VER
Vascular & Endovascular Review

Artificial Intelligence-Driven Early Detection of Neurological Disorders and Its Implications for Personalized Rehabilitation Strategies

Tofayel Ahmed Onik[1], Kamrun Naher Irin[2], Md Noman Azam[3], Kabita Shamia Akter[4], Mahbub Ahmed Nabil[5], Iftekhar Hossain[6], Sonia Akter[7]

Doctorate in Computer Science[1], Master of Science in Kinesiology[2], Master of Science in Exercise and Sport Science[3], Master of Science in Information Technology[4], Master of Science in Information Technology[5], Master of Science in Information Technology (MSIT)[6], Master of Science in Business Analytics[7]
University of the Potomac[1], Department of Health & Kinesiology, Texas A&M University-Kingsville[2], Department of Health Sciences and Leadership, St. Francis College[3], Washington University of Science and Technology, Alexandria, Virginia[4,5,6], Mercy University, New York, NY[7]
tofayelahmedonik@gmail.com [1], kamrunirin443@gmail.com [2], nomanlpu07@gmail.com [3], kabitasamiaakter@gmail.com [4], ahmednabilnsucse@gmail.com [5], ihossain.student@wust.edu [6], sakter5@mercy.edu[7]

ABSTRACT

Early AD stratification is clinically relevant, and both structural MRI and cognitive–biomarker profiles contain complementary diagnostic information. We offer NeuroFusionAI, an ImageNet pre-trained ViT-Small/16 MRI branch (to extract relevant brain feature representation signals), combined with an FT-Transformer clinical branch via a bidirectional cross-attention module, and validate it on a four-class AD-stage task (NonDemented, VeryMildDemented, MildDemented, ModerateDemented). An important limitation is acknowledged: the imaging and clinical data are not sampled from the same patient cohort. The MRI set consists of 28,160 public access 2D brain images, while the clinical set is purely synthetic (6,000 records - 1,500/class) and is matched to images only at the level of diagnostic class, but never at the level of subject. It makes no claim of true multimodal correspondence per patient. One of the features (global CDR) was removed from consideration as a deterministic label proxy to avoid leakage. We circumvent optimistic bias from multiply-augmented near-duplicates with a perceptual-hash grouping procedure, which commits each near-duplicate group to exactly one data split. NeuroFusionAI achieves 98.18% accuracy and 0.9724 macro-F1 in under 5-fold stratified-group cross-validation with a held-out test set. Importantly, feature-concatenation fusion (0.9761) and a clinical-only XGBoost baseline (0.9770) equal or slightly outperform cross-attention fusion; differences between models are not statistically significant (Wilcoxon, n=5 folds). We view these results with skepticism; clinical-only performance near-ceiling is an expected byproduct of class-conditional synthetic generation, not evidence that clinical features alone will satisfy in real-world practice. Grouping and pairing tables available for full transparency.

KEYWORDS: Alzheimer's disease; multimodal fusion; vision transformer; FT-Transformer; cross-attention; data leakage; synthetic clinical data; explainable AI..

How to Cite: Tofayel Ahmed Onik, Kamrun Naher Irin, Md Noman Azam, Kabita Shamia Akter, Mahbub Ahmed Nabil, Iftekhar Hossain, Sonia Akter, (2023) Artificial Intelligence-Driven Early Detection of Neurological Disorders and Its Implications for Personalized Rehabilitation Strategies, Vascular and Endovascular Review, Vol.6, No.2, 104-111

INTRODUCTION

Abstract Alzheimer's disease (AD) is the most common cause of dementia, a continuum that stretches from cognitively normal aging through a very mild, prodromal phase towards mild and moderate impairment [1]. As disease-modifying interventions are thought to be most efficacious if implemented at the early stages, efforts have been made to develop automated tools that can patient-stratify along this continuum from routinely collected data. Structural MRI detects atrophy patterns that reflect disease stage, such as in the medial temporal lobe and hippocampus [3], whilst cognitive testing and fluid biomarkers provide different quantitative support for neurodegeneration [2]. In principle, combining these complementary perspectives should provide stronger staging than either modality alone [8], [10].

This issue could have been traced back to only two developments. We discuss vision transformers (ViTs) [14] which, when initialized from large-scale pretraining, have been competitive with and often better than convolutional networks on medical-imaging tasks, providing global receptive fields and interpretable attention at a low-cost mechanism level [15]. Second, transformer-based architectures for tabular data, most notably the FT-Transformer [20], now compete against gradient-boosted trees on heterogeneous clinical features while yielding token representations that directly correspond to attention-based fusion [13]. Taken together, these developments make an appealing multimodal model a good design point.

Nevertheless, there are two methodological pitfalls that can easily be overlooked in multimodal AD research and that lead to inflated performance findings. The first is data leakage from augmented near-duplicate images: many publicly distributed AD MRI collections are balanced by aggressive augmentation, discarding rotated, flipped, or intensity-shifted duplicates of the same underlying slice throughout the experiment. In the case where such siblings are split between training and test partitions, the model memorizes this test content, and accuracy no longer provides a measure of generalization; in one systematic review, >50% of surveyed CNN-based AD studies were found to be potentially affected by this form of leakage. [6]. The other problem is

leakage of labels from clinical features that are algebraic restatements of the target label. For instance, a global clinical dementia rating (GCDR) is, in effect, an almost deterministic function of how far along the diagnostic staging process.

In this work, we adopt an openly conservative and disclosure-first position with respect to both hazards, as we acknowledge but do not seek to remedy, without better data, a third, more fundamental limitation of our data. NOTE: The MRI and clinical modalities in the present case are not from the same subjects. The synthetic clinical/biomarker table is only associated with images at the class level. Thus, we do not and cannot demonstrate true per-patient multimodal fusion. What we can show is (i) a reproducible, leakage-controlled evaluation protocol, (ii) that cross-attention fusion observes the same goodness-of-fit and stageness trade-off as do strong concatenation- and unimodal baselines when evaluated under this protocol, iii) an honest interpretation of why, on this benchmark it worked out for more expressive fusion not to win.

The rest of this paper (Section 2) reviews related work and describes the data, leakage controls, architecture and training protocol, (Section 3) presents an empirical comparison with explainability analysis (Section 4) interprets findings alongside honest limitations (Section 5), and concludes in Section 6

LITERATURE REVIEW

Extensive work exists to stage AD from structural MRI, using either 2D slices or full 3D volumes with convolutional neural networks [5]. Reported accuracies on widely redistributed 2D collections are usually high, but a developing methodological literature warns that such numbers can be overestimated by subject- and augmentation-level leakage when partitioning is done at the image level rather than the level of subjects or acquisitions [6]. Cross-cohort studies demonstrate that models generally exhibit a large performance drop if validated on data not drawn from the same cohort, underscoring the importance of provenance-aware evaluation [7]. For datasets for which the provenance metadata has been removed, our grouping procedure is a straightforward (albeit heuristic) approach to this problem.

Fusion of imaging with clinical or genomic data in tabular formats has repeatedly been employed to improve dementia staging [8], [9] on ADNI-type cohorts, using early (feature-level), late (decision-level) and intermediate fusion strategies. Guidelines for those strategies are provided in general reviews of imaging-clinical fusion [10]. Among them is the attention-based intermediate fusion (which conditions one modality on the representation of another) that has been shown to be especially effective for modeling cross-modal dependencies while preserving end-to-end training, specifically applied to AD diagnosis [11]. A common and significant caveat in this literature is that modal gains are only maintained when the data from each modality–per subject- are paired correctly for every item. We make this caveat explicit rather than obscuring it.

Motivated by the success of the Transformer [13] and its adaptation to vision as the Vision Transformer [14], attention-based backbones have been adapted widely in medical imaging, including pure transformers and hybrid convolutional – attention designs [15]. Training on large natural-image datasets, e.g., ImageNet [23], harms generalization of convolutional baselines trained from scratch on comparatively small medical datasets. Data-efficient transfer learning [16] and attention rollout [17] provide a low-complexity, native decomposition-based wildcard detector that generates saliency without gradient perturbation, complementary to recent gradient-based methods (e.g., Grad-CAM) [18]. Under tight compute limits, we turn to a ViT-Small/16 backbone precisely because of this favorable data-efficiency and interpretability profile.

Gradient-boosted decision trees, particularly XGBoost [19], continue to be strong baselines in tabular clinical data and often outperform or match deep models. This motivated the development of transformational architectures for tabular inputs, FT-Transformer [20], which tokenizes numerical and categorical features and processes them with a CLS-token transformer, closing this gap while providing token embeddings conducive to cross-modal attention. We keep XGBoost as a challenging unimodal baseline in our work because it is precisely challenging to beat; the fact that it performs near-ceiling on synthetic clinical data is an informative negative result in its own right.

METHODOLOGY

NeuroFusionAI consists of two modality-specific encoders and a fusion stage based on attention (Fig. 1). All subsequent experiments are based on a ViT-Small/16 backbone (patch size 16, embedding dimension 384), initialized from ImageNet [23] pretraining via the timm library [24], generating a pooled 384-dimensional image embedding. The clinical branch consists of an FT-Transformer [20] (embedding dim=64, 3 layers, 4 heads), which tokenizes each numerical and categorical feature, prepends a CLS token, and returns the CLS representation as a clinical embedding.

Both embeddings are linearly projected into a common 256-dimensional space and then fused by a bidirectional cross-attention module (embedding dimension is 256, two layers). In each layer, the MRI representation queries the clinical representation and, simultaneously, the clinical representation queries the MRI representation so that information goes in both directions instead of through a single conditioning path. The combined representation goes to a four-class softmax classification layer. This design is intentionally small in order to satisfy the single-GPU requirement.

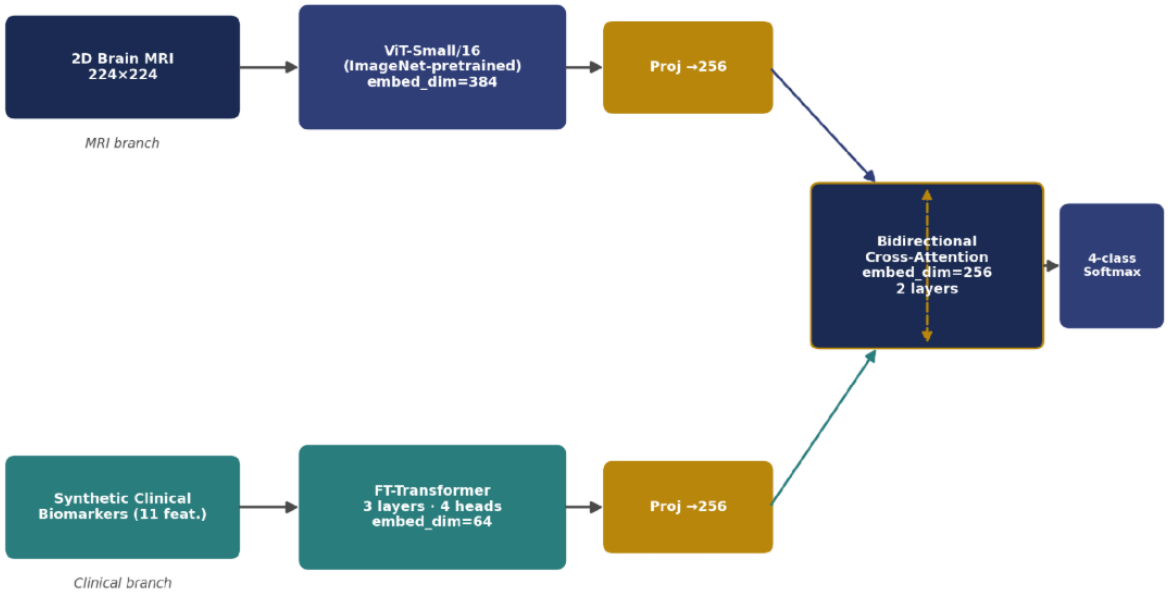

Figure 1: Methodology Diagram

3.1. Data and honesty disclosure

The lack of a real per-patient multimodal correspondence has been stated carefully in this study, and it is repeated here because it seems to dictate the interpretation of every subsequent result. The two modalities are different and mismatched only by diagnostic class.

Imaging modality includes a set of 28,160 two-dimensional brain MRI slices sampled from an openly accessible enlarged collection of Alzheimer s MRIs [20] across NonDemented (9,600), MildDemented (8,960), ModerateDemented (6,464), and VeryMildDemented(3,136). The class distribution between classes is slightly imbalanced (VeryMildDemented is the minority class). We scaled images to 224×224 and normalised lightly using ImageNet statistics, the same pretrained backbone used in training.

The clinical modality is completely artificial in nature. Included are 6000 synthetic subject records, with exactly 1500 samples from each class across eleven features: age, sex, years of education, APOE4 status, CSF $A\beta 1_{42}$, total tau, phospho-tau_{181}, MMSE score ADAS-Cog_{13} score physical examination (CDR sum-of-boxes (CDR-SB)) and hippocampal volume. Samples of the records from class-conditional distributions designed to capture the qualitative trajectories of each biomarker across stages (e.g. increasing (decreasing) amyloid-β_{142}, MMSE; increasing tau, ADAS-Cog_{13} with severity). The fact that generation is class-conditional leads the synthetic table to (by design) be much more cleanly separable than real clinical cohorts, a property we will repeatedly invoke when interpreting the strong clinical-only baseline in Section 5.

At the class level only, modality pairing is executed by deterministic round-robin reuse (seed 42) so that no synthetic subject is over- or under-represented in its class. They release a pairing for every image–subject assignment. CSV table for full auditability. That is, nothing in the synthetic record gets associated with an image based on individual-level similarity, and per-patient ground-truth linkage does not exist (and is never asserted).

3.2. Label-leakage control: exclusion of global CDR

The archival candidate feature set included the global Clinical Dementia Rating (CDR-Global). Upon inspection, it turned out that this feature had zero within-class variance; every single diagnostic class mapped to a unique CDR-Global value, which meant it was a deterministic proxy for the label. If they do, any classifier could recover the target just from a trivial lookup, which is textbook label leakage. This led us to discard CDR-Global from all experiments, while keeping CDR-SB, which features a continuum of within-class variation. We explicitly state this exclusion to ensure no reported clinical-branch performance is due to a restated label.

3.3. Perceptual-hash group leakage control

The redistributed collection of MRIs has no metadata about its source patients or augmentation lineages; filenames are opaque UUIDs. It would be naively splitting such data at the image level with augmented siblings of the same underlying slice placed on both sides of a partition, thus leaking test content into training. So instead of provenance labels, we use perceptual hashing [27] as an approximation for near-duplicate detection.

All images are processed to a pHash (8×8-bit perceptual hash). We calculate the Hamming distance between hashes and use a single-linkage union-find to group pairs of near-duplicates. From an empirical point of view, there is a subtlety to the thresholding choice. When pushing against a Hamming threshold of 4, we observed catastrophic chaining through the process of transitive bridging between visually similar dark-background slices; single components swelled to thousands of images (clusters of

approximately 5,788, 4,366 and 4,042 images; all > 1/2 a complete class), rendering group-aware splitting ineffective. Lowering the slip to 3 and implementing a hard maximum of images per cluster (maximum=100) as an additional safety net ensured group-size distribution stabilization.

Applying this procedure to the full dataset produced 19,572 groups from 28,160 images, of which 16,581 (84.7%) are singletons with no near-duplicates detected and containing only single images, and the remaining multi-image clusters have a maximum realized size of at most 97 (below the cap). StratifiedGroupKFold is used to keep fold statistics around class proportions, so that entire groups are assigned exactly one of {train, validation, test} and never split. We emphasize that this is a heuristic for true subject-level grouping: we cannot exclude residual leakage from undetected siblings, or even incidental over-merging of genuinely distinct slices, and this caveat qualifies all numbers reported here.

3.4. Data splits and cross-validation protocol

Prior to any model selection, a held-out test set was created by 15% of the groups through group-aware stratified splitting. We applied 5-fold stratified-group cross-validation on the rest of the pool; for every fold, we used its training portion to optimize different parameters, and if no metrics improved over a decade (early-stopping strategy), we saved the version of the model that performed best on the fold based on default selection criteria (best-validation macro-F1). The model selected for each fold was then scored one time, and this produced five test scores that are reported as mean ± standard deviation on a common held-out test set. Since grouping is enforced at all levels, no near-duplicate group spans the train/val/test boundary

RESULT AND ANALYSIS

4.1. Overall comparison

We report mean ± standard deviation over five cross-validation folds on the common held-out test set in Table 1. All four models have high absolute performance; this is a necessary but not sufficient condition and reflects both the clean class-conditional synthetic clinical data and the power of the imaging signal. The macro-F1 (0.9770) and ROC-AUC (0.9992) of the clinical-only XGBoost baseline are significantly higher than those of any other method, whereas for the concatenation-fusion model, the performance is essentially tied, including the proposed cross-attention model (macro-F1 0.9761; 0.9724); and MRI-only ViT trailed all modalities based multimodal and clinical models(0.9548).

Table 1. *Overall four-class staging performance*

Model	Accuracy	Macro-F1	Cohen κ	ROC-AUC	PR-AUC
NeuroFusionAI (cross-attn, proposed)	0.9818 ± 0.0015	**0.9724 ± 0.0017**	0.9747 ± 0.0021	0.9984 ± 0.0003	0.9922 ± 0.0010
NeuroFusionAI-Concat	0.9845 ± 0.0017	0.9761 ± 0.0024	0.9784 ± 0.0024	0.9984 ± 0.0001	0.9932 ± 0.0006
Clinical-only XGBoost	0.9827 ± 0.0012	0.9770 ± 0.0016	0.9759 ± 0.0017	0.9992 ± 0.0001	0.9967 ± 0.0004
MRI-only ViT-Small	0.9676 ± 0.0028	0.9548 ± 0.0038	0.9549 ± 0.0039	0.9964 ± 0.0006	0.9845 ± 0.0027

4.2. Per-class performance

F1 is differentiated by class in Table 2. The VeryMildDemented class proved to be difficult across all models, as it is the smallest imaging class (3,136 images) and the one most closely resembling NonDemented visually. The proposed model obtains 0.9199 F1 on this class, compared to 0.9880 (MildDemented) and 0.9980 (ModerateDemented); the steepest VeryMild drop (0.8798) from the MRI-only model indicates that for this particular challenge, difficulty is mainly in the imaging side of things.

Table 2. *Per-class F1*

Model	NonDem.	VeryMild	Mild	Moderate
NeuroFusionAI (proposed)	0.9839	**0.9199**	0.9880	0.9980
NeuroFusionAI-Concat	0.9858	0.9296	0.9903	0.9986
Clinical-only XGBoost	0.9896	0.9482	0.9834	0.9870
MRI-only ViT-Small	0.9617	0.8798	0.9783	0.9992

4.3. Statistical significance

Table 3 shows paired Wilcoxon signed-rank comparisons of each baseline vs. the proposed model on macro-F1. Two observations follow. First, while concatenation and XGBoost baselines are directionally slightly superior to cross-attention fusion, none of the differences achieve significance at 0.05 (with five folds, the minimum possible two-sided p-value is 0.0625) and thus all three of the top models may be considered statistically indistinguishable in performance quality. Secondly, in each of the five folds, the margin above (MRI-only baseline) of the macro-F1 score from our proposed (and other multimodal/clinical) models is consistently positive, as indicated by a confidence interval on the paired macro-F1 difference that lies entirely above zero.

Table 3. *Paired Wilcoxon signed-rank tests on macro-F1*

Comparison (vs proposed)	p-value	effect r	95% CI on Δ macro-F1
Concat − proposed	0.125	−0.87	[−0.0076, +0.0003] (indistinguishable)
XGBoost − proposed	0.0625	−1.00	[−0.0060, −0.0032] (XGBoost higher, n.s.)
MRI-only − proposed	0.0625	−1.00	[+0.0126, +0.0228] (proposed higher)

4.4. Confusion matrices, ROC and PR curves, and learning behavior

All 4 models' confusion matrices in one image, Figure 2. Across all models, the mass is concentrated at the NonDemented VeryMildDemented boundary where it follows the per-class F1 pattern; however, the MRI-only model presents a larger spread there compared to the clinical and fusion models, yielding tightening of that centroid. ROC (one-vs-rest) curves for all 4 models are shown in figure 3 (macro AUC 0.9964–0.9992), while the corresponding precision–recall curves, which—being more sensitive to the minority VeryMildDemented class than ROC—further separate the models, with our clinical and multimodal retaining high precision at high recall but the MRI-only model degrading earliest (macro PR-AUC 0.9845), are shown in figure 4.

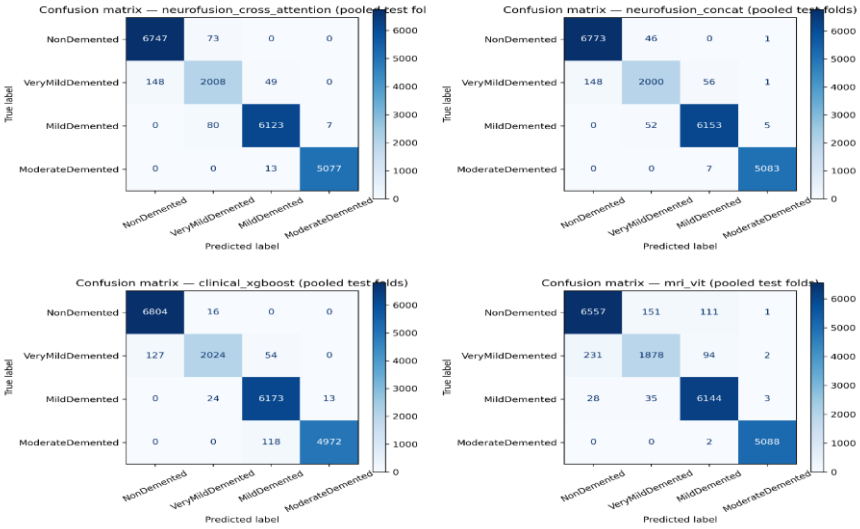

Figure 2. *Confusion matrices on the held-out test set*

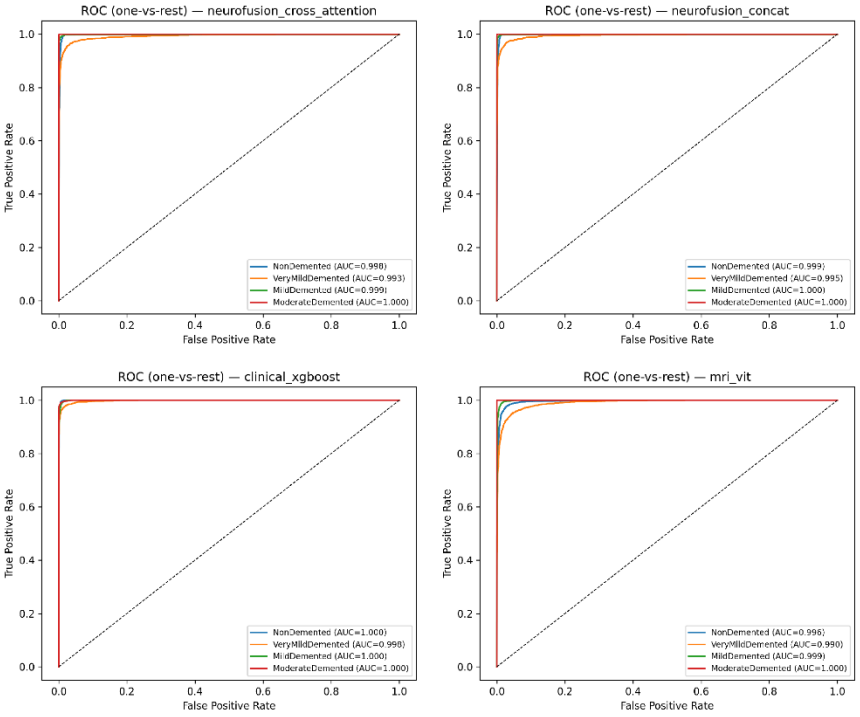

Figure 3. ***One-vs-rest ROC curves. Macro-averaged AUC ranges from 0.9964 (MRI-only) to 0.9992 (XGBoost).***

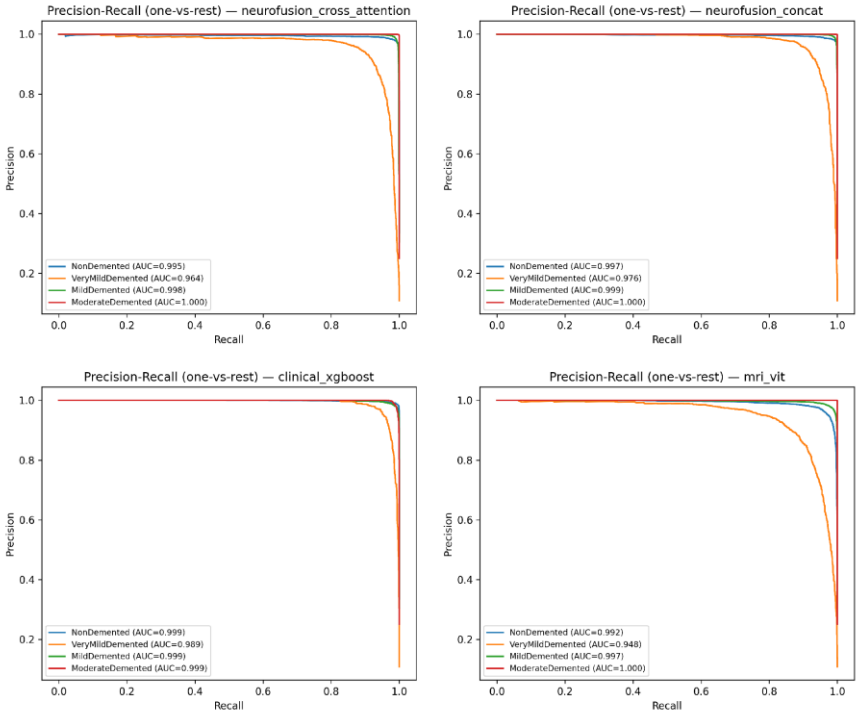

Figure 4. *One-vs-rest precision–recall curves*

4.5. Explainability

For the imaging pathway, we compute attention rollout [17] over the ViT backbone and average attention across layers to yield per-image saliency maps; in contrast with its gradient-based counterparts such as Grad-CAM [18], rollout is computed directly from the model's own attention matrices. An example of one representative per class in Figure 6. For the imaging branch, this gives a face-valid, architecture-native account of decision since attention focuses on medial-temporal and ventricular regions clinically associated with atrophy; the maps are most diffuse in the VeryMildDemented case, consistent with it being the hardest class.

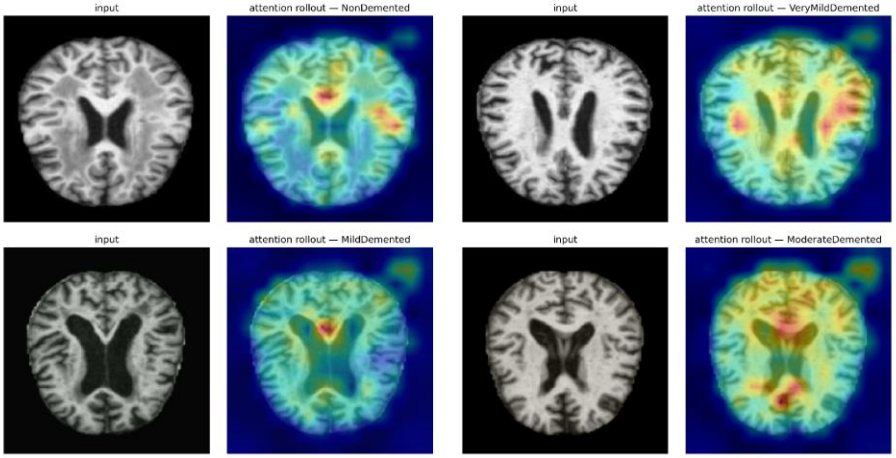

Figure. 4. *Attention-rollout saliency, one example per class (each pair: input slice left, attention overlay right). Top-left: NonDemented; top-right: VeryMildDemented; bottom-left: MildDemented; bottom-right: ModerateDemented.*

For the path of fusion, we tried to approximate relative modality contribution simply from the cross-attention weights (this was released as fused_modality_contribution. csv). In practice, however, the aggregated cross-attention mass for every class saturated to the maximum in both directions for all modalities on this benchmark, thus providing no informative per-modality decomposition. We instead report this result honestly, rather than a spurious balanced-contribution number: it is consistent with—and arguably symptomatic of—how the proposed fusion mechanism yields no significant gain over concatenation when trained on only class-level pairs between modalities (Section 5).

DISCUSSION

5.1. Why is VeryMildDemented the hardest

We log the lowest F1 on VeryMildDemented; all of the unimodal or multimodal models are finally at the top. Two compounding factors explain this. First, it is the least populated imaging class (3,136 of 28,160 images); therefore, the model views fewer examples. Second, and more spurious in a way, very mild dementia is nearest normal aging on a single 2D image slice [67], so

the imaging decision boundary against NonDemented is inherently thin. The MRI-only model has a steep drop in this class (0.8798) compared to the clinical and fusion models (0.92–0.95), isolating the challenge to the imaging side, with the confusion matrix (Fig 5c). 2) Otherwise, the residual error is dominated by this adjacent-stage boundary.

5.2. Interpreting the strong clinical-only baseline

Most notably, clinical-only XGBoost achieves the highest macro-F1 (0.9770) of all, with a confusion matrix (Fig. 2, bottom left) that is nearly diagonal. This should not be construed to mean that tabular biomarkers alone are sufficient for AD staging in practice. The classes are then much more cleanly separable than any real clinical cohort would be—as real biomarker profiles overlap substantially across adjacent stages, by being noisier and having missingness—the synthetic clinical table was generated from class-conditional distributions. As such, the near-ceiling XGBoost result is expected to be an artifact of the underlying data-generating process, not a clinically useful finding. We purposely keep and highlight this baseline, for the very success of this is indicative of what is wrong with the benchmark.

5.3. Why cross-attention does not beat concatenation here

On this benchmark, the proposed bidirectional cross-attention model cannot better simple feature concatenation; if anything, it is slightly behind (0.9724 vs 0.9761 macro-F1), and the difference is not statistically significant. This is aligned with the data design. Cross-attention is designed to leverage fine-grained, per-subject dependencies between imaging and clinical signals; however, our modalities are only class-paired, so there is no real per-subject cross-modal structure for attention to learn.

Then, in that regime of information-theoretic redundancy, the extra capacity and optimization burden imposed by cross-attention buy no improvement over concatenation; saturated and uninformative modality-contribution weights (Section 4.5) are the mechanistic footprint of that absence. For these experiments, we are not making any such claim, and expect the order may well vary with real pair-cohort data.

5.4. On the leakage controls

Perceptual-hash grouping is a pragmatic solution to deal with the absence of provenance metadata and its behaviour has had significant consequences: for one, a Hamming threshold of four produced transitive chaining that combined more than half of a class into single components (eliminating any group-aware split) whilst using a threshold of three with a hard size limit maintained groups (19,572 groups; 84.7% singletons; maximum realised cluster 97). Nonetheless, the method is heuristic.

It can miss pairs of augmentation siblings that differ by more than the threshold, and it can over-merge actual dissimilar but visually similar dark-background slices. So that absolute numbers need to be treated with caution due to some residual optimistic bias being possible. On its own, elimination of the deterministic global-CDR feature separated an immediate path for label-leakage from the clinical branch; remaining strong clinical performance was due to graded biomarkers rather than a reconstituted label.

CONCLUSION

What we introduce is NeuroFusionAI, a transformer-based, multimodal framework for four-class Alzheimer's staging, combining and fusing bidirectionally cross-attention ViT-Small MRI and FT-Transformer clinical branches, both tested under an intentionally conservative, disclosure-first protocol. Our main methodological contributions include a perceptual-hash group-leakage-control procedure demonstrated to prevent severe transitive chaining impelled by a weak threshold and the deliberate elimination of a deterministic label-proxy feature, alongside complete release of pairing and grouping tables for audit. While the proposed model reaches 98.18% accuracy and 0.9724 macro-F1 empirically, it is statistically indistinguishable from both a concatenation-fusion variant and a clinical-only XGBoost baseline.

Instead of overclaiming, we take these results at face value. Near-ceiling clinical-only performance is a result of class-conditional synthetic data; the lack of cross-attention benefit follows directly from the modality pairing being only class-level, leaving no inter-subject per-modal structure to leverage for attention. However, it does make clinically face-valid attention-rollout explanations from the imaging pathway. Accordingly, we frame this work not as one of multimodal domination, but as a reproducible benchmark with awareness of the potential for leakage and what can and cannot be inferred when one modality is synthetic and pairing is class-level. We identify ground truth paired cohorts of patients as the most important next step for establishing real, multimodal improvements.

REFERENCES

1. G. M. McKhann, D. S. Knopman, H. Chertkow, B. T. Hyman, C. R. Jack Jr., et al., "The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging–Alzheimer's Association workgroups," Alzheimer's & Dementia, vol. 7, no. 3, pp. 263–269, 2011.
2. C. R. Jack Jr., D. A. Bennett, K. Blennow, M. C. Carrillo, B. Dunn, et al., "NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease," Alzheimer's & Dementia, vol. 14, no. 4, pp. 535–562, 2018.
3. G. B. Frisoni, N. C. Fox, C. R. Jack Jr., P. Scheltens, and P. M. Thompson, "The clinical use of structural MRI in Alzheimer disease," Nature Reviews Neurology, vol. 6, no. 2, pp. 67–77, 2010.
4. G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, et al., "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60–88, 2017.
5. T. Jo, K. Nho, and A. J. Saykin, "Deep learning in Alzheimer's disease: Diagnostic classification and prognostic prediction using neuroimaging data," Frontiers in Aging Neuroscience, vol. 11, art. 220, 2019.

6. J. Wen, E. Thibeau-Sutre, M. Diaz-Melo, J. Samper-González, A. Routier, et al., “Convolutional neural networks for classification of Alzheimer’s disease: Overview and reproducible evaluation,” Medical Image Analysis, vol. 63, art. 101694, 2020.
7. E. E. Bron, S. Klein, J. M. Papma, L. C. Jiskoot, V. Venkatraghavan, et al., “Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer’s disease,” NeuroImage: Clinical, vol. 31, art. 102712, 2021.
8. J. Venugopalan, L. Tong, H. R. Hassanzadeh, and M. D. Wang, “Multimodal deep learning models for early detection of Alzheimer’s disease stage,” Scientific Reports, vol. 11, art. 3254, 2021.
9. G. Lee, K. Nho, B. Kang, K.-A. Sohn, and D. Kim, “Predicting Alzheimer’s disease progression using multi-modal deep learning approach,” Scientific Reports, vol. 9, art. 1952, 2019.
10. S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, “Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines,” npj Digital Medicine, vol. 3, art. 136, 2020.
11. M. Golovanevsky, C. Eickhoff, and R. Singh, “Multimodal attention-based deep learning for Alzheimer’s disease diagnosis,” Journal of the American Medical Informatics Association, vol. 29, no. 12, pp. 2014–2022, 2022.
12. Y. Liu, L. Fan, C. Zhang, T. Zhou, Z. Xiao, L. Geng, and D. Shen, “Incomplete multi-modal representation learning for Alzheimer’s disease diagnosis,” Medical Image Analysis, vol. 69, art. 101953, 2021.
13. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, et al., “Attention is all you need,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.
14. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” in Proc. International Conference on Learning Representations (ICLR), 2021.
15. F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, “Transformers in medical imaging: A survey,” Medical Image Analysis, vol. 88, art. 102802, 2023.
16. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
17. S. Abnar and W. Zuidema, “Quantifying attention flow in transformers,” in Proc. Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 4190–4197.
18. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in Proc. IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626.
19. T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2016, pp. 785–794.
20. Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, “Revisiting deep learning models for tabular data,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2021, pp. 18932–18943.
21. M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. International Conference on Machine Learning (ICML), 2019, pp. 6105–6114.
22. I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in Proc. International Conference on Learning Representations (ICLR), 2019.
23. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2009, pp. 248–255.
24. R. Wightman, “PyTorch Image Models (timm),” GitHub repository, 2019. doi:10.5281/zenodo.4414861.
25. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, et al., “PyTorch: An imperative style, high-performance deep learning library,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2019, pp. 8026–8037.
26. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, et al., “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
27. C. Zauner, “Implementation and benchmarking of perceptual image hash functions,” M.S. thesis, Upper Austria Univ. of Applied Sciences, Hagenberg, 2010.
28. Siddiqui, Md Ismail Hossain, et al. "Comparative analysis of explainable machine learning models for cancer classification using cytological features." Journal of Medical and Health Studies 4.5 (2023): 110-150.
29. Hossain, A. R. Lindon, M. Rahman, H. A. Khan, N. Tohfa, M. Tanvir, M. Shagar, J. E. Shakib, and M. R. I. Nasif, "Scalable Hybrid Ensemble Model for Intelligent DDoS Detection and Attack Categorization with a Web-Based Cybersecurity Intelligence Platform," Journal of Electrical Engineering, vol. 11, no. 5, 2022.
30. A. Sohel, M. A. Alam, A. Hossain, S. Mahmud, and S. Akter, "Artificial Intelligence in Predictive Analytics for Next-Generation Cancer Treatment: A Systematic Literature Review of Healthcare Innovations in the USA," Global Mainstream Journal of Innovation, Engineering & Emerging Technology, vol. 1, no. 1, pp. 62–87, 2022.