Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Interpretable multimodal deep learning for diagnosis and symptom-onset prediction in genetic frontotemporal dementia

Wenjie Kang

w.kang@erasmusmc.nl

Erasmus MC

Tine Swartenbroekx

Erasmus MC

Myrthe F. van Haaften

Erasmus MC

Lize C. Jiskoot

Erasmus MC

Stefan Klein

Erasmus MC

Harro Seelaar

Erasmus MC

Esther E. Bron

Erasmus MC

Research Article

Keywords:

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11483481/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Interpretable multimodal deep learning for diagnosis and symptom-onset prediction in genetic frontotemporal dementia

Wenjie Kang[a,*], Tine Swartenbroekx[b], Myrthe F. van Haaften[a], Lize C. Jiskoot[b], Stefan Klein[a], Harro Seelaar[b], Esther E. Bron[a]

[a]*Biomedical Imaging Group Rotterdam, Department of Radiology & Nuclear Medicine, Erasmus MC, Rotterdam, The Netherlands*
[b]*Department of Neurology, Erasmus MC, Rotterdam, The Netherlands*

Abstract

Background: Genetic frontotemporal dementia (FTD) is characterized by substantial clinical and neuroanatomical heterogeneity, making prediction of symptom onset challenging. We evaluated whether deep learning-derived MRI biomarkers could be combined with non-imaging biomarkers in an interpretable framework for diagnosis and phenoconversion prediction in genetic FTD.

Methods: We evaluated *Multi-ICNN*, a multimodal extension of the Interpretable Convolutional Neural Network, using data from the FTD Risk Cohort. Imaging biomarkers derived from T1-weighted MRI were combined with demographic, genetic, and neuropsychological biomarkers in an additive prediction model. We evaluated classification of symptomatic mutation carriers versus genetically confirmed non-carriers and phenoconversion prediction within 2- and 5-year time horizons. Biomarker contributions were

*Corresponding author: w.kang@erasmusmc.nl

examined at the group and participant levels.

Results: For diagnosis, *Multi-ICNN* achieved the highest mean area under the receiver operating characteristic curve (AUC) of 0.953, numerically outperforming the single-modality models. For phenoconversion prediction, the deep learning-based models outperformed the conventional volumetric and non-imaging models. The imaging-only *ICNN* achieved the highest mean AUCs of 0.798 and 0.780 for the 2- and 5-year horizons, respectively. Adding non-imaging biomarkers did not further improve *ICNN* performance. Neuropsychological measures contributed most strongly to diagnosis, whereas imaging-derived biomarkers dominated phenoconversion prediction, particularly biomarkers involving frontal and temporal regions. The additive model also provided participant-specific biomarker contributions.

Conclusions: Interpretable deep learning-derived MRI biomarkers showed promise for phenoconversion prediction in genetic FTD, while the value of multimodal integration depended on the clinical task. The main limitations were the modest sample size and lack of external validation, reflecting the rarity of genetic FTD and the limited availability of comparable longitudinal cohorts.

Keywords: Frontotemporal dementia, Magnetic resonance imaging, Convolutional neural network, Explainable artificial intelligence, Explainable Boosting Machine

1. Introduction

Frontotemporal dementia (FTD) is a heterogeneous neurodegenerative disorder characterized by progressive changes in behaviour, language, and

sometimes motor function. These clinical manifestations are associated with selective degeneration of the frontal and/or temporal lobes [1]. FTD accounts for approximately 5–10% of all dementia cases and up to 20% of cases of early-onset dementia [2, 3]. Around 30% of patients with FTD have a strong family history, highlighting the substantial contribution of genetic factors to the disease [3]. Most of the known heritability is attributable to autosomal dominant neuropathogenic variants in three major genes: chromosome 9 open reading frame 72 (*C9orf72*), progranulin (*GRN*), and microtubule-associated protein tau (*MAPT*) [4, 3].

Early and accurate diagnosis can help reduce healthcare costs and improve patient outcomes by enabling timely access to supportive and symptomatic therapies [5]. In genetic FTD, pathogenic variant carriers can be identified through presymptomatic genetic testing, often years before clinical symptoms emerge, providing an opportunity for prospective monitoring during the presymptomatic stage. Although the exact timing of symptom onset remains difficult to predict because of substantial inter-individual variability, identifying individuals at increased risk of so-called 'phenoconversion' is clinically relevant because it may support targeted monitoring and timely therapeutic intervention. In particular, identifying individuals at increased risk of near-term phenoconversion may facilitate patient stratification and recruitment for preventive clinical trials. Longitudinal studies in genetic FTD cohorts have identified cognitive [6, 7], imaging [8, 9, 10], and fluid biomarkers [11, 12] associated with presymptomatic disease stages and subsequent phenoconversion.

Given the substantial heterogeneity of FTD in clinical presentation, ge-

netic etiology, and disease progression [2], individual biomarker modalities may not fully capture the underlying pathophysiological changes. Structural MRI can reveal region-specific patterns of neurodegeneration, demographic information provides context for individual risk stratification, genetic markers capture mutation status and mutation-specific disease patterns, and neuropsychological assessments offer sensitive measures of early cognitive and behavioral impairment [13, 14, 15]. Studies integrating multiple biomarker domains have further demonstrated their complementary value for characterizing presymptomatic disease progression and the temporal sequence of disease-related changes [16, 17]. Integrating these complementary modalities may therefore improve diagnostic and predictive performance and provide a more comprehensive characterization of disease progression in genetic FTD.

Machine learning methods have shown considerable potential for diagnostic classification and prediction in FTD. Recent studies have primarily focused on cross-sectional classification using structural MRI, particularly for differentiating FTD subtypes and distinguishing FTD from other neurodegenerative disorders [18, 19, 20]. However, modelling complex relationships among high-dimensional multimodal biomarkers remains challenging, particularly in relatively small FTD cohorts. Although deep learning approaches have demonstrated promising predictive performance, they often function as black boxes, providing limited insight into their decision-making processes [21]. Furthermore, phenoconversion prediction in presymptomatic mutation carriers remains relatively understudied. In particular, few studies have integrated multimodal biomarkers within an interpretable framework for this task, despite its clinical relevance to early intervention and preven-

tive clinical trials. Therefore, interpretable approaches that can effectively leverage high-dimensional multimodal data for phenoconversion prediction are needed.

To address these challenges, we previously introduced an end-to-end interpretable framework named the Interpretable Convolutional Neural Network (*ICNN*), which integrates a Convolutional Neural Network (CNN)-based deep feature extractor with a Generalized Additive Model (GAM)-based classifier [22]. The *ICNN* enables learning directly from high-dimensional neuroimaging data while preserving transparent feature-level interpretability at both the individual and cohort levels. In Alzheimer's disease cohorts, the framework achieved competitive predictive performance and strong generalizability compared with state-of-the-art black-box models while providing clinically meaningful explanations of model decisions [22].

In this study, we extend the previously developed *ICNN* framework to integrate imaging, demographic, genetic, and neuropsychological biomarkers for the diagnosis and prediction of symptom onset in genetic FTD. We evaluate the resulting multimodal interpretable framework using data from the FTD Risk Cohort [23]. Two classification tasks are considered: (1) distinguishing symptomatic mutation carriers from non-carriers, and (2) predicting phenoconversion among presymptomatic mutation carriers. We assess its predictive performance and examine the contribution of each biomarker modality to the diagnosis and prediction of genetic FTD.

2. Methods

2.1. Study population

This study included participants from the Frontotemporal Dementia Risk Cohort (FTD-RisC) at the Erasmus MC University Medical Center (Rotterdam, the Netherlands) [23]. FTD-RisC follows individuals carrying a pathogenic variant in the *MAPT*, *GRN*, or *C9orf72* genes and their 50% at-risk family members through annual or bi-annual assessments. At each study visit, participants underwent a standardized assessment protocol that included a medical history review, neurological examination, neuropsychological testing, 3T MRI, and collection of blood and cerebrospinal fluid samples. Following each assessment, clinical status was determined during a multidisciplinary consensus meeting. Carriers were classified as phenoconverters if they met all of the following criteria: (i) evidence of progressive deterioration in behavior and/or language, (ii) functional decline defined as a CDR plus NACC FTLD score ≥ 0.5 that remained elevated at the subsequent follow-up assessment, and (iii) cognitive impairment based on neuropsychological evaluation [12]. Phenoconverters with more prominent motor deficits additionally had to meet: (i) mild motor symptoms consistent with ALS, and (ii) evidence of subsequent progression over time. Carriers who remained symptom-free until their last available follow-up assessment were classified as presymptomatic non-converters. The phenoconversion date was defined as the date of the clinical assessment at which these criteria were first satisfied.

Ethical approval for the study was granted by the Medical and Ethical Review Committee of the Erasmus MC University Medical Center (MEC-2009-409). All participants provided written informed consent prior to en-

rollment.

2.2. Imaging data

Longitudinal MRI data were acquired using three 3T scanner configurations: a Philips Achieva scanner with an 8-channel head coil (software release R4), a Philips Achieva scanner with a 32-channel head coil (software release R5), and a GE Healthcare MR750 scanner with a 48-channel head coil. Despite minor differences across scanner configurations and study periods, all scans were acquired using closely comparable whole-brain near-isotropic three-dimensional T1-weighted (T1w) protocols and were processed using the same preprocessing pipeline.

From the T1w scans, regional volumetric biomarkers and gray matter (GM) density maps were derived using the FreeSurfer–voxel-based morphometry (VBM) pipeline. As a first step, FreeSurfer version 6.0 [24] was used to segment supratentorial GM from the T1w MRI scans [25]. Regional cortical and subcortical volumes were derived using the Desikan–Killiany atlas [26] and normalized by dividing each regional volume by the corresponding intracranial volume (ICV). These ICV-normalized regional volumes were used as volumetric biomarkers in the subsequent model training and analyses.

As a second step, GM density maps were computed using an optimized VBM protocol [27, 28]. The GM maps were first nonlinearly registered to the Montreal Neurological Institute (MNI) GM probability template (ICBM152 Nonlinear Atlas, 2009 version) at a voxel resolution of $1 \times 1 \times 1$ mm. Spatial modulation was subsequently applied to account for local volume changes introduced by the nonlinear registration by multiplying the voxel values by the Jacobian determinants of the nonlinear transformations. The resulting

modulated GM density maps had a matrix size of 196 × 232 × 188. No spatial smoothing was applied because smoothing operations can, in principle, be learned by the convolutional filters. Quality control consisted of excluding scans with more than 5% outlier voxels, followed by manual inspection. The complete MRI preprocessing pipeline is publicly available at `https://gitlab.com/radiology/neuro/pipelines/freesurfer-fsl-vbm-structural-mri-pipeline`.

2.3. Non-imaging data

All non-imaging biomarkers were represented as tabular variables. They comprised demographic, genetic, and neuropsychological variables. Demographic variables included age, biological sex (female or male), and education level. Two levels of genetic information were considered: the gene affected in the family and the specific pathogenic variant identified in that family. The former indicated the gene in which the familial pathogenic variant was located (*MAPT*, *GRN*, *C9orf72*, or *TARDBP*), whereas the latter represented the specific sequence-level alteration within that gene. For non-carriers, the affected-gene variable referred to the gene implicated in their family, although they did not carry the corresponding familial pathogenic variant. Additional information on the coding and observed categories of the genetic variables is provided in Supplementary Table S1.

The neuropsychological assessment comprised six measures covering processing speed, executive functioning, language, verbal learning and memory, and visuospatial functioning (Table 1). Trail Making Test part A (TMT-A) was used to assess processing speed and visual attention, whereas part B (TMT-B) additionally assessed cognitive flexibility and set-shifting. Performance was quantified using completion time. Semantic fluency was mea-

sured as the number of animal names generated within one minute. Verbal learning and memory were assessed using the 15-Word Test (15WT), confrontational word retrieval using the 60-item Boston Naming Test (BNT), and visuospatial and executive functioning using the 14-point Clock Drawing Test (CDT). Except for the completion-time measures of TMT-A and TMT-B, higher scores indicated better performance. These measures capture cognitive domains known to be affected during the presymptomatic and symptomatic stages of genetic FTD [6, 7].

Table 1: Neuropsychological measures included in the multimodal model.

Measure	**Primary cognitive domain**	**Score included**
Trail Making Test A (TMT-A)	Processing speed and visual attention	Time required to complete part A
Trail Making Test B (TMT-B)	Executive functioning and cognitive flexibility	Time required to complete part B
Semantic animal fluency	Semantic verbal fluency	Number of animal names generated within one minute
15-Word Test (15WT)	Verbal learning and memory	Total number of words recalled across five learning trials (range: 0–75)
Boston Naming Test (BNT)	Confrontational word retrieval	Number of correctly named items (range: 0–60)
Clock Drawing Test (CDT)	Visuospatial and executive functioning	Total score (range: 0–14)

Note: Lower completion times indicate better performance for TMT-A and TMT-B; higher scores indicate better performance for all other measures.

2.4. Analysis cohorts and study tasks

The FTD Risk Cohort comprised 239 participants contributing 967 time-points with available T1w MRI scans. Following preprocessing and quality control as described above, 912 scans from 234 participants were available for cohort construction. Of these, 216 participants contributing 867 scans could be assigned to one of the three predefined participant groups, whereas 18 participants contributing 45 scans did not meet the criteria for group assignment and were excluded from the analysis cohorts. The derivation of the analysis cohorts is shown in Fig. 1, and the demographic and clinical characteristics of the participant groups are summarized in Table 2.

We investigated two clinical applications: (1) cross-sectional classification of symptomatic mutation carriers versus non-carriers (FTD vs. NC) and (2) prediction of phenoconversion among presymptomatic mutation carriers.

For the FTD vs. NC task, the baseline scan of each symptomatic carrier was included. Because the full NC group was significantly younger than the symptomatic carrier group, we constructed an age-matched NC group for this task. NC participants whose baseline age was above 45 years were eligible, and the most recent scan from each eligible participant was selected. The resulting age-matched NC group comprised 54 participants. The age distributions at the selected scans did not differ significantly between the symptomatic carrier and age-matched NC groups ($p > 0.05$).

For phenoconversion prediction, we formulated the problem as fixed-horizon binary classification because the primary clinical objective was to identify individuals at increased risk of conversion within clinically relevant time windows. Moreover, phenoconversion dates were determined at discrete

clinical assessments and were therefore subject to uncertainty related to the intervals between visits. Prediction was evaluated at 2- and 5-year time horizons. For each time horizon, scans from phenoconverters were labelled positive when they were acquired within the corresponding period before the phenoconversion date. Scans from non-converters were labelled negative when the participant remained presymptomatic for at least the corresponding period after the scan. Thus, the 2- and 5-year tasks estimated whether a presymptomatic mutation carrier would undergo phenoconversion within 2 or 5 years after a given scan, respectively.

Table 2: Baseline demographic characteristics of the study population. The age-matched non-carriers were used for the FTD vs. NC classification task. Statistics for years of education were calculated using available data only. For mutation carriers, genetic information refers to the pathogenic gene.

	Symptomatic ($n = 28$)	Presymptomatic ($n = 102$)	Non-carrier ($n = 86$)	Age-matched non-carriers ($n = 54$)
Age (years; mean ± SD)	62.0 ± 10.3	44.7 ± 11.9	46.1 ± 11.8	59.7 ± 8.2
Female, n (%)	13 (46.4)	61 (59.8)	48 (55.8)	29 (53.7)
MMSE (mean ± SD)	25.0 ± 5.7	29.2 ± 1.2	29.1 ± 1.1	29.2 ± 1.2
Years of education (mean ± SD)	13.1 ± 2.8	14.0 ± 2.7	14.1 ± 2.9	14.0 ± 2.9
Affected gene in the family, n	*MAPT*: 5 *GRN*: 8 *C9orf72*: 13 *TARDBP*: 2	*MAPT*: 20 *GRN*: 42 *C9orf72*: 38 *TARDBP*: 2	*MAPT*: 16 *GRN*: 42 *C9orf72*: 28 *TARDBP*: 0	*MAPT*: 9 *GRN*: 33 *C9orf72*: 12 *TARDBP*: 0

2.5. ICNN

We adopted the Interpretable Convolutional Neural Network (*ICNN*) proposed in our previous work [22]. The *ICNN* is designed for imaging data and

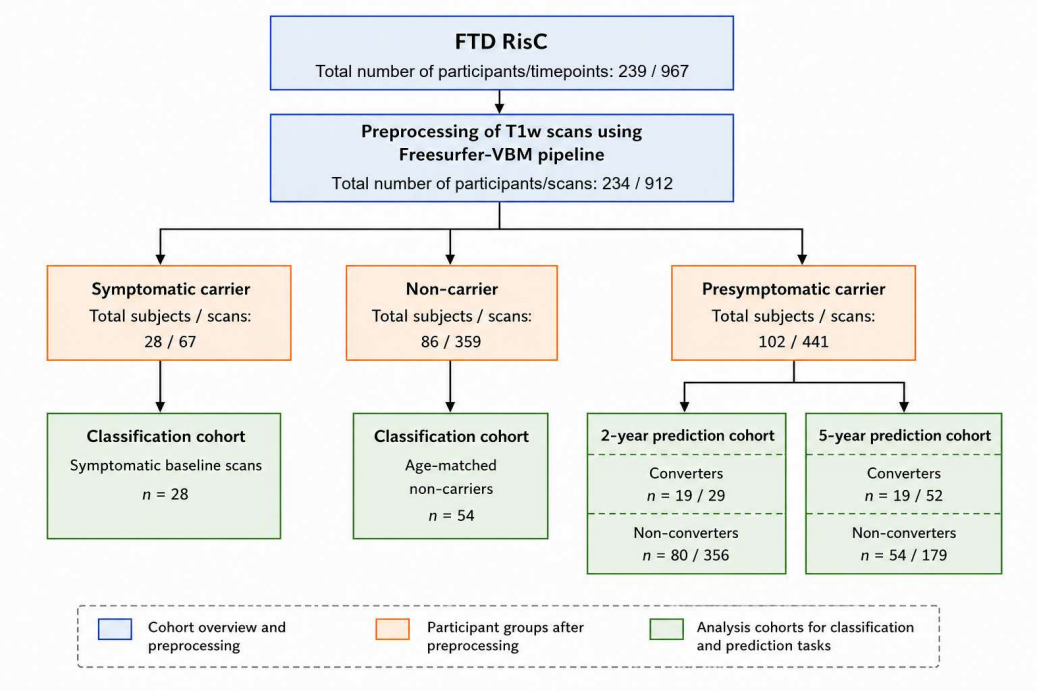


Figure 1: Overview of the participant selection pipeline and analysis cohorts. Numbers in each box indicate participants and timepoints or scans, as specified. Symptomatic mutation carriers and age-matched NC participants comprised the FTD vs. NC classification cohort, whereas presymptomatic mutation carriers were grouped into phenoconverters and non-converters for the 2- and 5-year phenoconversion-prediction cohorts. Abbreviations: NC, non-carrier; T1w, T1-weighted; VBM, voxel-based morphometry.

combines multi-branch deep learning-based feature extraction with a generalized additive model (GAM)-based output layer. Each imaging branch produces a feature representation that is passed to the GAM, yielding additive branch-level contributions to the final model output.

The *ICNN* represents a flexible interpretable modeling framework rather than a fixed architecture. In principle, different deep learning models can be employed as imaging feature extractors, and various white-box models can be used as the final classifier, provided that feature-level interpretability is preserved.

In this study, we instantiated the *ICNN* using a multi-branch CNN-based feature extractor and an Explainable Boosting Machine (EBM) as the classifier. The EBM is a tree-based implementation of a GAM that learns additive feature-contribution functions using gradient-boosted decision trees, thereby providing intrinsic interpretability while maintaining competitive model performance [29]. We selected this combination of a multi-branch CNN-based feature extractor and an EBM classifier based on our previous studies, in which this configuration demonstrated strong model performance and robust interpretability in Alzheimer's disease cohorts [22, 30].

For binary classification, the EBM models the additive predictor as

$$\eta = \beta_0 + \sum_j f_j(x_j),$$

where η denotes the additive predictor, β_0 is the intercept, and $f_j(\cdot)$ represents the learned shape function for feature x_j. The estimated probability of the positive class is obtained using the logistic inverse link function:

$$p = \sigma(\eta) = \frac{1}{1 + \exp(-\eta)}.$$

Consequently, each input feature contributes to the model output through its corresponding additive function, enabling probabilistic classification and transparent feature-level interpretation.

2.6. Multi-ICNN

The proposed *Multi-ICNN* extends the imaging-only *ICNN* by combining its imaging-derived *DL-biomarkers* with non-imaging biomarkers.

The training of *Multi-ICNN* follows a two-stage procedure. In the first stage, the *ICNN* is trained using imaging data only to learn region-level *DL-biomarkers*. In the second stage, the pretrained *ICNN* is used as a fixed feature extractor. The resulting *DL-biomarkers* are concatenated with tabular non-imaging features, including demographic variables, genetic information, and neuropsychological test scores, and jointly modeled using an EBM. The EBM applies the additive formulation described above to the combined imaging-derived and non-imaging features. Freezing the imaging feature extractor decouples imaging representation learning from multimodal classification, allowing the EBM to integrate the two feature types while preserving the additive interpretability of the final model. The overall workflow of *Multi-ICNN* is illustrated in Fig. 2.

2.7. Model training and validation

We evaluated five models: (1) *NonImage-EBM* (an EBM using non-imaging biomarkers only); (2) *Vol-EBM* (an EBM using regional volumetric biomarkers only); (3) *Multi-EBM* (an EBM combining regional volumetric and non-imaging biomarkers); (4) *ICNN* (an ICNN using GM density maps as input); and (5) *Multi-ICNN* as introduced in Section 2.6.

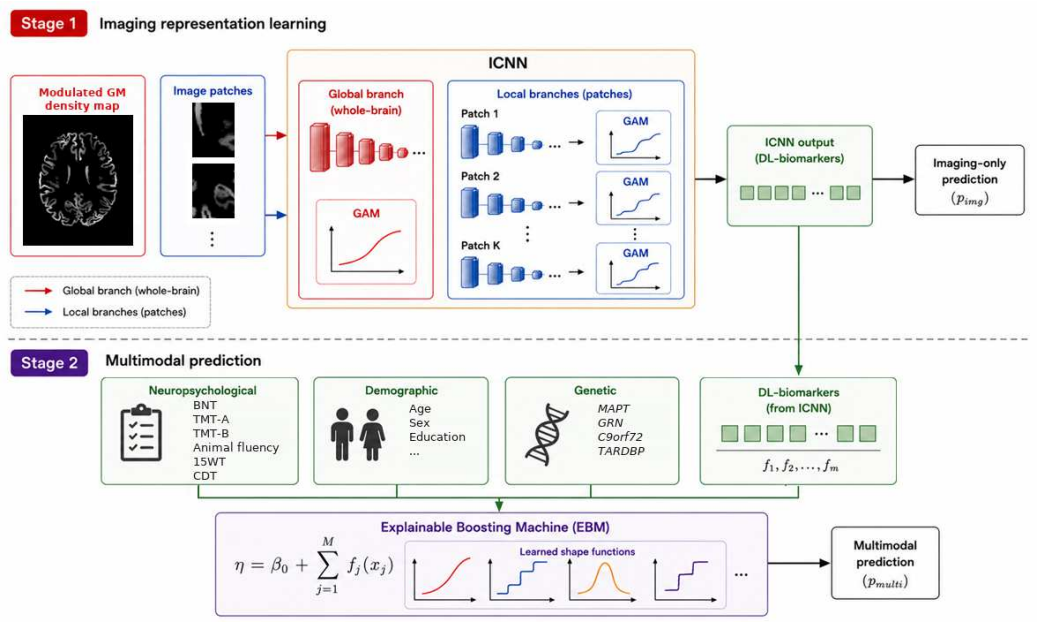


Figure 2: Overview of the proposed *Multi-ICNN* framework. Whole-brain imaging data (red), together with region-based image patches (blue), are processed by the *ICNN* to extract imaging-derived *DL-biomarkers*. Demographic features (age, sex, and education), genetic information (affected gene and, for the phenoconversion-prediction tasks only, the specific pathogenic variant), and neuropsychological test scores are represented as tabular non-imaging inputs. The *DL-biomarkers* and non-imaging features are concatenated and jointly modeled using an EBM, producing the final model output and feature-level contributions. p_{img} and p_{multi} denote the estimated probabilities of the positive class from the imaging-only and multimodal models, respectively.

For all models using non-imaging biomarkers, the affected-gene variable was included in all tasks, whereas the specific pathogenic-variant variable was included only in the phenoconversion-prediction tasks. The latter was excluded from the FTD vs. NC classification task because pathogenic-variant carrier status contributed to the definition of the outcome groups.

All models were trained and evaluated using 10 repeated stratified random train–validation–test splits with a ratio of 7:1:2. Data splitting was performed at the participant level. For the phenoconversion-prediction tasks, all longitudinal scans from the same participant were assigned to the same subset to prevent information leakage across the training, validation, and test sets. For the *ICNN*-based models, the validation set was used for early stopping and selection of the best-performing model checkpoint. Within each split, each neuropsychological score was residualized for age, sex, and years of education using a linear regression model fitted to the training set only and subsequently applied to the corresponding validation and test sets. The resulting residualized scores were used as inputs to all models incorporating non-imaging biomarkers.

To account for class imbalance, the *ICNN*-based models were trained using class-weighted cross-entropy loss, with class weights determined from the class frequencies in the training set. The initial learning rate and batch size were fixed a priori at 1×10^{-4} and 8, respectively, based on our previous ICNN study, which used the same model architecture and optimization procedure. The learning rate for the EBM components was set to 1×10^{-2}.

Model performance was evaluated using balanced accuracy (BA) [31] and the area under the receiver operating characteristic curve (AUC) [32]. Re-

ceiver operating characteristic (ROC) and precision–recall (PR) curves were additionally used to visualize model performance. Ninety-five percent confidence intervals (95% CI) for the mean performance metrics were estimated using the corrected resampled t-test across the 10 repeated splits, accounting for statistical dependence between splits [33]. Statistical differences between AUCs were assessed using DeLong's test [34].

2.8. Model outputs for interpretability

The models considered in this study provide interpretable outputs at the group, participant, and feature levels. All models employ a GAM, implemented using an EBM, as the output block. At the group level, feature importance was calculated separately for the training and test sets within each repeated train–validation–test split. For each feature, the absolute additive contributions were averaged across samples and subsequently normalized across all features. The resulting normalized feature importance scores were then averaged across the 10 repeated splits and summarized as the mean $\pm$ standard deviation. The 15 features with the highest mean importance in the training sets were displayed in the same order for the corresponding test-set results to facilitate direct comparison. At the participant level, the additive output block provides an individualized feature contribution profile for each model output, enabling inspection of participant-specific patterns of feature influence. This applies both to models using low-dimensional tabular inputs and to ICNN-based models, in which the GAM operates on imaging-derived *DL-biomarkers* extracted from high-dimensional inputs.

At the feature level, the EBM learns a shape function $f_j(x_j)$ for each input feature, representing the contribution of a given feature value to the

additive predictor. This applies to the volumetric and non-imaging biomarkers in *Multi-EBM*, as well as to the imaging-derived *DL-biomarkers* and non-imaging biomarkers in *Multi-ICNN*. Positive contribution values shift the model output toward the positive class, whereas negative values shift it toward the negative class. For each model, the 15 biomarkers with the highest mean group-level feature importance were selected for visualization. The feature values and corresponding additive contributions of the test samples were aggregated across the 10 repeated train–validation–test splits. Because the numerical scales of the *DL-biomarkers* in *Multi-ICNN* were not directly comparable across independently trained ICNN models, their values were converted to within-split percentiles using empirical cumulative distribution function ranks before aggregation. This transformation affected only the visualization and did not alter the model outputs or additive contribution values. For *Multi-EBM*, the original values of the volumetric and non-imaging biomarkers were retained.

3. Results

3.1. Model performance

Table 3 summarizes the model performance, and Fig. 3 shows the corresponding ROC and precision–recall (PR) curves. All performance estimates are reported as the mean [95% confidence interval (CI)] across the 10 repeated splits. For the FTD vs. NC task, all models achieved AUC values above 0.85. The proposed *Multi-ICNN* achieved the highest AUC of 0.953 [0.906, 0.999], followed by *ICNN* (0.926 [0.871, 0.981]) and *Multi-EBM* (0.911 [0.768, 1.000]). However, the improvements of *ICNN* and *Multi-ICNN* over

the other models were not statistically significant ($p > 0.05$).

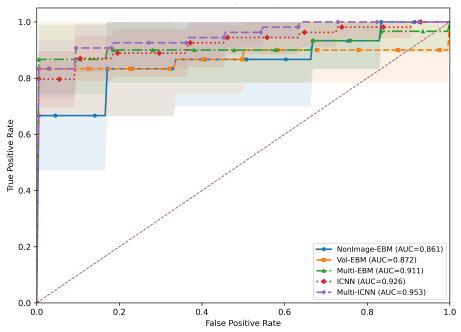
For the phenoconversion prediction tasks, the *NonImage-EBM* showed relatively low performance, with AUCs of 0.517 and 0.501 for the 2-year and 5-year phenoconversion-prediction tasks, respectively. The *Vol-EBM* achieved an AUC of 0.724 for the 2-year prediction task, but its performance decreased to 0.538 for the 5-year prediction task. In contrast, the deep learning–based *ICNN* and *Multi-ICNN* consistently achieved the highest performance across both prediction horizons. The *ICNN* achieved AUCs of 0.798 and 0.780 for the 2-year and 5-year prediction tasks, respectively, while the corresponding AUCs of the *Multi-ICNN* were 0.782 and 0.773. Both deep learning–based models significantly outperformed the other models ($p < 0.05$).

3.2. Model outputs for interpretability

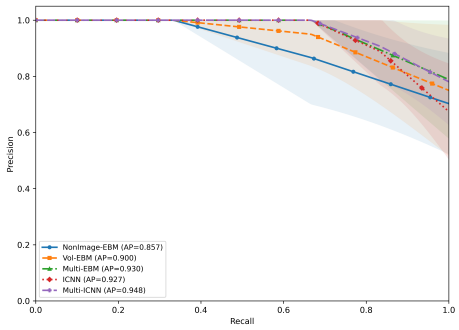
Fig. 4 presents the group-level feature importance estimated from the training sets, including *Multi-EBM* for the FTD vs. NC classification task and *Multi-ICNN* across all three tasks. The corresponding test-set results, displayed using the same features and feature order as in the training sets, are provided in Supplementary Fig. S1. Overall, the test-set results broadly reproduced the modality-level patterns observed in the training sets, although the relative importance of individual features varied across splits. Neuropsychological measures remained prominent for FTD vs. NC classification, whereas the imaging-derived *DL-biomarkers* selected for the phenoconversion-prediction tasks continued to show substantial contributions in the test sets. In the training sets, neuropsychological measures ranked among the most important features in both models for the FTD vs. NC classification task. In

Task	Model	BA (95% CI)	AUC (95% CI)
FTD vs. NC	NonImage-EBM	0.800 [0.706, 0.894]	0.861 [0.741, 0.981]
FTD vs. NC	Vol-EBM	0.817 [0.712, 0.921]	0.878 [0.742, 1.000]
FTD vs. NC	Multi-EBM	0.850 [0.695, 0.968]	0.911 [0.768, 1.000]
FTD vs. NC	ICNN	0.926 [0.882, 0.966]	0.926 [0.871, 0.981]
FTD vs. NC	Multi-ICNN	0.939 [0.893, 0.984]	0.953 [0.906, 0.999]
2-year prediction	NonImage-EBM	0.499 [0.497, 0.501]	0.517 [0.335, 0.755]
2-year prediction	Vol-EBM	0.511 [0.477, 0.544]	0.724 [0.506, 0.822]
2-year prediction	Multi-EBM	0.539 [0.469, 0.608]	0.676 [0.526, 0.826]
2-year prediction	ICNN	0.673 [0.567, 0.780]	0.798 [0.763, 0.833]
2-year prediction	Multi-ICNN	0.681 [0.573, 0.789]	0.782 [0.742, 0.822]
5-year prediction	NonImage-EBM	0.511 [0.444, 0.578]	0.501 [0.263, 0.649]
5-year prediction	Vol-EBM	0.541 [0.438, 0.642]	0.538 [0.327, 0.746]
5-year prediction	Multi-EBM	0.557 [0.476, 0.639]	0.520 [0.300, 0.740]
5-year prediction	ICNN	0.692 [0.602, 0.782]	0.780 [0.680, 0.879]
5-year prediction	Multi-ICNN	0.694 [0.601, 0.787]	0.773 [0.671, 0.847]

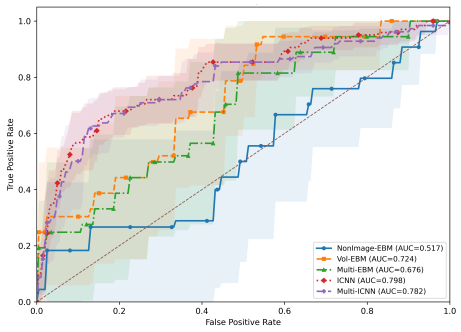
Table 3: Balanced accuracy (BA) and area under the receiver operating characteristic curve (AUC) for all models across tasks. Results are reported for the diagnosis task (FTD vs. NC) and the phenoconversion-prediction tasks evaluated at two time horizons (2- and 5-year prediction). Values represent the mean performance across 10 repeated train–validation–test splits. Ninety-five percent confidence intervals (95% CI) were estimated using the corrected resampled t-test. Abbreviations: FTD, frontotemporal dementia; NC, non-carrier; EBM, Explainable Boosting Machine; ICNN, Interpretable Convolutional Neural Network.



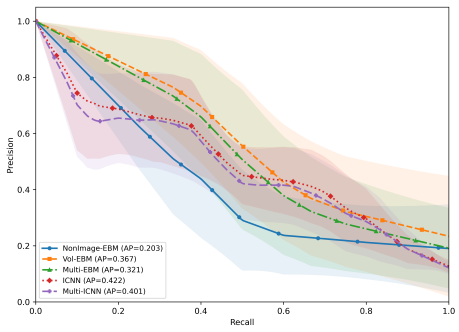
(a) FTD vs. NC (ROC)



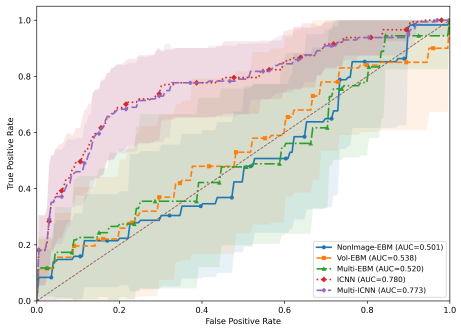
(b) FTD vs. NC (PR)



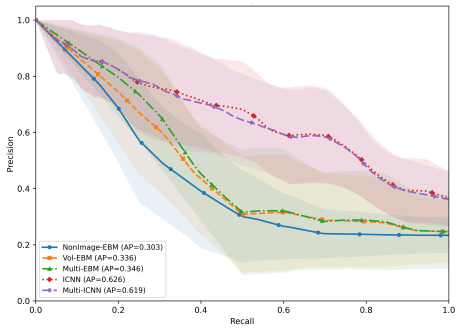
(c) 2-year prediction (ROC)



(d) 2-year prediction (PR)



(e) 5-year prediction (ROC)



(f) 5-year prediction (PR)

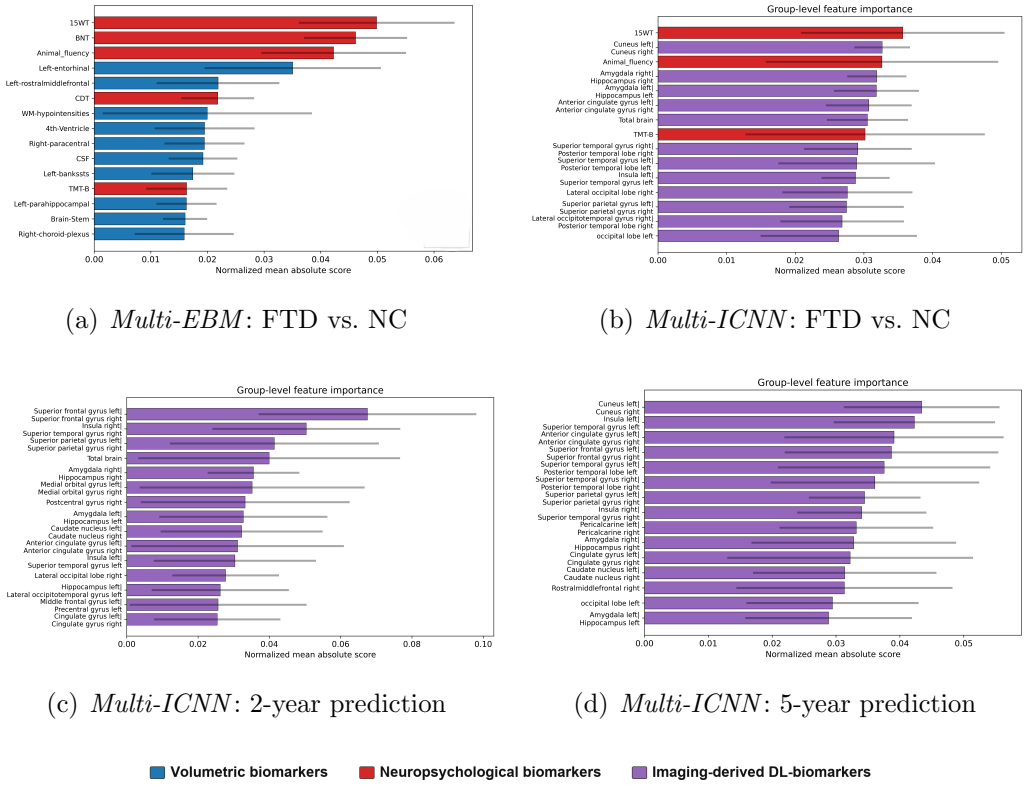
Figure 3: Receiver Operating Characteristic (ROC) and Precision–Recall (PR) curves for all evaluated models across tasks. The top row shows performance for the diagnosis task (FTD vs. NC), while the middle and bottom rows show phenoconversion prediction performance at the 2-year and 5-year time horizons, respectively. All curves are computed on held-out test sets and averaged over 10 repeated iterations. Shaded regions (matching the color of each curve) indicate the standard deviation across iterations.

contrast, all 15 highest-ranked features for the 2- and 5-year phenoconversion-prediction tasks were imaging-derived *DL-biomarkers*.

Among the imaging features, *Multi-EBM* assigned high importance to volumetric biomarkers, particularly those in frontal and temporal regions. In contrast, *Multi-ICNN* consistently assigned high importance to imaging-derived *DL-biomarkers* across all tasks. These *DL-biomarkers* comprised one global feature derived from the whole-brain input ("Total brain") and multiple region-specific features extracted from image patches, which were named according to the brain regions with the greatest spatial overlap. The highest-ranked imaging features were distributed across frontal and temporal regions, with additional contributions from the insula, anterior cingulate cortex, hippocampus, amygdala, and caudate. These regions broadly correspond to known patterns of neuroanatomical involvement in genetic FTD [35, 2, 13, 36]. Complete group-level feature-importance rankings and subgroup analyses are provided in the Supplementary Materials.

Figure 6 shows the learned shape functions for the FTD vs. NC classification task. For *Multi-ICNN*, the 15WT was the highest-ranked feature, while most of the remaining top-ranked biomarkers were imaging-derived *DL-biomarkers*. The resulting curves showed clear and predominantly monotonic relationships between the *DL-biomarkers* and their contributions to the predictions.

For *Multi-EBM*, the learned shape functions were generally consistent with established clinical patterns: worse neuropsychological performance, reduced regional gray matter volumes, larger ventricular volumes, and greater FreeSurfer-derived white matter hypointensity volumes shifted the model



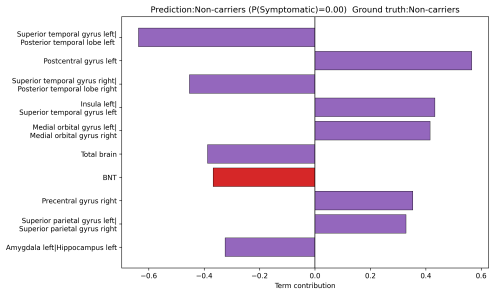
(a) *Multi-EBM*: FTD vs. NC

(b) *Multi-ICNN*: FTD vs. NC

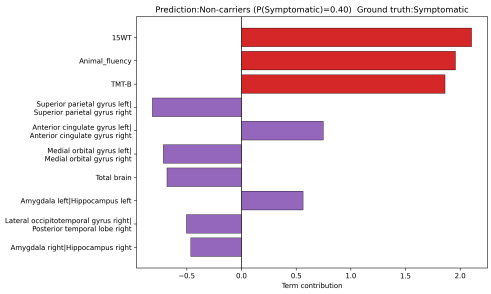
(c) *Multi-ICNN*: 2-year prediction

(d) *Multi-ICNN*: 5-year prediction

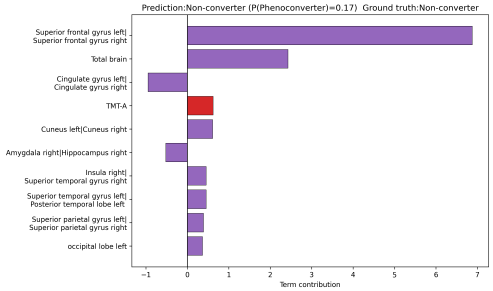
Figure 4: Group-level feature importance estimated from the training sets. Within each split, the absolute additive contributions of each feature were first averaged across samples and subsequently normalized across all features. The resulting normalized feature importance scores were then averaged across the 10 repeated train–validation–test splits. The 15 features with the highest mean normalized feature importance are shown for each model, with error bars representing the standard deviation across splits. Colors indicate the biomarker categories shown in the shared legend. Abbreviations: DL, deep learning; EBM, Explainable Boosting Machine; FTD, frontotemporal dementia; ICNN, Interpretable Convolutional Neural Network; NC, non-carrier.



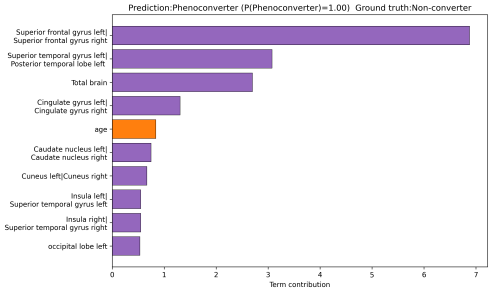
(a) Correctly classified non-carrier participant



(b) Misclassified symptomatic participant



(c) Correctly classified non-converter participant in the 2-year prediction task

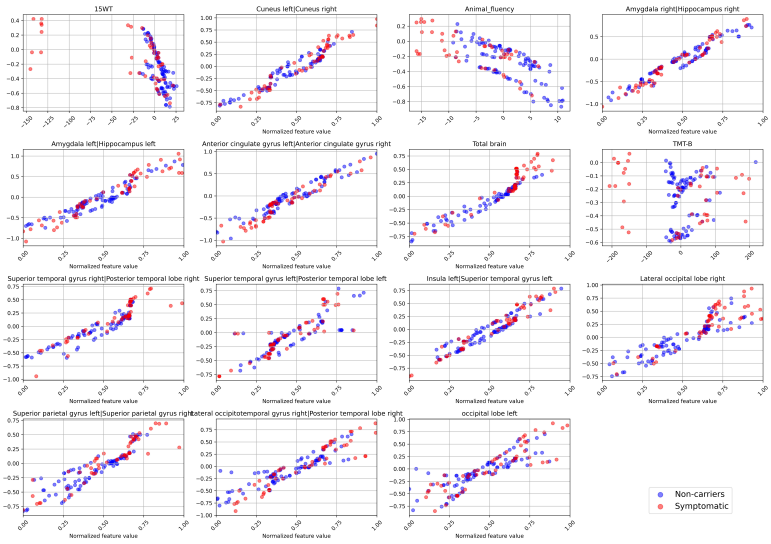


(d) Misclassified non-converter participant in the 2-year prediction task

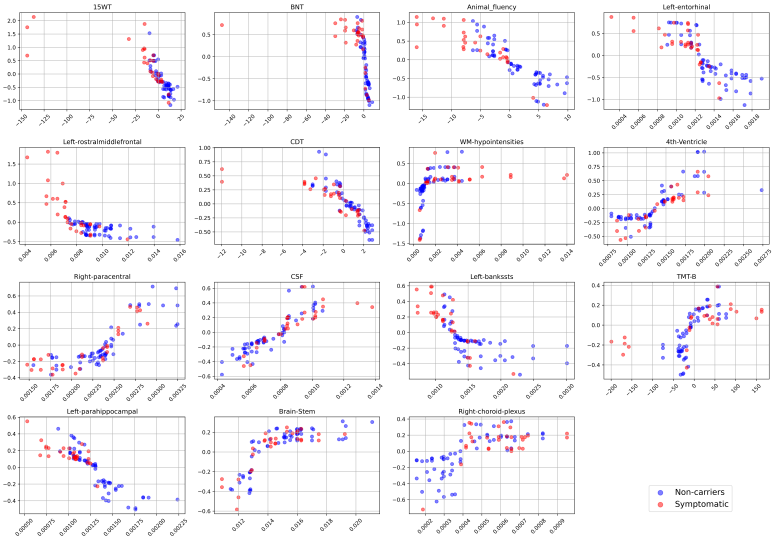


Figure 5: Individual-level feature contributions of *Multi-ICNN* for the FTD vs. NC classification task (top row) and the 2-year phenoconversion-prediction task (bottom row). For each task, one correctly classified participant and one representative misclassified participant are shown. The 10 features with the largest absolute contribution values are displayed for each prediction. Bars extending to the right increase the predicted probability of the positive class (FTD or phenoconversion), whereas bars extending to the left increase the predicted probability of the negative class (NC or non-conversion). Colors indicate imaging-derived, demographic, and neuropsychological biomarkers, as shown in the legend.

output toward FTD. Despite their different feature representations, both *Multi-ICNN* and *Multi-EBM* predominantly learned monotonic feature–response relationships.



(a) *Multi-ICNN*



(b) *Multi-EBM*

Figure 6: Learned shape functions for FTD vs. NC classification. The 15 biomarkers with the highest group-level feature importance are shown for each model. Each dot represents the observed feature value and its corresponding additive contribution for an individual test sample, aggregated across the test sets from the 10 repeated train–validation–test splits. Red dots denote symptomatic mutation carriers, and blue dots denote non-carriers. Positive contribution values shift the prediction toward FTD, whereas negative values shift it toward NC. For *Multi-ICNN*, the x-axis represents the within-split percentile of each *DL-biomarker*; for *Multi-EBM*, it represents the original feature value. Volumetric biomarkers used in *Multi-EBM* were corrected for intracranial volume.

4. Discussion

In this study, we investigated whether multimodal biomarkers could improve FTD vs. NC classification and phenoconversion prediction among presymptomatic mutation carriers. We also examined the relative contributions of imaging, demographic, genetic, and neuropsychological information to the model outputs.

Regarding biomarker importance, the relative contributions of the different modalities varied markedly between FTD vs. NC classification and phenoconversion prediction. For FTD vs. NC classification, neuropsychological measures were among the most informative features in both multimodal models. The 15-Word Test was the highest-ranked feature in both *Multi-EBM* and *Multi-ICNN*, while the Boston Naming Test and semantic animal fluency also ranked highly in *Multi-EBM*. These findings indicate that verbal learning, word retrieval, and semantic fluency provide substantial information for distinguishing symptomatic mutation carriers from non-carriers. Imaging biomarkers nevertheless remained important, particularly in *Multi-ICNN*, for which most of the remaining top-ranked features were imaging-derived *DL-biomarkers*. In contrast, all top-15 features for the 2-year and 5-year phenoconversion-prediction tasks were imaging-derived *DL-biomarkers*. This difference indicates that neuropsychological measures were particularly informative for distinguishing participants with clinically manifest FTD from non-carriers, whereas imaging-derived biomarkers contributed more strongly to prospective phenoconversion prediction among presymptomatic carriers. This pattern may reflect that cognitive changes become more pronounced around phenoconversion but remain relatively subtle and heterogeneous ear-

lier in the presymptomatic stage [13, 6].

Regarding model performance, the benefit of multimodal integration depended on the task. For FTD vs. NC classification, adding non-imaging biomarkers improved both model families: *Multi-EBM* achieved higher mean BA and AUC than either *Vol-EBM* or *NonImage-EBM*, and *Multi-ICNN* achieved the highest overall performance, with a BA of 0.939 and an AUC of 0.953. In contrast, adding non-imaging biomarkers provided little additional benefit for phenoconversion prediction. *ICNN* and *Multi-ICNN* achieved similar performance at both time horizons, with *ICNN* yielding slightly higher AUCs and *Multi-ICNN* slightly higher BAs. Both models substantially exceeded the mean performance of the conventional EBM-based models, particularly for 5-year prediction. Overall, the *ICNN*-based models, which use imaging-derived *DL-biomarkers*, achieved the highest mean phenoconversion-prediction performance, while the addition of non-imaging biomarkers provided little further improvement. This finding is consistent with the feature importance results, which showed that the phenoconversion-prediction tasks were driven predominantly by imaging-derived *DL-biomarkers*.

The results further demonstrate that phenoconversion can be predicted from presymptomatic structural MRI at both the 2-year and 5-year horizons. Whereas the predictive value of conventional volumetric and non-imaging biomarkers was limited, particularly for 5-year prediction, the models incorporating imaging-derived *DL-biomarkers* maintained AUCs of approximately 0.78–0.80 across both prediction horizons. This relatively stable performance suggests that whole-brain *DL-biomarkers* capture subtle and spatially distributed structural patterns associated with both near-term and

longer-term phenoconversion. Recent longitudinal MRI studies have similarly identified structural alterations associated with future phenoconversion before the emergence of clinical symptoms [37, 38]. Our findings extend this work by demonstrating that such prognostic information can be extracted from structural MRI within an intrinsically interpretable framework. Nevertheless, performance remained lower than that achieved for cross-sectional FTD vs. NC classification, and the uncertainty reflected by the confidence intervals highlights the need for validation in larger independent cohorts.

The group-level feature importance analyses revealed that the models relied on distributed neuroanatomical patterns characteristic of genetic FTD. Across tasks, the highest-ranking imaging-derived biomarkers were associated predominantly with image patches overlapping frontal and temporal regions, with additional contributions from the insula, anterior cingulate cortex, hippocampus, amygdala, and caudate. This distribution suggests that the models captured spatially distributed disease-related structural alterations rather than relying on a single anatomical region. The identified pattern is consistent with the heterogeneous neuroanatomical involvement reported in genetic FTD [2, 13, 36]. Ventricular biomarkers also contributed to model predictions, likely reflecting enlargement secondary to cortical and subcortical atrophy.

Compared with our previous findings in Alzheimer's disease, where imaging-derived biomarkers primarily emphasized medial temporal structures including the hippocampus, amygdala, and entorhinal cortex [22, 30], the present study highlighted greater contributions from frontal and temporal cortices, together with the insula, anterior cingulate cortex, and caudate. This broader

anatomical involvement agrees with previous structural MRI studies showing that genetic FTD preferentially affects distributed frontotemporal and subcortical regions, including the insula and anterior cingulate cortex, rather than being confined to medial temporal structures [2, 13, 36].

Regarding model stability, the high-ranking biomarkers identified by *Multi-EBM* were broadly consistent between the training and testing sets. Although the ranking of individual *DL-biomarkers* in *Multi-ICNN* varied across iterations and data splits, the anatomical regions represented by the highest-contributing features showed substantial overlap. This variability may reflect both the biological heterogeneity of genetic FTD and the distributed feature representations learned by the *ICNN* feature extractor. Nevertheless, the recurring involvement of similar anatomical regions across analyses supports the consistency of the overall neuroanatomical patterns identified by the model.

Recent explainable deep learning studies in FTD have primarily relied on post hoc attribution methods, such as Integrated Gradients, to visualize regions contributing to model predictions [21]. In contrast, the proposed framework is intrinsically interpretable, enabling subject-specific explanations directly through the prediction model rather than requiring a separate explanation algorithm. At the subject level, this enables transparent interpretation of participant-specific model outputs rather than relying solely on global feature rankings or black-box decision boundaries. Subject-specific explanations offer insights into the heterogeneity of genetic FTD by revealing how different imaging and non-imaging biomarkers contribute to individual predictions. This capability may facilitate clinical interpretation of model

outputs and support personalized assessment of disease risk. Nevertheless, systematic validation involving clinicians and other domain experts is required to establish the reliability, reproducibility, and clinical relevance of these subject-level explanations before clinical deployment.

For the visualization of feature-level trends, the shape functions of *Multi-EBM* revealed clear and predominantly monotonic relationships between individual biomarkers and model predictions. For most neuropsychological measures and cortical volumetric features, higher values were associated with the healthier class, whereas larger ventricular volumes and greater FreeSurfer-derived white matter hypointensity volumes shifted predictions toward the FTD class. These relationships are consistent with established structural and cognitive characteristics of neurodegeneration. Importantly, the shape functions not only identify which biomarkers are influential but also quantify how changes in individual biomarkers affect model predictions across their observed value ranges, thereby providing a more transparent representation of feature–outcome relationships.

For the imaging-derived *DL-biomarkers* in *Multi-ICNN*, the learned feature-response curves exhibited predominantly monotonic trends that were highly consistent across training iterations after alignment. Unlike conventional volumetric biomarkers, these *DL-biomarkers* represent latent features learned from whole-brain or patch-based image inputs and do not correspond directly to predefined anatomical measurements. Although their individual feature values are not directly interpretable, the consistent shape and direction of the feature–response curves indicate that the model learns stable and reproducible imaging representations associated with FTD classification.

This study has several limitations. First, T1-weighted structural MRI was the only imaging modality included. Although previous studies have demonstrated the predictive value of diffusion MRI and fluid biomarkers, such as neurofilament light chain, for disease progression and phenoconversion in genetic FTD [10, 12], these modalities were not consistently available across participants and visits in the current cohort. Restricting the analysis to T1-weighted MRI allowed us to maximize the number of available participants and avoid a substantial reduction in sample size, which is particularly important for training deep learning models. Second, the data were obtained from a single-center cohort. Although the proposed framework was evaluated using repeated subject-level train–validation–test splits, it has not yet been validated in an external cohort, limiting the assessment of its generalizability. Finally, despite the use of one of the largest available longitudinal genetic FTD cohorts, the rarity of the disease inevitably limited the sample size, particularly for phenoconversion prediction. The primary analyses pooled carriers of different pathogenic genes, and the available subgroup sample sizes were insufficient for a robust evaluation of gene-specific model performance. Differences in neuroanatomical patterns and disease trajectories across genetic subtypes may therefore have influenced the results.

Future work should validate the proposed framework in larger independent multicenter cohorts and investigate whether complementary imaging and fluid biomarkers can further improve phenoconversion prediction. Clinical evaluation of the subject-level explanations will also require an appropriate structured assessment framework, potentially involving expert evaluation of their plausibility, reproducibility, and utility for personalized disease mon-

itoring.

5. Conclusion

This study demonstrates that interpretable models achieved high performance for FTD vs. NC classification, whereas phenoconversion prediction remained more challenging. Deep learning-derived imaging representations nevertheless provided information beyond conventional volumetric measures, particularly for phenoconversion prediction. The relative contribution of multimodal information was task dependent: non-imaging biomarkers contributed prominently to FTD classification but had a more limited role in phenoconversion prediction, for which imaging-derived *DL-biomarkers* predominated. By integrating these feature representations within an additive framework, the proposed approach provides transparent group- and subject-level descriptions of how individual biomarkers contribute to model outputs. Although the predictive performance and clinical relevance of these explanations require validation in larger independent cohorts, our findings demonstrate the potential of interpretable deep learning for genetic FTD diagnosis and phenoconversion prediction.

List of abbreviations

ALS Amyotrophic lateral sclerosis

AUC Area under the receiver operating characteristic curve

BA Balanced accuracy

BNT Boston Naming Test

CDT Clock Drawing Test

CDR plus NACC FTLD CDR Dementia Staging Instrument plus National Alzheimer's Coordinating Center Frontotemporal Lobar Degeneration component

CI Confidence interval

CNN Convolutional neural network

DL Deep learning

EBM Explainable Boosting Machine

FTD Frontotemporal dementia

GAM Generalized additive model

GM Gray matter

ICNN Interpretable Convolutional Neural Network

ICV Intracranial volume

MMSE Mini-Mental State Examination

MRI Magnetic resonance imaging

NC Non-carrier

PR Precision–recall

ROC Receiver operating characteristic

ROI Region of interest

SD Standard deviation

T1w T1-weighted

TMT-A Trail Making Test part A

TMT-B Trail Making Test part B

VBM Voxel-based morphometry

15WT 15-Word Test

C9orf72 Chromosome 9 open reading frame 72

GRN Progranulin

MAPT Microtubule-associated protein tau

TARDBP TAR DNA-binding protein

Declarations

Ethics approval and consent to participate

Ethical approval for the study was granted by the Medical and Ethical Review Committee of Erasmus MC University Medical Center (MEC-2009-409). All participants provided written informed consent before enrolment.

Consent for publication

Not applicable.

Availability of data and materials

The FTD-RisC data supporting the findings of this study are not publicly available because of participant privacy, ethical restrictions, and applicable data-sharing agreements. Access may be requested from the FTD-RisC data custodian at FTDRisC@erasmusmc.nl, subject to approval and the establishment of an appropriate data-use agreement.

Competing interests

The authors declare that they have no known conflicting financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

Wenjie Kang was supported by the China Scholarship Council (grant number 202206540009). This work was supported by the European Union's Horizon Europe research and innovation programme under grant agreement No. 101156175 (PREDICTFTD). Esther E. Bron was supported by an Erasmus MC Fellowship 2022 and the TAP-dementia project funded by ZonMw (grant number 10510032120003) in the context of the Dutch National Dementia Strategy.

Authors' contributions

WK conceptualized the study, developed the methodology, performed the analyses, prepared the visualizations, and drafted the manuscript. TS, MFvH, and LCJ contributed to data acquisition and curation, clinical assessments, and critical revision of the manuscript. SK contributed to the methodology, supervision, and critical revision of the manuscript. HS contributed

to cohort development, clinical data acquisition, resources, supervision, and critical revision of the manuscript. EEB contributed to study conceptualization and methodology, supervision, funding acquisition, and critical revision of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Declaration of generative AI and AI-assisted technologies

During the preparation of this work, the authors used ChatGPT (OpenAI) for grammatical correction and to refine the visual styling of the flowchart in Figure 2. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

[1] A. Antonioni, E. M. Raho, P. Lopriore, A. P. Pace, R. R. Latino, M. Assogna, M. Mancuso, D. Gragnaniello, E. Granieri, M. Pugliatti, et al., Frontotemporal dementia, where do we stand? A narrative review, International journal of molecular sciences 24 (14) (2023) 11732.

[2] H. Seelaar, J. D. Rohrer, Y. A. Pijnenburg, N. C. Fox, J. C. Van Swieten, Clinical, genetic and pathological heterogeneity of frontotemporal dementia: a review, Journal of Neurology, Neurosurgery & Psychiatry 82 (5) (2011) 476–486.

[3] C. V. Greaves, J. D. Rohrer, An update on genetic frontotemporal dementia, Journal of neurology 266 (8) (2019) 2075–2086.

[4] H. Seelaar, W. Kamphorst, S. M. Rosso, A. Azmani, R. Masdjedi, I. de Koning, J. A. Maat-Kievit, B. Anar, L. D. Kaat, G. J. Breedveld, et al., Distinct genetic forms of frontotemporal dementia, Neurology 71 (16) (2008) 1220–1226.

[5] L. Robinson, E. Tang, J.-P. Taylor, Dementia: timely diagnosis and early intervention, Bmj 350 (2015).

[6] L. C. Jiskoot, J. L. Panman, L. van Asseldonk, S. Franzen, L. H. Meeter, L. Donker Kaat, E. L. van der Ende, E. G. Dopper, R. Timman, R. van Minkelen, et al., Longitudinal cognitive biomarkers predicting symptom onset in presymptomatic frontotemporal dementia, Journal of neurology 265 (6) (2018) 1381–1392.

[7] J. M. Poos, K. M. Moore, J. Nicholas, L. L. Russell, G. Peakman, R. S. Convery, L. C. Jiskoot, E. van der Ende, E. van den Berg, J. M. Papma, et al., Cognitive composites for genetic frontotemporal dementia: GENFI-Cog, Alzheimer's research & therapy 14 (1) (2022) 10.

[8] J. M. Papma, L. C. Jiskoot, J. L. Panman, E. G. Dopper, T. Den Heijer, L. Donker Kaat, Y. A. Pijnenburg, L. H. Meeter, R. Van Minkelen, S. A. Rombouts, et al., Cognition and gray and white matter characteristics of presymptomatic C9orf72 repeat expansion, Neurology 89 (12) (2017) 1256–1264.

[9] E. G. Dopper, V. Chalos, E. Ghariq, T. den Heijer, A. Hafkemeijer, L. C. Jiskoot, I. de Koning, H. Seelaar, R. van Minkelen, M. J. van Osch, et al., Cerebral blood flow in presymptomatic MAPT and GRN mutation carriers: a longitudinal arterial spin labeling study, NeuroImage: Clinical 12 (2016) 460–465.

[10] R. A. Feis, M. J. Bouts, F. de Vos, T. M. Schouten, J. L. Panman, L. C. Jiskoot, E. G. Dopper, J. van der Grond, J. C. van Swieten, S. A. Rombouts, A multimodal MRI-based classification signature emerges just prior to symptom onset in frontotemporal dementia mutation carriers, Journal of Neurology, Neurosurgery & Psychiatry 90 (11) (2019) 1207–1214.

[11] E. L. Van Der Ende, L. H. Meeter, J. M. Poos, J. L. Panman, L. C. Jiskoot, E. G. Dopper, J. M. Papma, F. J. De Jong, I. M. Verberk, C. Teunissen, et al., Serum neurofilament light chain in genetic frontotemporal

dementia: a longitudinal, multicentre cohort study, The Lancet Neurology 18 (12) (2019) 1103–1111.

[12] L. A. Giannini, H. Seelaar, E. L. van der Ende, J. M. Poos, L. C. Jiskoot, E. G. Dopper, Y. A. Pijnenburg, E. A. Willemse, L. Vermunt, C. E. Teunissen, et al., Clinical value of longitudinal serum neurofilament light chain in prodromal genetic frontotemporal dementia, Neurology 101 (10) (2023) e1069–e1082.

[13] J. D. Rohrer, et al., Presymptomatic cognitive and neuroanatomical changes in genetic frontotemporal dementia, The Lancet Neurology 14 (3) (2015) 253–262.

[14] L. H. Meeter, L. D. Kaat, J. D. Rohrer, J. C. Van Swieten, Imaging and fluid biomarkers in frontotemporal dementia, Nature Reviews Neurology 13 (7) (2017) 406–419.

[15] C. Duran-Aniotz, P. Orellana, T. Leon Rodriguez, F. Henriquez, V. Cabello, M. F. Aguirre-Pinto, T. Escobedo, L. T. Takada, S. D. Pina-Escudero, O. Lopez, et al., Systematic review: genetic, neuroimaging, and fluids biomarkers for frontotemporal dementia across Latin America countries, Frontiers in Neurology 12 (2021) 663407.

[16] J. L. Panman, V. Venkatraghavan, E. L. Van Der Ende, R. M. Steketee, L. C. Jiskoot, J. M. Poos, E. G. Dopper, L. H. Meeter, L. D. Kaat, S. A. Rombouts, et al., Modelling the cascade of biomarker changes in GRN-related frontotemporal dementia, Journal of Neurology, Neurosurgery & Psychiatry 92 (5) (2021) 494–501.

[17] A. M. Staffaroni, M. Quintana, B. Wendelberger, H. W. Heuer, L. L. Russell, Y. Cobigo, A. Wolf, S.-Y. M. Goh, L. Petrucelli, T. F. Gendron, et al., Temporal order of clinical and biomarker changes in familial frontotemporal dementia, Nature medicine 28 (10) (2022) 2194–2206.

[18] S. Dattola, A. Ielo, G. Varone, A. Cacciola, A. Quartarone, L. Bonanno, Frontotemporal dementia: a systematic review of artificial intelligence approaches in differential diagnosis, Frontiers in Aging Neuroscience 17 (2025) 1547727.

[19] S. Mirabian, F. Mohammadian, Z. Ganji, H. Zare, E. Hasanpour Khalesi, The potential role of machine learning and deep learning in differential diagnosis of Alzheimer's disease and FTD using imaging biomarkers: A review, The Neuroradiology Journal (2025) 19714009251313511.

[20] A. Rogeau, F. Hives, C. Bordier, H. Lahousse, V. Roca, T. Lebouvier, F. Pasquier, D. Huglo, F. Semah, R. Lopes, A 3D convolutional neural network to classify subjects as Alzheimer's disease, frontotemporal dementia or healthy controls using brain 18F-FDG PET, Neuroimage 288 (2024) 120530.

[21] D. Ma, J. Stocks, H. Rosen, K. Kantarci, S. N. Lockhart, J. R. Bateman, S. Craft, M. N. Gurcan, K. Popuri, M. F. Beg, et al., Differential diagnosis of frontotemporal dementia subtypes with explainable deep learning on structural mri, Frontiers in neuroscience 18 (2024) 1331677.

[22] W. Kang, L. Jiskoot, P. De Deyn, G. Biessels, H. Koek, J. Claassen,

H. Middelkoop, W. Flier, W. J. Jansen, S. Klein, et al., GL-ICNN: An End-To-End Interpretable Convolutional Neural Network for the Diagnosis and Prediction of Alzheimer's Disease, arXiv preprint arXiv:2501.11715 (2025).

[23] E. G. Dopper, S. A. Rombouts, L. C. Jiskoot, T. den Heijer, J. R. A. de Graaf, I. de Koning, A. R. Hammerschlag, H. Seelaar, W. W. Seeley, I. M. Veer, et al., Structural and functional brain connectivity in presymptomatic familial frontotemporal dementia, Neurology 83 (2) (2014) e19–e26.

[24] B. Fischl, FreeSurfer, Neuroimage 62 (2) (2012) 774–781.

[25] B. Fischl, D. H. Salat, A. J. Van Der Kouwe, N. Makris, F. Ségonne, B. T. Quinn, A. M. Dale, Sequence-independent segmentation of magnetic resonance images, Neuroimage 23 (2004) S69–S84.

[26] R. S. Desikan, F. Ségonne, B. Fischl, B. T. Quinn, B. C. Dickerson, D. Blacker, R. L. Buckner, A. M. Dale, R. P. Maguire, B. T. Hyman, et al., An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest, Neuroimage 31 (3) (2006) 968–980.

[27] C. D. Good, I. S. Johnsrude, J. Ashburner, R. N. Henson, K. J. Friston, R. S. Frackowiak, A voxel-based morphometric study of ageing in 465 normal adult human brains, Neuroimage 14 (1) (2001) 21–36.

[28] G. V. Roshchupkin, H. H. Adams, S. J. van der Lee, M. W. Vernooij, C. M. van Duijn, A. G. Uitterlinden, A. van der Lugt, A. Hofman, W. J.

Niessen, M. A. Ikram, Fine-mapping the effects of Alzheimer's disease risk loci on brain morphology, Neurobiology of aging 48 (2016) 204–211.

[29] H. Nori, S. Jenkins, P. Koch, R. Caruana, Interpretml: A unified framework for machine learning interpretability, arXiv preprint arXiv:1909.09223 (2019) 303–310.

[30] W. Kang, B. Li, L. C. Jiskoot, P. P. De Deyn, G. J. Biessels, H. L. Koek, J. A. Claassen, H. A. Middelkoop, W. M. van der Flier, W. J. Jansen, et al., An interpretable machine learning framework with data-informed imaging biomarkers for diagnosis and prediction of alzheimer's disease, Computerized Medical Imaging and Graphics (2026) 102722.

[31] V. García, R. A. Mollineda, J. S. Sánchez, Index of balanced accuracy: A performance measure for skewed class distributions, in: Iberian conference on pattern recognition and image analysis, Springer, 2009, pp. 441–448.

[32] J. N. Mandrekar, Receiver operating characteristic curve in diagnostic test assessment, Journal of Thoracic Oncology 5 (9) (2010) 1315–1316.

[33] C. Nadeau, Y. Bengio, Inference for the generalization error, Advances in neural information processing systems 12 (1999).

[34] E. R. DeLong, D. M. DeLong, D. L. Clarke-Pearson, Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach, Biometrics 44 (3) (1988) 837–845.

[35] W. W. Seeley, R. K. Crawford, J. Zhou, B. L. Miller, M. D. Greicius,

748 Neurodegenerative diseases target large-scale human brain networks,
749 Neuron 62 (1) (2009) 42–52.

750 [36] D. M. Cash, et al., Imaging endpoints for clinical trials in frontotemporal
751 dementia, Alzheimer's Research & Therapy 10 (2018) 37.

752 [37] K. van Veenhuijzen, H. H. Tan, A. D. Nitert, M. A. van Es, J. H. Veldink,
753 L. H. van den Berg, H.-J. Westeneng, Longitudinal magnetic resonance
754 imaging in asymptomatic c9orf72 mutation carriers distinguishes pheno-
755 converters to amyotrophic lateral sclerosis or amyotrophic lateral scle-
756 rosis with frontotemporal dementia, Annals of Neurology 97 (2) (2025)
757 281–295.

758 [38] T. Shen, J. Wuu, C. A. Olm, H. Radhakrishnan, Y. Li, P. M. Pattany,
759 V. Granit, N. Carberry, D. J. Irwin, L. Massimo, et al., Thalamic volume
760 as a susceptibility/risk biomarker for c9orf72 phenoconversion, medRxiv
761 (2025) 2025–09.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FTDRisCsupplementtherapy.pdf