these models improve diagnostic accuracy while providing human-interpretable outputs. Key innovations include graph-based representations of image–text relationships [5], contrastive image-text alignment for explanation generation [6], and the creation of synthetic training reports to overcome data scarcity [7]. Public datasets like ADNI (which provides longitudinal MRI and clinical data) and OASIS (open-access MRI scans) are commonly used to develop and evaluate these models [7, 8], and new resources such as the MINDSet registry are expected to further expand available data. Beyond accuracy, researchers are increasingly emphasizing model interpretability and trustworthiness, integrating mechanisms (graphs, attention maps, or text explanations) that align AI decisions with clinical reasoning [6]. These advances set the stage for our proposed project – an MRI-only VLM tailored for AD detection – which will build on current state-of-the-art by leveraging vision-language alignment on neuroimaging data. By focusing on brain MRI inputs and incorporating recent innovations (e.g., contrastive learning and report generation), our model aims to improve early AD prediction while delivering transparent, explainable diagnoses that can be validated by clinicians.

References

[1] Liu et al. 2014. *Early Diagnosis of Alzheimer’s Disease with Deep Learning*. IEEE.

[2] Rathore et al. 2022. *An MRI-based deep learning approach for accurate detection of Alzheimer’s disease*. ScienceDirect.

[3] Janghel and Rathore. 2021. *Deep CNN System for Early Diagnosis of Alzheimer’s*. ScienceDirect.

[4] Jo et al. 2019. *Deep Learning to Detect Alzheimer's Disease from Neuroimaging*. ScienceDirect.

[5] Lee et al. 2025. *Alzheimer’s Disease Recognition Using Graph Neural Network by Leveraging Image-Text Similarity from VLM*. Nature Scientific Reports.

[6] Can et al. 2024. *VisTA: Vision-Text Alignment Model with Contrastive Learning for Explainable Alzheimer’s Diagnosis*. arXiv:2502.01535.

[7] Chiumento and Liu. 2024. *Leveraging Multimodal Models for Enhanced Neuroimaging Diagnostics in Alzheimer’s Disease*. arXiv:2411.07871.

[8] Feng et al. 2023. *Large Language Models Improve Alzheimer’s Disease Diagnosis Using Multi-Modality Data*. arXiv:2305.19280.

[9] Wang et al. 2022. *MedCLIP: Contrastive Learning from Unpaired Medical Images and Text.* arXiv:2210.10163.

[10] Zhang et al. 2023. *BiomedGPT: A Generalist Vision-Language Foundation Model for Diverse Biomedical Tasks*. arXiv:2305.17100.

[11] Moor et al. 2023. *Med-Flamingo: a Multimodal Medical Few-shot Learner.* arXiv:2307.15189.