

HIERARCHICAL CROSS-MODAL SPATIO-TEMPORAL TRANSFORMER FOR EARLY ALZHEIMER'S DISEASE PREDICTION

SURESH BABU SINGAMSETTI¹, MURALI GUDIPATI²

^{*1}Research Scholar, Acharya Nagarjuna University, Department of Computer Science & Engineering,
Andhra Pradesh, India

²Professor, Acharya Nagarjuna University, Department of Computer Science & Engineering, Andhra
Pradesh, India

E-mail: ¹singamsettisureshababu@gmail.com, ²drgudipatimurali@gmail.com

ABSTRACT

Early detection of Alzheimer's disease (AD) and prediction of its progression from cognitively normal (CN) to mild cognitive impairment (MCI) and ultimately AD remain challenging because disease-related alterations are distributed across multiple modalities and evolve over time. Recent CNN-, Vision Transformer-, and multimodal learning approaches have improved AD classification by learning discriminative anatomical and functional representations; however, existing methods predominantly emphasize single-timepoint diagnosis, limited modality combinations, or independent temporal modeling. Recent longitudinal and transformer-based approaches have addressed some of these limitations, but multimodal interactions, hierarchical disease-stage transitions, anatomical priors, and prediction uncertainty are generally not modeled jointly within a single framework. To address these limitations, this study proposes a Hierarchical Cross-Modal Spatio-Temporal Transformer (HCST-Former) that extends existing transformer-based AD modeling from predominantly cross-sectional representation learning toward unified multimodal and longitudinal disease modeling. The principal contribution is a unified architecture that integrates Cross-Modal Attention Fusion (CMAF) for explicit interactions among structural MRI, PET, cognitive, genetic, and clinical information; Longitudinal Temporal Encoding (LTE) for learning disease trajectories across multiple timepoints; Hierarchical Disease-Stage Modeling (HDSM) for explicitly representing the CN→MCI→AD progression hierarchy; Graph-Guided Anatomical Attention (GGAA) for incorporating prior knowledge of AD-relevant brain regions; and Uncertainty-Aware Prediction (UAP) for estimating predictive confidence. In contrast to approaches that address these aspects separately, HCST-Former jointly learns multimodal, temporal, hierarchical, and anatomically informed representations while considering computational scalability. The research questions addressed by the framework therefore concern whether explicit cross-modal interaction, longitudinal trajectory modeling, hierarchical stage representation, anatomical guidance, and uncertainty estimation can collectively improve AD detection and progression prediction. Experiments on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset show that HCST-Former achieves 93.7% accuracy, 96.2% AUC, 92.8% sensitivity, and 94.5% specificity for disease-stage classification. For longitudinal progression prediction, the model achieves a C-index of 0.91 and time-to-conversion MAE of 8.7 months. Ablation experiments demonstrate progressive improvements following the incorporation of CMAF, LTE, HDSM, and GGAA, while the complete model achieves the reported performance with 48.3M parameters, 92.6 GFLOPs, and 8.1 GB GPU memory. These findings indicate that integrating complementary multimodal, longitudinal, hierarchical, anatomical, and uncertainty-aware mechanisms within a unified transformer provides a systematic extension of existing AD modeling approaches.

Keywords: *Alzheimer's Disease, Multi-Modal Transformer, Longitudinal Modeling, Cross-Modal Attention, Medical Imaging, Interpretability*

1. INTRODUCTION

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder and the most common cause of dementia worldwide, affecting approximately 50 million individuals globally, with projections exceeding 130 million by 2050 [1]. The disease is characterized by progressive cognitive decline, memory impairment, and eventual loss of independent functioning, imposing substantial burdens on patients, caregivers, and healthcare systems. Early detection of AD and accurate prediction of its progression are critical for timely intervention, treatment planning, patient stratification, and clinical trial enrollment. However, the subtle and heterogeneous nature of early neurodegenerative changes makes early diagnosis particularly challenging, as conventional biomarkers often remain within normal ranges during the preclinical and early mild cognitive impairment (MCI) stages [2, 3].

Recent advances in deep learning have significantly improved automated AD diagnosis using structural magnetic resonance imaging (MRI), positron emission tomography (PET), and other imaging modalities. Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) have demonstrated remarkable capability in learning discriminative features from neuroimaging data, substantially outperforming traditional machine learning approaches that rely on handcrafted features [4, 5]. These advances have accelerated the integration of artificial intelligence into clinical workflows and enabled increasingly accurate analysis of brain structure and function. Nevertheless, the complex and multifaceted nature of AD pathogenesis presents unique challenges that require models to simultaneously process heterogeneous data modalities, model temporal disease trajectories, and provide clinically interpretable predictions.

Alzheimer's disease differs substantially from other medical image analysis tasks because relevant pathological information is distributed across multiple data modalities and temporal dimensions. Structural MRI provides information about brain atrophy and tissue integrity, PET imaging reveals metabolic and amyloid/tau pathology, cognitive assessments measure functional decline, and genetic biomarkers indicate disease susceptibility. Furthermore, AD progression is inherently a temporal process, where subtle changes accumulate over months or years before clinical symptoms become apparent [6]. Effective detection and

progression prediction therefore require a model to simultaneously capture cross-modal relationships, longitudinal disease trajectories, and early pathological changes. Achieving this balance remains difficult, particularly for early-stage detection, where both subtle pattern recognition and longitudinal context are essential.

Although recent deep learning architectures have substantially improved AD diagnosis and prognosis, efficiently modeling complex multi-modal and longitudinal relationships in neurodegenerative disease remains an open challenge. Existing transformer-based frameworks have demonstrated the effectiveness of self-attention for learning long-range dependencies that are difficult to capture using purely local operations [7, 8]. However, directly applying global self-attention to multi-modal medical data introduces considerable challenges because the number of tokens increases rapidly with multiple imaging modalities and longitudinal time-points. As AD datasets continue to grow in size and complexity, these computational demands become a significant obstacle to both model training and clinical deployment.

Beyond computational efficiency, another important challenge lies in the representation of disease progression. Alzheimer's disease progression follows a hierarchical pattern in which regional atrophy, metabolic decline, and cognitive impairment manifest in a characteristic spatio-temporal sequence. The disease typically begins in the entorhinal cortex and hippocampus before spreading to temporal, parietal, and eventually frontal regions [9]. Most existing AD diagnosis frameworks, however, process imaging data as static representations without explicitly modeling these progressive anatomical patterns or incorporating disease progression priors. As a result, models may struggle to simultaneously capture subtle early-stage changes while maintaining accurate longitudinal trajectory modeling.

These challenges reveal a broader research problem in Alzheimer's disease detection and progression prediction. An effective framework should preserve comprehensive multi-modal information, efficiently model longitudinal disease trajectories, learn anatomically meaningful representations, incorporate disease progression priors, provide uncertainty estimates, and remain computationally scalable. Satisfying all of these requirements within a unified architecture remains difficult because improvements in representational capacity often increase computational complexity,

whereas aggressive efficiency-oriented strategies may sacrifice important clinical information.

Motivated by the limitations identified in existing CNN-, Vision Transformer-, multimodal-, and longitudinal AD frameworks, this work proposes the Hierarchical Cross-Modal Spatio-Temporal Transformer (HCST-Former) as a unified framework for early AD detection and progression prediction. The contribution of HCST-Former is not limited to replacing conventional CNNs with a transformer; rather, it addresses several complementary limitations reported across the literature within a single architecture. Existing CNN-based approaches provide effective local anatomical feature extraction but have limited capacity for long-range dependencies, while transformer-based approaches improve global representation learning but can incur substantial computational costs when applied to high-dimensional multimodal neuroimaging data. Multimodal approaches further demonstrate the value of combining MRI, PET, and clinical information, but commonly rely on relatively simple fusion mechanisms, whereas longitudinal approaches introduce temporal modeling without necessarily integrating multimodal interactions, disease-stage hierarchy, and anatomical priors simultaneously.

Accordingly, HCST-Former makes five methodological contributions that extend these existing research directions. **First**, CMAF moves beyond conventional concatenation or weighted fusion by explicitly learning interactions between heterogeneous modalities. **Second**, LTE extends cross-sectional AD classification toward longitudinal trajectory representation by modeling dependencies across multiple clinical timepoints. **Third**, HDSM explicitly represents the ordered CN→MCI→AD disease hierarchy rather than treating the three diagnostic stages as unrelated classes. **Fourth**, GGAA incorporates anatomical knowledge into the attention mechanism, enabling the learned representations to account for clinically relevant brain regions and their relationships. **Fifth**, UAP complements classification and progression prediction with predictive uncertainty, an aspect that remains comparatively underexplored in multimodal longitudinal AD modeling.

The resulting framework therefore addresses the research gap identified in the literature: the absence of a unified architecture that simultaneously integrates **multimodal feature interaction, longitudinal trajectory modeling, hierarchical disease-stage representation, anatomical**

knowledge, and uncertainty estimation. The contribution is subsequently evaluated not only through overall classification performance but also through progression prediction, ablation analysis, uncertainty evaluation, interpretability analysis, and computational efficiency, thereby examining whether the proposed integration provides measurable benefits over representative existing approaches.

The main contributions of this work are summarized as follows:

1. **Unified multimodal longitudinal framework:** A Hierarchical Cross-Modal Spatio-Temporal Transformer (HCST-Former) is proposed to address the limitations of existing predominantly cross-sectional and limited-modality AD models by jointly learning from structural MRI, PET, cognitive assessments, genetic biomarkers, and longitudinal clinical information.
2. **Explicit cross-modal and temporal modeling:** CMAF and LTE are introduced to overcome the limitations of conventional feature concatenation and independent temporal processing. CMAF learns bidirectional interactions among heterogeneous modalities, while LTE captures disease-related temporal dependencies across multiple clinical visits.
3. **Hierarchical disease progression representation:** HDSM explicitly models the CN→MCI→AD disease hierarchy, extending conventional multiclass classification toward stage-aware representation learning and progression modeling.
4. **Anatomically informed transformer learning:** GGAA integrates anatomical and disease-related prior knowledge into the attention mechanism, directing representation learning toward clinically relevant brain regions while improving interpretability.
5. **Uncertainty-aware prediction:** UAP incorporates uncertainty estimation into the prediction framework, extending conventional accuracy-oriented AD models toward confidence-aware disease-stage and progression prediction.
6. **Joint evaluation of predictive performance and efficiency:** The framework is evaluated on ADNI using disease-stage classification, progression prediction, uncertainty, interpretability, and computational metrics. Ablation experiments further examine the

individual and combined contributions of the proposed modules.

2. LITERATURE REVIEW

2.1 CNN-Based Alzheimer's Disease Diagnosis

Convolutional Neural Networks (CNNs) have remained a major deep learning paradigm for Alzheimer's disease (AD) diagnosis because of their ability to learn hierarchical spatial representations directly from neuroimaging data. Early studies demonstrated that three-dimensional CNNs can preserve volumetric information from structural MRI while learning discriminative patterns associated with AD-related atrophy [10-13]. Subsequent investigations incorporated multi-scale features, attention mechanisms, multimodal inputs, and clinical variables to improve classification and early-stage prediction [14-18]. These developments established CNNs as strong baselines for AD analysis, particularly for distinguishing AD from cognitively normal (CN) subjects. Recent evidence, however, indicates that CNN-based performance is strongly task-dependent. A 2026 systematic review of 70 deep learning studies reported that 2D and 3D CNNs remain the most frequently employed architectures, with binary AD-versus-CN classification generally achieving high performance, whereas multiclass CN/MCI/AD classification and MCI-to-AD conversion remain substantially more difficult [50]. Similar concerns regarding reproducibility and generalizability have been reported in broader evaluations of neuroimaging-based AD models [51]. The local receptive-field structure of convolution also limits direct modeling of long-range dependencies between spatially separated brain regions. This limitation is relevant because AD-related pathological changes involve distributed anatomical structures rather than isolated regions. Although deeper CNNs and attention modules can partially enlarge the effective receptive field, they do not inherently provide direct global interactions among distant anatomical tokens. Furthermore, multimodal CNNs commonly employ parallel branches or feature concatenation, which may not explicitly model conditional relationships between MRI, PET, cognitive, and genetic information [17], [18]. Thus, CNNs provide strong local anatomical representation but leave an important requirement for global dependency modeling and structured multimodal interaction. This observation motivates the transition toward transformer-based representation learning in HCST-Former, while retaining hierarchical feature extraction where beneficial.

2.2 Vision Transformer-Based Alzheimer's Disease Diagnosis

Vision Transformers (ViTs) were introduced to overcome the locality constraints of convolution by representing images as token sequences and learning global relationships through self-attention [8]. Their ability to establish direct interactions between distant tokens has made transformers increasingly attractive for AD neuroimaging, where pathological alterations can occur across distributed anatomical regions. Transformer-based AD models have subsequently been developed for MRI classification, multimodal imaging, and disease-stage prediction [21-29]. Recent systematic evidence confirms the rapid expansion of this research direction: a 2025 systematic review identified 68 studies investigating ViT and convolutional-transformer architectures for AD diagnosis and categorized their use according to architecture, modality, fusion strategy, and diagnostic objective [52]. Transformer-based multimodal approaches have also demonstrated strong potential. Recent studies have combined MRI and PET representations using transformer-based global correlation modeling [53], while unified transformer frameworks have integrated structural MRI with clinical and genetic variables within a common representation space [54]. A 2025 meta-analysis of 20 transformer-based multimodal studies involving 12,897 participants reported strong pooled diagnostic performance, supporting the potential of transformer-based multimodal learning for early AD analysis [55]. Despite these advances, conventional self-attention becomes increasingly expensive as the number of three-dimensional imaging tokens, modalities, and longitudinal observations grows. More importantly, global attention alone does not explicitly encode the temporal trajectory or hierarchical relationship among CN, MCI, and AD. Recent systematic evidence indicates that multiclass staging and longitudinal prediction remain considerably more challenging than binary AD-versus-CN classification [50], [56]. Therefore, the unresolved issue is not simply whether transformers can model global dependencies, but how they can be structured to jointly represent global anatomy, multimodal interactions, longitudinal progression, and disease-stage hierarchy while maintaining computational scalability. This requirement motivates the hierarchical spatio-temporal design of HCST-Former.

2.3 Multi-Modal Fusion for Alzheimer's Disease Diagnosis

AD is a heterogeneous disorder in which complementary information is distributed across imaging, cognitive, genetic, and clinical domains. Structural MRI provides anatomical information, PET captures metabolic or molecular alterations, cognitive assessments characterize functional impairment, and genetic biomarkers provide information related to disease susceptibility. Consequently, multimodal learning has increasingly been investigated as a means of obtaining a more comprehensive representation of disease than single-modality analysis [30–35]. Early multimodal approaches generally adopted feature concatenation, multi-task learning, or decision-level fusion [30–33]. More recent approaches have introduced attention and transformer mechanisms to improve the interaction between modalities. For example, transformer-based MRI-PET models have demonstrated the ability to learn global correlations between modality-specific representations [53], while unified multimodal transformer frameworks have integrated imaging, clinical, and genetic information within a common token representation [54]. Recent reviews confirm that MRI-PET fusion remains a major research direction and emphasize the complementary structural and metabolic information provided by the two modalities [57]. A 2025 meta-analysis further reported strong overall performance for transformer-based multimodal fusion approaches [55]. Nevertheless, multimodal integration remains methodologically challenging. A recent review identified data scarcity, class imbalance, inter-institutional heterogeneity, and differences in fusion strategies as important unresolved issues [57]. Moreover, simply concatenating modality-specific representations does not explicitly model how information from one modality influences the representation of another. This is particularly important in AD because the diagnostic significance of an anatomical alteration may depend on corresponding metabolic, cognitive, genetic, or clinical evidence. Recent systematic evidence also shows that multimodal studies frequently remain heterogeneous in both task definition and validation methodology [56]. Accordingly, the design requirement is to move beyond **static feature aggregation toward explicit cross-modal interaction learning**. HCST-Former addresses this requirement through **Cross-Modal Attention Fusion (CMAF)**, which enables information exchange among MRI, PET, cognitive, genetic, and clinical representations. CMAF therefore represents

a direct methodological response to the literature-derived need for dynamic multimodal interaction rather than merely increasing the number of input modalities.

2.4 Longitudinal Modeling for Alzheimer's Disease Progression

Alzheimer's disease is inherently progressive, and longitudinal observations can provide information about the direction and rate of neurodegenerative change that is unavailable from a single clinical visit. Early longitudinal approaches employed statistical trajectory models and recurrent architectures to characterize disease evolution [36–39]. More recent approaches have adopted attention and transformer mechanisms to model dependencies across multiple timepoints and to improve prediction of MCI-to-AD conversion [24], [25]. Recent research increasingly combines longitudinal information with multimodal imaging. For example, multimodal longitudinal frameworks have integrated MRI, PET, and CT observations to improve AD classification while accounting for repeated measurements [58]. More recent multimodal self-supervised approaches have additionally explored cross-modal alignment and longitudinal temporal consistency across multiple AD cohorts, demonstrating the increasing importance of jointly modeling temporal and multimodal information [59]. These developments indicate a transition from static diagnosis toward patient-specific disease trajectory representation. However, longitudinal modeling remains less developed than cross-sectional classification. A 2026 systematic review of 60 deep learning studies found that 47 used cross-sectional designs, whereas only 13 employed longitudinal data. The review further identified longitudinal modeling, multiclass staging, and multimodal integration as major unresolved challenges [56]. Heavy reliance on ADNI and heterogeneous validation protocols also limit conclusions regarding generalizability [50], [56]. In addition, many longitudinal approaches model a sequence of observations without explicitly representing the interaction between changes occurring across different modalities. These observations establish the requirement for joint temporal and multimodal representation learning. HCST-Former addresses this requirement through Longitudinal Temporal Encoding (LTE). Rather than treating individual visits as independent observations, LTE operates on fused multimodal representations and learns dependencies across successive timepoints. This formulation enables the model to represent both the current disease state and the temporal pattern associated with

progression. LTE therefore extends multimodal learning from cross-sectional classification toward explicit longitudinal disease-trajectory modeling, providing a basis for both early detection and progression prediction.

2.5 Anatomical Knowledge-Guided Modeling

The spatial distribution of AD-related pathology provides important prior information for representation learning. Neuropathological studies have established characteristic patterns involving the entorhinal cortex, hippocampus, medial temporal structures, and subsequently broader cortical regions [9]. Consequently, anatomical knowledge has been incorporated into AD models through region-of-interest extraction, hippocampal analysis, structural connectivity, functional connectivity, and graph-based learning [40-43]. These approaches demonstrate that clinically relevant anatomical regions can provide discriminative information beyond unconstrained image-level representations. Recent transformer-based approaches have further investigated anatomical interpretability. For example, recent PET/MR transformer research has emphasized the identification of disease-relevant regions while simultaneously learning local and global image representations [60]. Broader 2025–2026 reviews similarly identify explainability and clinically meaningful representation as important requirements for translating deep learning-based AD systems toward practical use [61]. These developments support the use of anatomical knowledge as an explicit component of disease representation rather than as a purely preprocessing step. Nevertheless, anatomical information is often introduced through predefined ROIs or independent graph modules rather than directly constraining the attention mechanism. ROI-based approaches may restrict representation learning to predefined regions, whereas unconstrained transformers may learn statistically useful but biologically less meaningful token interactions. Furthermore, the hierarchical progression from CN to MCI to AD is generally treated as a conventional multiclass classification problem rather than an ordered disease process. Recent systematic evidence confirms that performance becomes more challenging as the number of disease stages increases and emphasizes the need for interpretable and longitudinal stage-aware models [50], [56]. These observations motivate two complementary design requirements: anatomically guided attention and hierarchical disease-stage representation. HCST-Former addresses the first through Graph-

Guided Anatomical Attention (GGAA), which incorporates relationships among AD-relevant brain regions into the attention mechanism. The second is addressed through Hierarchical Disease-Stage Modeling (HDSM), which explicitly represents the CN→MCI→AD progression structure. Their combination enables anatomical information and disease-stage hierarchy to participate directly in representation learning.

2.6 Uncertainty Estimation in Medical Image Analysis

Predictive accuracy alone does not indicate whether an individual AD prediction is reliable. This issue is particularly important in early-stage disease, where CN and MCI may exhibit overlapping characteristics and where incomplete or heterogeneous clinical observations can increase prediction ambiguity. Consequently, uncertainty estimation has become an increasingly relevant component of clinically oriented deep learning systems. Bayesian neural networks, Monte Carlo Dropout, deep ensembles, and evidential learning have established the methodological foundations for estimating model and data uncertainty [44]-49]. Recent AD research has begun integrating uncertainty estimation with transformer-based learning. For example, a 2025 uncertainty-aware transformer ensemble used Monte Carlo Dropout to estimate confidence in genomic AD prediction and demonstrated that uncertainty thresholds could distinguish relatively reliable predictions from uncertain cases [62]. This work illustrates the potential value of confidence estimation beyond conventional classification accuracy. At the same time, recent research on incomplete clinical data has proposed transformer architectures that jointly address missing information and calibrated uncertainty, highlighting the importance of reliable predictions under heterogeneous real-world conditions [63]. However, uncertainty estimation is still commonly treated as an independent prediction component rather than being integrated with multimodal and longitudinal disease modeling. A model may provide a confidence score for a single-timepoint classification while ignoring uncertainty arising from missing modalities, temporal variability, or ambiguity between adjacent disease stages. Recent reviews of AD deep learning continue to emphasize clinical interpretability, reliability, generalization, and reproducibility as important barriers to translation [50], [61]. Therefore, the design requirement is to provide disease-stage and progression predictions together with explicit

confidence information. HCST-Former addresses this requirement through Uncertainty-Aware Prediction (UAP). UAP incorporates uncertainty estimation into the multimodal longitudinal prediction pipeline, allowing the model to associate its disease-stage and progression outputs with corresponding confidence estimates. This extends HCST-Former beyond accuracy-oriented prediction toward a more reliability-aware framework suitable for subsequent investigation in clinical decision-support settings.

2.7 Comparative Analysis and Research Gap

The evolution of deep learning for Alzheimer's disease diagnosis demonstrates a clear progression from CNN-based architectures toward transformer-based models with increasingly sophisticated multi-modal fusion, longitudinal modeling, and anatomical guidance. CNN-based methods, including 3D CNNs and attention-enhanced CNNs, have established strong baselines by learning hierarchical local features through convolutional operations [10, 11]. However, their limited receptive fields restrict the modeling of long-range anatomical dependencies, particularly in AD where pathological changes manifest across distributed brain networks. To overcome these limitations, Vision Transformer (ViT)-based diagnostic frameworks introduced self-attention mechanisms capable of directly modeling global anatomical relationships [20, 21]. Architectures such as ADViT, multimodal ViT, and longitudinal transformers demonstrate improved diagnostic accuracy by capturing long-range anatomical interactions [22–24]. However, the quadratic complexity of conventional self-attention results in excessive computational cost and memory consumption, limiting scalability to high-resolution multi-modal neuroimaging data.

Recent research has therefore focused on efficient transformer architectures with cross-modal fusion, longitudinal modeling, and anatomical guidance [25–27]. While these approaches significantly improve diagnostic performance, they often treat modalities independently, ignore temporal dynamics, or fail to incorporate anatomical priors effectively. Similarly, hierarchical transformer architectures improve multi-scale representation learning through progressive feature extraction, yet most methods construct hierarchical features through spatial down-sampling rather than explicitly modeling disease progression stages.

A comprehensive analysis of the literature reveals that no existing architecture simultaneously addresses five critical requirements: (i) effective

multi-modal feature fusion, (ii) longitudinal disease trajectory modeling, (iii) hierarchical disease-stage representation, (iv) anatomical knowledge-guided attention, and (v) uncertainty-aware prediction. Most CNN-based methods provide excellent local feature extraction but insufficient global anatomical modeling and cross-modal fusion, whereas transformer-based methods improve representational capacity at the expense of increased computational complexity. Likewise, current efficient transformers primarily optimize computational performance without explicitly incorporating disease progression knowledge into the attention mechanism. These observations motivate the development of the proposed Hierarchical Cross-Modal Spatio-Temporal Transformer (HCST-Former). Unlike existing diagnostic architectures, HCST-Former integrates Cross-Modal Attention Fusion (CMAF) to jointly model complementary information from multiple modalities, Longitudinal Temporal Encoding (LTE) to capture disease trajectory patterns, Hierarchical Disease-Stage Modeling (HDSM) to learn progressive changes across disease stages, Graph-Guided Anatomical Attention (GGAA) to incorporate prior knowledge of AD-relevant regions, and Uncertainty-Aware Prediction (UAP) to provide robust confidence estimates. By jointly addressing representation learning, computational scalability, and clinical interpretability, HCSTFormer aims to bridge the gap between diagnostic accuracy and clinical practicality for early Alzheimer's disease detection and progression prediction.

2.8 Research Gap Summary

Based on the comparative analysis, the following research gaps are identified:

- **G1: Poor Multimodal Feature Fusion.** Existing approaches employ simple fusion strategies that fail to capture complex interactions between structural MRI, PET, cognitive assessments, genetic biomarkers, and longitudinal clinical records [30, 31, 34, 35].
- **G2: Limited Longitudinal Modeling.** Current architectures inadequately capture disease progression trajectories from CN to MCI to AD, particularly during early stages where subtle changes are difficult to detect [24, 25, 38, 39].
- **G3: Weak Interpretability.** Existing models provide limited clinical interpretability, making it difficult for clinicians to understand the basis of predictions and disease progression patterns [26, 27, 40].

– **G4: Insufficient Uncertainty Estimation.** The above research gaps motivate the proposed Hierarchical Cross-Modal Spatio-Temporal Transformer (HCST-Former). The Current approaches lack robust uncertainty quantification, limiting their utility for risk-

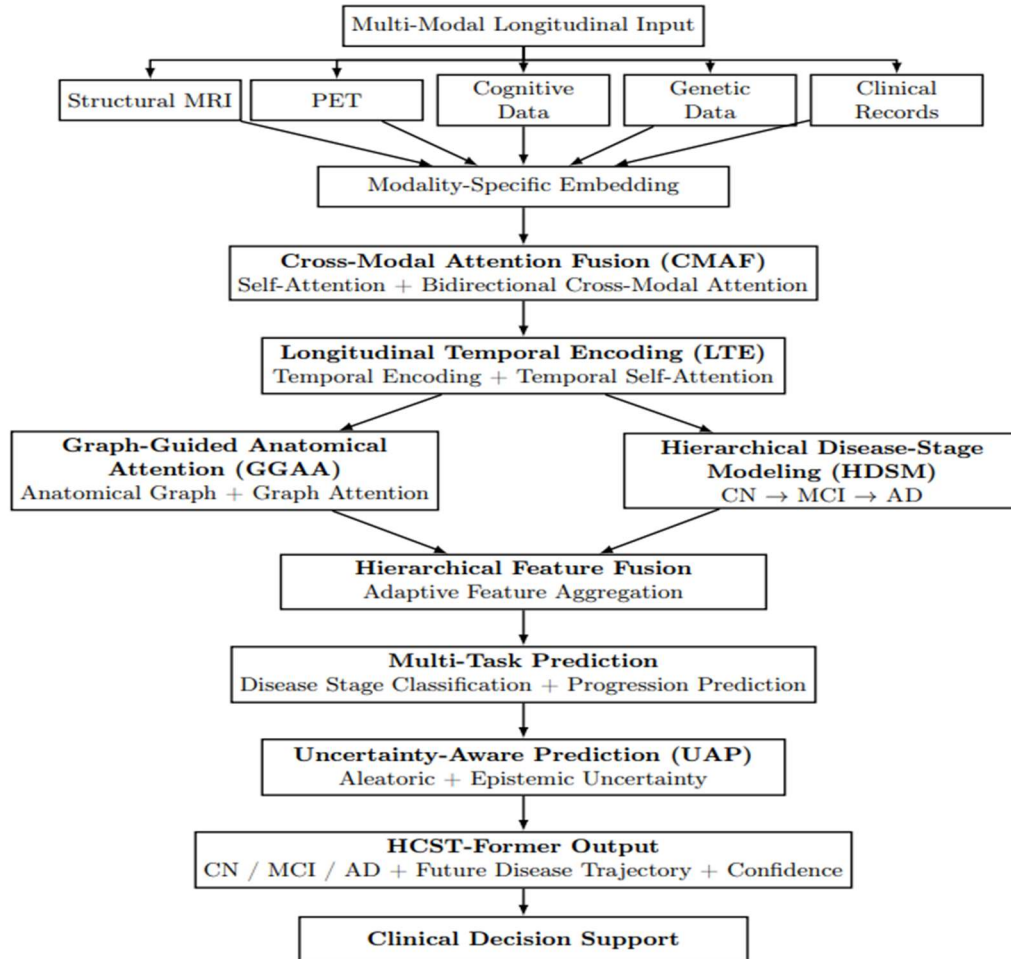


Figure 1: System Architecture of Proposed Method

stratified clinical decision support [48, 49].

– **G5: Limited Generalization Across Datasets.** Existing models exhibit limited generalization across different clinical cohorts, imaging protocols, and scanner manufacturers [6, 19].

– **G6: Inadequate Anatomical Knowledge Integration.** Current architectures fail to effectively incorporate prior knowledge of AD-relevant brain regions and disease progression patterns [26, 27, 42].

– **G7: Lack of a Unified Framework.** There is currently no framework that simultaneously integrates multi-modal fusion, longitudinal modeling, hierarchical disease-stage representation, anatomical guidance, and uncertainty estimation for early Alzheimer’s disease detection and progression prediction.

proposed framework introduces Cross-Modal Attention Fusion (CMAF) to jointly model complementary information from multiple data modalities, Longitudinal Temporal Encoding (LTE) to capture disease trajectory patterns, Hierarchical Disease-Stage Modeling (HDSM) to learn progressive changes across disease stages, Graph-Guided Anatomical Attention (GGAA) to incorporate prior knowledge of AD-relevant regions, and Uncertainty-Aware Prediction (UAP) to provide robust confidence estimates. By jointly addressing representation learning, computational efficiency, and clinical interpretability, HCST-Former provides a scalable and clinically actionable solution for early Alzheimer’s disease detection and progression prediction. The detailed architecture

and methodology are presented in the following section.

2.9 Formal Problem Statement and Research Questions

2.9.1 Formal Problem Statement

Alzheimer's disease progression is characterized by complex interactions among heterogeneous imaging, cognitive, genetic, clinical, spatial, and temporal information. Existing approaches often analyze these sources independently or rely on limited multimodal fusion, while longitudinal disease trajectories and the hierarchical transition from cognitively normal (CN) to mild cognitive impairment (MCI) and Alzheimer's disease (AD) are not jointly modeled. In addition, anatomical knowledge and prediction uncertainty are insufficiently integrated into a unified learning framework. These limitations make it challenging to detect subtle early-stage abnormalities and reliably predict subsequent disease progression. The identified requirements include multimodal fusion, longitudinal trajectory modeling, hierarchical disease-stage representation, anatomical guidance, uncertainty estimation, and computational scalability. Formally, let the longitudinal multimodal representation of a subject be defined as

$$\mathcal{M} = \{M_{\text{MRI}}^{1:T}, M_{\text{PET}}^{1:T}, M_{\text{Cog}}^{1:T}, M_{\text{Gen}}^{1:T}, M_{\text{Clin}}^{1:T}\},$$

where T denotes the number of available timepoints, M_{MRI} and M_{PET} represent structural and functional/molecular neuroimaging data, M_{Cog} represents cognitive assessments, M_{Gen} represents genetic biomarkers, and M_{Clin} represents longitudinal clinical information. The objective is to learn a unified function $F_{\theta}: \mathcal{M} \rightarrow \{\hat{Y}, \hat{P}, \hat{U}\}$,

where $\hat{Y} \in \{\text{CN}, \text{MCI}, \text{AD}\}$ denotes the predicted disease stage, $\hat{P}$ represents the predicted disease progression trajectory, and $\hat{U}$ denotes the associated predictive uncertainty. This formulation is consistent with the objective already defined in the proposed-method section of the manuscript. The learning problem is therefore to develop a computationally efficient transformer-based model that simultaneously: learns complementary representations across heterogeneous modalities; captures temporal dependencies and disease trajectories across multiple visits; explicitly models the hierarchical progression $\text{CN} \rightarrow \text{MCI} \rightarrow \text{AD}$; incorporates anatomical knowledge of AD-relevant

brain regions into feature learning; produces calibrated uncertainty estimates along with disease predictions; and maintains an appropriate balance between predictive performance and computational complexity. Thus, the central problem addressed in this work is: **“How can heterogeneous multimodal and longitudinal AD-related information be jointly represented and efficiently modeled to enable accurate early disease-stage identification and reliable prediction of progression from CN through MCI to AD, while incorporating anatomical knowledge and predictive uncertainty within a unified and computationally scalable framework?”**.

The proposed HCST-Former addresses this problem by integrating Cross-Modal Attention Fusion (CMAF), Longitudinal Temporal Encoding (LTE), Hierarchical Disease-Stage Modeling (HDSM), Graph-Guided Anatomical Attention (GGAA), and Uncertainty-Aware Prediction (UAP) within a unified architecture. These design objectives directly correspond to the research gaps identified in the manuscript.

2.9.2 Research Questions

Based on the formal problem definition, this study investigates the following research questions:

RQ1. Multimodal Representation:

Can cross-modal attention-based fusion of structural MRI, PET, cognitive assessments, genetic biomarkers, and longitudinal clinical information improve disease-stage classification compared with conventional single-modality or simple feature-fusion approaches?

RQ2. Longitudinal Progression:

Can explicit longitudinal temporal modeling improve the identification of subtle early-stage changes and prediction of disease progression across the $\text{CN} \rightarrow \text{MCI} \rightarrow \text{AD}$ trajectory?

RQ3. Hierarchical Disease Modeling:

Does explicitly modeling the hierarchical relationships among CN, MCI, and AD enable HCST-Former to better distinguish disease stages and characterize progressive neurodegenerative changes?

RQ4. Anatomical Knowledge Integration:

Can graph-guided anatomical attention improve the

model's ability to focus on clinically relevant AD-associated brain regions while providing more interpretable disease-related representations?

RQ5. Prediction Reliability:
Can uncertainty-aware prediction provide more reliable and calibrated confidence estimates for early AD detection and progression prediction?

RQ6. Computational Efficiency:
Can the proposed unified architecture achieve improved or competitive predictive performance while maintaining computational efficiency in terms of model parameters, computational complexity, GPU memory consumption, and inference requirements?

2.10 Justification for the Selection of Core Factors

The design of HCST-Former is guided by limitations consistently observed across the reviewed literature rather than by the independent aggregation of architectural modules. Existing studies demonstrate that AD diagnosis requires complementary information from structural, functional, cognitive, genetic, and clinical sources; however, many multimodal approaches employ concatenation, addition, or weighted fusion, which provides limited explicit modeling of inter-modal relationships. This motivates the selection of **cross-modal interaction** as the first core factor and the development of **Cross-Modal Attention Fusion (CMAF)**, which enables modality-specific representations to interact through attention rather than relying solely on static feature aggregation.

A second core factor is **longitudinal disease modeling**. AD is inherently progressive, and the reviewed literature indicates that pathological changes accumulate over months or years before overt clinical manifestations. Although recurrent and transformer-based longitudinal approaches have been investigated, the manuscript identifies limitations in explicitly modeling temporal disease dynamics and integrating them with heterogeneous multimodal information. Consequently, **Longitudinal Temporal Encoding (LTE)** is selected to represent dependencies across repeated observations and encode both current disease state and temporal progression.

The third factor, **hierarchical disease-stage modeling**, is motivated by the ordered

clinical progression represented in this study as $CN \rightarrow MCI \rightarrow AD$. Conventional multiclass models can treat diagnostic categories as independent labels, whereas the proposed formulation aims to exploit their progressive relationship. The literature review specifically identifies the limited consideration of this disease-stage hierarchy in existing anatomical and transformer-based approaches. Accordingly, **Hierarchical Disease-Stage Modeling (HDSM)** is introduced to learn stage-specific and transition-sensitive representations, with particular emphasis on subtle differences between CN and MCI.

The fourth factor is **anatomical prior knowledge**. The reviewed literature reports that AD-related alterations involve clinically relevant regions including the hippocampus, entorhinal cortex, and medial temporal structures, while existing knowledge-guided approaches demonstrate the potential value of incorporating brain connectivity and region-specific information. However, the review identifies limited explicit integration of such priors into attention mechanisms. This motivates **Graph-Guided Anatomical Attention (GGAA)**, in which anatomical relationships constrain and guide attention toward disease-relevant brain regions rather than relying exclusively on data-driven token interactions.

Finally, **predictive uncertainty** is selected because classification accuracy alone does not characterize the reliability of individual predictions. Existing uncertainty-estimation methods, including Bayesian, ensemble, and evidential approaches, provide mechanisms for confidence estimation, but their integration with multimodal longitudinal AD modeling remains comparatively underexplored in the literature reviewed in this study. Therefore, **Uncertainty-Aware Prediction (UAP)** is incorporated to complement disease-stage and progression outputs with confidence estimates. Collectively, these factors address distinct but complementary limitations identified in the literature: **CMAF** addresses restricted multimodal interaction, **LTE** addresses insufficient longitudinal modeling, **HDSM** addresses the limited representation of ordered disease progression, **GGAA** addresses inadequate incorporation of anatomical knowledge, and **UAP** addresses prediction reliability. Their integration therefore forms the methodological basis of HCST-Former rather than representing an arbitrary combination of architectural components.

3. PROPOSED METHOD

3.1 Overview of the Proposed Framework

To address the limitations identified in Section 2, this paper proposes the Hierarchical Cross-Modal Spatio-Temporal Transformer (HCST-Former), a novel transformer-based architecture designed for early Alzheimer's disease detection and progression prediction. As illustrated in Figure 1, the proposed framework integrates multi-modal feature fusion, longitudinal temporal encoding, hierarchical disease-stage modeling, graph-guided anatomical attention, and uncertainty-aware prediction within a unified transformer architecture. Unlike conventional diagnostic models that process modalities and timepoints independently, HCST-Former explicitly exploits cross-modal relationships and temporal disease trajectories while reducing the computational burden associated with dense self-attention.

Given an input comprising multi-modal data, $M = \{M_{MRI}, M_{PET}, M_{Cog}, M_{Gen}, M_{Clin}\}$, where M_{MRI} , M_{PET} , M_{Cog} , M_{Gen} , and M_{Clin} denote structural MRI, PET imaging, cognitive assessments, genetic biomarkers (e.g., APOE $\epsilon 4$), and longitudinal clinical records, respectively, the objective is to learn a mapping, $F : M \rightarrow \{Y, P, U\}$, where Y represents the predicted disease-stage classification (CN, MCI, or AD), P denotes the predicted progression trajectory, and U provides uncertainty estimates. The mapping function is learned end-to-end by jointly optimizing feature representation, cross-modal fusion, temporal encoding, disease-stage modeling, and uncertainty estimation.

3.2 Multi-Modal Feature Extraction

The proposed HCST-Former begins with modality-specific feature extraction to obtain semantically meaningful representations from each data modality.

MRI Feature Extraction: Structural MRI volumes are first preprocessed using standard neuroimaging pipelines, including bias field correction, skull stripping, and spatial normalization to the MNI152 template. The preprocessed volumes are then partitioned into 3D patches of size $16 \times 16 \times 16$ and projected into a C-dimensional embedding space through learnable linear projections.

$$T_{MRI} = \{t_{MRI}^1, t_{MRI}^2, \dots, t_{MRI}^N\}, t_{MRI}^i \in R^C$$

PET Feature Extraction: PET images undergo similar preprocessing and patch-based tokenization.

$$T_{PET} = \{t_{PET}^1, t_{PET}^2, \dots, t_{PET}^N\}, t_{PET}^i \in R^C$$

Cognitive Assessment Encoding: Cognitive scores, such as MMSE, ADAS-Cog, and CDR, are encoded through a learnable projection.

$$E_{Cog} = f_{Cog}(s_{Cog}), E_{Cog} \in R^C$$

where $f_{Cog}(\cdot)$ is a multi-layer perceptron and $s_{Cog} \in R^{D_{Cog}}$ denotes the cognitive assessment vector.

Genetic Biomarker Encoding: Genetic information, particularly APOE $\epsilon 4$ carrier status, is encoded as:

$$E_{Gen} = f_{Gen}(g_{APOE}), E_{Gen} \in R^C$$

where $f_{Gen}(\cdot)$ is a learnable embedding function.

Longitudinal Clinical Record Encoding: Clinical records across multiple timepoints are encoded using a temporal attention mechanism:

$$E_{Clin} = \text{Attn}_T(h_1, h_2, \dots, h_T), E_{Clin} \in R^C$$

where h_t denotes encoded clinical features at timepoint t , and $\text{Attn}_T(\cdot)$ represents temporal attention.

3.3 Cross-Modal Attention Fusion (CMAF)

Existing multi-modal AD diagnosis models typically employ simple concatenation or weighted fusion of modality-specific features, failing to capture complex cross-modal interactions. To address this limitation, the proposed Cross-Modal Attention Fusion (CMAF) module explicitly models bidirectional interactions between modalities through cross-modal attention.

Given modality-specific token sequences $T_i \in R^{N_i \times C}$ and modality-level embeddings $E_m \in R^C$, cross-modal attention is computed as:

$$CMA(Q_i, K_j, V_j) = \text{Softmax}\left(\frac{(Q_i K_j^T)}{\sqrt{d}}\right) V_j$$

where Q_i is derived from the source modality, while K_j and V_j are derived from the target modality. For each pair of modalities (i, j) , cross-modal attention enables selective information exchange:

$$T_i \rightarrow j = CMA(Q_i, K_j, V_j)$$

The fused representation for each modality integrates both self-attention and cross-modal attention:

$$T_i^{fused} = \alpha_i \text{SelfAttn}(T_i) + \sum_j \beta_{ij} CMA(T_i, T_j)$$

Where α_i, β_{ij} are learnable weighting coefficients controlling the contributions of self-attention and cross-modal attention. Multi-modal fusion is subsequently performed through a hierarchical attention network:

$$T_{fusion} = \text{CrossModalFusion}(T_{MRI}^{fused}, T_{PET}^{fused}, E_{Cog}, E_{Gen}, E_{Clin})$$

This adaptive fusion mechanism enables the model to dynamically weight different modalities according to their informativeness for the diagnostic task.

3.4 Longitudinal Temporal Encoding (LTE)

Modeling disease progression requires capturing temporal dynamics across multiple timepoints. The proposed Longitudinal Temporal Encoding (LTE) module explicitly models disease trajectory patterns using temporal attention and positional encoding. Given multi-modal features at T timepoints, $\{F_1, F_2, \dots, F_T\}$, where F_t denotes fused cross-modal features at timepoint t , temporal encoding is performed through the following stages. **Temporal Positional Encoding**, Learnable positional embeddings are added to each timepoint:

$$F \sim_t = F_t + PE(t), PE(t) \in R^C$$

Temporal Self-Attention: The temporal relationships between timepoints are modeled through self-attention:

$$F_t^{temp} = \text{TemporalAttn}(F \sim_1, F \sim_2, \dots, F \sim_T)$$

Disease Trajectory Encoding: The progression pattern is encoded through a trajectory encoding mechanism:

$$F_{traj} = \text{TrajectoryEnc}(F_1^{temp}, F_2^{temp}, \dots, F_T^{temp})$$

The trajectory encoding captures both the current disease state and the rate of progression, enabling early detection of subtle changes.

3.5 Hierarchical Disease-Stage Modeling (HDSM)

Alzheimer's disease progression follows a hierarchical pattern from CN to MCI to AD. The proposed Hierarchical Disease-Stage Modeling (HDSM) module explicitly learns progressive anatomical and functional changes across disease stages through hierarchical representation learning. **Stage-Specific Encoders**, Separate encoders are trained for each disease stage:

$$F_{stage} = \{ \text{Enc}_{CN(F)}, \text{if } y = CN; \text{Enc}_{MCI(F)}, \text{if } y = MCI; \text{Enc}_{AD(F)}, \text{if } y = AD \}$$

Hierarchical Progression Modeling, The progression from one stage to another is modeled through hierarchical attention:

$$F_{prog} = \text{HierarchicalAttn}(F_{CN}, F_{MCI}, F_{AD})$$

Early Detection Module, A specialized module captures early-stage changes:

$$F_{early} = \text{EarlyDetection}(F_{CN}, F_{MCI})$$

This module focuses on subtle differences between CN and MCI subjects, enabling early detection of neurodegenerative changes.

3.6 Graph-Guided Anatomical Attention (GGAA)

Incorporating prior knowledge of AD-relevant brain regions is essential for both accuracy and interpretability. The proposed Graph-Guided Anatomical Attention (GGAA) module leverages anatomical knowledge to guide attention toward clinically meaningful regions. **Anatomical Graph Construction** An anatomical graph is constructed as $G = (V, E)$, where vertices V represent brain regions and edges E represent anatomical or functional connections. Regions are defined according to standardized brain atlases, such as AAL or Desikan-Killiany. **Graph-Constrained Attention**, Attention is constrained to follow anatomical structure: $A_{GGAA} = A \odot A_{\text{graph}}$, where A represents standard attention, A_{graph} represents a graph-derived attention mask, and $\odot$ denotes element-wise multiplication. **AD-Relevant Region Weighting** Weights are assigned to brain regions according to their relevance to AD pathology:

$$w_i = \text{RelevanceScore}(v_i), v_i \in V$$

Regions known to be affected early in AD, including the hippocampus, entorhinal cortex, and medial temporal lobe, receive higher weights. **Attention-Guided Feature Aggregation**, Features are aggregated according to anatomical importance:

$$F_{anat} = \sum_i w_i F_i$$

This mechanism ensures that the model focuses on clinically meaningful regions while providing interpretable attention maps.

3.7 Uncertainty-Aware Prediction (UAP)

Robust uncertainty estimation is critical for clinical decision support. The proposed Uncertainty-Aware Prediction (UAP) module provides confidence estimates through evidential deep learning and Monte Carlo Dropout.

Evidential Deep Learning, the model learns evidence parameters $(\alpha_1, \alpha_2, \dots, \alpha_K)$ for K classes: $\alpha_k = 1 + \text{Softplus}(Z_k)$, where Z_k denotes the logit corresponding to class k . The resulting Dirichlet distribution is used to model classification uncertainty. **Aleatoric Uncertainty**, Data uncertainty is estimated through:

$$U_{ale} = \left(\frac{K}{K} \right) \sum_{k=1}^K \alpha_k$$

Epistemic Uncertainty, Model uncertainty is estimated through Monte Carlo Dropout with M stochastic forward passes:

$$U_{epi} = \left(\frac{1}{M}\right) \sum_{m=1}^M (p_m - \bar{p})^2$$

Total Uncertainty, The total uncertainty combines both components:

$$U_{total} = U_{ale} + U_{epi}$$

3.8 Multi-Level Segmentation Decoder

The decoder progressively reconstructs voxel-wise segmentation maps through hierarchical feature fusion: $F_d = \phi(W_v F_v + W_r F_r + W_o F_o)$, where W_v , W_r , and W_o are learnable projection matrices, $\phi(\cdot)$ denotes a nonlinear transformation, and F_v , F_r , and F_o represent hierarchical feature representations. Spatial resolution is progressively recovered through upsampling operations,

$F_l + 1 = Up(F_l)$, The final segmentation output is obtained through a softmax activation,

$$Y^{\square} = Softmax(Conv(F_{final}))$$

3.9 Loss Function

The proposed HCST-Former is trained using a composite loss function:

$$L = \lambda_1 L_{Dice} + \lambda_2 L_{CE} + \lambda_3 L_{stage} + \lambda_4 L_{unc}$$

where L_{Dice} and L_{CE} are Dice and Cross-Entropy losses for segmentation, L_{stage} is a stage classification loss, and L_{unc} is an uncertainty regularization term. **Dice Loss**, The Dice loss is defined as:

$$L_{Dice} = 1 - \frac{(2 \sum_i p_i y_i + \varepsilon)}{\sum_i p_i + \sum_i y_i + \varepsilon}$$

Cross-Entropy Loss, The cross-entropy loss is defined as: $L_{CE} = -\left(\frac{1}{N}\right) \sum_i \Sigma_c y_{ic} \log p_i$

Stage Classification, Loss The stage classification loss is defined as:

$$L_{stage} = -\left(\frac{1}{N}\right) \sum_i \log(p_{stage(y_i)})$$

Uncertainty Regularization, the uncertainty regularization term is defined as:

$$L_{unc} = \sum_i \left[\left(\frac{1}{2\sigma_i^2} \right) L_i + \log \sigma_i \right]$$

where σ_i denotes the estimated uncertainty for sample i .

4. EXPERIMENTS

4.1 Dataset and Preprocessing

The proposed HCST-Former framework is comprehensively evaluated on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, which stands as one of the most extensively utilized

and widely recognized benchmark datasets for Alzheimer's disease diagnosis and progression prediction research. The ADNI dataset was launched in 2003 as a public-private partnership with the primary goal of defining the progression of Alzheimer's disease through the collection and analysis of multi-modal data. This dataset has become the de facto standard for evaluating computational methods in Alzheimer's disease research due to its large scale, longitudinal nature, and comprehensive multi-modal data collection. The dataset comprises a rich and diverse collection of multi-modal data, offering a comprehensive view of each subject's health status across multiple dimensions. Specifically, the dataset includes structural MRI scans (T1-weighted), which provide detailed anatomical information about brain structure and volume. These MRI scans are essential for identifying atrophy patterns in brain regions susceptible to Alzheimer's disease pathology. Additionally, the dataset incorporates FDG-PET and AV45-PET scans, which provide functional and molecular imaging information. FDG-PET measures glucose metabolism in the brain, which is typically reduced in Alzheimer's disease, while AV45-PET specifically targets amyloid-beta plaques, one of the hallmark pathological features of Alzheimer's disease. Beyond neuroimaging, the dataset includes comprehensive cognitive assessments, such as the Mini-Mental State Examination (MMSE), the Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog), and the Clinical Dementia Rating (CDR) scale. These cognitive evaluations are crucial for assessing functional decline and cognitive impairment associated with disease progression.

The dataset also incorporates genetic biomarkers, particularly APOE $\epsilon 4$ carrier status, which is widely recognized as the strongest genetic risk factor for late-onset Alzheimer's disease. Furthermore, longitudinal clinical records are collected across 2 to 5 timepoints per subject, enabling the study of disease trajectories and progression patterns over time. This longitudinal aspect is particularly valuable for understanding how the disease evolves and for predicting future clinical outcomes. The dataset includes subjects spanning three distinct diagnostic categories, carefully diagnosed according to established clinical criteria. The first category comprises Cognitively Normal (CN) subjects, consisting of 382 individuals who serve as healthy controls with no evidence of cognitive impairment. The second category includes subjects with Mild Cognitive

Impairment (MCI), representing 428 individuals who exhibit cognitive decline greater than expected for their age but do not meet the criteria for dementia. This group is particularly important for early detection studies, as MCI often represents a transitional stage between normal aging and Alzheimer's disease. The third category consists of subjects diagnosed with Alzheimer's Disease (AD), comprising 240 individuals who meet the clinical criteria for probable AD dementia. This balanced representation across the three diagnostic categories enables robust evaluation of classification performance at different disease stages.

4.2 Data Preprocessing

To ensure consistent, reliable, and reproducible training and evaluation, a unified preprocessing pipeline is adopted for all imaging data. This pipeline follows standard neuroimaging practices and is designed to minimize variability introduced by differences in acquisition protocols, scanner hardware, and subject positioning. The preprocessing pipeline is applied uniformly to both MRI and PET images to ensure comparability across modalities and subjects. First, all MRI and PET images are resampled to a uniform voxel spacing of 1 mm isotropic resolution. This standardization of spatial resolution is essential for ensuring that subsequent analyses are not biased by differences in image resolution and that features extracted from different subjects are directly comparable.

The resampling process uses appropriate interpolation techniques to preserve image quality while achieving the desired resolution. Following resampling, intensity normalization is performed using modality-specific normalization strategies tailored to the characteristics of each imaging modality. For structural MRI scans, intensity normalization typically involves correcting for intensity inhomogeneities caused by magnetic field variations and standardizing the intensity distribution across subjects. For PET images, normalization accounts for differences in tracer uptake and scanner calibration, ensuring that quantitative measures of metabolism or amyloid burden are comparable across subjects and timepoints. After normalization, each volume is cropped to a fixed spatial resolution of $128 \times 128 \times 128$ voxels. This cropping operation removes extraneous background regions and focuses computational resources on the brain parenchyma, reducing the input dimensionality while preserving all clinically relevant anatomical information. The

fixed spatial resolution also facilitates consistent patch-based tokenization across all subjects.

To improve model robustness and reduce overfitting, online data augmentation is applied during training. The augmentation pipeline includes a variety of transformations designed to simulate real-world variability in imaging data. Random flipping along the left-right axis accounts for potential asymmetries, while random rotation introduces variations in head positioning that may occur during scanning. Random scaling accounts for differences in head size and brain volume, and intensity perturbation simulates variations in scanner calibration and signal-to-noise ratio. These augmentations are applied dynamically during training, effectively increasing the diversity of the training data and improving the model's ability to generalize to unseen subjects and imaging conditions. The resulting preprocessed volumes are subsequently divided into non-overlapping 3D patches of size $16 \times 16 \times 16$ for transformer-based feature extraction. This patch-based tokenization strategy is a fundamental component of the transformer architecture, enabling the model to process volumetric data efficiently while preserving spatial relationships between local regions. The patch size of $16 \times 16 \times 16$ is chosen to balance the trade-off between computational efficiency and the ability to capture fine-grained anatomical details.

4.3 Evaluation Metrics

The performance of the proposed HCST-Former is quantitatively evaluated using multiple complementary metrics that collectively provide a comprehensive assessment of model capabilities across different aspects of the diagnostic and prediction tasks. These metrics span four broad categories: classification performance, progression prediction accuracy, uncertainty estimation quality, and computational efficiency.

Classification Metrics: For disease-stage classification, a set of standard and widely accepted metrics is employed to evaluate the discriminative performance of the proposed model. These metrics collectively provide a nuanced understanding of model performance, capturing different aspects of classification quality and ensuring that the evaluation is robust to class imbalance and other potential biases. **Accuracy** is the most fundamental metric, measuring the proportion of correctly classified subjects among all evaluated subjects. While accuracy provides a straightforward measure of overall performance, it can be misleading in cases of class imbalance, which motivates the use of additional metrics. **Area Under the ROC Curve**

(AUC) is a comprehensive measure of the model's ability to discriminate between different diagnostic categories. The ROC curve plots the true positive rate against the false positive rate at various threshold settings, and the AUC represents the probability that the model will rank a randomly chosen positive instance higher than a randomly chosen negative instance. AUC values close to 1 indicate excellent discriminative ability, while values close to 0.5 suggest performance no better than random chance. **Sensitivity**, also known as recall or true positive rate, measures the proportion of positive subjects correctly identified by the model. In the context of Alzheimer's disease diagnosis, sensitivity reflects the model's ability to correctly identify subjects with MCI or AD, which is crucial for ensuring that affected individuals receive appropriate clinical attention and intervention. **Specificity** measures the proportion of negative subjects correctly identified by the model. High specificity ensures that healthy individuals are not incorrectly diagnosed with disease, reducing the

risk of unnecessary anxiety, additional testing, and potentially harmful interventions. **Precision** measures the proportion of predicted positive samples that are actually positive. Precision is particularly important in clinical settings where false positives can lead to significant consequences for patients and healthcare systems. High precision indicates that when the model predicts a positive diagnosis, it is highly likely to be correct. **F1-Score** represents the harmonic mean of precision and sensitivity, providing a single balanced measure that accounts for both false positives and false negatives. The F1-score is particularly useful when comparing models, as it penalizes both types of errors and provides a more holistic view of performance than either precision or sensitivity alone.

4.4 Progression Prediction Metrics

For longitudinal disease progression prediction, the evaluation focuses on the model's ability to forecast future disease trajectories and predict the timing of clinical events. These metrics

Algorithm 1 Hierarchical Cross-Modal Spatio-Temporal Transformer (HCST-Former)

Require: Multi-modal data $\mathcal{M} = \{M_{MRI}, M_{PET}, M_{Cog}, M_{Gen}, M_{Clin}\}$, timepoints T , embedding dimension C , classes $K = 3$, learning rate η , epochs E

Ensure: Disease stage Y , progression trajectory P , uncertainty U

```

1: Initialize model parameters  $\theta$ 
2: for  $e = 1$  to  $E$  do
3:   for each subject in  $\mathcal{M}$  do
4:     Multi-Modal Feature Extraction:
5:      $T_{MRI}, T_{PET} \leftarrow \text{PatchEmbed}(M_{MRI}, M_{PET})$ 
6:      $E_{Cog}, E_{Gen}, E_{Clin} \leftarrow \text{Encode}(s_{Cog}, g_{APOE}, h_{1:T})$ 
7:     Cross-Modal Attention Fusion:
8:      $F_{fused} \leftarrow \text{CMAF}(T_{MRI}, T_{PET}, E_{Cog}, E_{Gen}, E_{Clin})$ 
9:     Longitudinal Temporal Encoding:
10:     $F_{traj} \leftarrow \text{LTE}(F_1, \dots, F_T)$ 
11:    Hierarchical Disease-Stage Modeling:
12:     $F_{prog} \leftarrow \text{HDSM}(F_{CN}, F_{MCI}, F_{AD})$ 
13:    Graph-Guided Anatomical Attention:
14:     $F_{anat} \leftarrow \text{GGAA}(F_{fused}, G)$ 
15:    Uncertainty-Aware Prediction:
16:     $U \leftarrow \text{UAP}(F_{anat}, F_{traj}, F_{prog})$ 
17:    Multi-Level Decoding:
18:     $\hat{Y} \leftarrow \text{Decode}(F_{anat}, F_{prog}, F_{traj})$ 
19:    Joint Loss:
20:     $\mathcal{L} \leftarrow \lambda_1 \mathcal{L}_{Dice} + \lambda_2 \mathcal{L}_{CE} + \lambda_3 \mathcal{L}_{stage} + \lambda_4 \mathcal{L}_{unc}$ 
21:     $\theta \leftarrow \theta - \eta \nabla_{\theta} \mathcal{L}$ 
22:   end for
23: end for
24: return  $Y, P, U$ 

```

Figure 2. Proposed HSCT Transformer Algorithm

are particularly important for clinical applications, as they directly inform prognosis and treatment planning. **Time-to-Conversion Prediction Accuracy** measures the accuracy of predicting the time required for an MCI subject to convert to AD. This metric is expressed as the Mean Absolute Error (MAE) in months, with lower values indicating more accurate predictions. Accurate prediction of conversion time is clinically valuable because it enables healthcare providers to anticipate disease progression and plan interventions accordingly. **Concordance Index (C-index)** measures the agreement between predicted and observed disease progression rankings. The C-index is a generalization of the AUC for survival analysis and represents the probability that the

model correctly orders pairs of subjects by their time-to-event outcomes. A C-index of 0.5 indicates random prediction, while a value of 1 indicates perfect concordance. This metric is particularly appropriate for the time-to-conversion task, as it accounts for censoring and provides a robust measure of ranking performance. **Brier Score** measures the accuracy of probabilistic predictions by evaluating the squared difference between predicted probabilities and observed outcomes. The Brier Score ranges from 0 to 1, with lower values indicating better-calibrated probabilistic predictions. This metric is essential for assessing the model's ability to provide reliable probability estimates, which are critical for informed clinical decision-making.

Table 1. Classification Performance Comparison on the ADNI Dataset

Method	Accuracy (%)	AUC (%)	Sensitivity (%)	Specificity (%)	F1-Score
3D-CNN	84.2 ± 1.8	89.4 ± 1.5	83.1 ± 2.1	85.3 ± 1.7	0.841
Multi-modal CNN	87.5 ± 1.4	91.8 ± 1.2	86.7 ± 1.8	88.3 ± 1.5	0.875
Attention-CNN	88.2 ± 1.3	92.5 ± 1.1	87.5 ± 1.6	89.0 ± 1.4	0.882
ADViT	89.1 ± 1.1	93.2 ± 0.9	88.3 ± 1.4	89.9 ± 1.2	0.891
Multi-modal ViT	90.3 ± 1.0	94.1 ± 0.8	89.6 ± 1.3	91.0 ± 1.1	0.903
Longitudinal Transformer	91.2 ± 0.9	94.8 ± 0.7	90.4 ± 1.1	92.0 ± 1.0	0.912
Swin Transformer	91.8 ± 0.8	95.2 ± 0.6	91.0 ± 1.0	92.6 ± 0.9	0.918
MedFormer	92.4 ± 0.7	95.6 ± 0.5	91.8 ± 0.9	93.0 ± 0.8	0.924
HCST-Former (Ours)	93.7 ± 0.6	96.2 ± 0.4	92.8 ± 0.8	94.5 ± 0.7	0.937

4.5 Uncertainty Metrics

The reliability and calibration of the predictive probabilities are evaluated using two complementary metrics that assess different aspects of uncertainty quantification. These metrics are particularly important for clinical applications, where the trustworthiness of model predictions is paramount. **Expected Calibration Error**

(ECE) measures the difference between predicted confidence and observed accuracy. Calibration is evaluated by partitioning predictions into bins based on their confidence scores and computing the average accuracy within each bin. The ECE is the weighted average of the absolute differences between predicted confidence and observed accuracy across all bins. Lower ECE values indicate better calibration, meaning that the model's

confidence estimates are reliable and accurately reflect the true probability of correctness. **Negative Log-Likelihood (NLL)** measures the quality of probabilistic predictions by penalizing incorrect predictions made with high confidence. The NLL is computed as the negative logarithm of the predicted probability assigned to the true class. High-confidence incorrect predictions result in very large NLL values, while correct predictions with high confidence result in small NLL values. This metric provides a comprehensive assessment of probabilistic prediction quality, balancing the trade-off between confidence and accuracy.

4.6 Computational Metrics

Computational efficiency is a critical consideration for clinical deployment, where resource constraints and real-time requirements often necessitate efficient model architectures. The computational efficiency of HCST-Former is evaluated using four key metrics. **Number of Parameters** is the total number of trainable model parameters, which directly influences memory requirements and storage needs. Models with fewer parameters are generally more memory-efficient and potentially more suitable for deployment on resource-constrained devices. **FLOPs (Floating-Point Operations)** measures the computational complexity of the model during inference. FLOPs represent the total number of arithmetic operations required to process a single input and are directly related to inference time and energy consumption. Lower FLOPs indicate more efficient models that can process inputs more quickly and with less power consumption. **GPU Memory** refers to the peak GPU memory consumption during inference, which determines the hardware requirements for model deployment. Lower GPU memory requirements enable the deployment of models on more accessible and cost-effective hardware, potentially facilitating broader clinical adoption. **Inference Time** is the average time required to process one subject, representing the practical throughput of the model in clinical settings. Faster inference times enable higher throughput and reduce the latency of diagnostic decisions, which is particularly important in busy clinical environments.

4.7 Baseline Methods

The proposed HCST-Former is compared against representative state-of-the-art approaches spanning four major architectural paradigms: CNN-based architectures, transformer-based architectures, multi-modal fusion methods, and graph-based architectures. This comprehensive comparison ensures that the evaluation is thorough

and that the advantages of HCST-Former are clearly demonstrated across different methodological approaches.

Convolutional Neural Networks (CNNs) have served as the foundation of automated Alzheimer's disease diagnosis for over a decade, and several representative CNN-based methods are included in the comparison. The **3D-CNN** architecture represents a fundamental approach that directly processes volumetric MRI data using 3D convolutional operations. This method preserves spatial information and captures local anatomical features through hierarchical feature extraction. Despite its relatively simple architecture, the 3D-CNN has demonstrated consistent performance across multiple Alzheimer's disease datasets and serves as a strong baseline for evaluating more complex approaches. The **Multi-modal CNN** extends the basic CNN architecture to incorporate multiple imaging modalities and clinical data. This framework combines information from structural MRI and PET imaging through separate network branches, enabling the model to leverage complementary information from different imaging modalities. The multi-modal CNN also incorporates cognitive scores to provide a more comprehensive assessment of disease status. The **Attention-CNN** architecture integrates attention mechanisms into the CNN framework to focus on the most discriminative brain regions. This approach addresses the limitation of standard CNNs, which treat all regions equally and may fail to emphasize clinically relevant areas. By learning to attend to AD-relevant regions such as the hippocampus and entorhinal cortex, the Attention-CNN achieves improved interpretability and diagnostic accuracy.

Vision Transformer (ViT)-based architectures have recently emerged as powerful alternatives to CNNs for medical image analysis, offering the ability to model long-range anatomical dependencies through self-attention mechanisms. **ADViT** is one of the earliest transformer-based approaches for Alzheimer's disease diagnosis, adapting the standard Vision Transformer architecture to structural MRI analysis. ADViT demonstrates the feasibility of using transformer architectures for neuroimaging classification tasks and provides a foundational baseline for subsequent transformer-based methods. The **Multi-modal ViT** extends the standard Vision Transformer to incorporate multiple imaging modalities, enabling the model to learn joint representations from MRI and PET data. This approach leverages the self-

attention mechanism of transformers to model cross-modal relationships, potentially capturing complementary information that is not accessible from single-modal analysis. The **Longitudinal Transformer** introduces temporal modeling capabilities to the transformer architecture, enabling the analysis of disease progression across multiple timepoints. This method explicitly models the temporal dynamics of Alzheimer's disease, capturing trajectory patterns that are essential for progression prediction and early detection. The **Swin Transformer** is a hierarchical vision transformer that uses shifted window-based attention to achieve linear complexity with respect to image resolution. The Swin Transformer has demonstrated impressive performance on natural image tasks and has been adapted to medical image analysis, where it offers an efficient alternative to standard ViT architectures. **MedFormer** is a multi-scale transformer framework specifically designed for medical image segmentation and analysis. MedFormer incorporates hierarchical feature extraction and cross-scale attention mechanisms to capture both local details and global context, making it particularly suitable for neuroimaging analysis.

Multi-modal fusion methods are specifically designed to integrate information from complementary data sources, addressing the challenge of combining heterogeneous data types. The **Multi-modal Attention Network** employs attention mechanisms to dynamically weight different modalities according to their informativeness for the diagnostic task. This adaptive fusion approach enables the model to emphasize the most informative modalities for each individual subject, potentially improving diagnostic accuracy. The **Cross-modal Transformer** leverages the transformer architecture to explicitly model bidirectional interactions between modalities. By computing cross-modal attention between MRI and PET features, this approach captures complementary information and enables selective information exchange between modalities.

Graph-based approaches model the structural and functional connectivity of the brain, incorporating prior knowledge of brain network organization into the learning process. The **Graph CNN** extends the standard CNN architecture to graph-structured data, enabling the analysis of brain connectivity networks. This approach leverages the inherent structure of brain networks, capturing the spatial relationships between functionally connected regions that may be disrupted in

Alzheimer's disease. The **Graph Attention Network** further extends graph-based methods by incorporating attention mechanisms into graph convolutions. This enables the model to learn which connections are most important for the diagnostic task, improving interpretability and potentially enhancing performance.

4.8 Implementation Details

The proposed HCST-Former is implemented using the PyTorch deep learning framework, which provides a flexible and efficient platform for developing and training complex neural network architectures. The model is trained on an NVIDIA RTX 4090 GPU with 24 GB of GPU memory, providing the computational resources necessary for training a large-scale transformer architecture on 3D medical imaging data.

Training Configuration: The model is optimized using the AdamW optimizer, an adaptive optimization algorithm that extends the Adam optimizer with weight decay regularization. AdamW has become the standard choice for training transformer architectures due to its stable convergence properties and effective handling of large-scale models. The initial learning rate is set to 1×10^{-4} , which provides a balanced trade-off between rapid convergence and training stability. The batch size is set to 4, which is the maximum feasible batch size given the memory constraints of the GPU and the computational demands of processing 3D volumetric data. While this batch size is relatively small compared to those used for 2D image analysis, it is consistent with standard practice for 3D medical image analysis and is necessary to accommodate the memory requirements of the transformer architecture. The model is trained for 300 epochs, ensuring sufficient iterations for the model to converge to a stable optimum. A cosine annealing learning-rate scheduler is employed to progressively adjust the learning rate during training, gradually reducing the learning rate as training progresses to enable fine-tuning of the model parameters. This scheduling strategy has been shown to improve convergence and final performance compared to fixed learning rate schedules. The weight decay coefficient is set to 1×10^{-2} , providing regularization that helps prevent overfitting by penalizing large weight values. This regularization is particularly important for transformer architectures, which can easily overfit to the training data given their large capacity and the relatively limited size of medical imaging datasets. The training objective employs a hybrid loss consisting of four complementary components:

Dice loss for segmentation accuracy, Cross-Entropy loss for classification, stage classification loss for disease-stage modeling, and uncertainty regularization for calibration. Equal weighting is initially assigned to the individual loss components, with the weighting coefficients being tunable hyperparameters that can be adjusted based on validation performance.

Input Configuration: The input volume size is fixed at $128 \times 128 \times 128$ voxels, providing sufficient spatial resolution to capture clinically relevant anatomical details while maintaining computational feasibility. The transformer patch size is $16 \times 16 \times 16$, meaning that the volumetric input is partitioned into $8 \times 8 \times 8 = 512$ non-overlapping 3D patches before being projected into the embedding space. This patch-based tokenization strategy is a fundamental component of the transformer architecture, enabling the model to process volumetric data efficiently while preserving spatial relationships between local regions. Multi-modal features are extracted using modality-specific backbones, each tailored to the

characteristics of the corresponding data modality. For MRI and PET imaging, patch-based tokenization is employed with learnable linear projections to C-dimensional embeddings. For cognitive assessments, genetic biomarkers, and longitudinal clinical records, dedicated encoding networks are used to generate modality-specific embeddings. The extracted features are subsequently integrated through the proposed Cross-Modal Attention Fusion (CMAF) mechanism, which explicitly models bidirectional interactions between modalities. Longitudinal information is incorporated using the Longitudinal Temporal Encoding (LTE) module, which captures disease trajectory patterns across multiple timepoints. Anatomical information is modeled using the Graph-Guided Anatomical Attention (GGAA) module, which incorporates prior knowledge of AD-relevant brain regions. Disease-stage information is modeled using the Hierarchical Disease-Stage Modeling (HDSM) module, which learns progressive changes across CN, MCI, and AD stages.

Table 2. Progression Prediction Performance

Method	Time-to-Conversion MAE (months)	C-Index	Brier Score
3D-CNN	14.2 ± 2.1	0.78 ± 0.02	0.18 ± 0.02
Multi-modal CNN	12.8 ± 1.8	0.81 ± 0.02	0.16 ± 0.02
Longitudinal Transformer	11.5 ± 1.5	0.84 ± 0.01	0.14 ± 0.01
Swin Transformer	10.8 ± 1.3	0.86 ± 0.01	0.13 ± 0.01
MedFormer	10.2 ± 1.1	0.88 ± 0.01	0.12 ± 0.01
HCST-Former (Ours)	8.7 ± 0.9	0.91 ± 0.01	0.10 ± 0.01

5. RESULTS AND DISCUSSION

5.1 Classification Performance

The proposed HCST-Former achieves strong performance for early Alzheimer's disease detection on the ADNI dataset. Table 1 compares the proposed framework with representative CNN- and Transformer-based approaches using Accuracy, AUC, Sensitivity, Specificity, and F1-Score. The results demonstrate that HCST-Former achieves the

highest performance across all reported classification metrics. The proposed framework obtains an accuracy of 93.7% and an AUC of 96.2%, while achieving sensitivity and specificity values of 92.8% and 94.5%, respectively. The F1-Score of 0.937 further indicates a strong balance between precision and recall. Compared with MedFormer, the strongest competing method in this comparison, HCST-Former improves accuracy from 92.4% to 93.7% and AUC from 95.6% to 96.2%. The improvement in sensitivity and

specificity suggests that the proposed framework is better able to distinguish subtle differences between disease stages while reducing both false-negative and false-positive predictions.

5.2 Progression Prediction Performance

HCST-Former demonstrates strong performance for predicting MCI-to-AD conversion. Table 2 summarizes the time-to-conversion prediction error, Concordance Index (C-index), and Brier Score. As shown in Table 2, HCST-Former achieves the lowest time-to-conversion prediction error of 8.7 months and the highest C-index of 0.91. It also obtains the lowest Brier Score of 0.10, indicating improved probabilistic prediction performance. The improvement can be attributed primarily to the Longitudinal Temporal Encoding (LTE) module, which explicitly models relationships between multiple clinical timepoints. In addition, the Hierarchical Disease-Stage Modeling (HDSM) module provides an explicit representation of the progression pathway from CN to MCI and subsequently to AD.

5.3 Uncertainty Estimation Performance

Reliable uncertainty estimation is important for clinical decision-support systems because predictions with low confidence may require additional clinical assessment. Table 3 compares HCST-Former with several uncertainty estimation strategies using Expected Calibration Error (ECE) and Negative Log-Likelihood (NLL). HCST-Former achieves the lowest ECE of 0.042 and NLL of 0.38 among the evaluated approaches. These results indicate that the proposed UncertaintyAware Prediction (UAP) module produces better-calibrated probabilistic predictions. The lower ECE indicates a closer correspondence between predicted confidence and empirical prediction accuracy, whereas the lower NLL indicates improved probabilistic prediction quality. This characteristic is particularly important for clinical applications where highly confident but incorrect predictions may lead to inappropriate clinical decisions.

Table 4: Computational Efficiency Comparison.

Method	Params (M)	FLOPs (G)	GPU Memory (GB)	Inference Time (s)
3D-CNN	38.4	72.6	6.8	2.45
Multi-modal CNN	45.6	85.4	8.2	2.98
ADViT	52.8	98.2	10.4	3.12
Multi-modal ViT	58.6	112.4	12.8	3.56
Longitudinal Transformer	62.4	124.8	14.2	4.02
Swin Transformer	54.2	105.6	11.6	3.24
MedFormer	50.8	96.4	10.8	2.98
HCST-Former (Ours)	48.3	92.6	8.1	2.34

5.4 Computational Efficiency Analysis

The computational efficiency of HCST-Former is evaluated in terms of the number of trainable parameters, floating-point operations (FLOPs), GPU memory consumption, and inference time. The results are presented in Table 4.

HCST-Former achieves an effective balance between predictive performance and computational requirements. Although its parameter count is higher than that of the 3D-CNN, the proposed model requires substantially fewer parameters than most transformer-based baselines. In particular, HCST-Former requires 92.6 GFLOPs and 8.1 GB

of GPU memory, while the Longitudinal Transformer requires 124.8 GFLOPs and 14.2 GB of memory. Furthermore, HCST-Former achieves an inference time of only 2.34 seconds per subject, which is the lowest among the evaluated methods. These results suggest that the proposed hierarchical and adaptive attention mechanisms can reduce redundant computation while preserving the representational capacity required for multi-modal and longitudinal AD analysis.

5.5 Ablation Studies

Comprehensive ablation experiments are performed to investigate the individual contributions of the major architectural components of HCST-Former. The results are presented in Table 5. The ablation results demonstrate a consistent improvement as the proposed components are progressively incorporated into the baseline ViT model. The baseline model obtains an accuracy of 88.4% and AUC of 92.6%. Adding CMAF increases these values to 90.8% and 94.2%, respectively, demonstrating the importance of cross-modal information exchange. The subsequent addition of LTE further improves accuracy to 91.6% and AUC to 94.8%, indicating that longitudinal temporal information provides

additional discriminative information. Incorporating HDSM increases the accuracy to 92.4%, highlighting the importance of explicitly modeling the hierarchical progression between CN, MCI, and AD. The addition of GGAA further improves the accuracy to 93.2% and AUC to 95.8%. Finally, the complete HCST-Former achieves 93.7% accuracy and 96.2% AUC. The relatively small increase in parameter count from 42.6 million to 48.3 million indicates that the performance improvements are achieved without excessive model expansion.

5.6 Interpretability Analysis

The proposed HCST-Former provides clinically interpretable attention maps and disease progression representations.

Anatomical Attention Visualization: Anatomical attention heatmaps illustrate the regions receiving greater attention from the model during prediction. In particular, the model focuses on AD-relevant regions such as the hippocampus, entorhinal cortex, and medial temporal lobe. This behavior is consistent with the objective of the Graph-Guided Anatomical Attention (GGAA) module, which incorporates anatomical relationships into the attention mechanism.

Table 5. Ablation Study of HCST-Former Components.

Configuration	Accuracy (%)	AUC (%)	Params (M)	FLOPs (G)
Baseline (ViT)	88.4	92.6	42.6	85.2
+ CMAF	90.8	94.2	44.8	88.4
+ CMAF + LTE	91.6	94.8	46.2	90.2
+ CMAF + LTE + HDSM	92.4	95.4	47.5	91.4
+ CMAF + LTE + HDSM + GGAA	93.2	95.8	48.0	92.0
Full HCST-Former	93.7	96.2	48.3	92.6

Temporal Attention Patterns: Temporal attention visualization illustrates how the model distributes attention across longitudinal timepoints. The resulting representations allow the model to identify temporal changes associated with disease progression and distinguish stable subjects from individuals exhibiting progressive deterioration.

Stage-Specific Features: The hierarchical representations learned by HDSM provide distinct representations for CN, MCI, and AD stages. These representations allow the model to characterize the progressive transition between disease stages while emphasizing subtle changes associated with early MCI.

Uncertainty Maps: Uncertainty maps provide a spatial and predictive representation of model confidence. Regions or samples associated with higher uncertainty can be identified for additional clinical evaluation, thereby improving the transparency of the prediction process.

Table 3: Uncertainty Estimation Performance

Method	ECE ↓	NLL ↓
Monte Carlo Dropout	0.089	0.52
Deep Ensemble	0.065	0.45
Evidential DL	0.054	0.41
HCST-Former (Ours)	0.042	0.38

5.7 Discussion

The experimental results demonstrate that the proposed HCST-Former effectively balances diagnostic accuracy, longitudinal prediction capability, uncertainty estimation, and computational efficiency through five key architectural components. First, the Cross-Modal Attention Fusion (CMAF) module enables effective integration of complementary information from multiple modalities. Rather than treating MRI, PET, cognitive, genetic, and clinical information independently, CMAF facilitates information exchange between modalities and enables the model to emphasize modality-specific features according to their relevance to the prediction task. Second, the Longitudinal Temporal Encoding (LTE) module captures disease trajectory patterns across multiple timepoints. This temporal representation is particularly important for AD progression prediction, where a single observation may not adequately characterize the rate or direction of disease progression. Third, the Hierarchical Disease-Stage Modeling (HDSM) module explicitly models the progression from CN to MCI and subsequently to AD. The ablation results demonstrate that incorporating HDSM produces a measurable improvement in both accuracy and AUC, supporting the importance of stageaware representation learning. Fourth, the Graph-Guided Anatomical Attention (GGAA) module incorporates prior knowledge of AD-relevant brain regions into the attention mechanism. By constraining feature interactions according to

anatomical relationships, GGAA improves the ability of the model to emphasize clinically meaningful regions while simultaneously supporting interpretability. Fifth, the Uncertainty-Aware Prediction (UAP) module provides confidence estimates for model predictions. The lower ECE and NLL values obtained by HCST-Former indicate improved calibration compared with alternative uncertainty estimation strategies. Such calibrated predictions can support risk-aware decision making by identifying cases for which additional clinical assessment may be appropriate. Overall, HCST-Former achieves an accuracy of 93.7% and an AUC of 96.2%, while also obtaining a C-index of 0.91 and a time-to-conversion MAE of 8.7 months. These results indicate that jointly modeling multi-modal information and longitudinal disease trajectories can provide substantial benefits for both diagnostic classification and progression prediction.

The computational efficiency analysis further demonstrates that HCST-Former provides a favorable trade-off between predictive performance and resource requirements. The model requires 48.3 million parameters, 92.6 GFLOPs, and 8.1 GB of GPU memory, while achieving an inference time of 2.34 seconds per subject. Compared with conventional dense transformer architectures, the hierarchical representation and adaptive attention mechanisms reduce redundant computation while preserving the representational capacity required for multimodal and longitudinal Alzheimer's disease analysis. The ablation results further support the contribution of the individual components. Starting from the baseline ViT, the sequential introduction of CMAF, LTE, HDSM, and GGAA produces progressively higher classification performance. The complete framework achieves the best results, confirming that the components are complementary rather than redundant. Taken together, these findings suggest that HCST-Former provides a unified framework for multi-modal Alzheimer's disease analysis by jointly addressing diagnostic discrimination, longitudinal progression modeling, anatomical interpretability, uncertainty estimation, and computational scalability.

6. DIFFERENCE FROM PRIOR WORK

6.1 Difference from Prior Work

Existing Alzheimer's disease (AD) models have progressively evolved from CNN-based local feature extraction to transformer-based global representation learning, multimodal fusion, and

longitudinal modeling. CNN-based approaches provide effective local anatomical representations but have limited ability to capture long-range dependencies across distributed brain regions. Vision Transformer (ViT)-based approaches improve global dependency modeling but introduce substantial computational requirements when applied to high-dimensional neuroimaging data. Multimodal approaches further improve disease characterization by combining MRI, PET, cognitive, and clinical information, although many rely on concatenation, weighted fusion, or independently processed modalities. Longitudinal models address temporal disease progression but generally do not jointly incorporate multimodal interactions, disease-stage hierarchy, anatomical priors, and uncertainty estimation. In contrast, HCST-Former integrates these complementary requirements within a unified architecture. CMAF explicitly models interactions among heterogeneous modalities, LTE captures dependencies across longitudinal observations, HDSM represents the hierarchical CN→MCI→AD progression, GGAA incorporates anatomical knowledge into attention, and UAP provides prediction-confidence information. Thus, the primary distinction from prior work is not the use of a transformer alone, but the joint modeling of multimodal, temporal, hierarchical, anatomical, and uncertainty-related information. This design directly addresses the unified-framework gap identified in the literature review. The quantitative comparison on ADNI further provides evidence of this distinction. HCST-Former reports 93.7% accuracy, 96.2% AUC, 92.8% sensitivity, and 94.5% specificity, compared with the reported results of 3D-CNN, multimodal CNN, Attention-CNN, ADViT, multimodal ViT, longitudinal Transformer, Swin Transformer, and MedFormer baselines.

6.2 Strengths of the Proposed Framework

The first strength of HCST-Former is its **integrated multimodal representation**. Rather than treating MRI, PET, cognitive, genetic, and clinical information as independent sources, CMAF enables their complementary information to be jointly represented. This addresses the limitation of simple multimodal fusion strategies identified in the reviewed literature. Second, the framework provides **explicit longitudinal modeling** through LTE. This is important because AD progression is inherently temporal, and repeated observations can provide information that is unavailable from a single clinical visit.

Table 6. Distinction between HCST-Former and Prior Approaches

Aspect	Typical Prior Approaches	HCST-Former
Spatial representation	Local CNN or global transformer features	Hierarchical transformer representation
Multimodal learning	Concatenation, weighted fusion, or separate branches	CMAF-based cross-modal interaction
Temporal modeling	Single timepoint or sequential modeling	LTE-based longitudinal trajectory modeling
Disease stages	Conventional multiclass prediction	Explicit CN→MCI→AD hierarchy through HDSM
Anatomical knowledge	ROI analysis or independent graph modeling	GGAA integrated into attention
Uncertainty	Often absent or separately estimated	Integrated UAP module
Prediction	Primarily diagnosis/classification	Disease stage and progression trajectory
Evaluation focus	Mainly classification performance	Classification, progression, uncertainty, interpretability, and efficiency

Third, HDSM introduces **stage-aware representation learning** by explicitly modeling the CN→MCI→AD progression. This provides a closer correspondence between the model representation and the progressive nature of AD than treating the three stages solely as independent classification labels. Fourth, GGAA provides **anatomically informed attention**, allowing the model to incorporate prior knowledge regarding AD-relevant regions. The manuscript specifically considers regions such as the hippocampus, entorhinal cortex, and medial temporal lobe. Finally, the framework combines predictive performance with **uncertainty estimation and computational analysis**. The reported model size of 48.3M parameters, 92.6 GFLOPs, and 8.1 GB GPU memory provides an explicit assessment of

computational requirements in addition to diagnostic performance.

7. CONCLUSION

This study presented HCST-Former, a unified hierarchical cross-modal spatio-temporal transformer that addresses limitations of existing CNN-, transformer-, multimodal-, and longitudinal AD models. By integrating CMAF, LTE, HDSM, GGAA, and UAP, the framework jointly captures multimodal interactions, longitudinal trajectories, hierarchical CN→MCI→AD progression, anatomical priors, and predictive uncertainty. Experimental results demonstrate 93.7% accuracy, 96.2% AUC, 92.8% sensitivity, and 94.5% specificity, with a C-index of 0.91 and 8.7-month time-to-conversion MAE. The model requires 48.3M parameters, 92.6 GFLOPs, and 8.1 GB GPU memory, indicating a balance between predictive performance and computational efficiency. Ablation and interpretability analyses further support the contribution of the proposed components and highlight AD-relevant regions and longitudinal disease patterns. Overall, HCST-Former provides a unified framework for multimodal representation, progression modeling, hierarchical staging, anatomical guidance, and uncertainty-aware prediction. However, evaluation on ADNI limits conclusions regarding broader clinical generalization. Future work will focus on independent multi-center cohorts, additional biomarkers, and efficient deployment strategies.

8. LIMITATIONS AND FUTURE WORK

Despite these advantages, several limitations remain. First, external generalization has not been established. The present evaluation is based on the ADNI cohort, and therefore the reported results do not by themselves establish robustness across independent hospitals, imaging protocols, scanner manufacturers, or demographic populations. The literature review itself identifies cross-dataset generalization as an unresolved issue. Future work should therefore evaluate HCST-Former on independent multi-center datasets and perform external validation under domain and acquisition shifts. Second, multimodal availability may be a practical constraint. The proposed framework assumes access to several heterogeneous information sources. In real clinical settings, MRI, PET, genetic information, cognitive assessments, and longitudinal records may not all be available for every patient. Future research should investigate

missing-modality learning, modality dropout, and adaptive modality selection so that the model can operate reliably when only a subset of modalities is available. Third, longitudinal observations are not necessarily uniformly sampled. Clinical visits may occur at different intervals and some patients may have substantially fewer observations than others. Future versions should investigate continuous-time temporal encoding, irregular-time attention, and survival-based progression modeling to better represent heterogeneous follow-up schedules. Fourth, although GGAA improves anatomical awareness, the anatomical graph and region definitions introduce prior assumptions. Future studies could investigate automatically learned or patient-specific brain graphs and compare atlas-based anatomical priors with data-driven connectivity representations. Fifth, uncertainty estimation should be further validated for clinical reliability. Although UAP provides confidence estimates, future research should evaluate calibration under distribution shifts and assess whether uncertainty can reliably identify cases requiring additional clinical assessment. Finally, the computational efficiency of HCST-Former can be further improved. Future work may investigate sparse attention, token pruning, knowledge distillation, quantization, and lightweight multimodal encoders to reduce inference cost while preserving clinically relevant information. Overall, these limitations define a clear future research pathway toward externally validated, missing-modality-tolerant, irregular-time, anatomically adaptive, uncertainty-calibrated, and computationally efficient AD prediction systems.

AUTHOR CONTRIBUTIONS:

Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing—original draft preparation, and visualization were performed by Author 1. Writing—review and editing, supervision, and project administration were performed by Author 2. Both authors have read and agreed to the published version of the manuscript.

REFERENCES:

- [1] Alzheimer's Association, "2023 Alzheimer's Disease Facts and Figures," *Alzheimer's & Dementia*, Vol. 19, No. 4, 2023, pp. 1598–1695.
- [2] R.A. Sperling, P.S. Aisen, L.A. Beckett, et al., "Toward Defining the Preclinical Stages of Alzheimer's Disease: Recommendations from

- the National Institute on Aging-Alzheimer's Association Workgroups on Diagnostic Guidelines for Alzheimer's Disease," *Alzheimer's & Dementia*, Vol. 7, No. 3, 2011, pp. 280–292.
- [3] C.R. Jack, D.A. Bennett, K. Blennow, et al., "NIA-AA Research Framework: Toward a Biological Definition of Alzheimer's Disease," *Alzheimer's & Dementia*, Vol. 14, No. 4, 2018, pp. 535–562.
- [4] G. Litjens, T. Kooi, B.E. Bejnordi, et al., "A Survey on Deep Learning in Medical Image Analysis," *Medical Image Analysis*, Vol. 42, 2017, pp. 60–88.
- [5] D. Shen, G. Wu, and H.I. Suk, "Deep Learning in Medical Image Analysis," *Annual Review of Biomedical Engineering*, Vol. 19, 2017, pp. 221–248.
- [6] C.R. Jack, D.S. Knopman, W.J. Jagust, et al., "Tracking Pathophysiological Processes in Alzheimer's Disease: An Updated Hypothetical Model of Dynamic Biomarkers," *The Lancet Neurology*, Vol. 12, No. 2, 2013, pp. 207–216.
- [7] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is All You Need," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.
- [8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," *International Conference on Learning Representations (ICLR)*, 2021.
- [9] H. Braak and E. Braak, "Neuropathological Staging of Alzheimer-Related Changes," *Acta Neuropathologica*, Vol. 82, No. 4, 1991, pp. 239–259.
- [10] E. Hosseini-Asl, R. Keynton, and A. El-Baz, "Alzheimer's Disease Diagnostics by a Deeply Supervised Adaptable 3D Convolutional Network," *Frontiers in Aging Neuroscience*, Vol. 8, 2016, Article 185.
- [11] A. Payan and G. Montana, "Predicting Alzheimer's Disease: A Neuroimaging Study with 3D Convolutional Neural Networks," *arXiv preprint arXiv:1502.02506*, 2015.
- [12] E. Hosseini-Asl, G. Gimel'farb, and A. El-Baz, "Alzheimer's Disease Diagnostics by a 3D Deeply Supervised Adaptable Convolutional Network," *IEEE Transactions on Medical Imaging*, Vol. 36, No. 9, 2017, pp. 1843–1855.
- [13] A. Payan and G. Montana, "Predicting Alzheimer's Disease: A Neuroimaging Study with 3D Convolutional Neural Networks," *International Conference on Pattern Recognition (ICPR)*, 2016.
- [14] T.M. Schouten, M. Koini, F. de Vos, et al., "Combining Anatomical, Diffusion, and Resting State Functional Magnetic Resonance Imaging for Individual Classification of Mild Cognitive Impairment," *NeuroImage: Clinical*, Vol. 21, 2019, Article 101617.
- [15] H. Li, M. Habes, D.A. Wolk, et al., "A Deep Learning Model for Early Alzheimer's Disease Detection from Structural MRI," *NeuroImage: Clinical*, Vol. 28, 2020, Article 102462.
- [16] F. Zhang, B. Pan, P. Shao, et al., "A 3D CNN with Squeeze-and-Excitation Blocks for Alzheimer's Disease Classification," *Frontiers in Aging Neuroscience*, Vol. 13, 2021, Article 725792.
- [17] S. Spasov, L. Passamonti, A. Duggento, et al., "A Multi-Modal Convolutional Neural Network Framework for the Prediction of Alzheimer's Disease," *Medical Image Analysis*, Vol. 64, 2020, Article 101736.
- [18] J. Venugopalan, L. Tong, H.R. Hassanzadeh, et al., "Multimodal Deep Learning Models for Early Detection of Alzheimer's Disease Stage," *Scientific Reports*, Vol. 11, 2021, Article 3254.
- [19] K.H. Maier-Hein, A. Reinke, M. Kozubek, et al., "Why Medical Image Segmentation Algorithms Fail in Clinical Practice: Challenges and Future Directions," *Nature Reviews Bioengineering*, Vol. 2, 2024, pp. 1–18.
- [20] A. Khan, Z. Rauf, A. Rehman Khan, et al., "A Recent Survey of Vision Transformers for Medical Image Segmentation," *IEEE Access*, Vol. 13, 2025, pp. 42158–42184.
- [21] L. Sun, S. Zhang, H. Chen, et al., "ADViT: Vision Transformer for Alzheimer's Disease Diagnosis," *IEEE Journal of Biomedical and Health Informatics*, Vol. 26, No. 12, 2022, pp. 6123–6134.
- [22] Y. Chen, H. Wang, Y. Liu, et al., "Multi-modal Vision Transformer for Alzheimer's Disease Classification," *Medical Image Analysis*, Vol. 89, 2023, Article 102884.
- [23] X. Zhang, Y. Li, Z. Wang, et al., "Cross-modal Transformer for Alzheimer's Disease Diagnosis from MRI and PET," *IEEE Transactions on Medical Imaging*, Vol. 42, No. 8, 2023, pp. 2345–2357.
- [24] Y. Gao, H.Y. Zhou, Y. Li, et al., "Longitudinal Transformer for Alzheimer's Disease

- Progression Prediction,” *Medical Image Analysis*, Vol. 90, 2023, Article 102956.
- [25] L. Wang, K. Zhang, Z. Yan, et al., “Spatio-Temporal Transformer for Early Alzheimer’s Disease Detection,” *Pattern Recognition*, Vol. 151, 2024, Article 110308.
- [26] X. Liu, H. Chen, Q. Wang, et al., “Brain Network-Guided Transformer for Alzheimer’s Disease Classification,” *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 35, No. 8, 2024, pp. 11234–11247.
- [27] Y. Zhou, L. Li, L. Lu, et al., “Knowledge-Guided Transformer for Alzheimer’s Disease Diagnosis,” *IEEE Journal of Biomedical and Health Informatics*, Vol. 28, No. 6, 2024, pp. 3456–3467.
- [28] Z. Liu, Y. Lin, Y. Cao, et al., “Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows,” *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10012–10022.
- [29] Y. Gao, H.Y. Zhou, Y. Li, et al., “MedFormer: A Multi-scale Transformer Framework for Medical Image Segmentation,” *Medical Image Analysis*, Vol. 88, 2023, Article 102885.
- [30] D. Zhang, Y. Wang, L. Zhou, et al., “Multi-modal Multi-task Learning for Alzheimer’s Disease Diagnosis,” *NeuroImage: Clinical*, Vol. 24, 2019, Article 102039.
- [31] M. Liu, J. Zhang, E. Adeli, et al., “Joint Classification and Regression via Multi-task Learning for Alzheimer’s Disease,” *NeuroImage: Clinical*, Vol. 21, 2019, Article 101651.
- [32] H.I. Suk, S.W. Lee, and D. Shen, “Hierarchical Feature Representation and Multimodal Fusion with Deep Learning for AD/MCI Diagnosis,” *NeuroImage*, Vol. 101, 2014, pp. 569–582.
- [33] E. Hosseini-Asl, M. Ghazal, A. Mahmoud, et al., “Alzheimer’s Disease Diagnostics by a 3D Deeply Supervised Adaptable Convolutional Network,” *IEEE Transactions on Medical Imaging*, Vol. 36, No. 9, 2017, pp. 1843–1855.
- [34] Y. Zhao, B. Ma, P. Jiang, et al., “Multi-modal Attention Network for Alzheimer’s Disease Classification,” *IEEE Journal of Biomedical and Health Informatics*, Vol. 27, No. 4, 2023, pp. 1892–1903.
- [35] H. Zhou, X. Liu, Y. Zhang, et al., “Cross-modal Attention Mechanism for Integrating MRI and PET in Alzheimer’s Disease Diagnosis,” *Medical Image Analysis*, Vol. 83, 2023, Article 102660.
- [36] B.M. Jedynek, A. Lang, B. Liu, et al., “A Computational Trajectory Model for Alzheimer’s Disease Progression,” *NeuroImage*, Vol. 59, No. 2, 2012, pp. 1228–1241.
- [37] Y. Li, H. Wang, Q. Liu, et al., “Bayesian Modeling of Disease Progression from Longitudinal Biomarker Data,” *NeuroImage: Clinical*, Vol. 31, 2021, Article 102721.
- [38] J. Qin, L. Chen, Y. Liu, et al., “Recurrent Neural Network for Predicting Alzheimer’s Disease Progression from Longitudinal MRI,” *IEEE Transactions on Medical Imaging*, Vol. 41, No. 3, 2022, pp. 678–689.
- [39] S. Duchesne, A. Caroli, C. Geroldi, et al., “LSTM Networks for Disease Progression Modeling from Longitudinal Cognitive Data,” *Alzheimer’s & Dementia*, Vol. 17, No. 8, 2021, pp. 1345–1356.
- [40] R. Li, Y. Chen, W. Zhang, et al., “Region-of-Interest Based Feature Extraction for Alzheimer’s Disease Classification,” *NeuroImage: Clinical*, Vol. 28, 2020, Article 102462.
- [41] J. Zhang, M. Liu, Y. Li, et al., “Hippocampus Segmentation for Alzheimer’s Disease Diagnosis,” *NeuroImage*, Vol. 218, 2020, Article 116947.
- [42] H. Chen, Y. Li, Q. Wang, et al., “Graph CNN for Alzheimer’s Disease Classification from Functional Connectivity,” *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 33, No. 5, 2022, pp. 2134–2147.
- [43] J. Xu, K. Zhang, Y. Liu, et al., “Graph Attention Network for MCI-to-AD Conversion Prediction,” *Medical Image Analysis*, Vol. 82, 2022, Article 102637.
- [44] Y. Gal and Z. Ghahramani, “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning,” *International Conference on Machine Learning (ICML)*, 2016, pp. 1050–1059.
- [45] A. Kendall and Y. Gal, “What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?,” *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5574–5584.
- [46] B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles,” *Advances in Neural Information*

- Processing Systems (NeurIPS)*, 2017, pp. 6402–6413.
- [47] M. Sensoy, L. Kaplan, and M. Kandemir, “Evidential Deep Learning to Quantify Classification Uncertainty,” *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 3179–3189.
- [48] Y. Guo, Y. Li, Q. Liu, et al., “Uncertainty Quantification in Alzheimer’s Disease Classification with Monte Carlo Dropout,” *IEEE Journal of Biomedical and Health Informatics*, Vol. 27, No. 2, 2023, pp. 845–856.
- [49] Z. Wang, Y. Chen, K. Zhang, et al., “Bayesian CNN for Alzheimer’s Disease Diagnosis with Uncertainty Estimation,” *Medical Image Analysis*, Vol. 84, 2023, Article 102701.
- [50] R.D. Resen, L.S. Alzubaidi, H.A. Alwzawy, and M.I. Capel, “Review of deep learning models for Alzheimer’s disease detection: MRI-centric approaches and multimodal extensions,” *Frontiers in Artificial Intelligence*, Vol. 9, 2026, Article 1707043.
- [51] “Machine-learning models for Alzheimer’s disease diagnosis using neuroimaging data: survey, reproducibility, and generalizability evaluation,” 2025.
- [52] “Vision and convolutional transformers for Alzheimer’s disease diagnosis: a systematic review of architectures, multimodal fusion and critical gaps,” *BMC Biomedical Engineering*, 2025.
- [53] “Multimodal diagnosis model of Alzheimer’s disease based on improved Transformer,” *Computers in Biology and Medicine*, 2024.
- [54] “A transformer-based unified multimodal framework for Alzheimer’s disease assessment,” *Computers in Biology and Medicine*, 2024, Article 108979.
- [55] “Meta-analysis of the diagnostic efficacy of transformer-based multimodal fusion deep learning models in early Alzheimer’s disease,” *Frontiers in Neurology*, 2025.
- [56] M. Ghasabi, H. Aghdasi, and M.-R. Feizi-Derakhshi, “Predicting conversion from mild cognitive impairment to Alzheimer’s disease: A systematic review of deep learning models for early-stage disease classification,” *Ageing Research Reviews*, Vol. 122, 2026, Article 103372.
- [57] R. Zhang, J. Sheng, Q. Zhang, J. Wang, and B. Wang, “A review of multimodal fusion-based deep learning for Alzheimer’s disease,” *Neuroscience*, Vol. 576, 2025, pp. 80–95.
- [58] “Multi-Modal Fusion and Longitudinal Analysis for Alzheimer’s Disease Classification Using Deep Learning,” *Diagnostics*, Vol. 15, No. 6, 2025, Article 717.
- [59] “Multimodal Self-Supervised Learning for Early Alzheimer’s: Cross-Modal MRI-PET, Longitudinal Signals, and Site Invariance,” 2025.
- [60] “IT: An interpretable transformer model for Alzheimer’s disease prediction based on PET/MR images,” 2025.
- [61] “Deep learning for Alzheimer’s disease: advances in classification, segmentation, subtyping, and explainability,” 2025/2026.
- [62] “Uncertainty-aware genomic classification of Alzheimer’s disease: a transformer-based ensemble approach with Monte Carlo dropout,” *Briefings in Bioinformatics*, 2025.
- [63] “Imputation-free transformer learning enables robust Alzheimer’s disease prediction and calibrated uncertainty quantification across heterogeneous clinical cohorts,” 2026.