ADxClass: Multi-Domain Attention Fusion and Imputation of Missing Heterogeneous Tabular Data

Dhivyaa S P[1], Hyung-Jeong Yang*[,2], Sae-Ryung Kang[3], Soo-Hyung Kim[4]
[1]Master's Student, Department of AI Convergence, Chonnam National University
[2,4]Professor, Department of AI Convergence, Chonnam National University
[3]Professor, Department of Nuclear Medicine, Chonnam National University Hospital
*Corresponding Author

spdhivyaa@gmail.com, hjyang@jnu.ac.kr, campanella9@naver.com, shkim@jnu.ac.kr

ABSTRACT

Alzheimer's Disease (AD) is a neurodegenerative disorder characterized by a progressive decline in cognitive function. Accurate and early diagnosis of AD is crucial for effective management and treatment. Traditional machine learning models, though commonly applied, often fall short in capturing the intricate relationships between diverse tabular data. Furthermore, the missing data issue, typically addressed using conventional imputation techniques, leads to reduced accuracy and generalizability of AD classification models. This paper introduces ADxClass, a novel deep learning framework that enhances AD classification by leveraging multi-domain attention fusion and data type-based imputation techniques for handling missing heterogeneous tabular data. ADxClass integrates data from various domains, including demographic, cognitive, genetic, and biomarkers obtained from neuroimaging measurements, to improve the robustness and accuracy of AD classification models. The model's efficiency is validated via a 5-fold cross-validation on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, showing significant improvements in classification performance compared to traditional machine learning approaches.

Keywords: Alzheimer's Disease, missing data imputation, variational autoencoder, heterogeneous tabular data, multi-domain fusion

1. Introduction & Related Work

Alzheimer's Disease (AD) affects millions worldwide, making it a significant public health challenge. Early diagnosis is crucial for slowing disease progression and improving patient care [1]. The primary challenge in AD classification using tabular data is the presence of missing values, which can occur when patients miss appointments or fail to complete surveys. The common way is to disregard the missing data, but that can reduce the effectiveness of AI models. So, it is important to employ proper approaches for imputing the missing data. Traditional imputation methods, such as mean, median, mode and k-nearest neighbors (KNN) [2], a machine learning-based imputation technique, have been widely used in medical research to handle missing data. However, these methods often fall short in capturing the complex relationships among different variables.

Another challenge is AD classification methods, where several studies have explored different machine learning models, such as Support Vector Machines (SVM), Random Forest, and Gradient Boosting, often paired with imputation techniques to handle missing data. While SVM and Random Forest have shown promising results, their performance can degrade with high-dimensional or heterogeneous data, and their computational cost can be high when handling large datasets. More recent approaches, like denoising autoencoders combined with Random Forest, have improved imputation accuracy, but they still face limitations with overfitting and capturing the full complexity of comprehensive AD datasets [3]. This highlights the need for more robust, deep learning-based models for AD classification.

In this paper, we introduce Alzheimer's Disease Classification using X heterogeneous tabular data types (ADxClass), a novel framework that utilizes Variational Autoencoder-based imputation technique and multi-domain attention fusion to enhance the classification of AD. The key contributions are outlined as follows:

- We introduce a novel multi-domain attention fusion mechanism that systematically integrates cognitive, genetic, demographic, and biomarker data. This approach utilizes a tetrahedral structure to represent modality interactions, enabling the calculation of pair-wise attention scores between each modality pair. These scores dynamically assign weights, allowing the model to focus on the most relevant features from each modality, enhancing the robustness and accuracy of Alzheimer's Disease classification.
- To address missing data across different data types, we employ various imputation techniques. Specifically, for biomarker data, we utilize a Variational Autoencoder (VAE) based imputation technique and conduct a comprehensive comparison with traditional imputation methods to evaluate the effectiveness of the VAE-based approach.

2. Proposed Method

In this section, we propose a two-stage network. The first stage addresses the imputation of missing data from various tabular data types – Biomarker (Bio), Cognitive assessment scores (Cog), Demographic data (Demo/DG) and Genetic data (Gen), while the second stage incorporates a multi-domain attention fusion mechanism for AD classification task. The comprehensive framework is shown in Fig. 1.

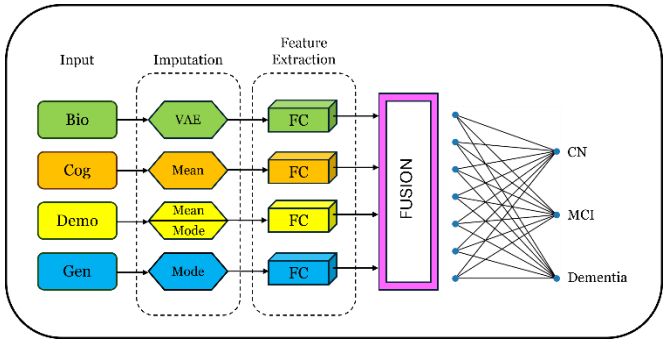

Fig. 1. Comprehensive Framework of the Proposed Model

2.1 Imputation Methods

Missing data in medical research, particularly in Alzheimer's Disease studies, is a common challenge that can significantly impact the quality and reliability of the analysis. The dataset used in this study exhibits varying degrees of missing data across different types. Variables with high missing data (over 30%) include critical biomarkers like hippocampal volume and ventricular size. Moderate missing data (10-15%) is observed in cognitive scores such as MMSE and ADAS13. Low missing data (less than 5%) occurs in demographic information and genetic data.

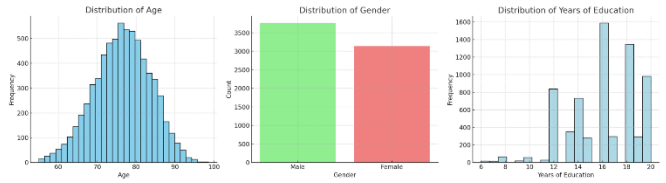

Fig. 2. Visualization of Demographic Information

Different data types, as shown in the Fig. 2,3 & 4, show unique characteristics that require specific imputation strategies.

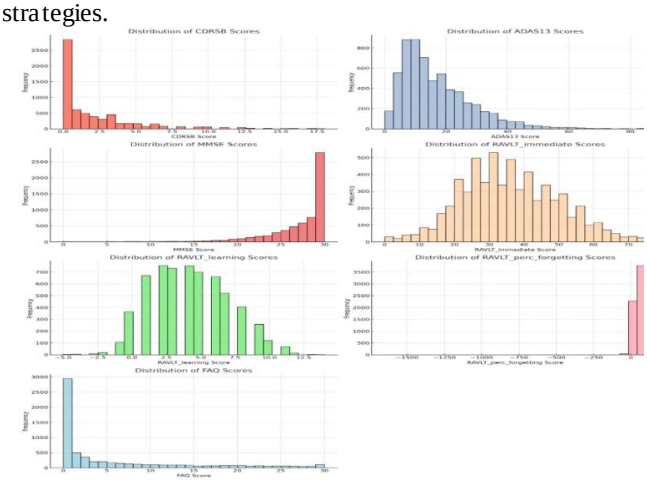

Fig. 3. Visualization of Cognitive Assessment Scores

- **Cognitive Scores:** These scores are imputed using mean imputation, which is suitable given their central tendency and the relatively moderate amount of missing data.
- **Genetic Data (APOE4 Variants):** For genetic data, mode imputation is primarily used, as these variables are categorical and binary.
- **Demographic Variables:** Age and education are imputed using mean imputation to preserve the central tendency. Gender, a categorical variable, is imputed using the mode, ensuring that the demographic profile remains accurate and unbiased.
- **Biomarker Data:** Due to the complex and interrelated nature of biomarker measurements, a Variational Autoencoder (VAE) [4] is employed for imputation. This advanced technique effectively captures the underlying relationships between brain regions, ensuring accurate and consistent imputations even in the presence of substantial amount of missing data.

A binary mask m is created, where each entry corresponds to whether a particular biomarker value is present (1) or missing (0). Missing values in the biomarker data are initially filled with zeros. This standardized initialization is crucial for the VAE to process the input data consistently across all samples. The input data, now with zeros for missing values, is passed through the VAE. The encoder maps the input into a latent space, producing a mean μ and a log-variance $\log(\sigma^2)$ for the latent variable. The latent variable z is sampled using the reparameterization trick and passed through the decoder to reconstruct the biomarker data.

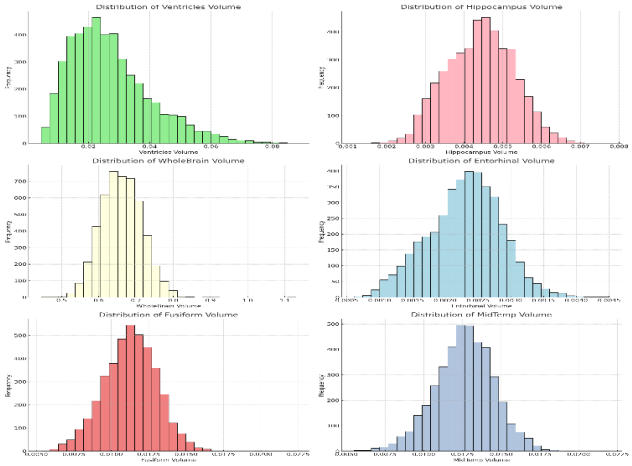

Fig. 4. Visualization of Brain Measurements (biomarkers)

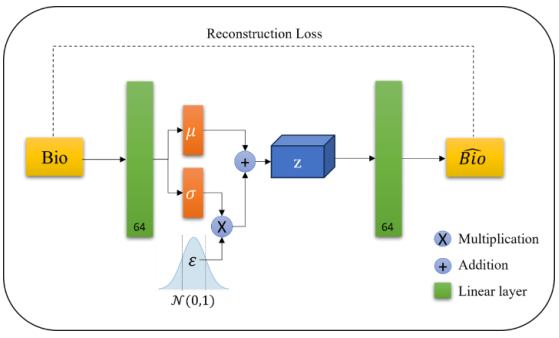

Fig. 5. Imputation of biomarker using VAE

The VAE's reconstruction loss is calculated only for the original (non-missing) values, as indicated by the mask m. Specifically, the loss is computed as the mean squared error (MSE) between the original biomarker data x and the reconstructed data $\hat{x}$:

$$\text{Reconstruction loss} = \sum_{i=1}^{n} m_i \cdot (x_i - \hat{x_i})^2 \quad (1)$$

where n is the number of biomarkers, m_i is the mask value for the i^{th} biomarker feature. The total loss is

computed by combining the reconstruction loss with the KL divergence, which regularizes the latent space to approximate a standard normal distribution.

Once the imputation is performed, each data type is processed through a fully connected (FC) layer specific to that datatype. This step transforms the raw input data into a higher-dimensional feature space, making it more suitable for subsequent fusion.

2.2 Multi-domain Attention Fusion

The multi-domain attention fusion mechanism utilizes a tetrahedral structure to model the interactions between four key modalities: cognitive assessments (Cog), demographic data (DG), genetic data (Gen), and biomarkers (Bio). Each vertex of the tetrahedron represents a modality, and the edges connecting these vertices represent pair-wise interactions between the modalities. The multi-domain attention fusion is represented in figure 6.

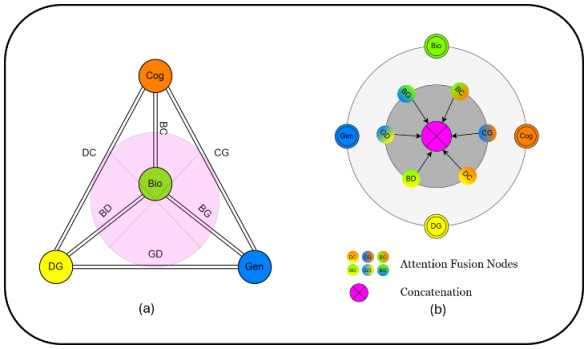

Fig. 6. Multi-domain Attention Fusion
(a) represents the 3-D tetrahedral representation.
(b) represents the 2D-view of the attention fusion

The fusion of modality-specific feature extraction, where each modality's input x_{mod} where mod refers to Cog, DG, Gen, and Bio are transformed into a high-dimensional feature representation h_{mod} through fully connected layers:

$$h_{mod} = f_{\theta_{mod}}(x_{mod}) \quad (2)$$

where $f_{\theta_{mod}}$ represents the transformation function for each modality. Then, pair-wise attention scores $\alpha_{i,j}$ are computed for each modality pair such as Bio-DG, Bio-Gen, Bio-Cog, Cog-Gen, Gen-DG, DG-Cog. These scores are calculated by concatenating the feature vectors of each pair and passing them through a linear transformation followed by a sigmoid activation:

$$\alpha_{i,j} = signmoid\left(W_{i,j} \cdot \left[h_i \oplus h_j\right]\right) \quad (3)$$

where $\left[h_i \oplus h_j\right]$ denotes the concatenated feature vectors from modalities i and j, while $W_{i,j}$ denotes the learnable parameters for that specific pair.

To ensure that the attention scores are properly weighted, they are normalized across all pairs using a SoftMax function:

$$\alpha'_{i,j} = \frac{\exp(\alpha_{i,j})}{\sum_{k,l} exp(\alpha_{k,l})} \quad (4)$$

where k and l are indices used to iterate over the set of all possible pairs of modalities. This type of iteration is known as combinatorial pairwise iteration [5]. The normalization assures that the scores sum to 1, effectively distributing the model's focus proportionately across all modality pairs. The normalized attention scores are then used to weight the contributions from each modality pair in a weighted sum to produce a fused feature representation:

$$Fused\ Features = \sum_{i,j} \alpha'_{i,j} \cdot \left[h_i \oplus h_j\right] \quad (5)$$

This fused representation emphasizes key interactions between modalities, integrating information to enhance the model's predictive power. Finally, the fused features are passed through an output layer to generate the final classification decision:

$$Output = f_{out}(Fused\ Features) \quad (6)$$

where f_{out} denotes the function that transforms the fused features into a probability distribution across the target classes, indicating the likelihood of each class for the given input.

3. Experiment Setup and Result Analysis

3.1 Dataset Overview

In our study, we utilized the ADNIMERGE data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [6]. The dataset includes biomarkers from T1-weighted MRI scans (Ventricles, Hippocampus, Whole Brain, Entorhinal, Fusiform, MidTemp), genetic profiles (APOE4, APOE4_0, APOE4_1, APOE4_2 alleles), cognitive assessment scores (CDRSB, ADAS13, MMSE, RAVLT_immediate, RAVLT_learning, RAVLT_perc_forgetting, FAQ), and demographic information (age, education, gender). After excluding patients with incomplete diagnoses and APOE4, the final dataset includes 6,558 subjects: 2,422 Cognitive Normal (CN), 2,717 Mild Cognitive Impairment (MCI), and 1,519 Alzheimer's Disease cases.

3.2 Experimental Setup

ADxClass model was implemented on a computational platform with the following specifications: The hardware included an NVIDIA A100 Tensor Core graphics processing unit featuring Multi-Instance GPU with 7 GB memory. The software environment consisted of the PyTorch library, utilizing the Adam optimizer with a learning rate set to 0.0001 for training the neural networks. To prevent exploding gradients, gradient clipping was applied during backpropagation, ensuring gradients did not exceed a maximum norm of 1.0. The model was trained using CrossEntropyLoss, with dropout (rate of 0.3) to prevent overfitting. For model validation, a 5-fold cross-validation approach was employed to rigorously evaluate the model's generalizability across different subsets of the data.

3.3 Result Analysis

Classification Task: Performance metrics, including accuracy, precision, recall, and macro-Area under Curve (mAUC), were calculated for each fold and averaged to ensure robust evaluation. Most previous studies have primarily focused on either binary classification of AD or relied on neuroimaging data for AD classification. In contrast, we have evaluated our model's classification performance by comparing it with a machine learning model, Support Vector Machine (SVM), and a simplest forms of deep learning model, Multilayer Perceptron (MLP). Importantly, we

applied the same settings mentioned above to all competing models to ensure a fair comparison. Table 1 compares the performance of SVM, MLP, and ADxClass models in Alzheimer's Disease classification task. ADxClass shows the highest mAUC of 0.9877, highlighting its superior ability to distinguish between classes across varying thresholds.

Table 1. Performance Comparison for AD Classification

Model	Accuracy	Precision	Recall	mAUC
SVM	0.8142	0.8168	0.8242	0.8194
MLP	0.8504	0.8606	0.8433	0.9593
ADxClass	**0.9243**	**0.9266**	**0.9213**	**0.9877**

Table 2 shows the results of an ablation study conducted to evaluate the contributions of the imputation method and multi-domain attention fusion to the ADxClass model's performance. The model's results were compared across three conditions: without imputation, without multi-domain attention fusion, and with both components. This analysis provides a comprehensive understanding of how each component impacts the classification results.

Table 2: Ablation Study Comparing Model Performance with and without Imputation and Multi-Domain Attention Fusion

Model	Accuracy	Precision	Recall	mAUC
w/o imputation	0.9109	0.9113	0.9104	0.9833
w/o multi-domain attention fusion	0.8672	0.8751	0.868	0.9695
With both	**0.9243**	**0.9266**	**0.9213**	**0.9877**

Effectiveness of VAE imputation for Biomarkers: The VAE-based imputation method demonstrates superior performance compared to other traditional imputation techniques like mean, median, and KNN. Figure 7 represents the histogram of before and after VAE imputation, where VAE imputation results in distributions that closely match the original data across various biomarkers, suggesting minimal distortion after imputation.

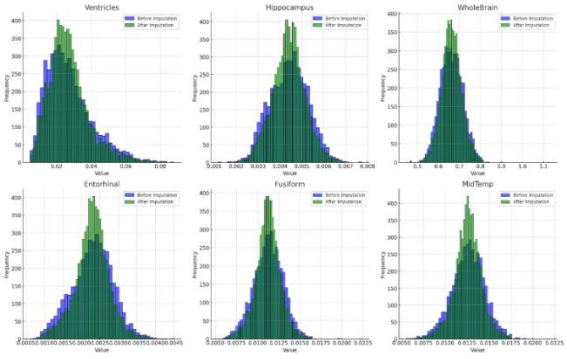

Fig. 7. Comparison of biomarker histograms before & after VAE imputation (Blue – Before; Green – After)

Additionally, the KS (Kolmogorov-Smirnov) test results show that VAE consistently achieves the lowest KS statistic, indicating that the distributions after VAE imputation are more similar to the original data distributions compared to those obtained using mean, median, or KNN methods. This highlights VAE's effectiveness in preserving the underlying data structure, making it a more reliable imputation technique.

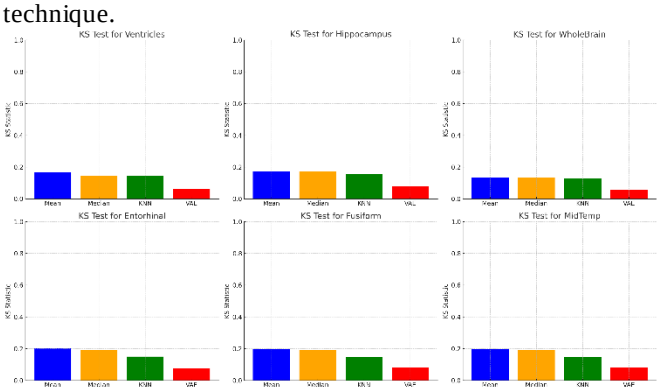

Fig. 8. KS Test Comparison of Different Imputation Techniques Across Biomarkers

4. Conclusion

The proposed ADxClass framework significantly enhances Alzheimer's Disease classification by effectively integrating heterogeneous data types through multi-domain attention fusion and advanced VAE-based imputation techniques. The model outperforms traditional methods, demonstrating superior accuracy and robustness in handling missing data, which is crucial for reliable disease diagnosis. Future work will focus on predicting the progression of Alzheimer's Disease, aiming to enhance early intervention and personalized treatment strategies.

Acknowledgement

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00256629) grant funded by the Korea government (MSIT), the ITRC (Information Technology Research Center) support program (IITP-2024-RS-2024-00437718) supervised by IITP, and a grant (HCRI 23038) from Chonnam National University Hwasun Hospital Institute for Biomedical Science.

References

[1] Z. Arvanitakis, R. C. Shah, and D. A. Bennett, Diagnosis and management of dementia," JAMA, vol. 322, no. 16, pp. 1589-1599, 2019.

[2] A. Jadhav, D. Pramod, and K. Ramanathan, "Comparison of performance of data imputation methods for numeric dataset," Applied Artificial Intelligence, vol. 33, no. 10, pp. 913-933, 2019.

[3] N. T. Haridas, J. M. Sanchez-Bornot, P. L. McClean, and K. Wong-Lin, "Autoencoder imputation of missing heterogeneous data for Alzheimer's disease classification," medRxiv, vol. 2024, no. 07, pp. 2024-07, 2024.

[4] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," arXiv preprint arXiv:1312.6114, 2013.

[5] Cormen, T. H., Leiserson, C. E., Rivest, R. L., Stein, C., "Introduction to Algorithms", Cambridge, MIT Press, 2009.

[6] Alzheimer's Disease Neuroimaging Initiative (ADNI). (n.d.). ADNI | Alzheimer's Disease Neuroimaging Initiative. https://adni.loni.usc.edu/.