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.

TreatmentFormer: Multimodal Transformer Modeling and Biomarker Prioritization Following Rapamycin Treatment in Cognitively Normal Adults with Alzheimer's Disease Risk

Chris Wang
University of Missouri
Carter Woods
University of Missouri
Thong Nguyen
University of Missouri
Jian Liu
University of Missouri
Nathan Bresette
University of Missouri
Ai-Ling Lin
ai-ling.lin@health.missouri.edu
University of Missouri
Jianlin Cheng
University of Missouri

Article

Keywords: Rapamycin, Multimodal deep learning, Transformer, Contrastive learning, Random forest

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11662487/v1

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

Additional Declarations: No competing interests reported.

TreatmentFormer: Multimodal Transformer Modeling and Biomarker Prioritization Following Rapamycin Treatment in Cognitively Normal Adults with Alzheimer's Disease Risk

Chris Wang[1,2], Carter Woods[2], Thong Nguyen[1,2], Jian Liu[1,2], Nathan Bresette[2-4], Ai-Ling Lin[2-5*], Jianlin Cheng[1-3*]

[1]Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA
[2]Roy Blunt NextGen Precision Health, University of Missouri, Columbia, MO, USA
[3]Institute for Data Science and Informatics, University of Missouri, Columbia, MO, USA
[4]Department of Radiology, University of Missouri, Columbia, MO, USA
[5]Division of Biological Sciences, University of Missouri, Columbia, MO, USA

[*]Corresponding authors:
Ai-Ling Lin (ai-ling.lin@health.missouri.edu)
Jianlin Cheng (chengji@missouri.edu)

Abstract

Rapamycin, an FDA-approved medication with potential neuroprotective effects, is being investigated to mitigate Alzheimer's disease (AD) risk in individuals with high genetic risk factor, such as the apolipoprotein e4 (APOE4) alleles. Although conventional analyses identify changes in individual biomarkers induced by rapamycin, they do not rank their contributions to treatment-status classification. We developed *TreatmentFormer*, a multimodal deep learning framework integrating brain imaging, microbiome profiles, blood metabolites, hematologic parameters, inflammatory markers, and clinical variables. Random Forest-based feature filtering reduces input dimensionality, supervised contrastive learning aligns modality-specific representations, and a transformer integrates these representations to distinguish baseline from post-treatment observations. In 23 participants with or without APOE4, *TreatmentFormer* achieved a mean accuracy of 71.25% across 10 participant-level splits, compared with 58.75–60.00% for Random Forest and 55.00–58.75% for multilayer perceptron baselines. Blood metabolites, complete blood count, and inflammatory markers had the highest modality-level SHAP importance and accounted for 26 of the 27 highest-ranked Random Forest features. These features included markers associated with aging, inflammation, and metabolism. These findings support the potential of interpretable multimodal AI to prioritize candidate biomarkers and guide future studies of interventions in populations at elevated AD risk.

Keywords: Rapamycin, Multimodal deep learning, Transformer, Contrastive learning, Random forest

Introduction

Rapamycin (sirolimus), an FDA-approved inhibitor of the mechanistic target of rapamycin (mTOR) pathway, has emerged as a promising therapeutic candidate because of its anti-aging and neuroprotective properties [1]. mTOR plays a central role in regulating cellular growth, metabolism, and autophagy. Its inhibition has been shown to extend lifespan in animal models and improve vascular function, mitochondrial activity, and protein homeostasis [2].

In preclinical studies, we demonstrated that rapamycin reduced Alzheimer's disease (AD) pathology and improved memory in AD mouse models [3]. We also found that rapamycin increased cerebral blood flow (CBF), glucose uptake, improved mitochondrial function and reduced cerebral amyloid-β (Aβ) accumulation in mice expressing human apolipoprotein E4 allele (APOE4) [1,4,5], the strongest common genetic risk factor for late-onset AD. Consistent with these preclinical findings, a pilot trial of low-dose rapamycin in cognitively normal, middle-aged adults stratified by APOE genotype showed increased CBF in APOE4 carriers [6]. Rapamycin has also shown potential cognitive and immunological benefits, including improved cognitive performance and enhanced immune responses in older adults [7].

A key challenge in early clinical studies of Alzheimer's disease interventions is deciding which measures to carry forward into larger trials. Rapamycin is being investigated for its potential to protect brain health, but characterizing changes after treatment remains challenging [8] because genetic, metabolic, inflammatory, and vascular factors may interact in nonlinear ways [7]. Conventional analyses can identify changes in individual markers, while multimodal machine learning can assess how those markers contribute individually and together to distinguishing baseline from post-treatment measurements [9]. Applying deep learning in this way extends its role in neurodegeneration research beyond diagnostic classification [10–15] and can help prioritize candidate biomarkers for future controlled trials.

Previously, we developed APOEFormer [16], a multimodal deep learning framework for classifying APOE4 carrier status. APOEFormer uses modality-specific encoders to process heterogeneous data and self-supervised contrastive pretraining to learn cross-modal relationships. The resulting latent representations are concatenated and passed to a transformer that models

interactions across modalities, followed by a classification head that predicts APOE4 status. APOEFormer achieved 75% classification accuracy and identified features, including blood metabolites, associated with APOE4 carrier status.

In the present study, we extend the APOEFormer framework to characterize differences between baseline and post-rapamycin measurements and identify features associated with those differences. We formulate this objective as a binary classification task: given a participant's multimodal data at a specific time point, predict whether the measurements were obtained before or after rapamycin treatment. Each participant contributes two samples, one at baseline and one after treatment. This paired design requires careful handling of within-participant dependence, temporal variation, and potential data leakage.

To address these challenges, we introduce *TreatmentFormer* (**Figure 1**), a multimodal deep learning framework that adds Random Forest-based feature filtering to the APOEFormer architecture [16], which incorporates contrastive representation learning [17] and transformer-based integration [18]. Deep learning methods offer complementary tools for biomedical data integration: convolutional neural networks have been widely used for MRI-based AD classification [19], whereas multilayer perceptrons (MLPs) and autoencoders have been applied to tabular clinical and biomarker data [20]. *TreatmentFormer* extends this foundation by introducing an initial Random Forest feature-filtering step to reduce noise in high-dimensional tabular data while retaining information across modalities. Supervised contrastive learning [17] then aligns modality-specific embeddings within a shared latent space, and a transformer-based model captures cross-modal interactions to classify baseline and post-treatment samples.

TreatmentFormer is designed for small-sample, high-dimensional biomedical settings, in which conventional deep learning approaches often face challenges [21]. By combining feature filtering with multimodal representation learning, the framework aims to improve the robustness and interpretability of treatment-status classification. It also seeks to identify biological features and quantitative measurements associated with changes observed after rapamycin administration. These findings may generate hypotheses about treatment-related biological responses and help inform future studies of targeted prevention and treatment strategies.

Results

TreatmentFormer was evaluated on 23 cognitively normal participants, each contributing a baseline and a post-treatment observation, across 10 repeated participant-level splits (16 training, 3 validation, and 4 test participants per run, giving 8 test samples per run). Random Forest feature selection, contrastive pretraining, and model training were performed on the training partition of each run only; full details are given in the Methods.

Results of Treatment Prediction with/without Feature Filtering

To assess the impact of feature filtering, the *TreatmentFormer* architecture was first evaluated without Random Forest preprocessing. Under this setting, the model achieved an average test accuracy of approximately 50% across 10 runs, with results distributed around chance-level performance. This outcome suggests that the model is unable to effectively learn from the full set of high-dimensional inputs. The presence of noisy or weak features likely overwhelms the model's ability to identify meaningful patterns, reinforcing the need for feature selection in this setting. After introducing Random Forest-based feature filtering, the model demonstrates a substantial improvement in performance. Across 10 experimental runs, the model achieves a much higher average test accuracy of 71.25% (**Figure 2**).

The distribution of results includes: 3 runs at 75% accuracy, 3 runs at 87.5% accuracy, 2 runs at 62.5% accuracy, and 2 runs at 50% accuracy. Given the small test size (8 samples) of each test run, these values correspond to discrete increments based on the number of correct predictions, with extremes at 50% and 87.5%. Across the 10 repeated test splits, *TreatmentFormer* achieved a mean accuracy of 71.25% ± 14.49% SD, with individual-run accuracies ranging from 50.0% to 87.5%. Pooling the existing out-of-sample predictions yielded 57 correct predictions out of 80 (71.25%). As accuracy is limited to discrete increments at this sample size, we also report AUROC, which *TreatmentFormer* achieved a mean AUROC of 79.38% ± 11.80% across the 10 runs. Participant-level bootstrap resampling produced a 95% confidence interval of 62.86–79.41% for the observed accuracy. To assess whether this repeated-participant structure inflates the reported confidence interval, we estimated the effective sample size using a participant-level cluster bootstrap over the full 23-participant cohort: design effect = 0.70, corresponding to an effective sample size of approximately 114, at or above the raw pooled N of 80.

Because participants appeared in multiple test splits, participant-level bootstrap resampling was used to account for this dependence when estimating the confidence interval; the extent of this overlap is described in the Discussion. Across the 10 runs, 18 of the 23 participants appeared in at least one test set, so evaluation does not depend on a single fixed partition. The observed improvement in performance highlights the importance of feature filtering in small-sample multimodal learning. By reducing input dimensionality and removing noisy features, the model is better able to focus on informative signals and learn meaningful cross-modal relationships.

These findings reinforce the hypothesis that controlling feature space is critical for enabling transformer-based models to perform effectively in high-dimensional, low-sample biomedical settings.

Random Forest-based Feature Ranking

Feature importance scores generated by the Random Forest feature filtering were aggregated across the 10 runs to identify consistently informative tabular features (see the top 27 consensus features and their modalities in **Figure 3**). For each run, Random Forest feature selection was performed separately using only the training partition for that run, and the top 27 features were retained subject to the minimum-feature-per-modality constraint. Because the Random Forest rankings can differ across runs, the specific features selected within the 27-feature set could also differ. Furthermore, because the test samples were not mutually exclusive across the repeated runs, the runs should not be interpreted as statistically independent evaluations.

Although each individual run maintained the minimum of three selected features per modality, the aggregated set of features shown in the plot does not necessarily contain exactly three features from each modality. To generate the plot, we collected the top 27 features from each run and calculated the mean feature-importance value for each feature using only the runs in which that feature was selected. We then plotted the resulting aggregated feature-importance values, illustrating the individual features that were selected and their average importance across the repeated runs. This method of aggregation was chosen as opposed to accounting for each feature in every run. Although this approach may favor features that appeared in only a few runs but had

strong contributions, it keeps the importance values on a comparable range. Conversely, averaging over every run would favor features that appeared more consistently but may not have as strong a contribution.

To assess the stability of this feature list, we counted how many of the 10 runs selected each feature and computed the Jaccard similarity between every pair of the 10 per-run feature sets. Two features, the coefficient of variation of red blood cell distribution width (RDW-CV, %) and macrophage inflammatory protein 1 alpha (MIP-1α) were selected in all 10 runs. Four additional features were selected in at least 8 of 10 runs—macrophage inflammatory protein 1 beta (MIP-1β), tyrosine kinase with immunoglobulin-like and epidermal growth factor–like domains 2 (Tie-2), interferon gamma (IFN-γ), and the standard deviation of red blood cell distribution width (RDW-SD, fL). This pattern indicates consistent selection of hematological and immune-related markers. Mean pairwise Jaccard similarity across the 10 runs was 0.25 (45 pairs, range 0.15–0.38). This is consistent with, and compares favorably to, stability levels reported for feature selection on comparably small, high-dimensional biomedical datasets: Kalousis et al. [22] showed that feature-selection stability depends strongly on sample size and dimensionality, and Díaz-Uriarte and Alvarez de Andrés [23], running a bootstrap stability analysis of Random Forest gene selection across nine microarray datasets, found that selected genes were rarely reproduced in more than 50% of resamples. Our most consistently selected features exceed that benchmark. Full per-feature selection counts are given in **Supplementary Table S1**.

Across the top 27 tabular features, the majority originated from blood-based biomarker modalities, with particularly strong representation from inflammatory markers and metabolite profiles (**Figure 3**). Although complete blood count (CBC) and inflammatory markers dominated the top 7 features, blood metabolites had the most features in the top 27 with 14 features, meaning that while individual features were not strong, the modality as a whole plays a large role. In contrast, microbiome and demographic features were not consistently selected, suggesting that their contribution to the short-term treatment prediction may be more subtle.

Interestingly, only one feature was selected from the Brain cerebral blood flow (CBF) modality (Putamen), but the feature exhibited moderate importance ranked at 21. This suggests that while this modality may be less dense in informative features, certain signals within it are highly relevant.

SHAP-based Feature Analysis

To quantify the contribution of each modality and feature used in the deep learning model, SHAP (SHapley Additive exPlanations) analysis was performed on the test data during each run (**Figure 4**). Modality-level importance scores were computed by aggregating mean absolute SHAP values across tokens in each modality and averaging results over 10 runs.

The box plots show the distribution of modality-level mean absolute SHAP values across the 10 repeated runs, providing a measure of the variability in each modality's contribution to the model predictions. Individual points represent the SHAP value obtained in each run, while the box summarizes the distribution across runs (**Figure 4**). The corresponding mean absolute SHAP values and standard deviations across runs are reported in **Supplementary Table S2**, providing the numerical summary of modality-level importance and its variability. Modalities with higher mean absolute SHAP values contributed more strongly to the model predictions, while greater spread indicates greater variation in modality-level importance across repeated model fits.

The results indicate that blood-based modalities, including complete blood count and metabolite profiles, consistently exhibit the highest importance. Inflammatory markers also contribute strongly to the model's predictions. These findings suggest that systemic biological signals play a dominant role in distinguishing baseline and post-treatment states, which is consistent with the feature analysis based on the Random Forest feature filtering.

Structural MRI and CBF measurements showed moderate contributions within the current model. For structural MRI and perfusion imaging, which were represented using the CLIP-based encoder, the corresponding SHAP values should be interpreted as model-specific contributions rather than direct measures of the intrinsic biological importance of these imaging modalities.

Microbiome features also exhibit meaningful contributions, indicating a potential importance of host-microbe interactions. Perfusion imaging showed a smaller contribution within the current model.

In contrast, demographic and clinical variables show considerably lower importance, which is expected given that these features are largely static and less reflective of short-term treatment effects.

Contextually, results presented in the SHAP analysis are dependent on the features that were passed through the Random Forest filter. This requires the framing of these results not as two independent analyses arriving at similar conclusions, but rather as data and importance of features at different stages of the pipeline. While Random Forest feature selection directly influences the input feature space, SHAP analysis reflects the behavior of the downstream transformer model. The consistency between Random Forest feature rankings and SHAP importance suggests that the filtering stage successfully preserves predictively informative signals that are subsequently used by the downstream transformer model.

Importantly, the transformer is still able to model interactions across modalities beyond the selected features, indicating that feature filtering does not limit the model's capacity but rather improves its ability to focus on relevant information.

Discussion

This study extends our previous clinical investigation of low-dose rapamycin by applying multimodal artificial intelligence to identify patterns that distinguish baseline from post-treatment measurements [6]. Our original study used conventional statistical analyses to establish the direction and significance of changes in individual hematologic, metabolic, inflammatory, and vascular markers. TreatmentFormer complements those analyses by ranking the contributions of individual features and modalities to classification and quantifying how accurately their combined information distinguishes baseline from post-treatment samples.

Figure 3 ranks the features by their importance in distinguishing samples collected before and after rapamycin treatment. Among the top 10 features, RDW, a hematological measure associated with aging, and inflammatory markers provided the strongest signals. RDW-CV ranked first and RDW-SD ranked fifth, highlighting the contribution of red blood cell size variability to classification. MIP-1α and IFN-γ ranked second and third, respectively. Together with IL-10 and MIP-1β, these markers suggest that immune cell recruitment and cytokine regulation may help characterize the response to rapamycin. The metabolites indoleacetate, ribose, and dihydroxyphenylalanine also contributed to classification. Their inclusion is biologically plausible given the role of mechanistic target of rapamycin complex 1 (mTORC1) in nutrient utilization and biosynthesis, although whether rapamycin alters these specific metabolites through mTORC1 inhibition remains unclear. Overall, the ranking highlights the importance of rapamycin effects on red blood cell size variability, inflammatory signaling, and metabolism, consistent with literature [24–26].

The present findings are consistent with the hematologic and inflammatory changes reported in our original study [6]. RDW-CV decreased significantly in both APOE4 carriers and non-carriers, whereas MIP-1α decreased significantly in non-carriers. The identification of these markers as informative features in the current analysis supports their relevance to the changes observed following rapamycin administration. Statistical significance and predictive importance, however, address different questions. A marker with a significant average change may not be the strongest discriminator of treatment status, whereas a marker with a smaller or more variable change may contribute useful predictive information through its relationships with other measurements. TreatmentFormer therefore adds a complementary perspective by assessing these contributions within an integrated multimodal model.

RDW is of particular interest because it reflects variability in erythrocyte size and is routinely assessed in clinical hematology. The reduction observed in our original study is consistent with a previous report of RDW changes following rapamycin administration in older adults [24]. RDW has also been associated with biological aging, chronic inflammation, and oxidative stress [27], making its predictive contribution biologically plausible.

Combining Random Forest-based feature filtering with multimodal deep learning improved classification performance in this small, high-dimensional dataset. Across 10 experimental runs, *TreatmentFormer* achieved a mean test accuracy of 71.25%. The improvement following feature filtering highlights the value of reducing input dimensionality when the number of measurements is large relative to the number of participants. Although the results suggest that the model captures recurring patterns within this cohort, the small test sets and overlap in participants across runs limit conclusions about performance stability and generalizability.

Blood metabolites, complete blood count measurements, and inflammatory markers were the most influential modalities. These modalities had the three highest SHAP importance values (**Figure 4**) and accounted for 26 of the 27 highest-ranked features in the aggregate Random Forest analysis (**Figure 3**). Their prominence is consistent with the metabolic, hematologic, and inflammatory changes identified in our original study. The convergence of the Random Forest and SHAP results at the modality level further supports the predictive relevance of these data sources within the current framework.

Tabular CBF measurements ranked fourth in modality-level SHAP importance and accounted for one of the 27 highest-ranked Random Forest features. This contribution is consistent with the cerebrovascular response reported in our original study. Imaging and microbiome data showed smaller contributions within the current model, although their incremental predictive value requires evaluation through modality-ablation analyses. Demographic and clinical variables had the lowest SHAP values, which may partly reflect the paired study design: characteristics that remain unchanged between baseline and follow-up would be expected to contribute little to distinguishing the two time points. Low predictive importance in this setting therefore does not imply low clinical or biological relevance.

To evaluate whether the observed performance extended beyond feature selection alone, we compared *TreatmentFormer* with Random Forest and multilayer perceptron (MLP) classifiers. Both classifiers used the same Random Forest-selected features and participant-level training, validation, and test partitions as *TreatmentFormer* within each of the 10 runs. Each baseline was evaluated using tabular features alone and with CLIP imaging embeddings included in the

candidate feature pool. Full performance metrics, including precision, recall, specificity, F1 score, and AUROC, are reported in **Table 1**. Random Forest achieved mean test accuracies of 58.75% with tabular features alone and 60.00% with imaging embeddings, whereas the corresponding MLP accuracies were 58.75% and 55.00%. All were lower than *TreatmentFormer*'s mean accuracy of 71.25%.

These comparisons suggest that Random Forest-based feature selection alone does not fully explain TreatmentFormer's performance. However, they do not isolate the contributions of contrastive pretraining, transformer-based fusion, and encoder fine-tuning. The baseline classifiers used frozen, off-the-shelf CLIP embeddings, whereas *TreatmentFormer* incorporated encoder fine-tuning. Consequently, part of the performance difference may reflect adaptation of the imaging encoder. Component-level ablation studies are needed to determine the contribution of each architectural and training choice.

We have previously used Random Forest models to estimate brain age from structural MRI and medical data [28] and to classify clinical stages of Alzheimer's disease and rank amyloid, tau, and neurodegeneration biomarkers [29]. In this study, we developed TreatmentFormer to distinguish baseline from post-rapamycin measurements and identify the biomarkers and data modalities that contributed most to this distinction. TreatmentFormer builds on Random Forest by using it to reduce the number of tabular features while preserving representation across modalities. It then uses supervised contrastive learning to learn modality-specific representations and transformer attention to model relationships across modalities. This approach combines the ability of Random Forest to rank individual biomarkers with the ability of the transformer to integrate complementary information from imaging, microbiome, blood, and clinical measures. Modality-level SHAP analysis further shows which data sources contributed most to the final predictions. These results support the use of multimodal integration to characterize pre- to post-treatment differences and prioritize candidate biomarkers for evaluation in larger, controlled trials.

This study had three main limitations. First, without an untreated or placebo control group, treatment status coincided with visit timing. Held-out testing assessed classification in unseen participants from the same cohort, but the model could also capture changes related to time,

scanner variation, seasonality, or assay batch. The findings therefore describe differences between baseline and post-treatment measurements, with uncertain attribution to rapamycin. Second, CLIP ViT-B/32 was pretrained on natural image–text pairs, which differ from structural MRI and arterial spin labeling (ASL) perfusion images. Imaging SHAP values may therefore reflect both the information in the images and how well the encoder represents it. Comparisons with MRI-specific encoders would help clarify the contribution of imaging. Third, the 10 runs used overlapping training and test sets. Only 18 of 23 participants were tested at least once, and some appeared in up to six test sets. Training sets shared an average of 11.1 of 16 participants (69.2%). Within each run, feature selection, data-dependent preprocessing, and model training were confined to the training partition. The overlap across runs means that mean accuracy summarizes repeated evaluations within a single cohort. Larger, independent cohorts are needed to assess the reproducibility of classification performance and feature rankings.

Future studies should validate TreatmentFormer in larger, placebo-controlled cohorts with broader racial, ethnic, and APOE genotype representation. Additional follow-up visits could establish when the biomarker patterns emerge and how long they persist, while placebo comparisons could help distinguish rapamycin-associated changes from other changes over time. Comparisons with MRI-specific encoders and regional imaging analyses could clarify the contribution of imaging, while systematically removing individual modalities or model components could determine which inputs and modeling steps improve classification. Linking reproducible biomarker patterns to clinically meaningful outcomes would help determine whether TreatmentFormer can prioritize endpoints and support treatment monitoring in populations at elevated risk for Alzheimer's disease.

In conclusion, we showed that TreatmentFormer could distinguish measurements collected before and after low-dose rapamycin treatment in cognitively normal adults. Across 10 participant-level splits, the model achieved a mean test accuracy of 71.25%, outperforming the Random Forest and multilayer perceptron baselines. Blood metabolites, complete blood count measures, and inflammatory markers provided the strongest signals, with red blood cell distribution width and several cytokines among the top-ranked features. These results build on our earlier findings by showing how biomarkers measured together can help distinguish the two visits. If validated in

larger, placebo-controlled studies, this approach could help identify which biomarkers to measure in future trials aimed at reducing Alzheimer's disease risk.

Materials and Methods

Model Overview

The workflow of TreatmentFormer, a three-stage multimodal framework for modeling rapamycin treatment effects from heterogeneous biomedical data, is illustrated in **Figure 1**. The model builds upon the APOEFormer architecture, with key modifications to incorporate Random Forest-based feature filtering and new treatment-specific prediction.

The first stage focuses on reducing noise in the high-dimensional tabular data using Random Forest feature selection. The most informative features are retained while ensuring representation across modalities. In the second stage, modality-specific encoders transform raw inputs into latent embeddings. These embeddings are aligned in a shared latent space using a supervised contrastive learning objective, enabling the model to capture relationships across modalities at the subject level. In the third stage, the pretrained embeddings are treated as tokens and passed into a transformer-based model, which integrates information across modalities through self-attention mechanisms to produce a subject-level prediction of treatment status.

Stage 1: Random Forest-based Feature Filtering

When applying multimodal deep learning models to small biomedical datasets, a key challenge is the combination of high-dimensional inputs and limited sample size. In this setting, models are highly susceptible to overfitting and may struggle to distinguish meaningful patterns from noise. In our initial experiments, directly applying the APOEFormer base architecture to the treatment classification task without feature filtering resulted in near-random performance, with an average test accuracy of approximately 50%. This indicates that the model was unable to extract a reliable signal from the full set of input features due to high data dimensionality and small sample size. This failure is largely due to the presence of many of weak or noisy features, particularly within tabular modalities. These features increase the dimensionality of the input space, making it difficult for the transformer to effectively learn cross-modal interactions [30]. As a result, the model's

attention mechanism becomes diluted across irrelevant inputs, limiting its ability to focus on informative signals.

To address this, we add a feature filtering stage using Random Forest, designed to reduce noise while preserving the most informative features across modalities, which is described below in detail.

Random Forest-based Feature Filtering

Random Forest (RF) is an ensemble learning method composed of multiple decision trees trained using bootstrap sampling and feature subsampling [31]. At each split, a random subset of features is considered, and splits are selected to maximize the reduction in impurity, typically measured using Gini impurity. Feature importance is computed by aggregating the total impurity reduction contributed by each feature across all trees. This provides a natural and interpretable ranking of features based on their contribution to predictive performance.

In the context of TreatmentFormer, a Random Forest classifier with 800 trees was fitted independently to the training data for each run. The trees were grown using the default maximum depth setting, with no explicit depth constraint, and class weighting was set to `balanced_subsample` to account for class imbalance within each bootstrap sample. Feature importance was quantified using Mean Decrease in Impurity (MDI), also referred to as Gini importance, which measures the average reduction in node impurity attributable to each feature across the forest. The highest-ranked features according to MDI were then selected for downstream model training, subject to the minimum-per-modality constraint.

RF is particularly well-suited for high-dimensional biomedical tabular data, where features are numerous, heterogeneous, and often noisy [30]. Its use of bagging reduces variance and improves stability, making it robust in small-sample settings [31]. In addition, feature importance provides valuable insight to what passes through the filter, and helps see what individual features the filter deems important to the classification task [31].

Alternative feature selection methods, such as LASSO or principal component analysis (PCA), were considered but not adopted due to their limitations in this context [32]. LASSO enforces sparsity through linear regularization, which may fail to capture nonlinear relationships between features that are common in biomedical data. PCA, while effective for dimensionality reduction, produces transformed features that are difficult to interpret biologically and may obscure modality-specific information. Gradient boosting methods such as XGBoost offer strong predictive performance but can produce unstable feature importance rankings in small datasets due to their sequential learning process and sensitivity to hyperparameters [33]. In contrast, Random Forest provides a balance of robustness, interpretability, and stability. Its feature importance estimates are less sensitive to overfitting and better reflect generalizable patterns, making it well-suited for feature filtering in small-sample multimodal settings.

Feature Selection Strategy

In TreatmentFormer, the tabular features (i.e., non-imaging features such as microbiome profiles, blood-based biomarkers, cerebral blood flow measurements, and clinical variables) are fed through Random Forest, and following computation of feature importance scores, the top 27 tabular features were selected for downstream modeling. To ensure that the multimodal nature of the dataset is preserved, an additional constraint was imposed: at least three features from each modality must be included in the selected set. This prevents the model from collapsing onto a single dominant modality and encourages learning of cross-modal interactions. This design choice is particularly important in multimodal transformer architectures, where the ability to model relationships across modalities is central to performance [34]. By enforcing modality diversity, the model is better positioned to leverage complementary information from heterogeneous data sources.

This 27 feature threshold was determined solely from the validation performance plateau observed in the range of 25-35 features, where 27 sits in the said stable range and fits the constraint of 3 minimum features per modality. **Table 2** shows the validation performance observed, which was considered for choosing a candidate feature count. Notably, 27 was not the best-performing count in this range: 30 features achieved the highest validation accuracy (75.00%), compared to 68.33% at 27, which in fact was the median and aligned to the selection criteria that the number of features

be representative of the range. Had the threshold been chosen by performance alone, 30 would have been selected instead. We selected 27 because it falls within the observed plateau and satisfies the minimum-three-features-per-modality constraint, prioritizing a parsimonious feature count over the single highest-scoring one. Validation accuracy remained within a relatively narrow range across 25–35 selected features, with no monotonic improvement as the number of features increased. Given the limited validation sample size, these fluctuations likely reflect sampling variability rather than a statistically meaningful performance difference, supporting the use of a feature range rather than interpreting a single feature count as optimal.

Importantly, the Random Forest feature-selection procedure is independently refitted for each training partition. Thus, the top 27 features are newly determined from the training participants in each run rather than using a fixed feature set calculated from the full dataset. The selected features are then applied to the corresponding validation and test participants.

Baseline Model Comparison

To assess whether TreatmentFormer's performance is attributable to its architecture rather than to Random Forest-based feature selection alone, we compared it against two conventional classifiers trained on the same selected features. For each of the 10 experimental runs, a Random Forest classifier (800 trees, `class_weight="balanced_subsample"`) and a multilayer perceptron (two hidden layers of 128 and 64 units, ReLU activation, Adam optimizer) were trained on the same RF-selected features and the same patient-level training, validation, and test split used for TreatmentFormer in that run, evaluated both with the original tabular feature set and with CLIP imaging embeddings added to the candidate feature pool. Random Forest hyperparameters (maximum depth, minimum samples per leaf) were selected by grid search on each run's validation set; MLP training used early stopping based on validation accuracy (maximum 200 epochs, patience 20). We report accuracy, balanced accuracy, precision, recall (sensitivity), specificity, F1 score, and AUROC (summed across the 80 out-of-sample predictions from the 10 runs) for each baseline and feature configuration in **Table 1**.

SHAP-based Feature Importance

To characterize the contribution of each data modality to TreatmentFormer predictions, we performed a SHAP (SHapley Additive exPlanations) analysis using the trained attention-based model from each repeated test run. SHAP values were computed with `shap.GradientExplainer`, using a small background set drawn from the corresponding training dataset and then generating explanations for all test samples in that run. Because the model operates on multimodal token embeddings, SHAP values were initially obtained at the token and embedding-dimension level. For each token, the SHAP values were averaged across embedding dimensions to obtain a signed token-level contribution, while the absolute values were averaged across embedding dimensions to quantify the magnitude of the contribution independently of direction. Tokens were then grouped according to their corresponding modality, and the mean absolute SHAP values of the tokens within each modality were averaged to obtain a sample-level modality importance score. These sample-level scores were subsequently averaged across all test samples within each run to obtain a run-level mean absolute SHAP value for each modality.

SHAP values were calculated independently for each repeated train/test split, thereby preventing the interpretation from relying on a single model fit. To characterize the stability of modality importance across repeated runs, we summarized each modality using its mean absolute SHAP value across the 10 runs and the corresponding between-run standard deviation. We additionally visualized the distribution of modality-level SHAP values across runs using boxplots, with individual points representing individual runs. Because SHAP values quantify the contribution of model inputs to the model's predictions rather than causal effects, these results were interpreted as measures of model reliance on each modality and not as evidence that a modality causally influences treatment classification. For the imaging modalities, these SHAP values specifically quantify the contribution of the representations extracted by the CLIP-based encoder and therefore should not be interpreted as direct measures of the intrinsic biological information content of structural MRI or perfusion imaging.

Impact on Downstream Transformer Learning

Feature filtering provides several key benefits for the downstream transformer model. First, reducing the number of input features minimizes the influence of irrelevant or weak signals [30].

Without filtering, the model must implicitly learn to ignore noise, which can hinder training and reduce generalization performance. Second, high-dimensional inputs increase model variance, particularly in small datasets. This increases the risk of overfitting to spurious correlations rather than learning meaningful patterns [35]. Third, transformers compute pairwise interactions across all input tokens. As the number of tokens increases, the attention space grows quadratically, making it more difficult for the model to identify salient relationships [18]. By restricting the input to the most informative features, the model can more effectively allocate attention and learn meaningful cross-modal interactions. Together, these effects lead to the substantial improvement in performance observed when incorporating Random Forest feature filtering into the deep learning pipeline.

Stage 2: Modality Encoding and Alignment

Image feature encoding

Structural MRI and perfusion image volumes provide rich anatomical and functional information but present challenges due to their high dimensionality and variability across subjects. To address these challenges, a slice-wise encoding strategy leverages large-scale pretrained visual representations generated by a pretrained vision-language model, CLIP ViT-B/32 [36], while preserving depth-wise anatomical structure. Applying a readily accessible pretrained vision encoder offers the advantage of incorporation into our multimodal framework without requiring additional large-scale medical-imaging pretraining. Each 3D volume is decomposed along the axial dimension into an ordered sequence of 2D slices. This enables the use of pretrained 2D vision models while preserving depth-wise anatomical structure. Each slice is normalized to account for low-contrast regions, then converted into a three-channel image format. Feature embeddings are generated from the 2D slices using the pretrained CLIP ViT-B/32 model [36].

Because CLIP was pretrained on natural images rather than grayscale structural MRI or ASL perfusion data (see Discussion), the vision encoder is allowed to update fully during training rather than kept frozen, so it can adapt to this domain shift. The implementation includes a mechanism intended to progressively unfreeze additional layers over training; however, all vision-encoder parameters are made trainable at initialization, so this mechanism has no effect in practice, and the encoder is fully fine-tuned from the first training step. Nevertheless, the use of a natural-image

pretrained encoder introduces a domain gap, and the resulting imaging representations should be interpreted as model-specific features rather than as an optimal representation of the underlying medical images.

The resulting 512-dimensional embeddings generated by the CLIP ViT-B/32 model are projected into a compact latent space using a lightweight projection head consisting of linear transformations, normalization, and nonlinear activation. The final embeddings (hidden features) are normalized to ensure consistent scale and organized as a depth-ordered sequence, with padding applied to enable batch processing.

Tabular feature encoding

In addition to imaging data, the model incorporates multiple tabular modalities, including microbiome profiles, blood-based biomarkers, cerebral blood flow measurements, and clinical variables (e.g., gender, age, and body mass index [BMI]) [6]. There were no missing values in the six tabular modalities for the 23 patients included in the study, and all available tabular measurements for each patient were included in the analysis. Therefore, no imputation, missing-value exclusion, or other missing-data handling procedure was required for the tabular data.

As described in Stage 1, 27 features are selected from these modalities by the Random Forest module, with at least three features for each tabular modality. These tabular modalities differ in dimensionality and statistical distribution, which can lead to scale imbalance and suboptimal feature interactions. To address this, each modality is given a dedicated multi-layer perceptron (MLP) encoder that maps its raw input features into a shared latent embedding space.

Each encoder consists of stacked transformation blocks composed of linear projection, layer normalization, nonlinear activation, and dropout. A final normalization step ensures that all modality embeddings lie on a comparable scale, enabling stable interactions during multimodal fusion [37].

Supervised Contrastive Representation Learning for Modality Alignment

A key challenge in multimodal biomedical data integration is aligning heterogeneous modalities that differ in scale, noise characteristics, and semantic meaning. To encourage treatment-discriminative multimodal representations, we employed a supervised contrastive objective across the eight modality-specific embeddings generated for each observation (**Figure 1**). Each observation was defined by a participant and timepoint, with the corresponding treatment label assigned to all modality embeddings from that observation. Within each training batch, the eight modality embeddings from each observation were concatenated to form the contrastive embedding set. Pairwise similarities were calculated using the dot product scaled by a temperature parameter of $\tau = 0.1$. Embeddings associated with the same treatment label contributed to the positive term, whereas embeddings associated with different treatment labels contributed to the denominator as negative examples. Self-pairs were excluded. The contrastive objective therefore encouraged embeddings associated with the same treatment class to occupy similar regions of the representation space while separating embeddings from different treatment classes.

The sampling unit was an individual participant-timepoint observation, rather than an individual modality or image slice. Each participant contributed separate baseline and post-treatment observations, and both observations retained the same participant identity but could have different treatment labels. Importantly, positive pairs were defined solely by equality of the treatment labels and did not require the embeddings to originate from the same participant or timepoint. Thus, repeated measurements from the same participant were not explicitly paired as positives; they were treated according to their respective treatment labels. For a batch containing B observations, the eight modality embeddings per observation produced $N = 8B$ embeddings for the contrastive objective. The loss was calculated as

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\sum_{j\neq i,\ y_j=y_i}\exp\left(e_i^{\top}e_j/\tau\right)}{\sum_{j\neq i}\exp\left(e_i^{\top}e_j/\tau\right)}\right)$$

where e_i denotes an embedding, y_i its treatment label, and $\tau = 0.1$. Because participant identity and timepoint were not used to construct the positive or negative masks, the contrastive objective did not explicitly encourage representations to encode subject identity or timepoint. Participant-level partitioning was maintained separately during dataset splitting so that baseline and post-

treatment observations from a given participant remained within the same training, validation, or test partition.

Stage 3: Transformer-based Data Integration and Treatment Prediction

Multimodal Transformer Integration

Following supervised contrastive pretraining in Stage 2, modality-specific embeddings are treated as tokens and passed into a transformer-based architecture for treatment classification. Transformer models are particularly well-suited for this task due to their ability to capture complex dependencies through attention mechanisms [18]. By dynamically weighting features and modeling long-range interactions, transformers can effectively integrate heterogeneous information without relying on fixed receptive fields.

Each transformer layer consists of multi-head self-attention followed by residual connections, dropout, and layer normalization. There were 4 Transformer Layers, and 2 total Attention Heads with a Dropout rate of 0.3 (**Table 3**). Through stacked attention layers, the model learns higher-order interactions across modalities and adaptively reweights their contributions based on relevance to the prediction task.

To obtain a fixed-dimensional representation, the variable-length token sequence is aggregated using a learned attention pooling mechanism [38]. This approach assigns an importance score to each token and computes a weighted sum of token embeddings, enabling the model to focus on the most informative modalities for each subject.

Supervised Learning for Treatment Prediction

The transformer model is fine-tuned in a supervised manner to predict treatment classification. Each subject at a time point (prior to treatment or post-treatment) is represented as a sequence of modality embeddings and processed by the transformer to produce a classification output.

To prevent data leakage, training is conducted using patient-level splits, ensuring that samples from the same individual do not appear across training and evaluation sets. Optimization is performed using the AdamW optimizer with weight decay, gradient clipping, and mixed-precision

training to improve stability and computational efficiency. AdamW learning rate is set at 1×10^{-4}, and weight decay at 5×10^{-3} (**Table 3**).

A focal loss objective was employed [39] to down-weight well-classified examples and place greater emphasis on difficult or misclassified samples. The positive- and negative-class weighting parameters were set to 0.9 and 0.1, respectively, before sigmoid transformation in the loss function, corresponding to effective coefficients of approximately 0.71 and 0.52. The slightly larger positive-class weight was intended to increase sensitivity to post-treatment samples rather than to correct for class imbalance, as baseline and post-treatment observations were approximately balanced by study design. The focusing parameter $\gamma = 7.0$ was selected empirically by trial and error during preliminary model development to place greater emphasis on difficult examples; no formal hyperparameter search was performed for this parameter.

Experimental Setup

Dataset

The multimodal biomedical dataset used in this work consists of heterogeneous data collected from 23 patients in Columbia, Missouri [6]. They were cognitively normal adults aged 45-65 years, with MoCA scores ≥ 26 and Clinical Dementia Rating scores of 0; individuals with MCI, dementia, or AD were excluded. The cohort included nine APOE4 carriers (eight APOE3/APOE4 and one APOE4/APOE4) and 14 non-carriers (12 APOE3/APOE3 and two APOE2/APOE3). The dataset reflects the multifactorial nature of neurodegenerative disease progression and includes both imaging and tabular modalities [6].

A total of eight modalities were used, including two imaging modalities (structural MRI, perfusion imaging) and six tabular modalities (inflammatory markers [40 features], blood metabolites [104 features], cerebral blood flow (CBF) measurements [18 features], microbiome data [728 features], complete blood count (CBC) [46 features], and clinical/demographic variables [5 features]). Structural MRI provides detailed anatomical information, while perfusion imaging (obtained using arterial spin labeling (ASL)) captures functional blood flow dynamics. Structural MRI has been widely used for quantifying neurodegeneration and regional atrophy patterns in aging and neurological disorders [40]. ASL is a non-invasive MRI technique that quantifies cerebral blood

679 flow (CBF) by magnetically labeling inflowing arterial blood water, enabling voxel-wise perfusion
680 mapping without contrast agents [41]. The tabular modalities capture systemic biological
681 processes, including metabolism, inflammation, and host-microbiome interactions.
682
683 Participants received oral rapamycin (sirolimus; 1 mg/day) for four weeks [6]. Adherence was
684 assessed by medication accountability, diary review, and circulating rapamycin concentrations.
685 All 23 enrolled participants completed the intervention, with no post-enrollment attrition. Each
686 patient contributes two samples corresponding to pre-treatment (Baseline) and post-treatment
687 (Post) timepoints. As a result, the dataset contains paired samples per subject, enabling analysis of
688 treatment-induced changes over time. This is a main difference in data preprocessing compared to
689 the APOEFormer model for APOE4 allele prediction in our previous study [16]. In the APOE4
690 classification task, the genotype is fixed per patient, which means the patient's APOE classification
691 does not change over time or as a result of the rapamycin treatment. However, in this work, the
692 same patient has two treatment statuses at two different time points (prior to rapamycin treatment
693 and post-treatment).
694
695 ***Ethical Considerations***
696 This study involved a secondary analysis of data collected under a previously approved clinical
697 trial. The original clinical trial was approved by the University of Missouri Institutional Review
698 Board (IRB #2091042; approved August 12, 2022) and conducted in accordance with Good
699 Clinical Practice guidelines. The trial was registered at ClinicalTrials.gov (NCT05386914) on May
700 20, 2022. All participants provided written informed consent prior to participation. The present
701 analysis used de-identified data collected under this approved protocol. All methods were
702 performed in accordance with the relevant guidelines and regulations.
703
704 ***Data Partitioning and Data Leakage Prevention***
705 To evaluate the proposed framework rigorously, experiments were conducted using patient-level
706 data splits to prevent data leakage. The dataset was partitioned into 16 patients for training, 3 for
707 validation, and 4 for testing. All timepoints (Baseline and Post-Treatment) from a given
708 participant remain in the same partition. Model training, including contrastive pretraining and
709 CLIP fine-tuning, is performed using the training partition, while the validation partition is used

for checkpoint selection and early stopping. The complete order of feature selection, preprocessing, representation learning, model selection, and final testing is illustrated in **Figure 5**, with each stage restricted to the data partition available at that point. The test partition remains unavailable until final evaluation and is not used for feature selection, preprocessing, model training, or model/threshold selection.

Because each patient contributes both baseline and post-treatment samples, the test set contains 8 total samples per run. Patient-level partitioning was necessary because having same patient data split between the training and testing set can lead to inflated results due to same patient data memorization and the model learning patterns that are not intrinsic to the treatment impact.

To further improve robustness, 10 experimental runs were performed with different randomized patient splits. Each of the 23 participants was eligible for selection into the test set in each of the 10 runs. This repeated evaluation strategy provides a more reliable estimate of model performance in small-sample settings. Model performance was measured using classification accuracy. Due to the small size of the test set, accuracy changes in discrete increments of 12.5%, corresponding to a single misclassification. This granularity is important when interpreting variability across runs. To quantify uncertainty in model performance across the repeated participant-level test splits, we performed 10,000 participant-level bootstrap resamples of the existing out-of-sample predictions. Participants, rather than individual observations, were resampled to account for the repeated appearance of participants across test splits, with all corresponding out-of-sample predictions retained for each sampled participant. The 95% confidence interval for pooled accuracy was calculated from the resulting bootstrap distribution.

Data availability

The code for TreatmentFormer is available at https://github.com/jianlin-cheng/TreatmentFormer. The participant data analyzed in this study are available from the corresponding author (A.-L.L.) on reasonable request.

Funding

This work was supported by a Curators' professorship and the Paul K. and Diane Shumaker Professorship to J.C. and the MU Radiology and NextGen start-up funding to A.-L.L.

Author contributions

J.C. and A.-L.L. conceptualized the study. J.C. and C.W. developed the AI methodology. A.-L.L. and C.Wo. provided patient data. C.W., C.Wo., T.N., and J.L. developed the software. C.W., T.N., J.L., and C.Wo. performed the computational experiments and collected data. C.W., T.N., C.Wo., J.L., N.B., A.-L.L., and J.C. validated the results and contributed to writing the manuscript. All authors reviewed the manuscript.

Additional information

Competing interests: The authors declare no competing interests.

References

1. Lin, A.-L. *et al.* APOE genotype-dependent pharmacogenetic responses to rapamycin for preventing Alzheimer's disease. *Neurobiol. Dis.* **139**, 104834 (2020).
2. Harrison, D. E. *et al.* Rapamycin fed late in life extends lifespan in genetically heterogeneous mice. *Nature* **460**, 392–395 (2009).
3. Lin, A.-L. *et al.* Chronic rapamycin restores brain vascular integrity and function through NO synthase activation and improves memory in symptomatic mice modeling Alzheimer's disease. *J. Cereb. Blood Flow Metab.* **33**, 1412–1421 (2013).
4. Lin, A.-L. *et al.* Rapamycin rescues vascular, metabolic and learning deficits in apolipoprotein E4 transgenic mice with pre-symptomatic Alzheimer's disease. *J. Cereb. Blood Flow Metab.* **37**, 217–226 (2017).
5. Sanganahalli, B. G. *et al.* mTOR inhibition enhances synaptic and mitochondrial function in Alzheimer's disease in an APOE genotype-dependent manner. *J. Cereb. Blood Flow Metab.* **44**, 1745–1758 (2024).
6. Aware, C. *et al.* Rapamycin increases cerebral blood flow and modulates metabolic, inflammatory, and microbiome profiles in healthy middle-aged APOE4 carriers: A pilot single-arm trial. *J. Cereb. Blood Flow Metab.* https://doi.org/10.1177/0271678X261490342 (2026).
7. Mannick, J. B. *et al.* mTOR inhibition improves immune function in the elderly. *Sci. Transl. Med.* **6**, 268ra179 (2014).
8. Long, J. M. & Holtzman, D. M. Alzheimer disease: An update on pathobiology and treatment strategies. *Cell* **179**, 312–339 (2019).
9. Esteva, A. *et al.* A guide to deep learning in healthcare. *Nat. Med.* **25**, 24–29 (2019).
10. Toumaj, S., Heidari, A., Shahhosseini, R. & Jafari Navimipour, N. Applications of deep learning in Alzheimer's disease: A systematic literature review of current trends, methodologies, challenges, innovations, and future directions. *Artif. Intell. Rev.* **58**, 44 (2024).
11. Afifi, I., Elgendy, M., Abdelfatah, M. & El-Sappagh, S. Vision and convolutional transformers for Alzheimer's disease diagnosis: A systematic review of architectures, multimodal fusion and critical gaps. *Brain Inform.* **13**, 1 (2025).
12. Ali, S., Piana, M., Pardini, M. & Garbarino, S. Graph neural networks in Alzheimer's disease diagnosis: A review of unimodal and multimodal advances. *Front. Neurosci.* **19**, 1623141 (2025).
13. Tripathy, R. K. *et al.* Effective integration of multi-omics with prior knowledge to identify biomarkers via explainable graph neural networks. *npj Syst. Biol. Appl.* **11**, 43 (2025).
14. Guo, H., Yang, Z., Zhang, G., Lv, L. & Zhao, X. Meta analysis of the diagnostic efficacy of transformer-based multimodal fusion deep learning models in early Alzheimer's disease. *Front. Neurol.* **16**, 1641548 (2025).
15. Jasodanand, V. H. *et al.* AI-driven fusion of multimodal data for Alzheimer's disease biomarker assessment. *Nat. Commun.* **16**, 7407 (2025).
16. Nguyen, T. *et al.* APOEFormer: A multimodal AI model for APOE4 carrier status prediction and phenotype characterization in Alzheimer's disease. *Bioengineering* **13**, 1113 (2026).
17. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. in *37th International Conference on Machine Learning (ICML)* vol. 119 1597–1607 (PMLR, 2020).
18. Vaswani, A. *et al.* Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 5998–6008 (2017).

19. Hosseini-Asl, E., Gimel'farb, G. & El-Baz, A. Alzheimer's disease diagnostics by a deeply supervised adaptable 3D convolutional network. Preprint at https://arxiv.org/abs/1607.00556 (2016).
20. Venugopalan, J., Tong, L., Hassanzadeh, H. R. & Wang, M. D. Multimodal deep learning models for early detection of Alzheimer's disease stage. *Sci. Rep.* **11**, 3254 (2021).
21. Varoquaux, G. Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage* **180**, 68–77 (2018).
22. Kalousis, A., Prados, J. & Hilario, M. Stability of feature selection algorithms: A study on high-dimensional spaces. *Knowl. Inf. Syst.* **12**, 95–116 (2007).
23. Díaz-Uriarte, R. & Alvarez de Andrés, S. Gene selection and classification of microarray data using random forest. *BMC Bioinform.* **7**, 3 (2006).
24. Kraig, E. *et al.* A randomized control trial to establish the feasibility and safety of rapamycin treatment in an older human cohort: Immunological, physical performance, and cognitive effects. *Exp. Gerontol.* **105**, 53–69 (2018).
25. Mannick, J. B. & Lamming, D. W. Targeting the biology of aging with mTOR inhibitors. *Nat. Aging* **3**, 642–660 (2023).
26. Saxton, R. A. & Sabatini, D. M. mTOR signaling in growth, metabolism, and disease. *Cell* **168**, 960–976 (2017).
27. Yadav, S., Deepika & Maurya, P. K. A systematic review of red blood cells biomarkers in human aging. *J. Gerontol. A Biol. Sci. Med. Sci.* **79**, glae004 (2024).
28. Woods, C., Xing, X., Khanal, S. & Lin, A.-L. Machine learning-driven prediction of brain age for Alzheimer's risk: APOE4 genotype and gender effects. *Bioengineering* **11**, 943 (2024).
29. Hammond, T. C. *et al.* β-amyloid and tau drive early Alzheimer's disease decline while glucose hypometabolism drives late decline. *Commun. Biol.* **3**, 352 (2020).
30. Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3**, 1157–1182 (2003).
31. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
32. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58**, 267–288 (1996).
33. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
34. Ngiam, J. *et al.* Multimodal deep learning. in *28th International Conference on Machine Learning (ICML)* 689–696 (2011).
35. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. (Springer, New York, NY, USA, 2009). doi:10.1007/978-0-387-84858-7.
36. Radford, A. *et al.* Learning transferable visual models from natural language supervision. in *38th International Conference on Machine Learning (ICML)* vol. 139 8748–8763 (PMLR, 2021).
37. Ba, J. L., Kiros, J. R. & Hinton, G. E. Layer normalization. Preprint at https://arxiv.org/abs/1607.06450 (2016).
38. Ilse, M., Tomczak, J. & Welling, M. Attention-based deep multiple instance learning. in *35th International Conference on Machine Learning (ICML)* vol. 80 2127–2136 (PMLR, 2018).
39. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. in *IEEE International Conference on Computer Vision (ICCV)* 2999–3007 (2017). doi:10.1109/ICCV.2017.324.

40. Jack, C. R. *et al.* Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurol.* **9**, 119–128 (2010).
41. Alsop, D. C. *et al.* Recommended implementation of arterial spin-labeled perfusion MRI for clinical applications: A consensus of the ISMRM perfusion study group and the European consortium for ASL in dementia. *Magn. Reson. Med.* **73**, 102–116 (2015).

Table 1. Random Forest and MLP baseline performance (10 runs, mean ± SD)

Model	Accuracy	Precision	Recall	Specificity	F1	AUROC
RF, tabular only	58.75 ± 9.76	0.612	0.675	0.500	0.598	0.738
RF, tabular + CLIP	60.00 ± 9.35	0.632	0.650	0.550	0.616	0.694
MLP, tabular only	58.75 ± 8.00	0.529	0.575	0.600	0.538	0.594
MLP, tabular + CLIP	55.00 ± 10.00	0.577	0.525	0.575	0.487	0.644

Table 2. Validation accuracy across the evaluated feature-selection range.

Number of Features	Validation Accuracy (%)
25	70.00
26	70.00
27	68.33
28	63.33
29	68.33
30	75.00
31	73.33
32	65.00
33	65.00
34	60.00
35	61.67

Table 3. Hyperparameters and training settings for the TreatmentFormer model.

Parameter	Value
Transformer Layers	4
Attention Heads	2
Maximum Fine-tuning Epochs	1000
Early Stopping Patience	500
Batch Size	128
Embedding Dimension	128
Dropout	0.3
CLIP/Projection Learning Rate	1×10^{-3}
AdamW Learning Rate	1×10^{-4}
Weight Decay	5×10^{-3}
Focal Loss raw α_{pos}	0.9
Focal Loss raw α_{neg}	0.1
Focal Loss γ	7.0
Classification Threshold	0.5
Contrastive Temperature τ	0.1
Gradient Clipping	Applied
Mixed-Precision Training	Applied
Validation-Based Early Stopping	Applied

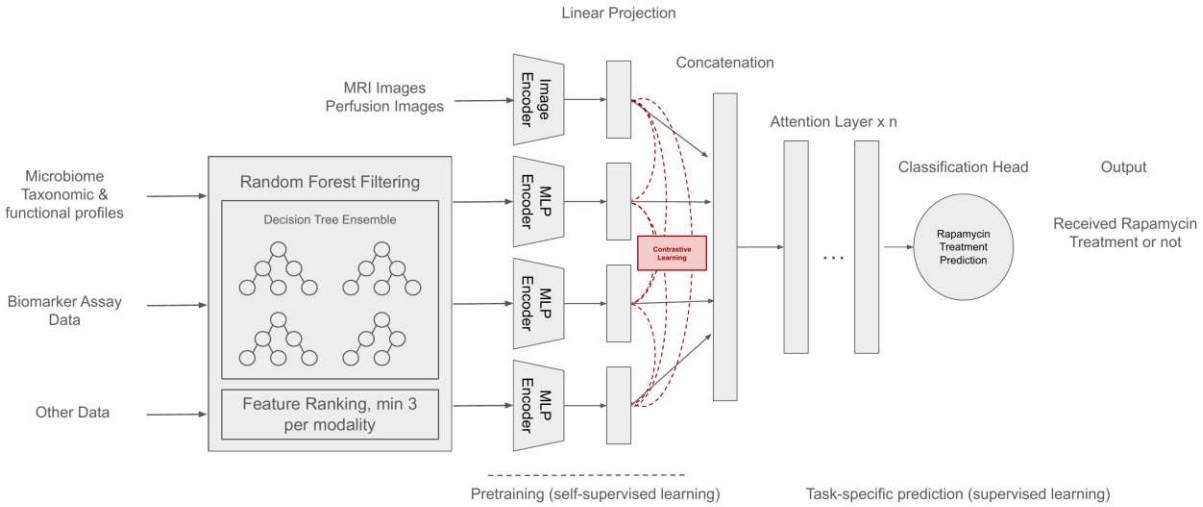

Figure 1. ***Overview of TreatmentFormer***. A three-stage multimodal AI framework for rapamycin treatment classification with Random Forest-based feature filtering.

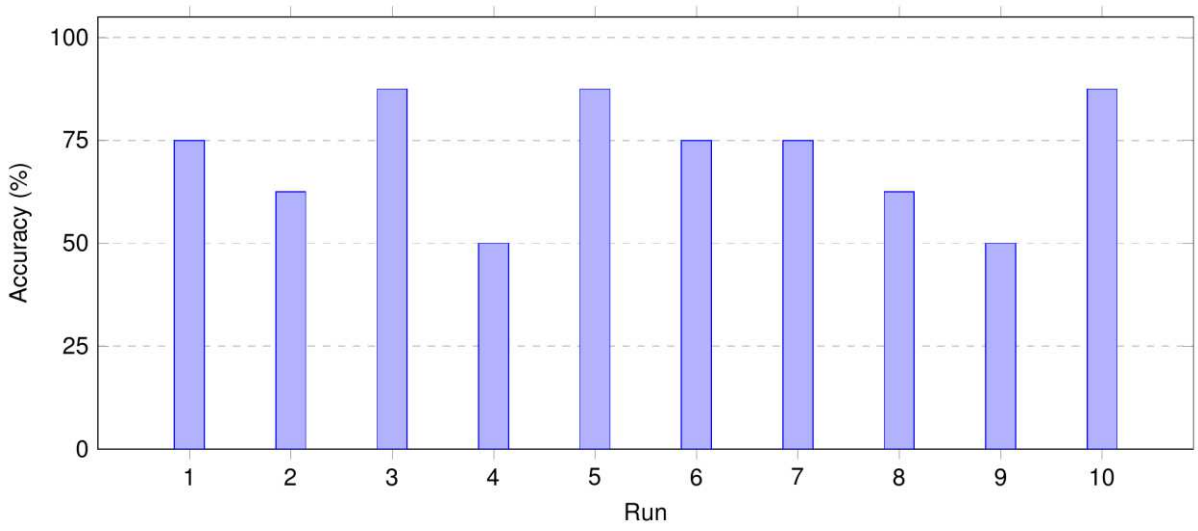

Figure 2. Rapamycin treatment classification accuracy across 10 experimental test runs. Performance variability reflects differences in data partitioning and sample composition across runs.

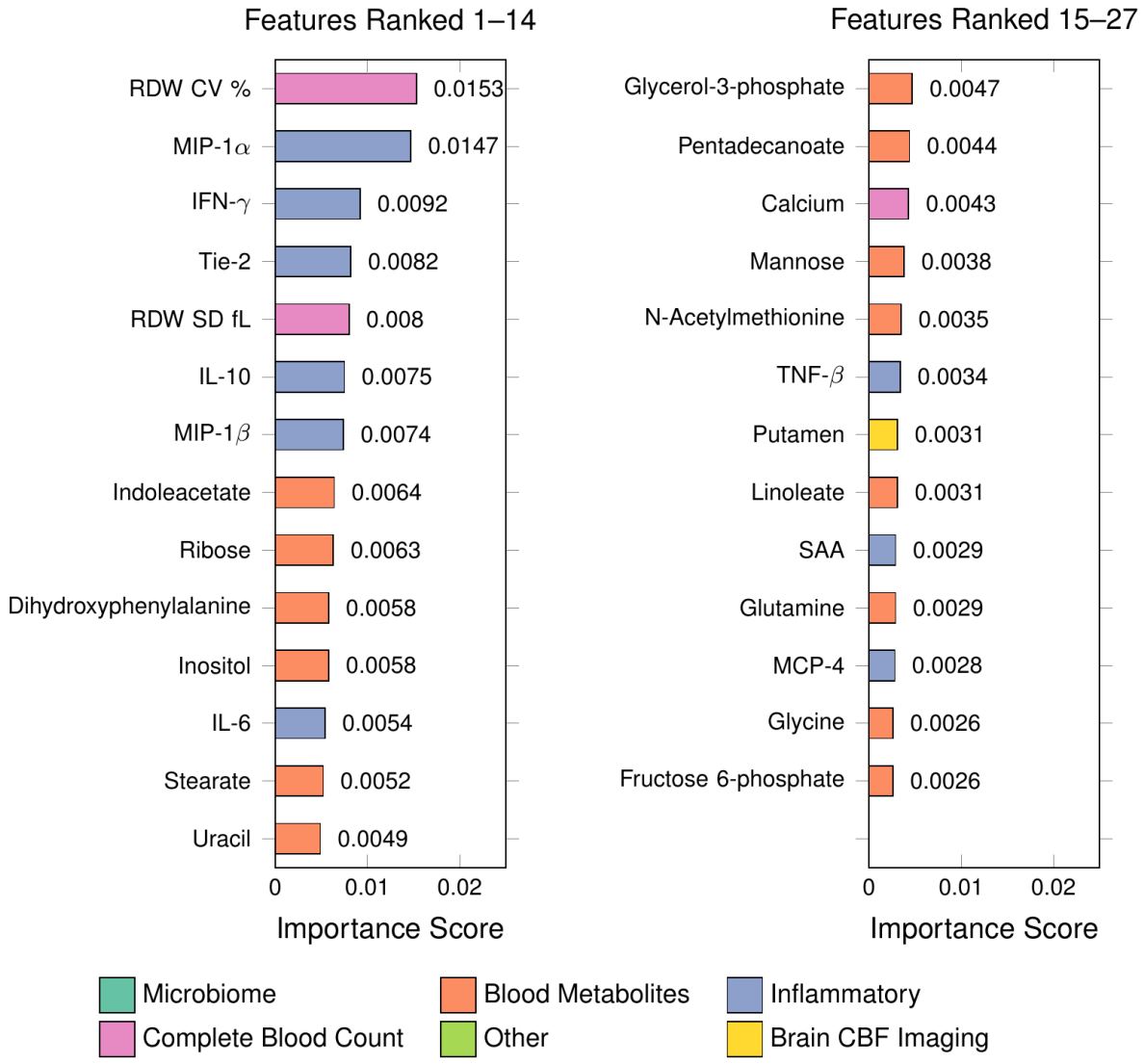

Figure 3. Top 27 features aggregate summary ranked by importance, colored by modality. Features generated through collecting top 27 features from each run. Mean importance for each feature calculated by using only the runs in which that feature was selected. Abbreviations: CBF, cerebral blood flow; fL, femtoliters; IFN-γ, interferon gamma; IL-6, interleukin 6; IL-10, interleukin 10; MCP-4, monocyte chemoattractant protein 4; MIP-1α, macrophage inflammatory protein 1 alpha; MIP-1β, macrophage inflammatory protein 1 beta; RDW CV %, red blood cell distribution width–coefficient of variation, expressed as a percentage; RDW SD fL, red blood cell distribution width–standard deviation, measured in femtoliters; SAA, serum amyloid A; Tie-2, tyrosine kinase with immunoglobulin-like and epidermal growth factor–like domains 2; TNF-β, tumor necrosis factor beta (lymphotoxin alpha).

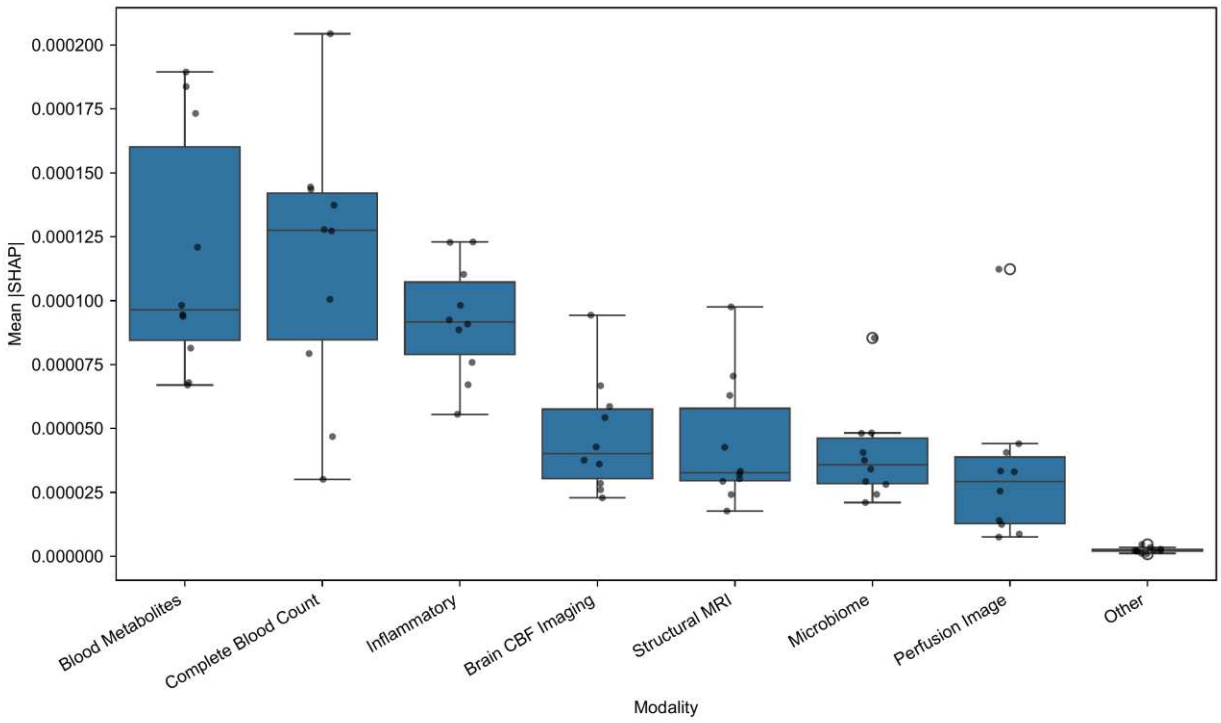

Figure 4. Variability of modality-level SHAP values across repeated runs. Boxplots show the distribution of mean absolute SHAP values across 10 repeated runs for each modality, with individual points representing individual runs. Modalities are ordered by their average mean absolute SHAP value across runs. Higher values indicate greater contribution to model predictions.

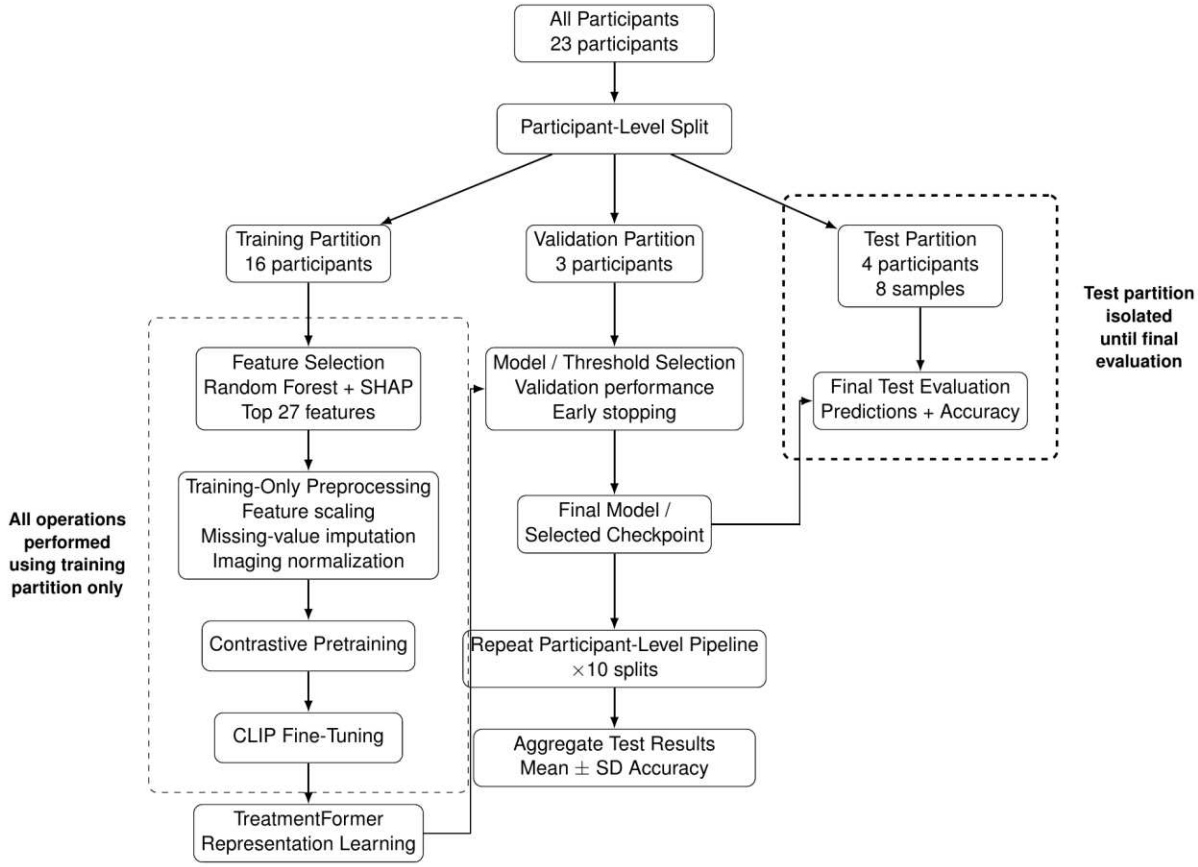

Figure 5. Participant-level evaluation pipeline. In each run, participants are first partitioned into training, validation, and test sets. Feature selection (including selection of the 27 features), feature scaling, missing-value imputation, imaging normalization, contrastive pretraining, and CLIP fine-tuning are performed using the training partition only. The validation partition is used for model selection, early stopping, and threshold selection. The test partition remains isolated until the final evaluation. The complete procedure is repeated across 10 participant-level splits.

Supplementary Files

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

- 2SupplementaryInformation.pdf