VERONECA: A Multimodal Boosting Framework for Alzheimer's Disease Classification with Uncertainty Quantification and Calibrated Clinical Decision Making

Michele Minervini

Big Data Project - Academic Year 2025/2026

Abstract

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder whose early and accurate diagnosis remains a critical challenge in clinical neurology. This work presents VERONECA, a multimodal machine-learning framework that integrates three-dimensional structural Magnetic Resonance Imaging (3D MRI) with Electronic Health Record (EHR) clinical variables to support the multiclass classification of patients into Cognitively Normal (CN), Mild Cognitive Impairment (MCI), and Alzheimer's Disease (AD) classes.

The architecture combines a fine-tuned 3D ResNet-18 convolutional neural network for volumetric brain imaging with a Random Forest classifier for tabular clinical features, fused through `IRBoostSH` (an Implicit Randomization Boosting with Shared Weights algorithm based on multi-armed bandit modality selection). Beyond classification, we implement a comprehensive uncertainty quantification pipeline using Monte Carlo Dropout for epistemic uncertainty estimation, clinically-grounded Test-Time Augmentation for aleatoric uncertainty, and Inductive Conformal Prediction providing distribution-free coverage guarantees.

Probability calibration via Isotonic Regression is followed by a cost-sensitive Bayesian decision rule that accounts for the asymmetric clinical severity of misdiagnosis errors. Extensive experiments conducted under stratified 5-fold cross-validation on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset demonstrate that the *Baseline + Calibration* strategy (standard boosting training followed by post-hoc Isotonic Regression calibration and argmax prediction) yields superior overall performance compared to cost-sensitive training, which tends to over-correct decision boundaries at the expense of global accuracy. Explainability is addressed through SHAP values for clinical feature attribution and Grad-CAM heatmaps for spatial localisation of discriminative brain regions.

Contents

1 Introduction **4**

2 Business Understanding **4**
2.1 Problem Definition . . . 4
2.2 Success Criteria . . . 5
2.3 Risk Assessment . . . 5

3 Data Understanding **6**
3.1 The ADNI Dataset . . . 6
3.2 Exploratory Analysis . . . 7
3.3 Data Quality and Missing Modalities . . . 11

4 Data Preparation **11**
4.1 Clinical Data Pipeline . . . 11
4.2 MRI Image Pipeline . . . 12
4.3 Data Augmentation for MRI . . . 12
4.4 Cross-Validation and Data Splits . . . 13

5 Modeling **13**
5.1 Architecture Overview . . . 13
5.2 Clinical Branch: Random Forest . . . 13
5.3 Imaging Branch: VeroResNet . . . 14
5.3.1 Architecture Details . . . 14
5.3.2 Transfer Learning and Fine-Tuning . . . 14
5.4 Multimodal Boosting: IRBoostSH . . . 14
5.4.1 Algorithm Description . . . 14
5.4.2 Handling Missing Modalities . . . 15
5.4.3 Prediction Aggregation . . . 15
5.4.4 Cost-Sensitive Boosting . . . 15
5.5 Uncertainty Quantification . . . 16
5.5.1 Epistemic Uncertainty: Monte Carlo Dropout . . . 16
5.5.2 Aleatoric Uncertainty: Test-Time Augmentation . . . 16
5.5.3 Distribution-Free Uncertainty: Inductive Conformal Prediction . . . 17
5.6 Probability Calibration . . . 18
5.6.1 Isotonic Regression . . . 18
5.6.2 Temperature Scaling . . . 18
5.6.3 Calibration Metrics . . . 18
5.7 Cost-Sensitive Decision Making . . . 19
5.8 Explainability . . . 19
5.8.1 SHAP for Clinical Features . . . 19
5.8.2 Grad-CAM for Brain Imaging . . . 19

6 Evaluation **19**
6.1 Experimental Setup . . . 19
6.2 Classification Performance . . . 20
6.3 Calibration Analysis . . . 20
6.4 Baseline + Calibration vs. Cost-Sensitive Training . . . 22
6.5 Uncertainty Quantification Results . . . 24
6.5.1 MC Dropout (Epistemic) . . . 24
6.5.2 TTA (Aleatoric) . . . 24
6.5.3 Conformal Prediction . . . 25

6.6 Explainability Analysis 25
6.6.1 SHAP Feature Importance 25
6.6.2 Grad-CAM Visualisation 27
6.7 Modality Contribution Analysis 27

7 Conclusions and Future Work **29**
7.1 Summary 29
7.2 Limitations 29
7.3 Synthesis and Future Directions 29

1 Introduction

Alzheimer's Disease (AD) is the most prevalent form of dementia, accounting for approximately 60/80% of all dementia cases worldwide [Alzheimer's Association, 2023]. It is characterised by a progressive deterioration of cognitive functions (including memory, language, and executive abilities) that ultimately leads to complete functional dependency. The disease follows a clinical continuum: patients typically transition from a Cognitively Normal (CN) state through Mild Cognitive Impairment (MCI) ("an intermediate stage with measurable cognitive decline but preserved daily functioning") before progressing to overt AD dementia.

Early and accurate identification of the disease stage is paramount for several reasons. First, emerging disease-modifying therapies have shown the greatest efficacy when administered during prodromal stages [van Dyck et al., 2023]. Second, correct staging informs prognosis and enables timely care planning for patients and families. Third, the differential diagnosis between MCI and early AD remains inherently challenging even for experienced clinicians, given the substantial overlap in neuropsychological profiles and imaging biomarkers [Petersen, 2014].

Current clinical diagnosis relies on a combination of cognitive assessments (*e.g.*, Mini-Mental State Examination "MMSE", Alzheimer's Disease Assessment Scale "ADAS"), cerebrospinal fluid biomarkers, and structural neuroimaging. However, each individual modality offers only a partial view: cognitive scores are susceptible to floor and ceiling effects, biomarker assays exhibit non-negligible measurement variability, and visual inspection of MRI scans is subjective and time-consuming. A multimodal approach that systematically integrates heterogeneous data sources is therefore a natural path toward more robust and reliable diagnostic systems.

Contributions. Building upon the VERONECA architecture originally developed in [Buttaro et al., 2025], this work extends the framework along three principal axes:

1. **Uncertainty Quantification (UQ).** We implement and compare three complementary UQ paradigms: Monte Carlo (MC) Dropout for epistemic uncertainty, clinically-grounded Test-Time Augmentation (TTA) for aleatoric uncertainty, and Inductive Conformal Prediction (ICP) for distribution-free prediction sets with guaranteed coverage.

2. **Probability Calibration and Cost-Sensitive Decision Making.** We introduce a post-hoc calibration pipeline based on Isotonic Regression, followed by a Bayesian decision rule that minimises expected clinical cost according to a domain-specific misclassification cost matrix.

3. **Explainability.** We integrate SHAP-based feature attribution for the clinical modality and Gradient-weighted Class Activation Mapping (Grad-CAM) for spatial localisation of discriminative brain regions in 3D MRI inputs.

The following sections of this report are organised according to the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology [Chapman et al., 2000]. Section 2 presents the Business Understanding phase. Sections 3 and 4 describe Data Understanding and Data Preparation, respectively. Section 5 details the Modeling phase, covering the multimodal architecture, boosting algorithm, uncertainty quantification, and calibration pipeline. Section 6 reports the experimental Evaluation. Finally, Section 7 provides Conclusions and directions for Future Work.

2 Business Understanding

2.1 Problem Definition

The overarching goal is to develop a *clinically-aware* Computer-Aided Diagnosis (CAD) system that:

- Performs three-class classification of subjects into CN, MCI, or AD categories by jointly leveraging structural 3D MRI and tabular EHR data.
- Provides *calibrated probability estimates* that faithfully reflect the true likelihood of each diagnostic outcome.
- Quantifies prediction uncertainty at the individual patient level, enabling clinicians to identify cases requiring further review.
- Incorporates domain knowledge about the *asymmetric costs* of different misclassification types in a principled decision-making framework.
- Supports interpetability of the model's predictions, through visual explanations (SHAP plots and Grad-CAM heatmaps).

2.2 Success Criteria

The project defines the following measurable success criteria:

1. **Prediction quality.** Achieve competitive macro-averaged accuracy and F1-score on 5-fold cross-validation, matching or exceeding the baseline VERONECA system.
2. **Calibration.** Reduce the Expected Calibration Error (ECE) through post-hoc calibration, ensuring that model confidence is aligned with empirical accuracy.
3. **Uncertainty reliability.** Demonstrate a statistically significant positive correlation between model uncertainty and prediction errors, and achieve empirical conformal coverage $\geq 1 - \alpha$ (target 90% for $\alpha = 0.1$).
4. **Clinical cost reduction.** The calibration + decision pipeline should not increase the mean clinical cost per sample compared to the uncalibrated baseline.
5. **Explainability.** Produce interpretable explanations (SHAP plots, Grad-CAM heatmaps) that are consistent with known AD neuroanatomy (*e.g.*, hippocampal and temporal lobe atrophy).

2.3 Risk Assessment

The principal risks identified at project inception are:

- **Data scarcity.** The ADNI dataset, while the gold standard for AD research, provides a limited number of subjects with matched MRI and clinical data, especially for the MCI class.
- **Class imbalance.** The three diagnostic classes are not equally represented, potentially biasing the classifier toward the majority class.
- **Missing modalities.** Not all subjects have available MRI scans; the boosting framework must gracefully handle missing imaging data.
- **Cost of errors.** In a clinical setting, a missed AD diagnosis (false negative) carries substantially greater consequences than a false positive, necessitating an asymmetric cost framework.

3 Data Understanding

3.1 The ADNI Dataset

The data used in this project originate from the Alzheimer's Disease Neuroimaging Initiative (ADNI), a multi-site longitudinal study launched in 2003 with the primary objective of developing clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of AD [Jack et al., 2008]. The specific data release employed is the `ADNIMERGE` file dated November 29, 2024, which consolidates key summary variables from multiple ADNI data tables into a single participant-level dataset.

Clinical Data (EHR). The tabular modality consists of demographic, cognitive, and clinical variables extracted from the ADNIMERGE file. After removing columns with >90% missing values, baseline-suffixed duplicates (columns ending in `_bl`), and domain-irrelevant fields (see Table 1 for the full exclusion list), the retained clinical features include:

- **Demographics:** Age, gender, years of education, ethnicity.
- **Cognitive scores:** MMSE (Mini-Mental State Examination), ADAS11 and ADAS13 (Alzheimer's Disease Assessment Scale), RAVLT (Rey Auditory Verbal Learning Test) sub-scores (immediate, learning, forgetting, percent forgetting), LDELTOTAL (Logical Memory delayed recall), DIGITSCOR (digit span), TRABSCOR (trail-making test).
- **Functional assessments:** FAQ (Functional Activities Questionnaire), EcogPt and EcogSP scores (Everyday Cognition "patient and study partner versions") across Memory, Language, Visuospatial, Planning, Attention, and Organization domains.
- **Neuroimaging-derived volumetric measures:** Ventricles and Entorhinal volumes are retained in the clinical dataset after filtering, while other volumetric measures (Hippocampus, WholeBrain, Fusiform, MidTemp, ICV) are excluded.

Categorical variables (*e.g.*, gender, ethnicity) are one-hot encoded. Continuous variables with remaining missing values are imputed as part of the data preparation phase.

Imaging Data (3D MRI). The imaging modality comprises T1-weighted MPRAGE (Magnetisation Prepared Rapid Gradient Echo) structural MRI scans in NIfTI format (`.nii`). Each scan captures a full three-dimensional volume of the subject's brain. The image legend file (`ADNI_T1_MPRAGE_12_17_2024.csv`) provides the mapping between Image Data IDs and subject identifiers (`PTID`), enabling the linkage of imaging and clinical records. 76 unique subjects have available MRI scans. Among these, the baseline class distribution (`DX_bl`) is: CN (23), MCI (38), AD (15).

Diagnostic Labels. The target variable is the clinical diagnosis (`DX`) extracted from ADNIMERGE, mapped to a numerical encoding: CN $\rightarrow$ 1, MCI $\rightarrow$ 2, AD (Dementia) $\rightarrow$ 3. Records with missing diagnosis are excluded.

Table 1: Columns excluded from the ADNIMERGE clinical dataset.

Category	Excluded columns
Administrative	`RID`, `SITE`, `COLPROT`, `ORIGPROT`, `EXAMDATE`, `update_stamp`
Biomarkers	`ABETA`, `TAU`, `PTAU`, `FDG`, `PIB`, `AV45`
Genetics	`APOE4`, `PTRACCAT`, `PTETHCAT`
Neuroimaging	`Hippocampus`, `WholeBrain`, `Enthorhinal`, `Fusiform`, `MidTemp`, `ICV`
Scales	`CDRSB`, `EcoPtMem`, `FSVERSION`
Months	`M`
Baseline suffixed	All columns with `_bl` / `_BL` suffix

3.2 Exploratory Analysis

Dataset composition and class distribution. After preprocessing, the ADNI dataset comprises 16,410 subjects with clinical records (CN: 6,389; MCI: 8,270; AD: 1,751), of which 76 have available structural MRI scans. Figure 1 shows the diagnostic distribution among subjects with multimodal data: the prevalence of MCI cases (50%, $n = 38$) relative to CN (30%, $n = 23$) and AD (20%, $n = 15$) reflects the natural epidemiological pattern of Alzheimer's disease progression, where MCI represents an extended intermediate prodromal stage between normal cognition and overt dementia.

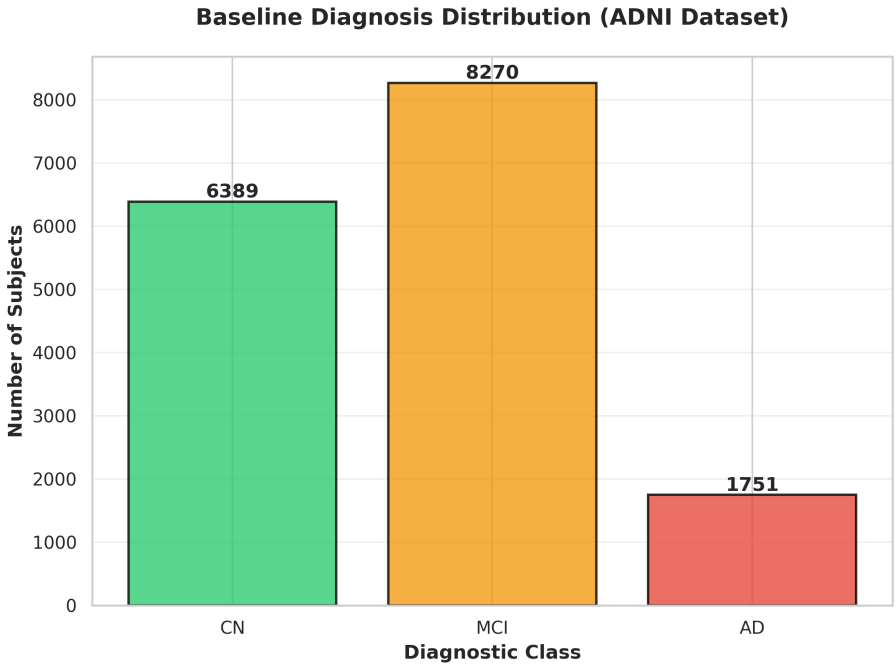


Figure 1: Distribution of baseline diagnostic labels among the 76 subjects with available MRI scans. The dataset exhibits natural class imbalance characteristic of AD progression studies, with MCI as the most prevalent category (50%), followed by CN (30%) and AD (20%).

Missing values and feature quality. Following the 90% missingness threshold filtering, the retained clinical features exhibit varying but acceptable levels of missing data. Higher missingness rates (40/60%) are observed primarily in neuropsychological test subscales (*e.g.*, RAVLT components, EcogSP domain scores) and composite cognitive indices (*e.g.*, mPACCdigit, mPACCtrailsB), reflecting the longitudinal study design and the gradual evolution of ADNI assessment protocols across cohorts and phases. Core demographic and functional measures (age, education, MMSE, FAQ) are nearly complete (<10% missing). The Random Forest base learner handles

residual missingness natively through surrogate splits, eliminating the need for explicit imputation.

Clinical feature distributions across diagnostic classes. Figure 2 presents the distribution of four key clinical features (MMSE, ADAS13, FAQ, Age) stratified by diagnostic class. The violin plots reveal distinct separation trends with varying degrees of overlap:

- **MMSE**: The CN class exhibits a strong ceiling effect, with scores heavily concentrated at the maximum (median $\approx 29/30$). MCI subjects show broader variability (median ≈ 27), while the AD class displays a marked cognitive decline with a significantly lower and more dispersed distribution (median ≈ 22).
- **ADAS13 and FAQ**: Both metrics show a monotonic increase with disease progression. While ADAS13 shows a steady shift in medians (CN ≈ 9, MCI ≈ 18, AD ≈ 32), the **FAQ** scores are almost exclusively zero for the CN group, beginning to disperse slightly in MCI (median ≈ 2) before showing a drastic and highly variable increase in AD patients (median ≈ 16).
- **Age**: The age distributions remain remarkably consistent across the three groups, with medians stable around **73/75 years**. This confirms that observed cognitive and volumetric differences are not confounded by age but purely reflect the pathological status.

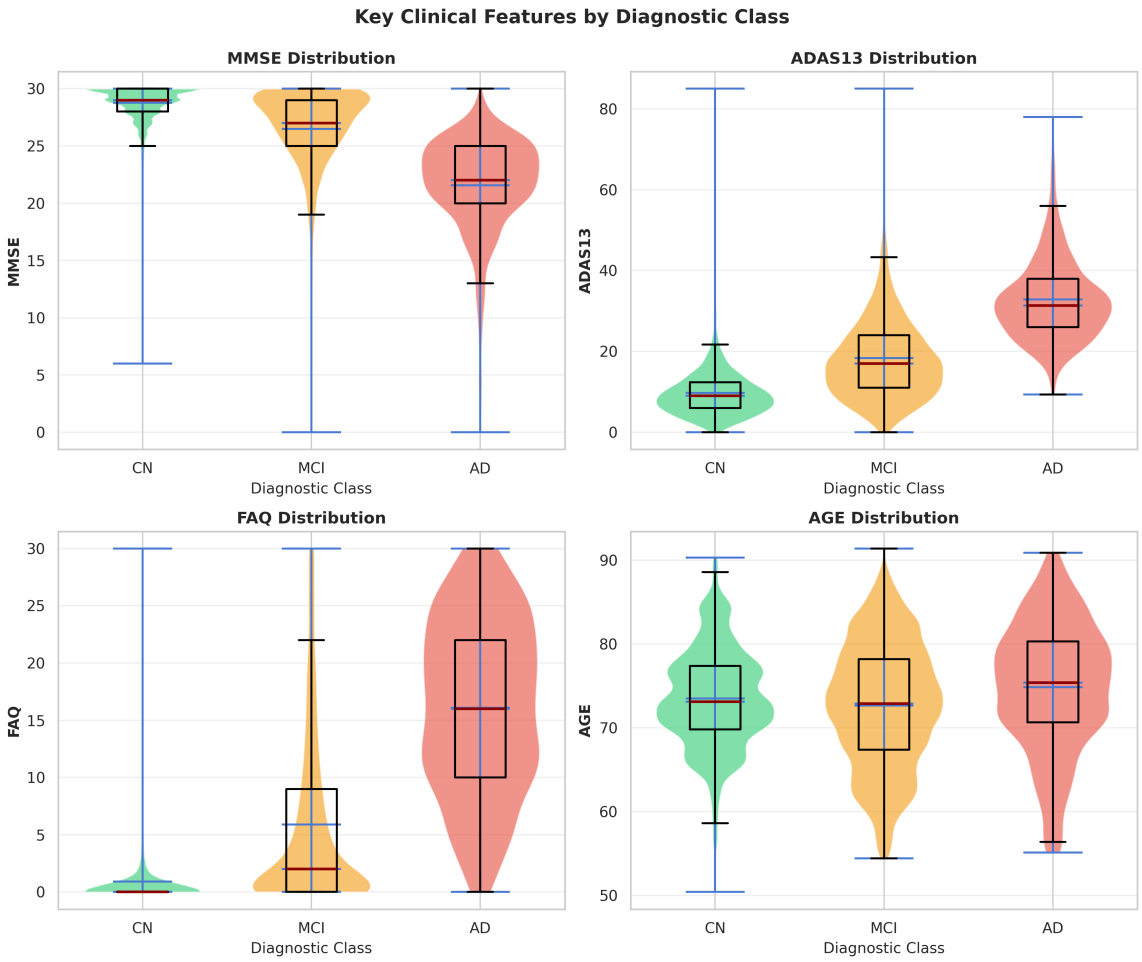


Figure 2: Distribution of key clinical features (MMSE, ADAS13, FAQ, Age) across diagnostic classes. Violin plots show probability density, overlaid with box plots indicating quartiles and medians. Note the ceiling effect in MMSE for CN and the sharp increase in FAQ variance for the AD group.

Brain volumetric biomarkers. Figure 3 displays regional brain volumes (Hippocampus, Ventricles, Entorhinal cortex) stratified by diagnosis, confirming hallmark neuroimaging signatures:

- **Medial Temporal Lobe Atrophy**: A significant and progressive volumetric reduction is observed in the hippocampus and entorhinal cortex. The median hippocampal volume drops from approximately **7200** mm^3 in CN to roughly **5500** mm^3 in AD (a $\approx 23\%$ reduction), with MCI subjects occupying an intermediate range (≈ 6800 mm^3). A similar trend is visible for the entorhinal area, though the separation between CN and MCI is more nuanced.

- **Ventricular Enlargement**: Ventricular volume shows a clear expansion. The median volume increases from $\approx$ 32,000 mm^3 (CN) to over 50,000 mm^3 (AD), representing an expansion exceeding 50%. The high number of outliers and large variance in the AD class suggest a heterogeneous structural response in late-stage disease.

- **MCI Heterogeneity**: MCI subjects exhibit distributions that broadly overlap both CN and AD ranges, reinforcing the status of this class as a clinical "limbo" and highlighting the necessity of multimodal integration to identify high-risk converters.

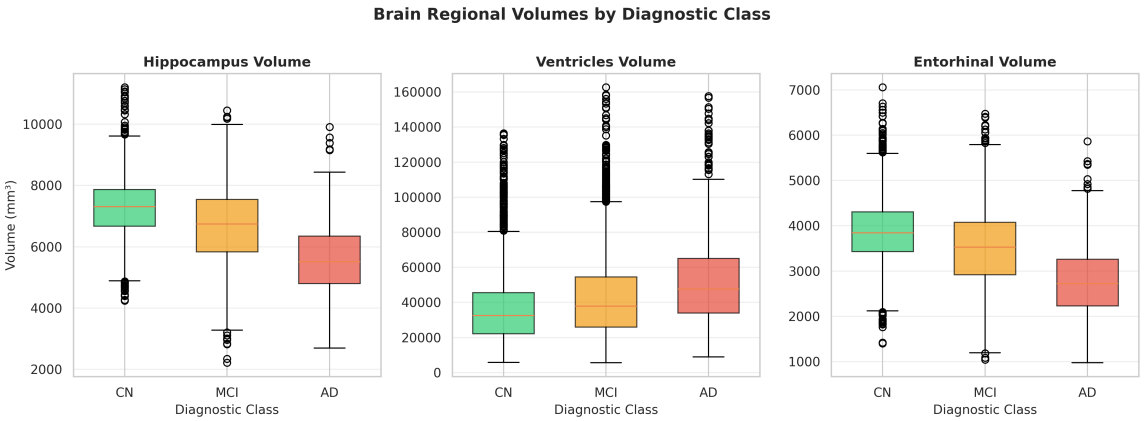


Figure 3: Regional brain volumes by diagnostic class. Box plots show median, interquartile range, and outliers. Progressive hippocampal/entorhinal atrophy and significant ventricular expansion characterize the transition from CN to AD, with MCI representing a highly heterogeneous intermediate stage.

Feature correlations. Figure 4 reveals expected relationships between cognitive and structural biomarkers. As anticipated, global cognition measures exhibit strong inverse correlations with impairment scales; specifically, MMSE shows a strong negative correlation (deep blue) with ADAS11 and ADAS13, as higher ADAS scores indicate greater impairment. MMSE is also positively correlated with memory and executive function assessments, such as LDELTOTAL, mPACCdigit, and mPACCtrailsB, as well as memory-related scores like RAVLT_immediate and RAVLT_learning. Conversely, functional impairment (FAQ) demonstrates consistent negative correlations with cognitive performance (MMSE, LDELTOTAL) and a positive correlation with impairment severity (ADAS). Furthermore, the matrix highlights the link between structural biomarkers and cognition: Entorhinal volume is positively associated with cognitive scores, while Ventricular volume shows an inverse relationship, increasing as cognitive performance declines. These patterns confirm the construct validity of the selected features in capturing the multifaceted nature of AD.

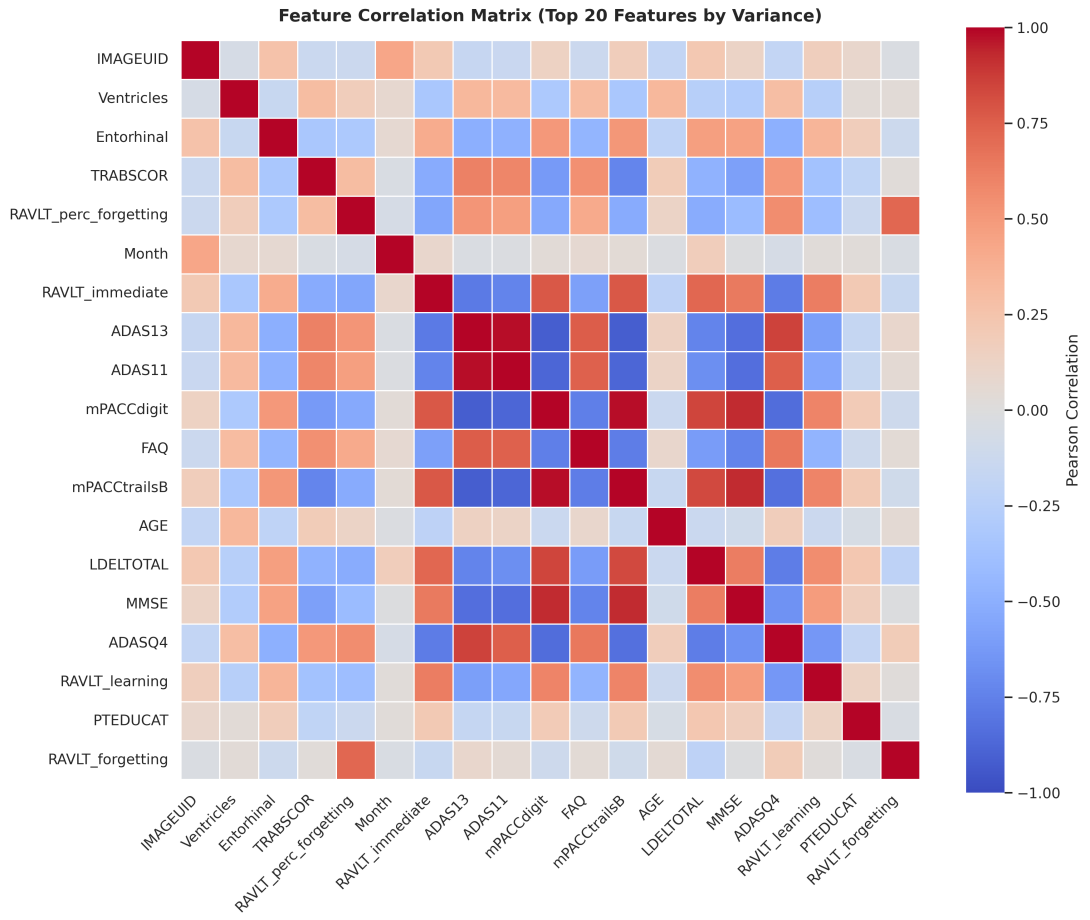


Figure 4: Pearson correlation matrix of the top 20 clinical and volumetric features by variance. Warm colors (red) indicate positive correlation, cool colors (blue) indicate negative correlation. The clustering shows strong associations between global cognitive assessments (MMSE, ADAS), memory tests (RAVLT, LDELTOTAL), and structural measures (Entorhinal, Ventricles), confirming expected clinical patterns in Alzheimer's Disease.

3.3 Data Quality and Missing Modalities

A key characteristic of the ADNI dataset is the incomplete overlap between modalities: not all subjects with clinical records have a corresponding MRI scan. The VERONECA boosting framework is inherently designed to handle this situation, as each boosting iteration operates on the subset of subjects for which data is available for the selected modality (Section 5.4). This flexibility is critical for maximising sample utilisation in a real-world clinical setting where multimodal completeness cannot be guaranteed.

4 Data Preparation

4.1 Clinical Data Pipeline

The clinical data preparation proceeds through the following stages:

1. **Column filtering.** Administrative, biomarker, and high-missingness columns are removed (Table 1). Specifically, all columns with more than 90% missing values are discarded, as are all baseline-specific columns (suffix `_bl`).

2. **Missing value handling.** After column filtering, remaining missing entries are *not imputed* during preprocessing. Instead, they are handled natively by the base learner: scikit-learn's `RandomForestClassifier` supports missing values directly through surrogate split mechanisms during tree construction, allowing the model to learn optimal branching strategies in the presence of incomplete data without requiring explicit imputation.

3. **Categorical encoding.** Categorical variables (*e.g.*, gender, ethnicity) are transformed via scikit-learn's `OneHotEncoder` to produce a fully numeric feature matrix.

4. **Merging.** Clinical features and image paths are merged with diagnostic labels on the composite key (`Patient ID`, `VISCODE`) using left joins, ensuring that all labelled subjects are retained even if one modality is missing.

4.2 MRI Image Pipeline

The imaging pipeline transforms raw NIfTI volumes into standardised, compact tensor representations:

1. **Loading.** Raw `.nii` files are loaded via the NiBabel library and converted to floating-point tensors.

2. **Central cropping.** Each volume is centre-cropped to a fixed spatial resolution of $128 \times 128 \times 50$ voxels. This removes non-informative peripheral regions (skull boundary, air) while preserving the central brain structures relevant for AD diagnosis (*e.g.*, hippocampus, temporal cortex, ventricles).

3. **Min-max normalisation.** Voxel intensities are linearly rescaled to the $[0, 1]$ range:

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x) + \epsilon}, \quad \epsilon = 10^{-8}. \tag{1}$$

4. **Caching.** Preprocessed tensors are serialised as `.pkl` files in the `images_post/` directory to avoid redundant computation during training.

4.3 Data Augmentation for MRI

The limited number of available MRI scans ($\sim$76 unique images) presents a significant challenge for CNN training, as deep networks are prone to overfitting on small datasets. To address this, we implement a *pre-computed* data augmentation strategy that generates five augmented variants per original image, yielding a $5\times$ expansion of the imaging dataset.

The augmentation pipeline, implemented using the `albumentations` library, applies a composition of clinically-grounded transformations applied independently to each 2D slice of the 3D volume:

- **Rotation** (± 10, $p = 0.6$): simulates natural head movement during acquisition.
- **Affine translation and scaling** ($\pm 5\%$ each, $p = 0.4$): models patient positioning variability.
- **Elastic deformation** ($\alpha = 30$, $\sigma = 5$, $p = 0.3$): mimics scanner field inhomogeneity and inter-subject anatomical variation.
- **Gaussian noise** ($\sigma \in [0.005, 0.015]$, $p = 0.3$): emulates signal-to-noise ratio degradation.
- **Random brightness and contrast** ($\pm 10\%$, $p = 0.4$): accounts for scanner calibration differences.
- **Random gamma correction** ($\gamma \in [0.9, 1.1]$, $p = 0.3$): introduces non-linear intensity variation.

Anti-leakage strategy. When using augmented images, it is essential to prevent data leakage: augmented copies of the same original image must not appear in both training and test sets. This is enforced through `GroupKFold` cross-validation, where the group key is the original image identifier. This guarantees that all augmented variants of a given subject are confined to the same fold.

4.4 Cross-Validation and Data Splits

All experiments employ stratified 5-fold cross-validation (`StratifiedKFold`, $K_{\text{splits}} = 5$, `random_state`= 42). For experiments involving Conformal Prediction and post-hoc calibration, the original training fold is further subdivided into a *training* set (60% of total data), a *calibration* set (20%), and a *test* set (20%). The calibration set is used for:

- Fitting the Isotonic Regression calibrator.
- Computing conformal non-conformity score quantiles.

This three-way split ensures that calibration and conformal thresholds are estimated on data unseen during training, preventing overly optimistic calibration estimates.

5 Modeling

5.1 Architecture Overview

The VERONECA system follows a *late-fusion* multimodal paradigm: each data modality is processed by a dedicated base learner, and their outputs are combined through an ensemble boosting algorithm. Figure 5 provides a schematic overview.

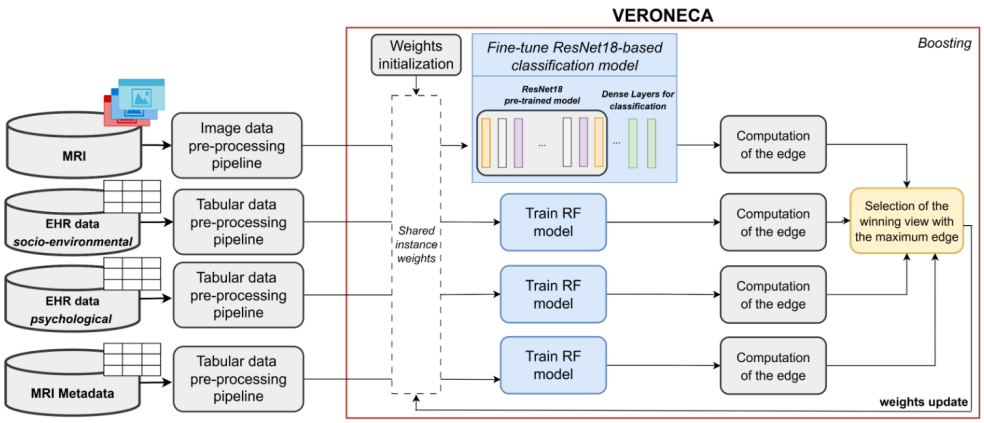


Figure 5: High-level architecture of the VERONECA multimodal classification system. Clinical tabular data is processed by a Random Forest, while 3D MRI volumes are processed by a fine-tuned ResNet-18. The IRBoostSH algorithm fuses the two modalities through boosting with multi-armed bandit modality selection.

5.2 Clinical Branch: Random Forest

The clinical modality is processed by a `RandomForestClassifier` from scikit-learn, configured with 10 trees and a minimum samples-per-split threshold of 60. Although relatively small in ensemble size, this configuration has been empirically chosen to balance expressiveness with sample-weight responsiveness, which is critical in the boosting setting where sample weights are re-distributed at each iteration.

5.3 Imaging Branch: VeroResNet

The imaging branch employs a 3D ResNet-18 architecture as the backbone feature extractor, provided by the MONAI (Medical Open Network for Artificial Intelligence) library. The network is initialised with weights pre-trained on the MedicalNet dataset [Chen et al., 2019] and fine-tuned on the ADNI images.

5.3.1 Architecture Details

The `VeroResNet` model extends the standard 3D ResNet-18 with a classification head comprising:

1. A fully connected layer: $512 \to 512$ with ReLU activation.
2. Dropout layer ($p = 0.5$), serving dual purpose for regularisation during training and MC Dropout during inference.
3. A fully connected layer: $512 \to 256$ with ReLU activation.
4. Dropout layer ($p = 0.5$).
5. Output layer: $256 \to 3$ (three classes: CN, MCI, AD).

5.3.2 Transfer Learning and Fine-Tuning

At `freeze_level`$= 2$ (the default configuration), the entire ResNet backbone is frozen and only the dense classification head is trained. This strategy mitigates overfitting given the limited number of MRI samples, as the convolutional layers retain generalised volumetric feature representations learned from a much larger medical imaging corpus.

The network is trained with:

- **Loss:** Cross-entropy with per-sample weighting (from the boosting distribution).
- **Optimiser:** Adam (lr $= 10^{-4}$).
- **Mini-batch size:** 8 images.
- **Epochs:** 50 per boosting iteration.

5.4 Multimodal Boosting: IRBoostSH

The two modality-specific base learners are fused through the **Implicit Randomization Boosting with Shared Weights (IRBoostSH)** algorithm, an extension of AdaBoost to the multimodal setting that incorporates a multi-armed bandit strategy for modality selection.

5.4.1 Algorithm Description

At each boosting iteration $t = 1, \ldots, T$ (with $T = 10$ by default):

1. **Weight normalisation.** The sample weight distribution $w^{(t)}$ is normalised to sum to one.
2. **Modality selection.** A modality m_t is sampled from a probability distribution $q^{(t)}$ over modalities, computed as:

$$q_m^{(t)} = (1 - \gamma) \frac{p_m^{(t)}}{\sum_{m'} p_{m'}^{(t)}} + \frac{\gamma}{K}, \tag{2}$$

where K is the number of modalities, $\gamma = 0.3$ is an exploration parameter, and $p_m^{(t)}$ is the arm probability (reward-based weight) for modality m.

3. **Weak learner training.** The base estimator associated with modality m_t is fitted on the corresponding training data with the current sample weights $w^{(t)}$.

4. **Edge computation.** The weighted edge (advantage over random guessing) is computed as:

$$\text{edge}_t = \sum_i w_i^{(t)} \cdot 2\left(\mathbb{1}[\hat{y}_i = y_i] - 0.5\right). \tag{3}$$

5. **Learner weight.** The contribution weight α_t is computed as:

$$\alpha_t = \frac{\eta}{2} \ln \frac{1 + \text{edge}_t}{1 - \text{edge}_t}, \tag{4}$$

where $\eta = 1.0$ is the learning rate.

6. **Sample weight update.** Weights are updated via:

$$w_i^{(t+1)} \propto w_i^{(t)} \exp(-\alpha_t \cdot 2\left(\mathbb{1}[\hat{y}_i = y_i] - 0.5\right)). \tag{5}$$

Misclassified samples receive increased weight, focusing subsequent iterations on difficult examples.

7. **Arm probability update.** The multi-armed bandit reward r_{m_t} updates p_m to bias future selections toward better-performing modalities:

$$p_m^{(t+1)} = p_m^{(t)} \exp\left(\frac{\gamma}{3K}\left(r_m + \frac{\sigma}{q_m^{(t)}\sqrt{TK}}\right)\right), \tag{6}$$

where $\sigma = 0.15$ and $r_{m_t} = (1 - \sqrt{1 - \text{edge}_t^2})/q_{m_t}^{(t)}$.

5.4.2 Handling Missing Modalities

A distinctive feature of `IRBoostSH` is its native support for missing modalities. At each iteration, the weak learner operates only on the subset of samples for which the selected modality is available. Weight updates and edge calculations are accordingly restricted to this subset, while samples lacking the selected modality retain their weights unchanged. This enables the framework to exploit all available information without requiring complete multimodal data for every subject.

5.4.3 Prediction Aggregation

Final predictions are obtained by weighted combination of all T weak learner outputs:

$$P(y = c \mid x) = \frac{\sum_{t=1}^{T} \alpha_t \cdot \hat{p}_t(y = c \mid x_{m_t})}{\sum_{t=1}^{T} \alpha_t \cdot \sum_{c'} \hat{p}_t(y = c' \mid x_{m_t})}, \tag{7}$$

where only iterations for which the subject has the corresponding modality data contribute to the sum.

5.4.4 Cost-Sensitive Boosting

An alternative training regime incorporates a clinical cost matrix $C \in \mathbb{R}^{3\times3}$ directly into the boosting objective. Instead of maximising classification accuracy, the edge is redefined based on weighted misclassification cost:

$$\text{edge}_t^{\text{cost}} = 1 - 2 \cdot \frac{\sum_i w_i^{(t)} \cdot C(y_i, \hat{y}_i)}{\max(C) \cdot n}, \tag{8}$$

and sample weights are updated proportionally to the normalised cost incurred. The clinical cost matrix used is:

Table 2: Clinical misclassification cost matrix $C(i, j)$ where rows represent true classes and columns represent predicted classes.

	Pred: CN	**Pred: MCI**	**Pred: AD**
True: CN	0.0	0.3	0.9
True: MCI	0.5	0.0	0.7
True: AD	1.0	0.8	0.0

The highest cost is assigned to the AD→CN error (missed dementia), reflecting the clinical imperative of not missing a progressive neurodegenerative condition.

5.5 Uncertainty Quantification

We decompose predictive uncertainty into three complementary components, each captured by a dedicated method.

5.5.1 Epistemic Uncertainty: Monte Carlo Dropout

Epistemic (model) uncertainty reflects the model's lack of knowledge and can, in principle, be reduced with additional data. We estimate it via MC Dropout [Gal and Ghahramani, 2016]: at inference time, the dropout layers in VeroResNet (with $p = 0.5$) are kept active, and $N_{\text{mc}} = 25$ stochastic forward passes are performed for each test sample.

At the boosting level, MC Dropout is applied by iterating the full ensemble prediction N_{mc} times, each time sampling different dropout masks in the CNN models. The resulting distribution of predicted probability vectors is summarised as:

$$\bar{p}(y = c \mid x) = \frac{1}{N_{\text{mc}}} \sum_{n=1}^{N_{\text{mc}}} p_n(y = c \mid x), \tag{9}$$

$$\text{Confidence} = \max_c \bar{p}(y = c \mid x), \tag{10}$$

$$\text{Epistemic UQ} = \text{Std}\Big[\max_c p_n(y = c \mid x)\Big]_{n=1}^{N_{\text{mc}}}. \tag{11}$$

High epistemic uncertainty indicates that the model's predictions are sensitive to its internal stochastic parameters, suggesting that the sample lies in a region of the feature space that is poorly covered by the training data.

To obtain an epistemic uncertainty signal from the clinical Random Forest, we explored a tree subsampling strategy: at test time, for each sample, we randomly selected 70% of the trees in the forest and computed predictions over multiple such random subsets. The variability across these sub-forests was intended as an epistemic uncertainty estimate for the clinical modality.

5.5.2 Aleatoric Uncertainty: Test-Time Augmentation

Aleatoric (data) uncertainty captures inherent noise in the input data and cannot be reduced by collecting more training data. We estimate it through Test-Time Augmentation (TTA) applied to the *clinical* modality with feature-specific, clinically-grounded noise levels:

Table 3: Feature-specific noise levels for TTA on clinical data, based on published measurement variability literature.

Feature	**Noise type**	**Magnitude**
MMSE	Absolute	$\pm$2.0 points
ABETA (CSF Aβ42)	Relative	$\pm$8.0%
TAU	Relative	$\pm$8.5%
PTAU	Relative	$\pm$12.0%
Age	None	0 (stable)
APOE4	None	0 (genetic)
Other features	Relative	$\pm$5.0% (default)

For $N_{\text{tta}} = 10$ augmented versions of the clinical data, the ensemble prediction is repeated and aleatoric uncertainty is estimated as the standard deviation of the maximum predicted probabilities across augmentations:

$$\text{Aleatoric UQ} = \text{Std}\Big[\max_c p_n^{\text{tta}}(y = c \mid x)\Big]_{n=1}^{N_{\text{tta}}}. \tag{12}$$

Imaging data was not perturbed during TTA. Empirically, the boosting modality-weight analysis showed that the imaging branch contributed less than 1% to the ensemble's cumulative prediction weight across boosting iterations; therefore, applying TTA to images would have negligible impact on ensemble uncertainty estimates while incurring substantial compute cost. For this reason we concentrated TTA on the clinical modality.

5.5.3 Distribution-Free Uncertainty: Inductive Conformal Prediction

Conformal Prediction (CP) provides a principled, distribution-free method for constructing *prediction sets* with guaranteed finite-sample coverage [Vovk et al., 2005]. Unlike MC Dropout and TTA, which produce heuristic uncertainty scores, CP offers the rigorous guarantee:

$$\Pr[y_{\text{new}} \in \mathcal{C}(x_{\text{new}})] \geq 1 - \alpha, \tag{13}$$

where α is the user-specified miscoverage rate and $\mathcal{C}(x)$ is the prediction set.

Procedure. Following the Inductive Conformal Prediction framework [Papadopoulos et al., 2002]:

1. **Non-conformity scores.** On the calibration set, compute the non-conformity score for each sample as $s_i = 1 - \hat{p}(y_i \mid x_i)$, where $\hat{p}(y_i \mid x_i)$ is the model's predicted probability for the true class.

2. **Quantile threshold.** Compute the conformal quantile:

$$\hat{q} = \text{Quantile}\left(\{s_1, \ldots, s_{n_{\text{cal}}}\},\ \frac{\lceil (n_{\text{cal}} + 1)(1 - \alpha) \rceil}{n_{\text{cal}}}\right). \tag{14}$$

3. **Prediction sets.** For a test sample x, the prediction set is:

$$\mathcal{C}(x) = \{c \in \{1, \ldots, K\} : \hat{p}(y = c \mid x) \geq 1 - \hat{q}\}. \tag{15}$$

4. **Conformal uncertainty.** Defined as the normalised set size:

$$u_{\text{conf}}(x) = \frac{|\mathcal{C}(x)| - 1}{K - 1}, \tag{16}$$

yielding $u_{\text{conf}} = 0$ for singleton sets (high certainty) and $u_{\text{conf}} = 1$ for full sets (complete uncertainty).

In our setting, $\alpha = 0.1$ provides a target coverage of 90%. The clinical interpretation is intuitive: a singleton prediction set (*e.g.*, {AD}) indicates a confident diagnosis, while a multi-element set (*e.g.*, {MCI, AD}) signals diagnostic ambiguity warranting further clinical investigation.

5.6 Probability Calibration

Raw ensemble probabilities from boosting classifiers are often poorly calibrated [Niculescu-Mizil and Caruana, 2005]: they may be systematically overconfident or underconfident, meaning that a predicted probability of 0.8 does not correspond to an 80% empirical frequency of correctness. Calibration is essential for the subsequent cost-sensitive decision rule, which requires accurate probability estimates to compute expected costs.

5.6.1 Isotonic Regression

We apply **Isotonic Regression (IR)** as the primary calibration method for the final ensemble output [Zadrozny and Elkan, 2002]. IR fits a non-parametric, monotonically non-decreasing mapping for each class:

$$p_c^{\text{cal}}(x) = f_c\big(p_c^{\text{uncal}}(x)\big), \quad c \in \{1, \ldots, K\}, \tag{17}$$

where f_c is the isotonic regression function fitted on the calibration set via the pair-adjacent-violators algorithm. Calibrated probabilities are then renormalised to sum to one:

$$\tilde{p}_c^{\text{cal}}(x) = \frac{p_c^{\text{cal}}(x)}{\sum_{c'} p_{c'}^{\text{cal}}(x)}. \tag{18}$$

5.6.2 Temperature Scaling

For individual CNN outputs, **Temperature Scaling (TS)** [Guo et al., 2017] is available as an additional calibration option. TS learns a single scalar parameter $T > 0$ by minimising the negative log-likelihood on the calibration set:

$$p^{\text{cal}}(y = c \mid x) = \text{softmax}\Big(\frac{z_c}{T}\Big), \tag{19}$$

where z_c are the raw logits. The optimal T is found via L-BFGS optimisation (50 iterations, $\text{lr} = 0.01$, initial $T = 1.5$).

5.6.3 Calibration Metrics

Calibration quality is assessed through:

- **Expected Calibration Error (ECE):** $\text{ECE} = \displaystyle\sum_{b=1}^{B} \frac{n_b}{n} \left|\text{acc}_b - \text{conf}_b\right|$, averaged over $B = 10$ confidence bins.
- **Maximum Calibration Error (MCE):** $\text{MCE} = \max_b \left|\text{acc}_b - \text{conf}_b\right|$.
- **Brier Score:** $\text{BS} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \sum_{c=1}^{K} \big(p_c^i - y_c^i\big)^2$.
- **Negative Log-Likelihood (NLL):** $\text{NLL} = -\dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \ln p_{y_i}^i$.

5.7 Cost-Sensitive Decision Making

Given calibrated probabilities, the final prediction is determined by a Bayesian decision rule that minimises expected clinical cost [Elkan, 2001]:

$$\hat{y}(x) = \arg\min_{j} \sum_{i=1}^{K} C(i,j) \cdot \tilde{p}^{\text{cal}}(y = i \mid x), \tag{20}$$

where $C(i,j)$ is the cost matrix defined in Table 2. This rule shifts decision boundaries to be more conservative for high-cost errors: for instance, a patient whose calibrated probabilities slightly favour CN over MCI may be classified as MCI if the cost of missing impairment is sufficiently high.

Sensitivity analysis. To assess the robustness of the decision rule, we perform a sensitivity analysis by perturbing the cost matrix elements by factors in $[0.8, 1.2]$ and measuring the fraction of predictions that remain unchanged. High prediction stability (>90%) across perturbations indicates that the decision boundaries are not overly sensitive to the specific cost values chosen.

5.8 Explainability

5.8.1 SHAP for Clinical Features

For the clinical Random Forest models, we employ SHAP (SHapley Additive exPlanations) [Lundberg and Lee, 2017] via `TreeExplainer`. SHAP values decompose each prediction into additive feature contributions grounded in cooperative game theory.

Given that the ensemble contains multiple RF models from different boosting iterations, we compute a weighted aggregation:

$$\phi_j^{\text{agg}}(x) = \frac{\sum_{t \in \mathcal{T}_{\text{clin}}} \alpha_t \cdot \phi_{j,t}(x)}{\sum_{t \in \mathcal{T}_{\text{clin}}} \alpha_t}, \tag{21}$$

where $\mathcal{T}_{\text{clin}}$ is the set of boosting iterations that selected the clinical modality and $\phi_{j,t}(x)$ are the SHAP values for feature j from the RF model at iteration t.

5.8.2 Grad-CAM for Brain Imaging

For the CNN models, we apply Gradient-weighted Class Activation Mapping (Grad-CAM) [Selvaraju et al., 2017] to generate spatial heatmaps highlighting the brain regions most influential for each prediction. The target layer is the last convolutional block (`layer4`) of the ResNet backbone, and the heatmap is computed as:

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k^c A^k \right), \quad \alpha_k^c = \frac{1}{Z} \sum_{d,h,w} \frac{\partial y^c}{\partial A^k_{d,h,w}}, \tag{22}$$

where A^k is the k-th feature map activation and y^c is the logit for class c. The 3D heatmap is upsampled to the original input resolution via trilinear interpolation and overlaid on axial brain slices for visualisation.

6 Evaluation

6.1 Experimental Setup

All experiments are evaluated under stratified 5-fold cross-validation. The key experimental configurations are summarised in Table 4.

Table 4: Summary of experimental configurations. "CS Train" = cost-sensitive boosting during training; "IR" = Isotonic Regression post-hoc calibration; "CS Infer" = cost-sensitive Bayesian decision rule at inference.

Exp	Description	Train	MCDO	TTA	CP	CS Train	IR	CS Infer
exp1	Baseline	✓						
exp8	Baseline + MCDO		✓					
exp11	Baseline + TTA			✓				
exp13	Full pipeline (Calib + CP + CS)				✓		✓	✓
exp15	Aggressive CS Inference				✓		✓	✓
exp16	CS Training + Calib + CP	✓			✓	✓	✓	✓
exp17	CS Training + Calib (argmax)				✓	✓	✓	
exp18	Standard + Calib (argmax)				✓		✓	

6.2 Classification Performance

Table 5 reports the classification performance for the key experimental configurations that differ in training strategy and produce measurable performance variation. Note that configurations that only modify inference-time behaviour without retraining (*e.g.*, MC Dropout in exp8, cost-sensitive decision rule in exp13 vs. exp18) yield identical classification metrics to their corresponding baseline, since the final prediction is still based on argmax of the same underlying probability distribution; their distinct contributions are captured through uncertainty quantification (Section 5.5) and calibration analysis (Section 5.6).

Table 5: Classification performance (macro-averaged, 5-fold CV). Configurations with inference-only modifications (*e.g.*, MC Dropout, CS decision) are omitted as they produce identical metrics to their base training configuration. Best results in bold.

Configuration	Accuracy	Precision	Recall	F1-score
Baseline (exp1)	0.8142	0.8226	0.8176	0.8198
+ TTA (exp11)	**0.8170**	**0.8261**	**0.8197**	**0.8226**
+ Calibration + CP + CS (exp13)	0.8091	0.8183	0.8116	0.8147
CS Training + Calib + CP (exp16)	0.7822	0.7975	0.7838	0.7899

Note on Data Augmentation. Given that preliminary analysis of the boosting algorithm's modality weights revealed minimal contribution from the imaging component (Section 6.7), we explored whether augmenting the training set with additional MRI samples could enhance the model's ability to leverage visual features. Specifically, we applied clinically-grounded data augmentation techniques as detalied in section 4.3 to the MRI training data, effectively increasing the number of unique imaging samples available for training the ResNet branch. However, this strategy did not yield measurable improvements in classification performance. The lack of benefit suggests that the limited contribution of imaging features to the ensemble's predictions is not primarily due to insufficient training data, but rather reflects either inherent limitations in the 3D ResNet architecture's capacity to extract discriminative volumetric patterns from relatively small medical imaging datasets, or the dominant predictive power of clinical tabular features which already capture much of the variance relevant to disease staging. Consequently, the augmented training configuration is not included in the comparative analysis below.

6.3 Calibration Analysis

Table 6 presents the calibration metrics averaged across all 5 folds, demonstrating the substantial improvement achieved through Isotonic Regression post-hoc calibration. The Expected

Calibration Error (ECE) is reduced by 89%, from 0.1799 to 0.0195, indicating that the calibrated probabilities are highly aligned with empirical accuracy. Figure 6 visualises the ECE reduction for each individual fold, showing consistent improvements across all cross-validation splits with reductions ranging from 83% to 96%.

Table 6: Calibration metrics before and after Isotonic Regression (5-fold average).

Metric	**Before IR**	**After IR**	**Improvement**	**Improv. %**
ECE	0.1799	0.0195	0.1604	89.1%
MCE	0.3405	0.1334	0.2071	56.8%
Brier	0.1120	0.0899	0.0221	19.7%
NLL	0.5442	0.4479	0.0963	17.7%

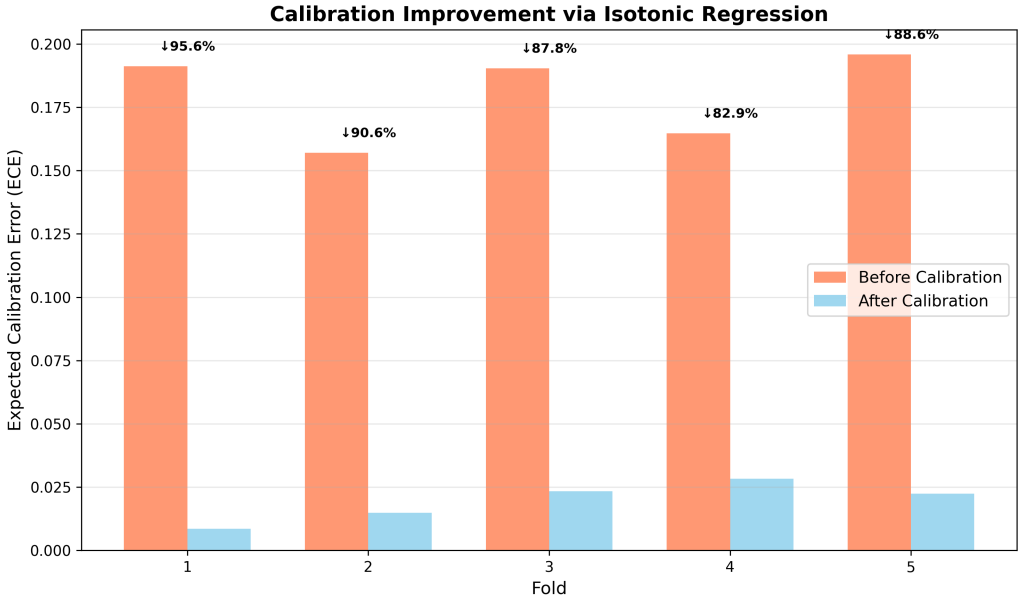
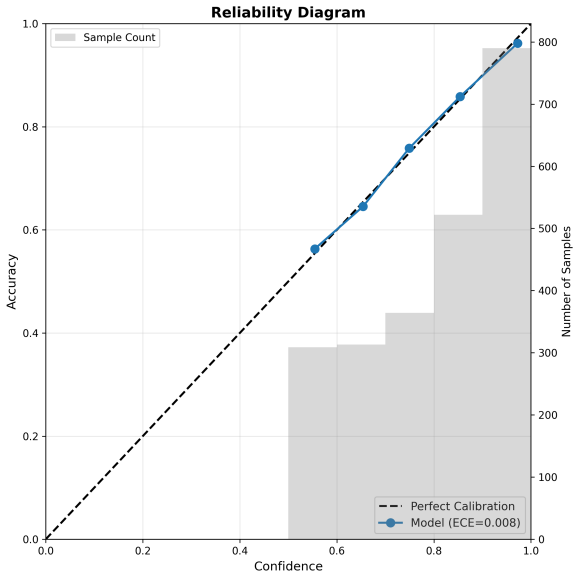
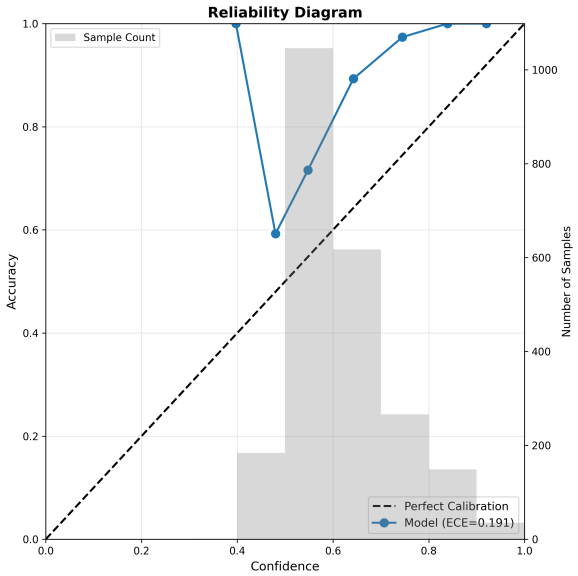


Figure 6: Expected Calibration Error (ECE) before and after Isotonic Regression across all 5 folds. The percentage reduction is displayed above each fold's bars. Calibration consistently improves across all folds, with ECE reductions ranging from 82.9% (Fold 4) to 95.6% (Fold 1).



(a) Before Isotonic Regression (ECE = 0.191).



(b) After Isotonic Regression (ECE = 0.008).

Figure 7: Reliability diagrams for Fold 1 before and after Isotonic Regression calibration. The diagonal dashed line represents perfect calibration (predicted confidence matches empirical accuracy). Grey bars indicate the number of samples per confidence bin. Before calibration (left), the model shows systematic overconfidence in the mid-to-high confidence range. After calibration (right), the predicted probabilities closely align with the perfect calibration line, reducing ECE from 0.191 to 0.008 (95.6% improvement).

6.4 Baseline + Calibration vs. Cost-Sensitive Training

A central research question of this work is whether it is more effective to incorporate clinical cost awareness during *training* (cost-sensitive boosting, Section 5.4.4) or during *post-processing* (calibration + cost-sensitive decision, Sections 5.6–5.7).

Our experiments compare the following strategies:

1. **Baseline + Calibration (exp18):** Standard accuracy-maximising boosting → Isotonic Regression calibration → argmax prediction.

2. **Baseline + Calibration + CS Inference (exp13):** Standard boosting → IR calibration → cost-sensitive Bayesian decision.

3. **Cost-Sensitive Training + Calibration (exp16):** Cost-minimising boosting → IR calibration → cost-sensitive decision.

4. **Cost-Sensitive Training + Calibration + Argmax (exp17):** Cost-minimising boosting → IR calibration → argmax prediction.

Table 7: Comparison of training strategies: standard vs. cost-sensitive boosting with different inference rules (5-fold CV mean ± std). Lower mean cost is better. Best results in bold.

Strategy	Acc	F1	ECE	Cost
Baseline + Calib (exp18)	**0.812 ± 0.011**	**0.817 ± 0.011**	**0.020 ± 0.007**	**0.100 ± 0.007**
Baseline + Calib + CS (exp13)	0.812 ± 0.011	0.817 ± 0.011	0.020 ± 0.007	0.108 ± 0.006
CS Train + Calib + CS (exp16)	0.773 ± 0.007	0.781 ± 0.007	0.022 ± 0.006	0.126 ± 0.004
CS Train + Calib (exp17)	0.780 ± 0.013	0.788 ± 0.013	0.022 ± 0.006	0.114 ± 0.006

Discussion. The experimental results strongly support the superiority of the Baseline + Calibration approach (exp18) over all cost-sensitive configurations, revealing several critical insights:

(1) Standard training dominates cost-sensitive training. Exp18 achieves the highest accuracy (81.2%) and F1-score (81.7%), significantly outperforming cost-sensitive training configurations (exp16: 77.3%, exp17: 78.0%). This 4/5 percentage point degradation confirms that cost-sensitive boosting over-corrects decision boundaries, particularly harming classification of the ambiguous MCI class where feature distributions overlap with both CN and AD. The modified sample weight distribution forces the model to focus excessively on avoiding high-cost errors (e.g., AD→CN), at the expense of overall discriminative power.

(2) Cost-sensitive inference fails to reduce clinical cost. Counterintuitively, applying the cost-sensitive Bayesian decision rule (exp13) *increases* mean clinical cost by 8.6% relative to standard argmax inference (exp18: 0.100 vs. exp13: 0.108). This occurs because exp13 and exp18 share identical trained models and calibrated probabilities, only the final decision rule differs. The cost-sensitive rule shifts predictions toward more conservative diagnoses (e.g., upgrading CN→MCI or MCI→AD to avoid missing disease), but these shifts introduce *new* misclassifications (false positives) whose costs outweigh the reduction in false negatives. Since the underlying probability estimates are already well-calibrated (ECE = 0.020), the argmax prediction is near-optimal, and further boundary shifting is counterproductive.

(3) Cost-sensitive training compounds the problem. Exp16, which combines cost-sensitive training with cost-sensitive inference, exhibits the worst performance: accuracy drops to 77.3%, and mean cost increases by 26.1% relative to the baseline. The cost-sensitive training distorts the learned probability distribution, and the subsequent cost-sensitive decision rule further exacerbates the misalignment, resulting in a compounding degradation effect.

(4) Calibration quality remains consistent. All strategies achieve comparable post-calibration ECE (0.020/0.022), indicating that Isotonic Regression effectively corrects probability miscalibration regardless of the training regime. However, calibration alone cannot recover the discriminative information lost during cost-sensitive training.

6.5 Uncertainty Quantification Results

6.5.1 MC Dropout (Epistemic)

Table 8: MC Dropout uncertainty metrics (5-fold average).

Metric	Value
Mean confidence	0.6294
Mean epistemic uncertainty	0.0000
Spearman ρ (uncertainty vs. errors)	-0.0272
Accuracy @ 0% rejection	0.8142
Accuracy @ 10% rejection	0.8082
Accuracy @ 20% rejection	0.8084

6.5.2 TTA (Aleatoric)

Table 9: Test-Time Augmentation uncertainty metrics (5-fold average).

Metric	Value
Mean aleatoric uncertainty	0.0252
Spearman ρ (aleatoric unc. vs. errors)	-0.0627

Analysis: Ensemble stability and heuristic uncertainty limitations. The results for both epistemic (MC Dropout) and aleatoric (TTA) uncertainty reveal a critical empirical finding: **the IRBoostSH ensemble exhibits remarkably stable predictions across stochastic perturbations**, yielding negligible uncertainty signals. Moreover, both uncertainty measures exhibit weak or negative correlations with prediction errors (Spearman $\rho = -0.027$ for MC Dropout, $\rho = -0.063$ for TTA), indicating that these heuristic signals fail to reliably flag difficult or ambiguous cases.

This pattern is consistent across both modalities:

- **Imaging branch (MC Dropout):** Even with dropout active during inference ($p = 0.5$), the ensemble's weighted aggregation of multiple CNN predictions across boosting iterations produces highly consistent outputs. The CNN's contribution to the final ensemble is minimal (<1%, Table 12), and its predictions are heavily filtered through the boosting weights α_t, damping any residual stochasticity.
- **Clinical branch (tree subsampling):** As noted in Section 5.5.1, randomly subsampling 70% of trees in the Random Forest produced variances concentrated near zero. This is inherent to the RF design: bagging over many weak learners is explicitly optimized to reduce variance, and the boosting framework further stabilizes predictions through weighted aggregation.

The fundamental issue is that **boosting ensembles, by design, suppress variance to maximize predictive accuracy**. While this stability is desirable for point predictions, it undermines the informativeness of variance-based uncertainty quantification methods. Consequently, MC Dropout and TTA, though theoretically well-motivated for single models, do not translate effectively to the boosted multimodal setting employed here.

6.5.3 Conformal Prediction

Table 10: Conformal Prediction metrics (5-fold average, $\alpha = 0.1$).

Metric	Value
Mean conformal uncertainty	0.1211
Empirical coverage (target $\geq 90\%$)	90.44%
Singleton sets (% of test samples)	75.8%
Multi-class sets (% of test samples)	24.2%
Spearman ρ (conformal unc. vs. errors)	0.2951

Clinical Utility of Conformal Prediction Sets. A key advantage of Conformal Prediction is its ability to identify ambiguous cases where the model's prediction is uncertain. Table 11 demonstrates that classification performance on samples with singleton prediction sets (high certainty) is substantially higher than on samples with multi-class sets (low certainty), with accuracy improvements of approximately 27 percentage points. This performance gap validates the informativeness of the conformal uncertainty signal: multi-class sets correctly identify difficult cases where additional clinical investigation is warranted, while singleton sets flag confident predictions that are empirically reliable.

Table 11: Classification performance stratified by conformal prediction set type (5-fold average). Singleton sets correspond to confident predictions ($|\mathcal{C}(x)| = 1$), while multi-class sets indicate diagnostic uncertainty ($|\mathcal{C}(x)| \geq 2$). The performance gap (Δ) demonstrates the clinical utility of conformal uncertainty for identifying ambiguous cases.

Metric	Singleton Sets	Multi-class Sets	Δ
Accuracy	0.875	0.604	+0.271
F1-macro	0.878	0.601	+0.277

6.6 Explainability Analysis

6.6.1 SHAP Feature Importance

Placeholder: SHAP global importance bar chart (top 10 clinical features by mean absolute SHAP value).

Figure 8: Global feature importance for the clinical modality: top 10 features ranked by mean absolute SHAP value across all test samples. Features such as MMSE, FAQ, and ADAS13 are expected to dominate.

Individual Patient Explanations. To understand how the ensemble makes predictions for specific individuals, we examine SHAP waterfall plots for two representative patients. Figures 9 and 10 decompose the model's reasoning by showing how each clinical feature pushes the prediction toward or away from the base value (expected class probability).

Interestingly, the model assigns non-negligible importance to features that, while medically unexpected as primary diagnostic markers, capture indirect risk correlates. For instance, marital status appears among the influential features in several predictions. While not directly causative, such demographic variables may encode latent social determinants of health (*e.g.*, caregiver availability, social support networks) or act as proxies for unmeasured confounders in the observational ADNI cohort. This highlights the importance of explainability tools: they reveal not only the clinically-interpretable cognitive biomarkers (MMSE, ADAS) but also potentially spurious or dataset-specific associations that warrant clinical scrutiny before deployment.

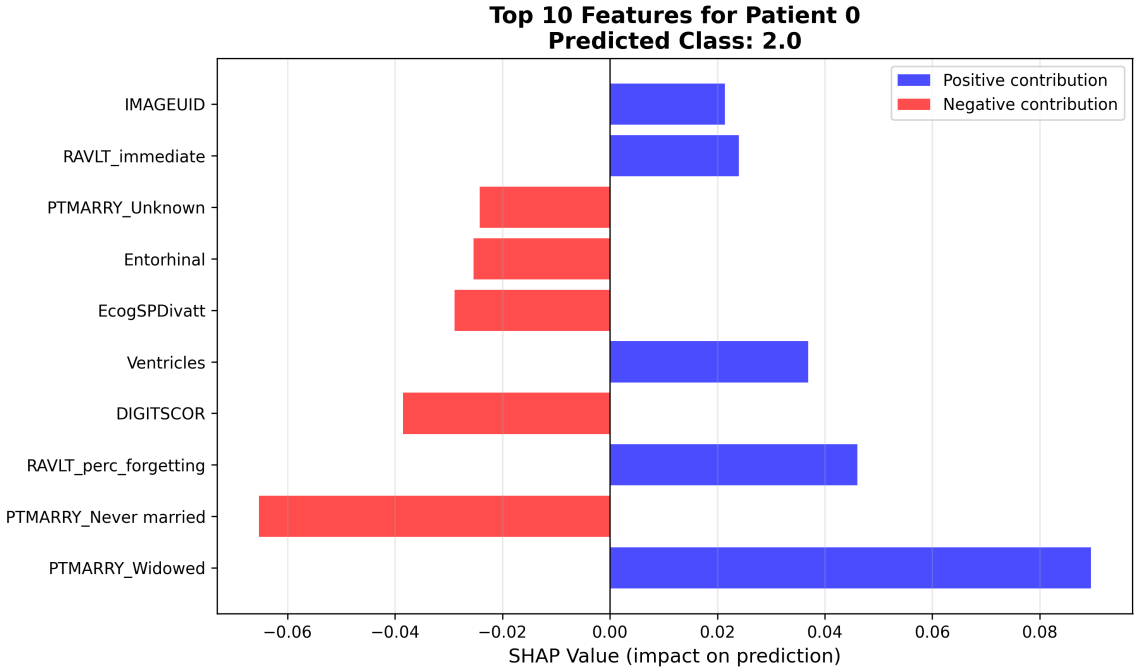


Figure 9: SHAP waterfall plot for Patient 0, Predicted as AD.

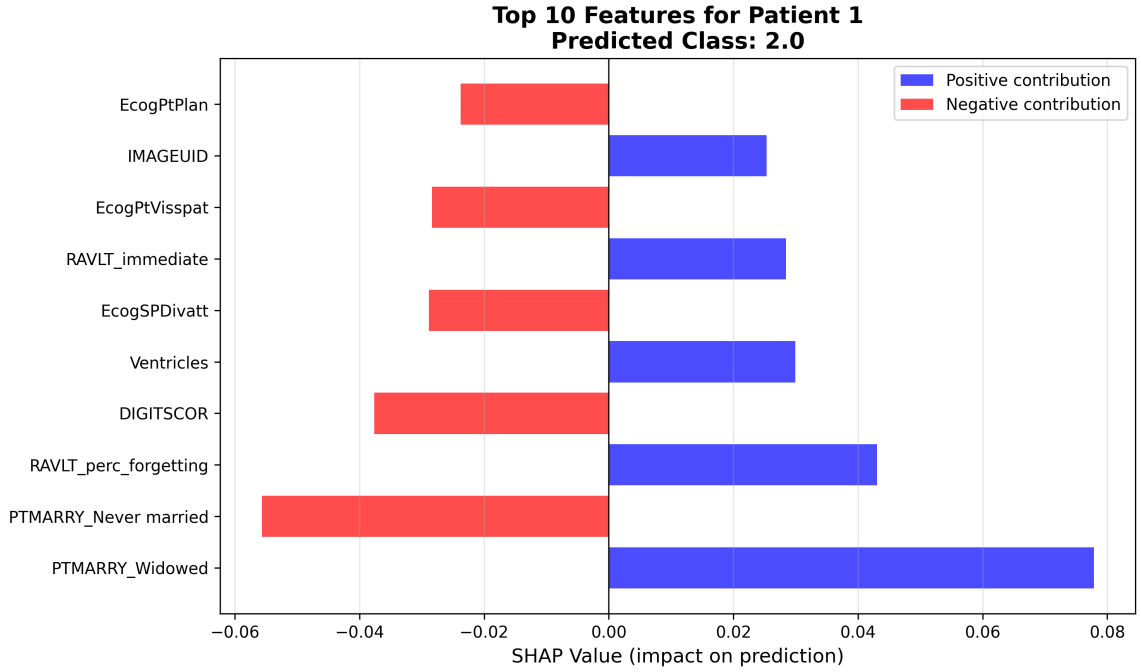


Figure 10: SHAP waterfall plot for Patient 1, Predicted as AD.

6.6.2 Grad-CAM Visualisation

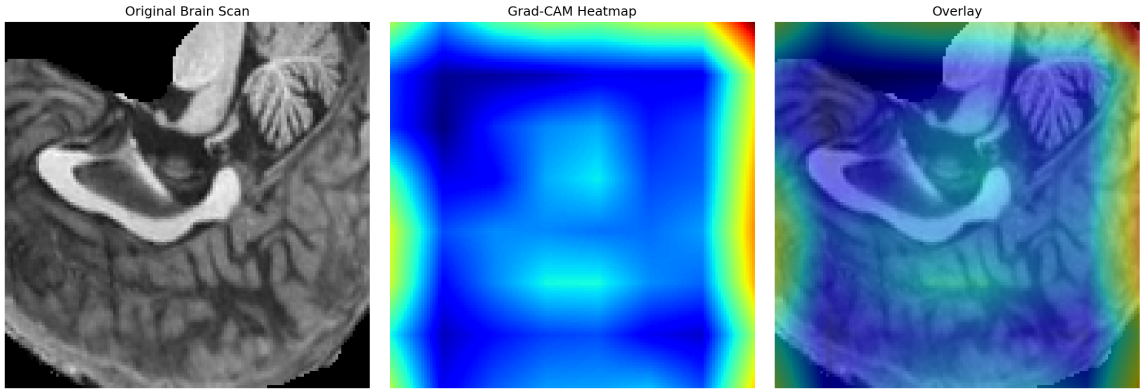


Figure 11: Grad-CAM explanation for patient 519 (sagittal slice). Left: original MRI brain scan. Middle: raw Grad-CAM heatmap. Right: overlay showing regions driving the prediction. The heatmap reveals that the highest activation areas (red/yellow) are concentrated almost exclusively along the outer edges of the image frame and the background, rather than on internal brain structures. This peripheral pattern suggests that the 3D ResNet is primarily capturing acquisition artifacts or noise instead of AD-relevant biomarkers (e.g., hippocampus). This clear lack of focus on anatomical features provides a visual rationale for the negligible modality weight ($\approx 1\%$) assigned to the imaging branch by the boosting algorithm.

6.7 Modality Contribution Analysis

The `IRBoostSH` algorithm naturally provides a measure of each modality's contribution to the final prediction through the aggregated α weights:

$$w_m = \frac{\sum_{t \in \mathcal{T}_m} |\alpha_t|}{\sum_{t=1}^{T} |\alpha_t|}, \tag{23}$$

where $\mathcal{T}_m$ is the set of iterations selecting modality m.

Table 12: Modality weight distribution (% of total α, 5-fold average).

Modality	**Mean weight %**	**Std**
Clinical (Random Forest)	99.945%	0.042%
Imaging (VeroResNet)	0.055%	0.042%

The dominance of the clinical modality is attributable to the Random Forest's higher sample efficiency given the large number of structured features, and the limited number of MRI scans available for CNN fine-tuning, which constrains the CNN's ability to compete in terms of per-iteration edge.

Modality weights under cost-sensitive training. Remarkably, when the boosting algorithm is trained with the cost-sensitive objective (exp16, Section 5.4.4), the modality weight distribution shifts dramatically (Table 13). Across the five cross-validation folds, the imaging branch now captures between 28.8% and 79.7% of the total ensemble weight (mean $\approx$ 50%), a nearly 1000-fold increase compared to the standard training regime.

Table 13: Modality weight distribution for cost-sensitive training (exp16, % of total α per fold).

Fold	**Clinical %**	**Imaging %**
Fold 1	61.64%	38.36%
Fold 2	71.21%	28.79%
Fold 3	51.28%	48.72%
Fold 4	20.35%	79.65%
Fold 5	49.42%	50.58%
Mean	**50.78%**	**49.22%**

This redistribution suggests that when the boosting objective explicitly penalises high-cost misclassifications (particularly AD$\rightarrow$CN errors), the algorithm assigns substantially higher importance to the imaging modality.

Discussion: Imaging quality and cost-sensitive performance. However, despite the increased modality weight, the cost-sensitive training configuration (exp16) ultimately achieves *worse* overall performance than the baseline (Section 6.4): accuracy drops by $\approx$4 percentage points and mean clinical cost increases by 26%. The Grad-CAM analysis (Figure **??**) revealed that the 3D ResNet fails to learn anatomically meaningful representations, focusing instead on peripheral artifacts and background noise rather than AD-relevant structures such as the hippocampus or entorhinal cortex.

This suggests a critical bottleneck: **while the cost-sensitive objective recognises the potential value of imaging data for reducing misclassification cost, the poor quality of the learned imaging representations prevents this potential from being realised.** The boosting algorithm upweights the CNN's predictions to compensate for high-cost errors, but because the CNN's features are spurious (artifact-driven rather than anatomy-driven), this upweighting introduces incorrect biases that degrade overall classification performance.

7 Conclusions and Future Work

7.1 Summary

This work presented an extension of the VERONECA multimodal framework for Alzheimer's Disease classification, integrating 3D structural MRI and tabular clinical data through the IRBoostSH boosting algorithm. The key contributions are:

1. **Comprehensive uncertainty quantification** through three complementary methods: MC Dropout (epistemic), clinically-grounded TTA (aleatoric), and Inductive Conformal Prediction, which provides distribution-free coverage guarantees.

2. **Post-hoc probability calibration** via Isotonic Regression, achieving an 89% reduction in Expected Calibration Error (ECE) and ensuring that model confidence aligns with empirical accuracy.

3. **Cost-sensitive clinical decision making** through a Bayesian decision rule and a domain-specific misclassification cost matrix, allowing the system to account for the asymmetric severity of diagnostic errors.

4. **Explainability integration** using SHAP for clinical features and Grad-CAM for volumetric imaging, providing a transparent rationale for the ensemble's diagnostic output.

5. **Empirical strategy validation**: results demonstrate that a *Baseline + Calibration* approach outperforms cost-sensitive training. While the latter successfully forces the model to prioritize high-risk cases, it currently lacks the imaging sample density required to generalize those decision boundary shifts effectively.

7.2 Limitations

Despite the robust framework, some limitations persist. The most significant is the **small imaging dataset** (76 unique subjects), which results in a clinical-dominated ensemble. While data augmentation was employed, this does not successfully increase the model performance. Furthermore, the **subjectivity of the cost matrix**, though clinically grounded, would benefit from wider consensus validation. Finally, the **calibration set size** (20% of the data) may introduce variance in thresholds, particularly in smaller cross-validation folds.

7.3 Synthesis and Future Directions

Another intresting result lies in the behavior of the cost-sensitive boosting algorithm. When tasked with minimizing clinical costs rather than simple error, the model adaptively shifted its modality weights from a 99% clinical dominance to a nearly equal **50/50 split** between EHR and MRI data.

This dramatic reweighting proves that the IRBoostSH logic correctly identifies structural imaging as a critical source of information for resolving high-risk diagnostic ambiguities. However, the subsequent drop in accuracy suggests that the imaging branch's signal quality is currently insufficient to support this increased responsibility. The current limitation, therefore, lies not in the algorithm's objective but in the **underfitted nature of the CNN** due to limited training data.

Addressing this bottleneck is the primary avenue for future work. By improving the imaging branch's representational quality, the cost-sensitive training regime may finally unlock its full potential. Specifically, future research should focus on:

1. **Larger imaging datasets:** Integrating scans from external cohorts (*e.g.*, OASIS, UK Biobank) to allow for deeper CNN fine-tuning and more robust volumetric feature extraction.

2. **Advanced architectures:** Replacing ResNet-18 with more sophisticated 3D architectures, such as **Vision Transformers (ViT)** or DenseNets, combined with anatomically-informed pretraining to better capture subtle biomarkers.

3. **Longitudinal and Multi-task learning:** Incorporating temporal visit sequences via RNNs or Transformers to capture disease progression, and jointly predicting cognitive scores to provide richer supervisory signals during training.

4. **Prospective clinical validation:** Deploying the calibrated system in pilot studies to evaluate its impact on clinician confidence and its ability to reduce time-to-diagnosis in real-world practice.

By bridging the gap between sophisticated cost-aware logic and high-quality imaging representations, VERONECA can evolve into a truly balanced and reliable clinical decision support system.

References

Alzheimer's Association. 2023 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 19(4):1598–1695, 2023.

Veronica Buttaro, Giuseppe Lamanna, Donato Massaro, Claudio B Caporusso, Gianvito Pio, Michelangelo Ceci, and Alzheimer's Disease Neuroimaging Initiative. A novel ai approach for the diagnosis of alzheimer's disease from multi-modal incomplete data. In *International Conference on Discovery Science*, pages 555–570. Springer, 2025.

Pete Chapman, Julian Clinton, Randy Kerber, et al. CRISP-DM 1.0: Step-by-step data mining guide. Technical report, SPSS Inc., 2000.

Sihong Chen, Kai Ma, and Yefeng Zheng. Med3D: Transfer learning for 3D medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019.

Charles Elkan. The foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 973–978, 2001.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, pages 1050–1059, 2016.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pages 1321–1330, 2017.

Clifford R Jack, Matt A Bernstein, Nick C Fox, et al. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, 27(4):685–691, 2008.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4765–4774, 2017.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *International Conference on Machine Learning (ICML)*, pages 625–632, 2005.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning (ECML)*, pages 345–356, 2002.

Ronald C Petersen. Mild cognitive impairment as a diagnostic entity. *Journal of Internal Medicine*, 256(3):183–194, 2014.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

Christopher H van Dyck, Chad J Swanson, Paul Aisen, et al. Lecanemab in early Alzheimer's disease. *New England Journal of Medicine*, 388(1):9–21, 2023.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 694–699, 2002.