Statistically valid explainable black-box machine learning: applications in sex classification across species using brain imaging

Tingshan Liu[†1], Jayanta Dey[†1], Beiya Xu[†1], Eric W. Bridgeford[2], Samuel Alldritt[3], Karl-Heinz Nenning[3], Kyoungseob Byeon[3], Ting Xu[3], and Joshua T. Vogelstein[*4]

[1]Department of Biomedical Engineering, Johns Hopkins University
[2]Stanford University
[3]Child Mind Institute
[4]Department of Biomedical Engineering, Institute for Computational Medicine, Kavli Neuroscience Discovery Institute, Johns Hopkins University

ABSTRACT

Sex classification using neuroimaging data has the potential to revolutionize personalized diagnostics by revealing subtle structural brain differences that underlie sex-specific disease risks. Despite the promise of machine learning, traditional methods often fall short in providing both high classification accuracy and interpretable, statistically validated feature importance scores for high-dimensional imaging data. This gap is particularly evident when conventional techniques such as random forests, LIME, and SHAP are applied, as they struggle with complex feature interactions and managing noise in large datasets. We address this challenge by developing an integrated framework that combines Oblique Random Forests (ORFs) with a novel, permutation-based feature importance testing algorithm. ORFs extend traditional random forests by employing oblique decision boundaries through linear combinations of features, thereby capturing intricate interactions inherent in neuroimaging data. Our feature importance testing method, NEOFIT, rigorously quantifies the significance of each feature by generating null distributions and corrected p-values. We first validate our approach using simulated datasets, establishing its robustness and scalability under controlled conditions. We then apply our method to classify sex from both voxel-wise structural MRI and cortical thickness data in humans and macaques, facilitating direct cross-species comparisons. ORFs achieves AUC > 0.80 on human data, and > 0.70 on macaque data, while NEOFIT identifies statistically significant features aligned with sex-dimorphic neuroanatomy. Our results demonstrate that the proposed framework not only enhances classification performance but also provides clear, interpretable insights into the neuroanatomical features that distinguish sexes. These methodological advancements pave the way for improved diagnostic tools and contribute to a deeper understanding of the evolutionary basis of sex differences in brain structure.

1 INTRODUCTION

The opportunity to harness high-dimensional medical imaging data to improve diagnostic precision and treatment outcomes is immense. Prior studies in radiography and ultrasound have reported clinically relevant diagnostic performance from Machine Learning-based pipelines[1, 2]. In neuroimaging, structural Magnetic Resonance Imaging (MRI) measurements offer a wealth of information that can reveal subtle brain differences linked to neurodevelopmental, cognitive, and behavioral traits, as well as susceptibility to disorders such as autism, schizophrenia, and Alzheimer's disease [3–9]To realize this promise in clinical settings, methods are needed that not only deliver high predictive accuracy but also provide interpretable, statistically validated insights into which features are truly significant.

A major challenge in neuroimaging classification is ensuring biological interpretability. Feature selection and rigorous statistical validation are crucial when models handle thousands of features. While popular post-hoc explanation techniques such as Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) offer valuable insights [10, 11], they face challenges in high-dimensional settings. LIME

[*]corresponding: jvogels3@jh.edu; [†] denotes equal contribution.

relies on local linear approximations, which can lead to variability in the explanations when the feature space is large and complex. Similarly, although SHAP's Shapley values are theoretically robust, their computation in high-dimensional data can be computationally intensive and may result in noisy estimates. A critical gap remains in the absence of a robust method that consistently discriminates genuine signal from noise in high-dimensional neuroimaging data.

We address this gap by developing an integrated framework that combines Oblique Random Forests (ORFs) with a permutation-based feature importance testing algorithm. ORFs, which use linear combinations of features at each split [12], are particularly adept at capturing complex interactions inherent in neuroimaging data. Our feature importance testing framework statistically validates feature contributions by generating null distributions and applying Bcorrected p-values, thereby reducing overfitting and enhancing interpretability.

To systematically evaluate our approach, we first benchmark it on simulated datasets to assess its effectiveness in feature selection under controlled conditions. We then apply it to human and macaque neuroimaging data to classify biological sex using both voxel-wise MRI features and parcellated cortical thickness (CTh) data. This workflow allows us to validate our method's robustness before interpreting real neuroanatomical differences across species. Our results highlight both conserved and species-specific patterns of sex-related structural variation, with key differentiating regions emerging in the limbic system, sensory processing areas, and motor-associated regions. By integrating machine learning with statistical validation, our framework provides an interpretable and scalable approach for neuroimaging classification, with broader applications beyond sex classification.

2 METHODS

2.1 Overview

To evaluate our classification framework, we first apply it to simulated datasets, providing a controlled benchmark for assessing feature selection and classification accuracy. This establishes the effectiveness of our approach before transitioning to structural MRI data. We then analyze human and macaque MRI datasets, using oblique random forests and feature importance testing to classify biological sex. MRI preprocessing ensures cross-species comparability, with both structural MRI volume and cortical thickness data serving as primary feature spaces for classification. This two-stage workflow—benchmarking on simulations followed by real-data application—allows a rigorous assessment of our method's reliability in identifying key brain regions relevant to sex classification.

2.2 Oblique random forest

We employed Oblique Random Forests (ORF) for classification tasks, specifically leveraging two variants from the `treeple` library: SPORF (Sparse Projection Oblique Random Forest) [13] and MORF (Manifold Oblique Random Forest) [14]. These methods are beneficial for addressing high-dimensional data with complex relationships, as they extend traditional axis-aligned decision trees by allowing each decision tree to split data using random projections, enhancing the ability to capture non-linear patterns and interactions within the feature space. ORFs are particularly advantageous in neuroimaging analysis, where data are often high-dimensional, and the relationships between features are non-linear and complex and may not align with traditional grid-like partitions of the data.

We specifically selected SPORF and MORF for their suitability in handling cortical thickness and MRI volume data, respectively. SPORF has advantages over other oblique random forests primarily because it integrates sparsity-inducing regularization [13]. This feature is particularly beneficial when working with datasets that may contain a large number of irrelevant or redundant features, as it helps to focus the model on the most important variables. For our cortical thickness data, SPORF's ability to select a subset of relevant features improves interpretability and reduces overfitting, which is crucial for our sex classification task. On the other hand, MORF was chosen for MRI volume data due to its robust handling of correlated features [14]. Since MRI data may have intricate spatial correlations between features, incorporating feature locality could improve model performance by capturing these spatial dependencies. Both SPORF and MORF handle high-dimensional data efficiently, making them well-suited for the relatively high-dimensional nature of our data [13, 14].

To validate model performance, we used the out-of-bag (OOB) score, which provides an internal validation metric.

The OOB score is computed by evaluating each tree on the data points that were not used during its construction (i.e., those not included in the bootstrap sample). This technique helps estimate the generalization ability of the model without requiring a separate validation dataset. The OOB score is particularly useful in neuroimaging studies, where cross-validation can be computationally expensive. Using OOB scoring, we were able to efficiently assess model performance and avoid overfitting by providing a robust estimate of classification accuracy. Figure S1 depicts a flow diagram for our classification framework.

2.3 Feature importance testing

Feature importances from RF provide insights into how strongly each feature contributes to the model's predictions. In RFs, feature importances are typically computed using the Gini impurity-based measure: each feature's importance is proportional to the total reduction in Gini impurity it achieves when used to split nodes across all trees in the forest. For SPORF, this process is enhanced by the sparse projection mechanism, where only a subset of features is considered at each split, ensuring that the feature importance estimates focus on the most relevant variables while mitigating noise from irrelevant ones. In MORF, feature importances are computed similarly but are influenced by the method's consideration of feature locality, allowing it to capture spatial correlations in MRI volume data. This leads to a more nuanced understanding of how spatially distributed voxel values contribute to classification tasks.

To further contextualize the feature importance analysis, we compare our approach with two popular model-agnostic algorithms for feature importance: LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (Shapley Additive Explanations). LIME, introduced by Ribeiro et al. [10], aims to explain individual predictions by approximating the complex model locally with a simpler, interpretable model. It computes feature importances by perturbing the input data and observing the changes in model predictions, providing insight into the local behavior of the classifier. SHAP, on the other hand, is grounded in cooperative game theory and computes feature importance by calculating the Shapley values, which represent the average contribution of each feature across all possible permutations of feature combinations [11]. Both LIME and SHAP are widely adopted in machine learning for interpreting black-box models and provide global and local feature importance measures. While these methods are powerful, they can be computationally expensive and may not scale well to high-dimensional neuroimaging data. Our feature importance testing Algorithm, which adapts the tree permutation method, offers a more scalable solution for assessing the statistical significance of features in such high-dimensional settings.

To evaluate the significance of feature importances, we adapt and extend the tree permutation method proposed by Coleman et al. [15] to address the unique challenges of high-dimensional neuroimaging data. This method enables scalable and efficient hypothesis testing in random forests by generating null distributions of feature importance values through tree permutations [15]. Rather than generating n independent forests for permutation testing, it constructs a null distribution by randomly shuffling trees within an ensemble, preserving the dependency structure while drastically reducing computional overhead. To make it feasible for neuroimaging, we modify the framework by leveraging a two-forest approach: one baseline forest and another with permuted features. Additionally, we incorporate Bonferroni-Holm correction to control for multiple comparisons across millions of features, ensuring robust significance assessment. This allows us to assess which cortical thickness or voxel features have statistically significant contributions to sex classification. The details of this method are outlined in Algorithm 1.

2.4 Simulated datasets

To demonstrate the efficacy of our feature importance testing algorithm, NEOFIT, we performed simulations using the MNIST and Trunk datasets. These simulations allowed us to evaluate the algorithm's performance in controlled settings before applying it to real-world human and macaque MRI data.

For the MNIST dataset, we followed the setup described in Li et al. [14], specifically Figure 5, where the goal was to classify digits 3 and 5 out of the full MNIST set. The MNIST dataset contains images of handwritten digits (28×28 pixels), resulting in a high-dimensional feature space of 784 features per image. In our simulation, we used a subset of the dataset with only 3 and 5 for a binary classification task. This allowed us to test our algorithm's ability to identify significant features in a relatively simple, yet high-dimensional, dataset. The MNIST dataset is commonly used to benchmark classification methods, especially when assessing the performance of algorithms designed to handle nonlinearities in data.

Algorithm 1 Training an Explainable Classifier using NEOFIT: Neuro-Explainable Optimal Feature Importance Testing.

Input: Dataset $\mathcal{D}$ with features and labels, number of trees n_{trees}, number of permutations n_{perm}, number of bootstraps B.

Output: Feature importance scores $\mathbf{I}$, adjusted p-values.

for $b \in B$ **do**
  $\mathcal{D}_b \leftarrow$ BOOTSTRAP($\mathcal{D}$)
  $\mathbf{I}_b \leftarrow$ FITDECISIONTREE($\mathcal{D}_b$) ▷ Train tree, compute feature importances
  $\mathcal{D}'_b \leftarrow$ SHUFFLELABELS($\mathcal{D}_b$)
  $\mathbf{I}'_b \leftarrow$ FITDECISIONTREE($\mathcal{D}'_b$)
end for
function COMPUTETESTSTATISTIC($\mathbf{I}, \mathbf{I}'$)
  $\mathbf{R} \leftarrow$ RANK($\mathbf{I}$), $\mathbf{R}' \leftarrow$ RANK($\mathbf{I}'$) ▷ Rank importances
  $T_i \mathrel{+}= \mathbb{I}(R'_i > R_i)$ for all $i \in \{1, \ldots, d\}$ ▷ Update statistics
  $T_i \leftarrow T_i / n_{\text{trees}}$
end function
for $p \in \{1, \ldots, n_{\text{perm}}\}$ **do**
  for $t \in \{1, \ldots, n_{\text{trees}}\}$ **do**
    $(\mathbf{I}_t^{\text{perm}}, \mathbf{I}_t'^{\text{perm}}) \leftarrow$ PERMUTEPAIRS($\mathbf{I}_t, \mathbf{I}'_t$)
  end for
  $\mathbf{T}_p^{\text{null}} \leftarrow$ COMPUTETESTSTATISTIC($\mathbf{I}^{\text{perm}}, \mathbf{I}'^{\text{perm}}$) ▷ Compute null statistic
end for
$p_i \leftarrow \frac{1+\sum_{p=1}^{n_{\text{perm}}} \mathbb{I}(T_i^{\text{null}} > T_i)}{1+n_{\text{perm}}}$ for all i ▷ Compute p-values
$\mathbf{P} \leftarrow$ BONFERRONI-HOLM($\mathbf{p}$)

- BOOTSTRAP($\mathcal{D}$): Draw a bootstrap sample from dataset $\mathcal{D}$.
- FITDECISIONTREE($\mathcal{D}$): Train a decision tree on sample $\mathcal{D}$ and compute feature importances.
- SHUFFLELABELS($\mathcal{D} = (\mathbf{X}, \mathbf{y})$): Apply a random permutation σ such that $\mathbf{y}' = \sigma(\mathbf{y})$.
- RANK($\mathbf{I}$): Rank feature importances $\mathbf{I}$ from the largest. Assign rank d to missing features.
- PERMUTEPAIRS($\mathbf{I}, \mathbf{I}'$): Shuffle the feature importance values in $\mathbf{I}$ and $\mathbf{I}'$ together while maintaining their pairwise associations.
- BONFERRONI-HOLM($\mathbf{p}$): Apply the Bonferroni-Holm correction to p-values $\mathbf{p}$.

For the Trunk simulation, we replicated the setup from Tomita et al. [13], specifically Figure 4, which involves comparing two Gaussian distributions. Both distributions have an identity covariance matrix, with the first distribution having a mean vector starting at -1 and decreasing by a factor of $\sqrt{i}$ for each dimension i, while the second distribution's mean vector is the negative of the first. As the dimensionality d of the data increases, the two distributions become closer, making the task of classification progressively more difficult. The mathematical formulation of these two distributions is provided in Appendix 4.2. The Trunk dataset is specifically designed to simulate high-dimensional, sparse data and tests the ability of algorithms like SPORF to efficiently identify relevant features in a sparse and increasingly challenging classification task. The simplicity of the Trunk dataset makes it an ideal setting for evaluating the feature importance and selection abilities of our algorithm in a controlled, synthetic environment.

Both of these simulations allowed us to rigorously test the performance of our feature importance testing Algorithm in scenarios that mimic real-world challenges in feature selection and classification. By applying this method to both the MNIST and Trunk datasets, we can demonstrate its ability to reduce computational overhead while maintaining classification accuracy, showcasing its suitability for handling high-dimensional neuroimaging data.

2.5 Simulated performance

In our simulation experiments, we aimed to compare the performance of four feature selection algorithms—RF, LIME, SHAP, and NEOFIT (Neuro-Explainable Optimal Feature Importance Testing)—across two datasets: MNIST, a widely used benchbark dataset, and Trunk, a simulated dataset. The main objective was to evaluate how effectively each algorithm selects the most informative features for classification tasks when varying the number of features, ranging from 1 to 512 features.

For each feature dimension d, the dataset was partitioned into three subsets: training (60%), validation (20%), and testing (20%). In each iteration, we trained a RF on the training set and ranked the features by their importance using the Gini impurity measure. Then, we selected the top d features based on this ranking, trained a new RF on the validation set using only these features, and evaluated the resulting model on the testing set. This procedure was repeated for 50 iterations to ensure stability and robustness in the results.

We also tested the performance of LIME, SHAP, and our feature importance testing algorithm on the same feature importances derived from RF. LIME and SHAP are popular techniques for model interpretability: LIME approximates the model locally, and SHAP uses Shapley values to attribute contributions to individual features. Our algorithm, NEOFIT, applied the tree permutation method to generate null distributions of feature importances and selected features based on statistically significant p-values, providing a more rigorous approach to feature selection.

In both MNIST and Trunk simulations, our algorithm consistently outperformed the others. In the MNIST dataset, we achieved over 97% accuracy using fewer features than any other method, demonstrating the efficiency of our approach (Figure 1). For the Trunk dataset, we obtained over 80% accuracy, which was higher than the best performance of all other algorithms, even with a relatively small number of features (Figure 1). These results underscore the superiority of our method, particularly in high-dimensional data, where both accuracy and feature efficiency are crucial.

2.6 Structural MRI dataset

The human MRI volume data were sourced from publicly available repositories, including the 1000 Functional Connectomes (FCON1000), Healthy Brain Network (HBN), and several other initiatives. A full list of the datasets can be found in Table 1 in the supplementary materials. We included data from unrelated participants across multiple cohorts to ensure diversity and replicability. Specifically, T1-weighted anatomical MRI data were analyzed. Detailed preprocessing protocols, as well as demographic and acquisition parameters, can be found on the respective dataset portals. There are 14,380 subjects in total, with 7,416 females and the remaining 6,964 males, having a mean age of 39 for females and 31 for males. S2 provides the full age and sex distribution for human data.

Macaque MRI data were sourced from the PRIME-DE consortium [16] and included 592 rhesus macaques (*Macaca mulatta*) scanned under anesthesia, including 265 females and 327 males, scanned under anesthesia. Throughout this manuscript, NHP and macaque are used interchangeably to refer to rhesus macaques. The

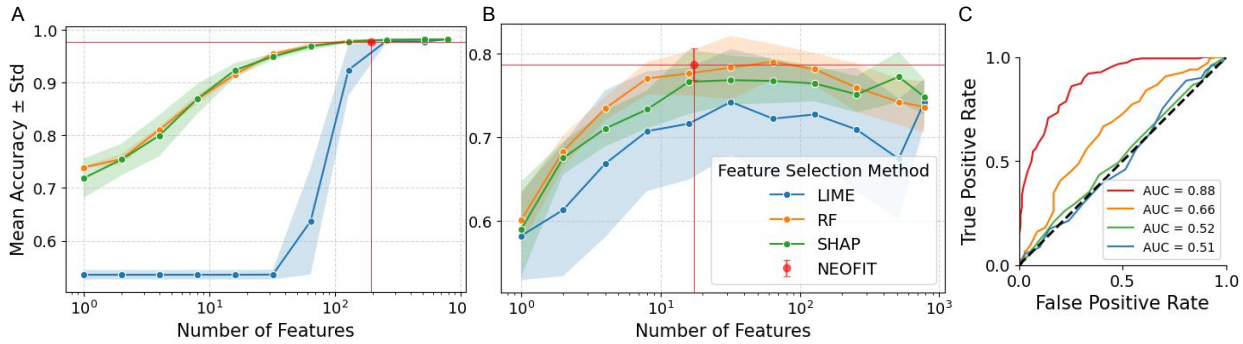

Figure 1: Simulation results for feature selection performance. A: Classification accuracy (mean ± standard deviation) for MNIST digits 3 vs. 5 across different numbers of selected features. B: Results for the Trunk dataset. 50 repetitions were run per condition. Our method, NEOFIT, is indicated by the red dot, with red horizontal and vertical lines marking the corresponding accuracy and feature count. C: ROC curve for Trunk simulation including $n = 17$ top features where n is determined by NEOFIT.

mean age is 2.0 for females and 1.8 for males. Final volumes were resamplmed to match the resolution of the human data, facilitating direct comparisons across humans and macaques. For further details, please refer to the source of the data at PRIME-DE: UW-Madison. Detailed age and sex distributions for macaque data are also shown in S3.

2.7 MRI preprocessing

The MRI preprocessing pipeline was designed to ensure high-quality input data for subsequent analyses. Following established preprocessing protocols [17], the raw T1-weighted images underwent several key steps, including denoising, brain extraction, and tissue segmentation. First, noise reduction was applied to enhance image quality while preserving anatomical details. Brain extraction (or skull stripping) was then performed to remove non-brain tissues, facilitating accurate cortical and subcortical measurements. The images were subsequently segmented into different tissue classes, such as gray matter, white matter, and cerebrospinal fluid. To ensure data reliability, an initial quality control step was conducted to screen for artifacts, motion-related distortions, or structural anomalies. Preprocessed data were further reviewed through visual inspection, and scans that failed at any stage of processing were either corrected or excluded from the final dataset.

To generate gray and white matter density maps, we employed voxel-based morphometry (VBM) using the Statistical Parametric Mapping (SPM) framework. For human data, preprocessing was performed using CAT12, an advanced and optimized pipeline for high-resolution human brain imaging [18]. The CAT12 toolbox enabled improved tissue segmentation and bias field correction, ensuring robust gray and white matter probability maps [18]. For macaque data, a customized template was used in conjunction with methods adapted from SPM-mouse [19]. This approach accounted for structural differences between human and macaque brains, improving segmentation accuracy and ensuring valid cross-species comparisons.

2.8 Cortical thickness dataset

Cortical thickness (CTh) measurements were derived from surface-based morphometric analyses tailored for both human and macaque brains. For human subjects, CTh was extracted using the CAT12 toolbox, an advanced framework optimized for surface reconstruction and thickness estimation [18]. This pipeline includes topology correction, spherical mapping, and projection-based thickness estimation to ensure accurate and reliable measurements. For macaques, CTh was computed using a species-specific pipeline based on a customized surface template, following methodologies from the Macaque CHART framework [17], which improves precision in non-human primate surface morphometry.

To enable cross-species comparisons, cortical thickness measurements were parcellated using an established alignment framework [20], ensuring that homologous cortical regions were systematically mapped between humans and macaques. This approach minimizes species-specific biases and enhances the biological interpretabil-

ity of interspecies analyses.

CTh data were processed under two parcellation schemes: Schaefer and Markov. The Schaefer parcellation, based on functional connectivity gradients, provides a data-driven, hierarchical atlas designed to balance spatial resolution and functional specificity [21]. We used the 200-region version to optimize computational efficiency while retaining detailed cortical representation. In contrast, the Markov parcellation is derived from anatomical connectivity patterns in macaques, based on tracer studies that define 182 structurally connected cortical regions [22]. While originally designed for different species, both atlases were adapted for cross-species analysis through mapping transformations, enabling direct comparison between human and macaque cortical organization.

The dataset includes cortical thickness measurements from 10,608 human and 572 macaque subjects, each analyzed under both parcellation schemes, yielding four datasets: $\{\text{Human, Macaque}\} \times \{\text{Schaefer, Markov}\}$. These datasets provide a comprehensive framework for investigating species-specific and evolutionarily conserved cortical patterns, facilitating robust statistical analysis of sex-related cortical differences and their biological significance. A detailed depiction of the age and sex distributions for CTh data in both species is provided in S3.

2.9 Variable specification and relationship identification for within- and cross-species MRI analysis

To make our structural assumptions underpinning our analysis explicit, we employ a directed acyclic graph (DAG) framework [23, 24]. DAGs formalize causal relationships between variables and provide graphical criteria for whether particular effects can be identified within the context of a given analysis. Figure 2 illustrates our assumed causal structure, where our primary interest is the effect of biological sex (the *exposure*, green) on brain structural features (the *outcomes*, blue): voxel-wise gray and white matter density maps, or parcellated cortical thickness. The total effect includes pathways mediated by intracranial volume (ICV) and hormonal status (the *mediators*, tan). Since the relationship between ICV and brain structure is well-established and represents a nuisance scaling factor rather than our primary interest [3], we employed species-specific spatial normalization, which effectively controls for ICV differences (indicated by the boxed border in Figure 2). Hormonal status, conversely, represents a key biological mechanism through which sex differences manifest; controlling for hormones would remove the effect we aim to study. Therefore, our within-species analyses identify brain regions where biological sex has a significant effect, conditional on ICV normalization but preserving hormone-mediated pathways. Age (white) is a common cause of hormonal status and brain features. Finally, our cross-species analysis examines whether species (the *effect modifier*, pink) modifies the magnitude of sex effects on brain features, testing for evolutionary conservation versus species-specific adaptation.

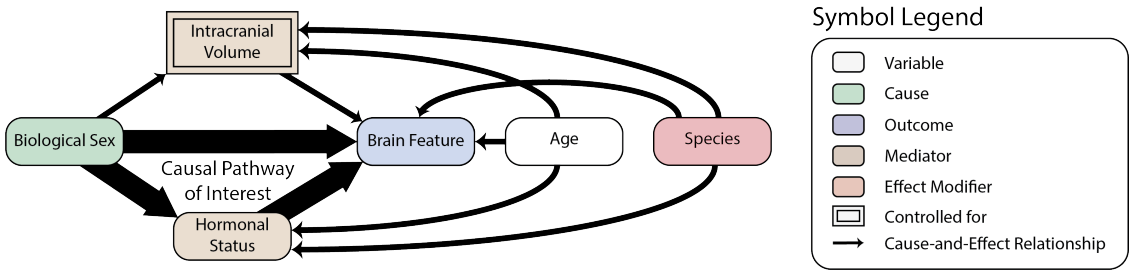

Figure 2: Directed acyclic graph illustrating the assumed causal structure for sex classification analysis. Nodes represent variables, with biological sex (green) as the exposure and brain features (blue) as outcomes. Mediators (tan) include intracranial volume (ICV, boxed to indicate that our analysis controls for ICV differences) and hormonal status. Age (white) affects hormonal status and brain features. Species (pink) acts as an effect modifier. Arrows indicate causal relationships.

3 RESULTS

3.1 Hyper-parameter tuning

We fine-tuned several hyper-parameters for both SPORF and MORF implemented with `treeple` (https://treeple.ai/):

- `max_features`: Determines the maximum number of features considered for each split. A lower value encourages diversity among trees, preventing overfitting; a higher value can lead to stronger splits and higher accuracy for individual trees while too high a value may lead to less randomization and poorer generalization.
- `max_patch_dim` (for MORF): Defines the dimensionality of the patches used for splitting the data. Adjusting this parameter allows the model to explore different feature subspaces and balance model complexity and accuracy.
- `feature_combinations` (for SPORF): Controls the number of linear combinations of features evaluated at each split in the tree. A higher value leads to a less sparse model, as it incorporates more features into each decision. Increasing the number of combinations improves the flexibility of the model but may reduce interpretability and increase computational complexity.

Neither `max_features` nor `max_patch_dim` affect the OOB score for MRI volume data (results not shown). For CTh data, `feature_combinations` and `max_features` influence human data far more than macaque data (Figure 3). Given that SPORF applies linear combinations of features at each split, we tested three types of feature normalization and evaluated their impact on classification performance. Based on these comparisons (Figure S4), we selected z-score normalization as it yielded the most stable and consistent performance across datasets.

3.2 Performance of the sex classifier

After hyper-parameter optimization, we evaluated the performance of our random forest models using the area under the curve (AUC) from the receiver operating characteristic (ROC) curve, which quantifies classifier performance across various decision thresholds. Figure 4 illustrates the ROC curves and corresponding AUC scores for both MRI volume and CTh data. Our model demonstrates an especially high AUC for human MRI volume data, while exhibiting moderate classification performance for macaque data. Overall, the human data consistently outperforms the macaque data across both imaging modalities.

3.3 Performance of the feature importantce testing

To compare our method with previous interpretable feature importance testing approaches on the sex-classification task, we replicated the procedure described in the Materials and Methods section on CTh data.

Features were selected using LIME and SHAP applied to trained Random Forests, with the number of trees matched across species ($5,000$ for human and $20,000$ for macaque). We then trained classifiers using only the selected significant features and reported accuracy as a proxy for the reliability of each selection method. For LIME and SHAP, we varied the number of top features to examine performance as a function of the selected-feature count; our methods, NEOFIT, yields a single, data-driven estimate of the number of significant features. Fig 5 reports results on the Markov parcellation data; corresponding results for Schaefer parcellation are provided in S5.

On the human CTh data, NEOFIT reaches around an accuracy of 0.70 with 29 features, whereas LIME requires a comparable subset and SHAP requires a much larger subset to match the same level. On the NHP CTh data, NEOFIT achieves an accuracy of approximately 0.66 using 44 features, which is above the standard deviation bands of LIME and SHAP across most feature-subset sizes. Unlike the post hoc methods, NEOFIT does not require tuning the feature-subset size and shows lower variance. Taken together, these results indicate that NEOFIT delivers comparable accuracy with fewer features, improving the interpretability of sex classification.

We recorded importance–building times for each feature-selection method under a unified protocol. On the Markov CTh dataset, the human cohort required LIME: 288.57 seconds, SHAP: 447.71 seconds, and NEOFIT: 107.04 seconds to build feature importances; for the macaque cohort the corresponding times were 1064.00 seconds, 100.62 seconds, and 194.47 seconds. Times are the median over 5 stratified resamples, and all

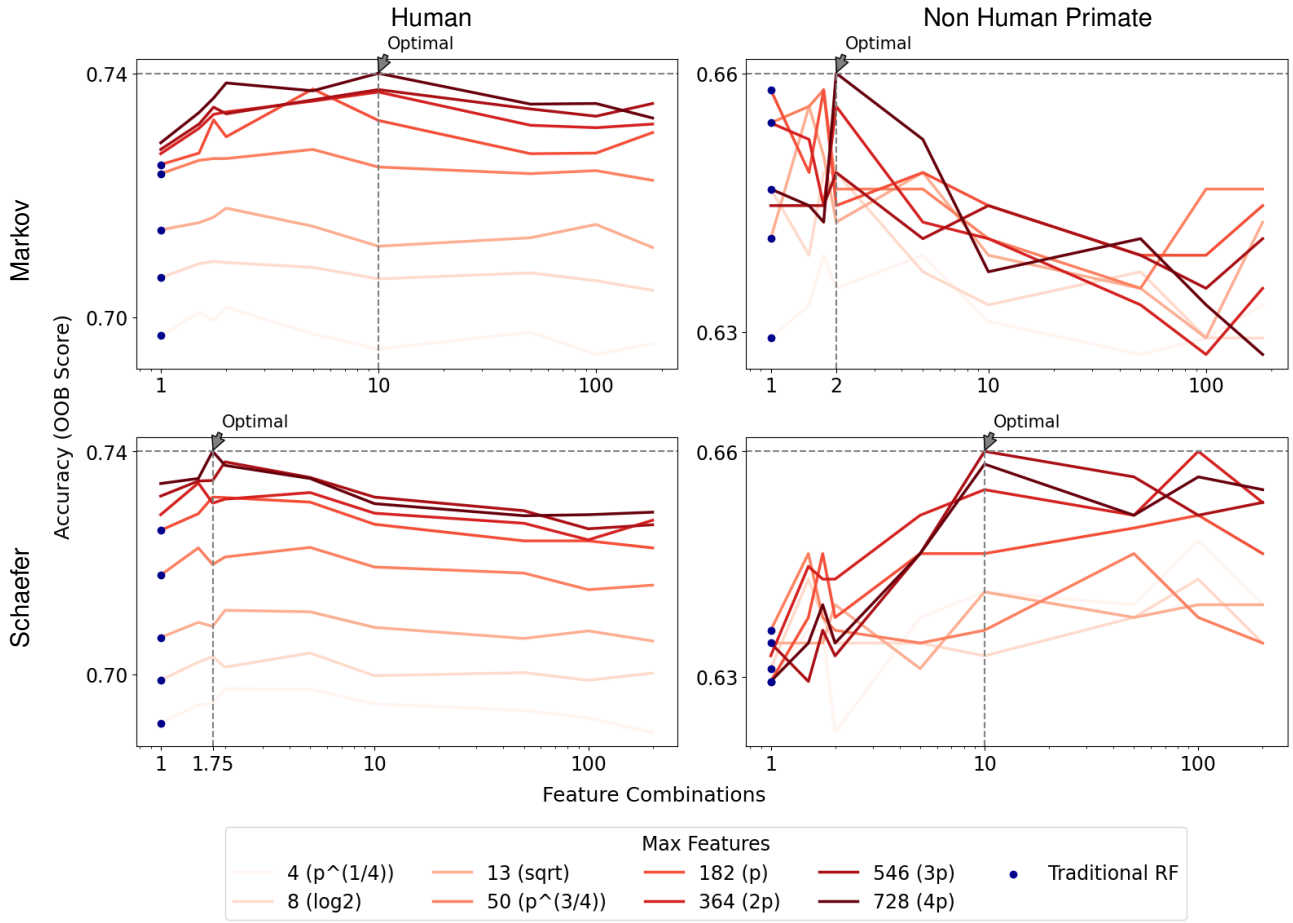

Figure 3: Performance comparison of oblique random forest (SPORF) using cortical thickness data across various hyperparameters. The figure displays the impact of different model configurations, including `feature_combinations`, and `max_features` on performance. Blue dots are achievable by traditional random forest models. In subsequent experiments, we set `max_features` to $4p$ for all CTh data. For `feature_combinations`, we used 10 for human Markov, 2 for macaque Markov, 1.75 for human Schaefer, and 10 for macaque Schaefer. The number of estimators was set to 5000 for human data and 20,000 for macaque data. All other parameters followed the default settings of the `treeple` implementation (`treeple.ai`).

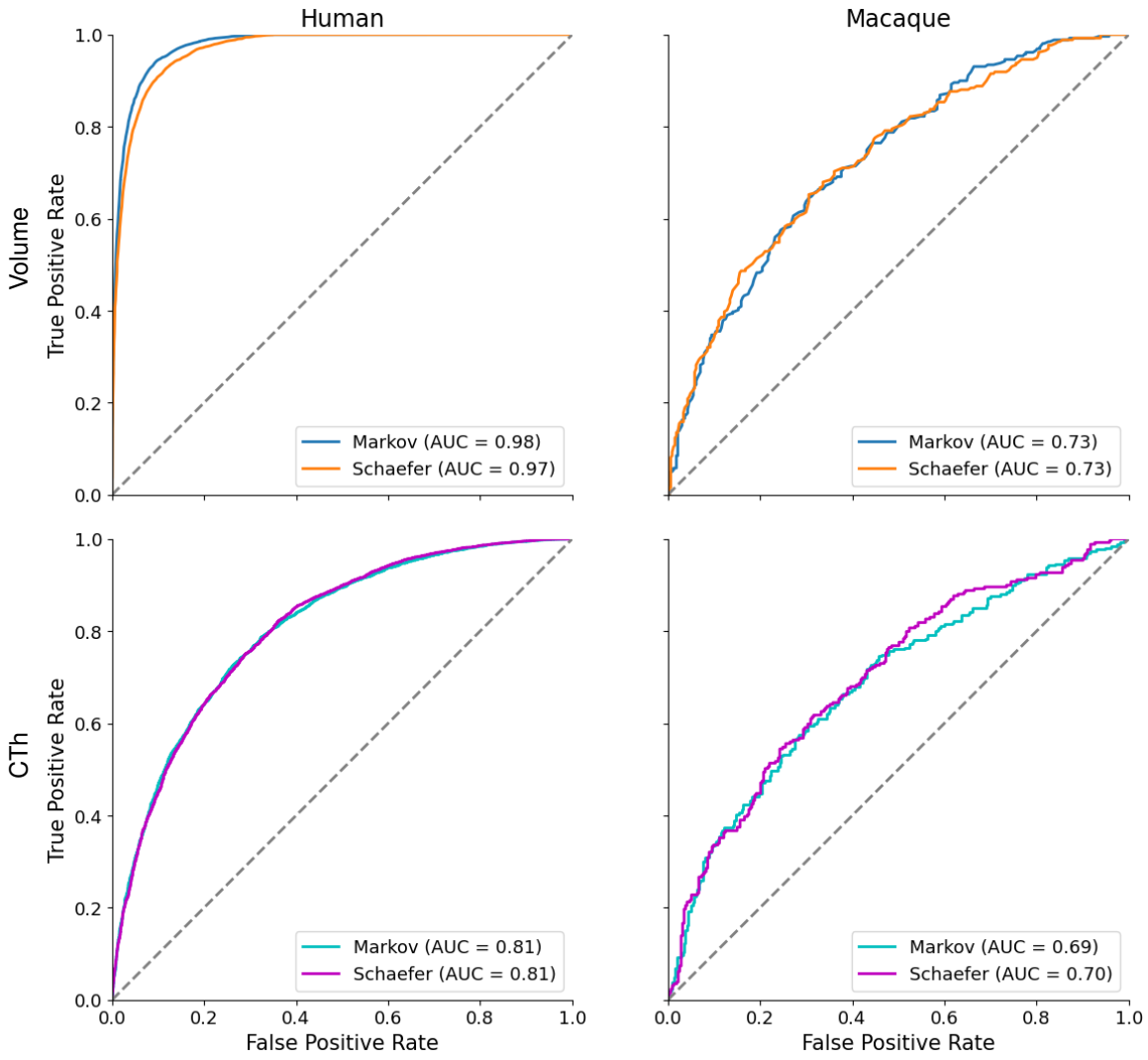

Figure 4: Comparison of True Positive Rate versus False Positive Rate for sex classification models: (top row) MRI volume data, (bottom row) cortical thickness (CTh) data. Panels show results for (left) humans and (right) NHPs. Curves are color-coded by tissue type (gray/white matter) for MRI volume data and by parcellation scheme (Markov/Schaefer) for CTh data. The area under the curve (AUC) is annotated for each category.

experiments were executed on a dual-socket Intel Xeon Gold 6248R system (2×24 physical cores, 96 hardware threads; base 3.0 GHz, turbo 4.0 GHz), using 50 parallel workers for both training and inference.

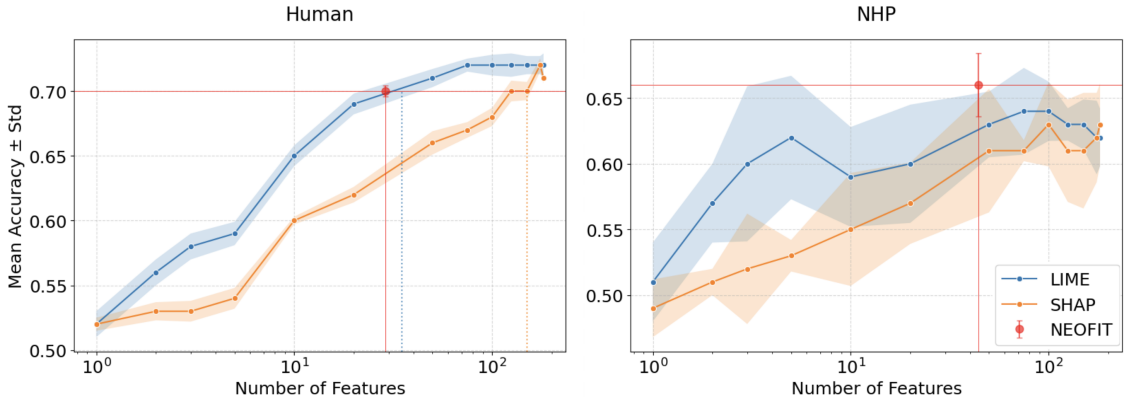

Figure 5: Comparison of feature selection performance on Markov CTh data. Classification accuracy on human Markov CTh data (left) and macaque Markov CTh data (right) versus the number of selected features for each method. Mean and standard deviation are computed over five repetitions per condition. NEOFIT is shown as a red point at its chosen feature count.

3.4 Voxel-wise feature importance maps

We applied MORF to gray matter and white matter MRI volume data for both humans and macaques separately, generating feature importance values for each voxel along with their corresponding p-values. Figure 6 illustrates the p-values generated using our feature importance testing algorithm 1, NEOFIT, while figures (S6, S7, S8, S9) displaying the corresponding raw feature importance values are provided in the supplementary material.

Regions with significant feature importance for sex classification in humans were identified primarily in the limbic system, including the amygdala, hippocampus, and thalamus, as well as in occipital regions (Figure 6). The involvement of the limbic structures, which are implicated in motor coordination, sensory integration, and emotion regulation [25], aligns with previous findings that link sexual dimorphism in humans to emotional regulation and memory processing, functions predominantly associated with the amygdala and hippocampus [3]. Similarly, the occipital regions, responsible for visual processing, have shown structural differences between sexes, potentially reflecting sex-specific adaptations in sensory processing [3].

In macaques, feature importance was also primarily concentrated in the limbic system, encompassing the superior temporal gyrus, dentate gyrus (part of hippocampus), putamen, and caudate nucleus (Figure 6). Additional contributions were observed in regions directly connected to the limbic system, such as the orbital gyrus, insula, and claustrum, and in regions with indirect interactions, including the precentral gyrus which is primarily associated with motor function (Figure 6).

3.5 Parcel-wise feature importance maps

For the CTh data, we applied SPORF to the parcellated data, utilizing both the Schaefer and Markov parcellation schemes for humans and macaques. The feature importance maps, showing raw feature importance values across all parcels, as well as their corresponding p-values, are provided in the supplementary materials (Figures S10, S11, S12, S13). To facilitate comparisons of significance patterns between the species, we merged the p-value maps for humans and macaques into a single surface, displayed with distinct color schemes (Figure 7). For parcels where both species exhibit significant p-values, we use a mosaic coloring pattern to highlight the significance levels for each species (Figure 7). This approach emphasizes regions where both species show consistent patterns of significance, as well as where differences exist.

The Markov parcellation reveals that human-specific regions are concentrated in networks associated with higher-order cognitive processes, including the dorsal and ventral attention networks and the default mode network (DMN). These cortical thickness differences likely reflect species-specific traits related to social cognition and complex behaviors. In contrast, macaque-specific regions overlap primarily with the posterior ventral attention

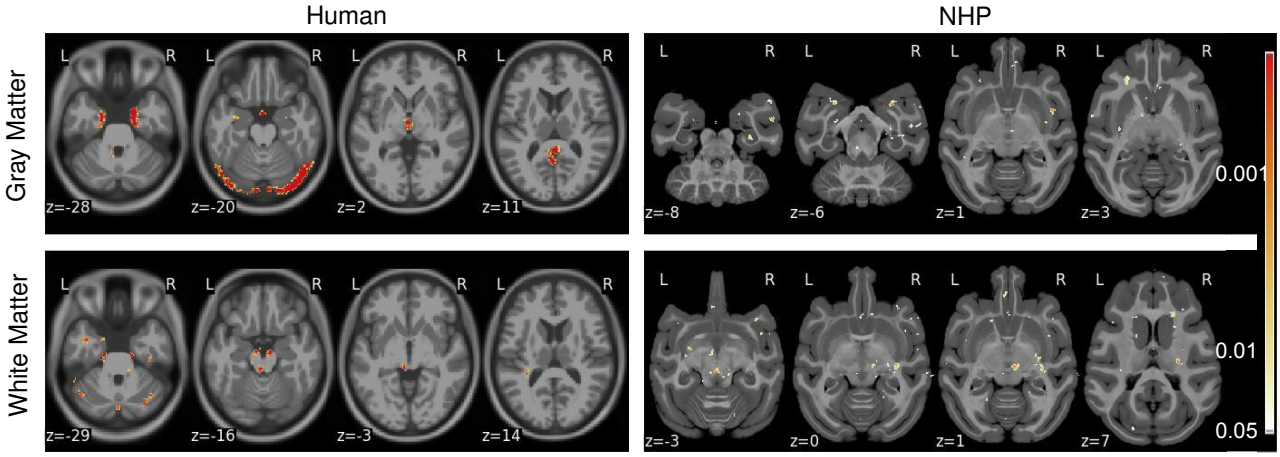

Figure 6: Voxel-wise feature importance values and corrected p-values for human (left column) and NHP (right column) computed using our feature importance testing algorithm with $n_{\text{permutations}} = 50{,}000{,}000$. Selected horizontal slices of structural MRI volume on the MNI template show regions with at least 20 voxels and significant feature importance values $(p < 0.05)$. All slices were visually examined, and only those with visibly large regions are presented here.

network (Figure 7), indicating sex-related differences predominantly in sensory-driven or attentional systems. The orbitofrontal–limbic network emerges as a significant shared region across species (Figure 7), consistent with its conserved role in emotion and reward processing among primates.

For the Schaefer parcellation, shared significant regions are largely part of the posterior ventral attention network (Figure 7), underscoring the evolutionary conservation of attentional mechanisms underlying sex-related cortical differences. In humans, a broader array of networks, including somatomotor, frontoparietal, DMN, and visual networks, is implicated (Figure 7), highlighting the functional and structural complexity of human cortical organization. This diversity is likely linked to advanced motor control, cognition, and social interaction. Similar to the Markov parcellation, macaque-specific regions are confined to posterior networks, such as the posterior dorsal and ventral attention networks and the DMN (Figure 7), suggesting a more limited scope of sex-based cortical differences compared to humans.

We conducted a network-level analysis of the CTh data by aggregating parcels according to the 7-Yeo brain network scheme [26]. A Wilcoxon signed-rank test was applied to compare feature importances at the network level between humans and macaques across the seven Yeo networks: Visual, Somatomotor, Dorsal Attention, Ventral Attention, Limbic, Frontoparietal, and Default. The Frontoparietal network showed a significant difference in feature importances between species in the Schaefer parcellation, with humans exhibiting higher feature importance values (Figure 8).

4 DISCUSSION

Our study demonstrates that the integration of oblique random forests (ORFs) with permutation-based feature importance testing fills a critical gap in neuroimaging analysis. By capturing complex, non-linear interactions through oblique splits and rigorously validating feature importance, our method reliably distinguishes true signal from noise in high-dimensional data. This advancement is evident in our simulation studies, where our approach outperforms conventional methods such as LIME and SHAP, and is further validated in the classification of sex from both voxel-wise MRI and parcellated CTh data.

Our analysis estimated the overall effect of biological sex on brain structure, representing sex effects averaged across the age distributions observed in our sample. Under the causal framework outlined in (Fig 2), this can be conceptualized as an average treatment effect (ATE) [27]. Our human data exhibited an age imbalance between sexes (females: mean 39 years, males: mean 31 years) arising from aggregation of datasets with different demographic compositions. If sample inclusion depended jointly on both sex and age, conditioning on the selected sample could induce collider bias [28]. While we found no explicit evidence of such joint selection

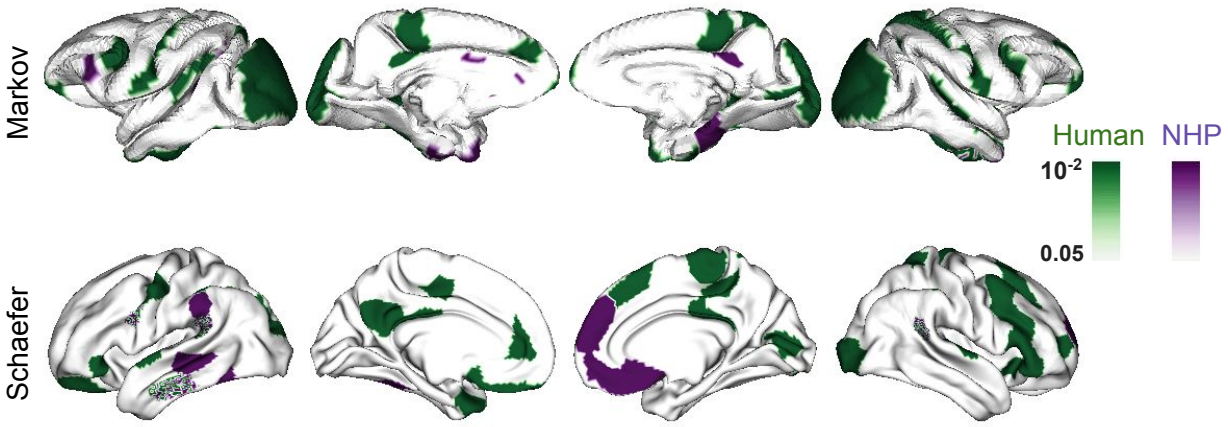

Figure 7: Parcel-wise feature importance values and corrected p-values mapped onto brain surfaces for human (green) and NHP (purple), computed using our feature importance testing algorithm with $n_{\text{permutations}} = 5000$. Rows show parcellations: top - Markov, bottom - Schaefer. Significance is determined using a threshold of $p < 0.05$ for either species. Parcels where both species meet the significance threshold are highlighted with a mosaic pattern, indicating overlap of significant results while preserving species-specific contributions. This visualization enables direct comparison of significant brain regions across humans, non-human primates, or both.

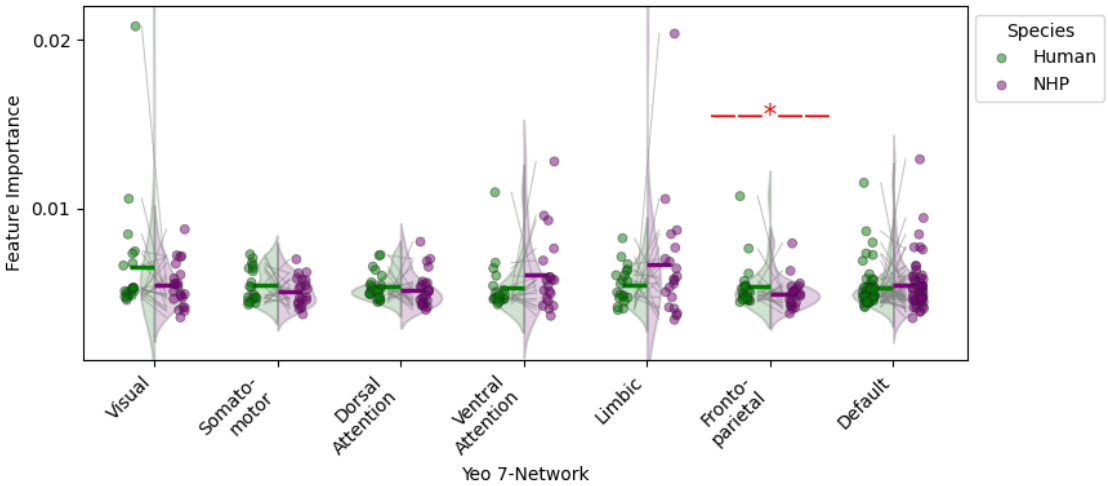

Figure 8: Comparison of network-wise classification feature importances across species for the Schaefer parcellation (a similar plot for the Markov parcellation can be found in the appendix S14). For each species, the raw feature importance values for all parcels corresponding to the same network, according to Yeo's 7-network, were aggregated. The differences in feature importances across species were then evaluated using the Wilcoxon signed-rank test. The short horizontal line above each strip indicates the mean, and the red star highlights the pair with a significant difference ($p < 0.05$).

mechanisms in the data collection protocols, we cannot definitively rule out this possibility in large-scale multi-site aggregations.

Structural MRI volume and CTh data introduce distinct challenges for classification. Voxel-wise MRI data is inherently high-dimensional, leading to challenges such as feature redundancy, increased measurement noise, and the curse of dimensionality. While MORF addresses local feature dependencies by considering spatially correlated feature patches, distinguishing meaningful biological signals from noise remains difficult, particularly given the variability in MRI acquisition protocols across datasets.In contrast, CTh data is lower-dimensional but relies on template-driven parcellation schemes, which may not optimally capture fine-grained anatomical differences across individuals. Parcellation choices can constrain classification performance, as predefined regions may not align with the most relevant biological variations. This limitation can be exacerbated when parcellations are derived from standard population-based templates rather than subject-specific cortical features.

Our sex classification results provide important insights into the underlying neuroanatomical patterns. Specifically, the identification of both shared and species-specific regions suggests a balance between evolutionary conservation and species adaptation. In humans, significant regions—primarily within the limbic system and higher-order cognitive areas—underscore the role of emotion and memory processing in sex differentiation. Conversely, in macaques, regions linked to sensory processing and motor control appear more prominent. This dichotomy aligns with prior evidence of conserved neural mechanisms across primates while also reflecting species-specific adaptations [3, 5, 20]. While both species demonstrate significant sex-related differences in the limbic system, the specific regions implicated differ, reflecting species-specific neurobiological adaptations. Humans show pronounced involvement of the amygdala and hippocampus, emphasizing emotional and memory-related sex differences, whereas macaques exhibit a broader network involving motor, sensory, and integrative regions such as the precentral gyrus and insula. These findings may reflect evolutionary divergence in sex-specific neural functions, with humans demonstrating more specialization in cognitive and emotional domains. Collectively, our findings highlight the potential of our method to reveal meaningful structural markers that can inform our understanding of neurobiological sex differences and their evolutionary origins.

4.1 Limitations

Despite these promising results, several limitations remain. First, although our permutation-based feature importance testing improves statistical validation, voxel-wise data—with a vast number of features—require an extremely high number of permutations to reliably correct p-values, resulting in significant computational cost. Second, our current classification accuracy for human MRI volume and cortical thickness data is moderate, which raises concerns regarding the robustness of the identified features as reliable markers. Finally, while our approach outperforms established methods like LIME and SHAP in our simulations, further validation on independent datasets and exploration of computational efficiency in extremely high-dimensional settings remain important avenues for future research.

4.2 Future work

Building on the above limitations, we outline several directions for future work. First, incorporating additional MRI-derived features (e.g., cortical curvature and white-matter connectivity) could improve classification and yield a more comprehensive view of sex-linked structural variation. Second, extending cross-species analyses to additional primates would help clarify evolutionary patterns and the extent to which observed differences are conserved across species. Third, transfer-learning approaches—such as training on human data and adapting to macaque data—may leverage shared structure while accounting for species-specific differences, thereby improving cross-species comparability. Furthermore, evaluating our method on a population-scale dataset such as the UK Biobank would provide a stringent test of out-of-sample generalization and scalability [29, 30].

In addition, future work could examine conditional average treatment effects (CATEs) across age to assess whether sex differences vary across developmental stages and aging [23, 27, 31], providing a temporal perspective on sexual dimorphism while potentially mitigating selection-bias concerns potentially introduced by conditioning on ICV.

We also plan to incorporate honest tree procedures (sample-splitting trees in which one subsample determines splits and a separate subsample estimates leaf predictions and related quantities), as proposed in [32, 33].

Coupling NEOFIT with honest splitting is expected to reduce adaptive bias in importance estimates and improve the reliability of importance estimators.

To reduce NEOFIT build time, we leveraged Google's Yggdrasil Decision Forests (YDF; https://github.com/google/yggdrasil-decision-forests) to accelerate SPORF training, yielding substantial speedups on cortical-thickness (CTh) classification tasks. As a next step, we will integrate a YDF-backed ORF backend into the NEOFIT implementation to further reduce training and inference times.

Beyond fundamental structural differences, future work should examine sex-related brain variation in the context of neurological and psychiatric disorders, where sex is a known risk factor (e.g., autism, schizophrenia, depression). Understanding how these differences manifest in both healthy and clinical populations may refine diagnostic tools and inform personalized treatment strategies. More broadly, integrating machine learning with rigorous statistical validation has the potential to reshape neuroimaging analysis, paving the way for more interpretable and biologically meaningful discoveries in neuroscience.

DATA AND CODE AVAILABILITY

The data used in this study for human and macaque brain analysis are obtained from publicly available, open-access datasets, as detailed in the METHODS 2 section of this paper.

The code used to perform the analysis and generate results for sex classification an be accessed at https://github.com/neurodata/sex_classification.

AUTHOR CONTRIBUTIONS

T.L.: Conceptualization, Methodology, Software, Formal Analysis, Data Curation, Writing—Original Draft, Writing—Review & Editing; J.D.: Conceptualization, Methodology, Software, Formal Analysis, Data Curation, Writing—Review & Editing; B.X.: Formal Analysis, Software, Writing—Review & Editing; E. W. B.: Writing—Review & Editing; S.A.: Data Curation; K.N.: Data Curation; K.B.: Data Curation; J.T.V.: Conceptualization, Methodology, Supervision, Writing—Review & Editing, and Funding Acquisition; T.X.: Conceptualization, Methodology, Supervision, Data Curation, Writing—Review & Editing, and Funding Acquisition.

DECLARATION OF COMPETING INTEREST

The authors declare no competing interests.

ACKNOWLEDGEMENTS

The authors would like to thank Yuxin Bai and Haoyin Xu for their valuable insights and advice on this work.

Supplementary Material

Mathematical Formulation of the Trunk Simulation

For the Trunk simulation, we compare two Gaussian distributions. The mean of the first distribution starts at -1 and decreases by a factor of $\sqrt{i}$ for each dimension i. The mean of the second distribution starts at 1 and also decreases by a factor of $\sqrt{i}$ for each dimension i. The mathematical formulation is as follows:

$$\mathbf{x}_1 \sim \mathcal{N}(\mu_1, \mathbf{I}_d), \quad \mathbf{x}_2 \sim \mathcal{N}(\mu_2, \mathbf{I}_d)$$

where:

- $\mathbf{x}_1$ and $\mathbf{x}_2$ are random vectors drawn from two multivariate normal distributions.
- $\mathcal{N}(\mu, \Sigma)$ denotes a multivariate Gaussian distribution with mean μ and covariance matrix Σ.
- $\mu_1 = (-1, -\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{3}}, \ldots, -\frac{1}{\sqrt{d}})$ is the mean vector for the first distribution, where each component decreases by a factor of $\sqrt{i}$.

Table 1: List of human structural MRI volume datasets

Project	Dataset	Description	# subjects after quality control
ABIDE [34]	ABIDE	Autism Brain Imaging Data Exchange	180
Age Risk [35]	sub-AgeRisk	Subset of the Age Risk dataset	187
AOMIC [36]	AOMIC-ID1000_ds003097	Amsterdam Open MRI Collection	928
Berlin Mind-Brain [37]	BERLIN_MINDBRAIN	Berlin Mind-Brain Project	49
Cam-CAN [38]	CANCAM	The Cambridge Centre for Ageing and Neuroscience	653
COBRE [39]	COBRE	Center for Biomedical Research Excellence	146
Dallas Lifespan [40]	DALLAS_LIFESPAN	Lifespan study in Dallas	315
FCON1000 [41]	Ann Arbor (a, b), Atlanta, Baltimore, Bangor, Beijing, Berlin, Cambridge, ICBM, Milwaukee (b), Munich, New York (a, b), NYU_TRT, Oxford, Palo Alto, Saint Louis	Functional Connectomes Project datasets from multiple sites	900
General Imaging Datasets [42]	ds002382, ds003381, ds003592, ds003639, ds004173, ds004215	Various functional/structural imaging datasets	539
HBN [43]	HBN_CBIC, HBN_CUNY, HBN_RU, HBN_SI	Healthy Brain Network datasets	3817
Human Connectome Project [44]	HCPAging, HCPYoungAdults	HCP Aging and Young Adults datasets	1697
IXI [45]	IXI	Brain imaging research dataset	563
Narratives Dataset [46]	Narratives_ds002345	Dataset focusing on narrative comprehension	345
NEO [47]	NEO2012_Adelstein	Neuroimaging dataset from 2012	39
NKI Rockland [48]	NKI_ROCKLAND	Nathan Kline Institute: Rockland dataset	810
OASIS [49]	OASIS-1, OASIS-2, OASIS-3	Open Access Series of Imaging Studies: Phases 1–3	1998
ON[50]	ds-0057522	NIMH Healthy Research Volunteer Dataset on OpenNeuro	154
PETfrog [51]	PETfrog	Dataset combining PET and MRI studies	149
Quirón-Valencia [52]	QUIRON_VALENCIA	Structural studies dataset	44
SALD [53]	SALD	Southwest Autism Research & Resource Center dataset	494
TRAIN [54]	TRAIN-39	Video-game training study	39
Wayne State [55]	WayneState_10_longitudinal, WayneState_11, WayneState_EF	Longitudinal and executive function datasets	319
Yale Low-Res Imaging [56]	YALE_LOWRES	Yale Low-Resolution Imaging dataset	100

- $\mu_2 = (1, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{3}}, \ldots, \frac{1}{\sqrt{d}})$ is the mean vector for the second distribution, where each component decreases by a factor of $\sqrt{i}$.
- $\mathbf{I}_d$ is the $d \times d$ identity matrix, representing independent unit-variance features.

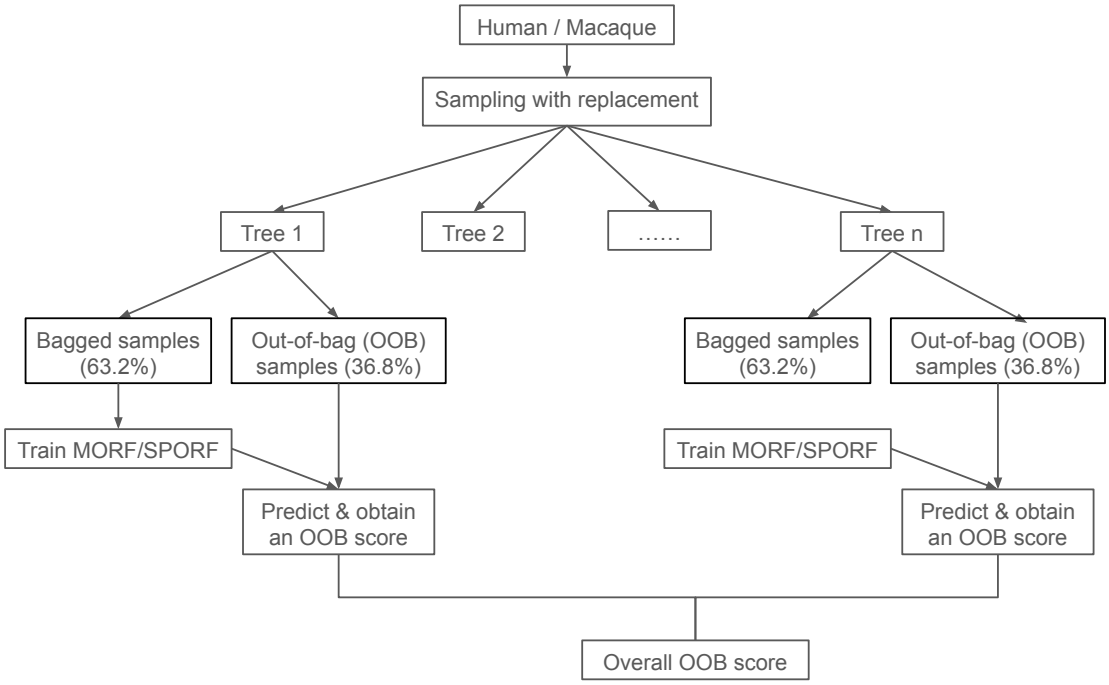

Figure S1: Flow diagram for the classification framework.

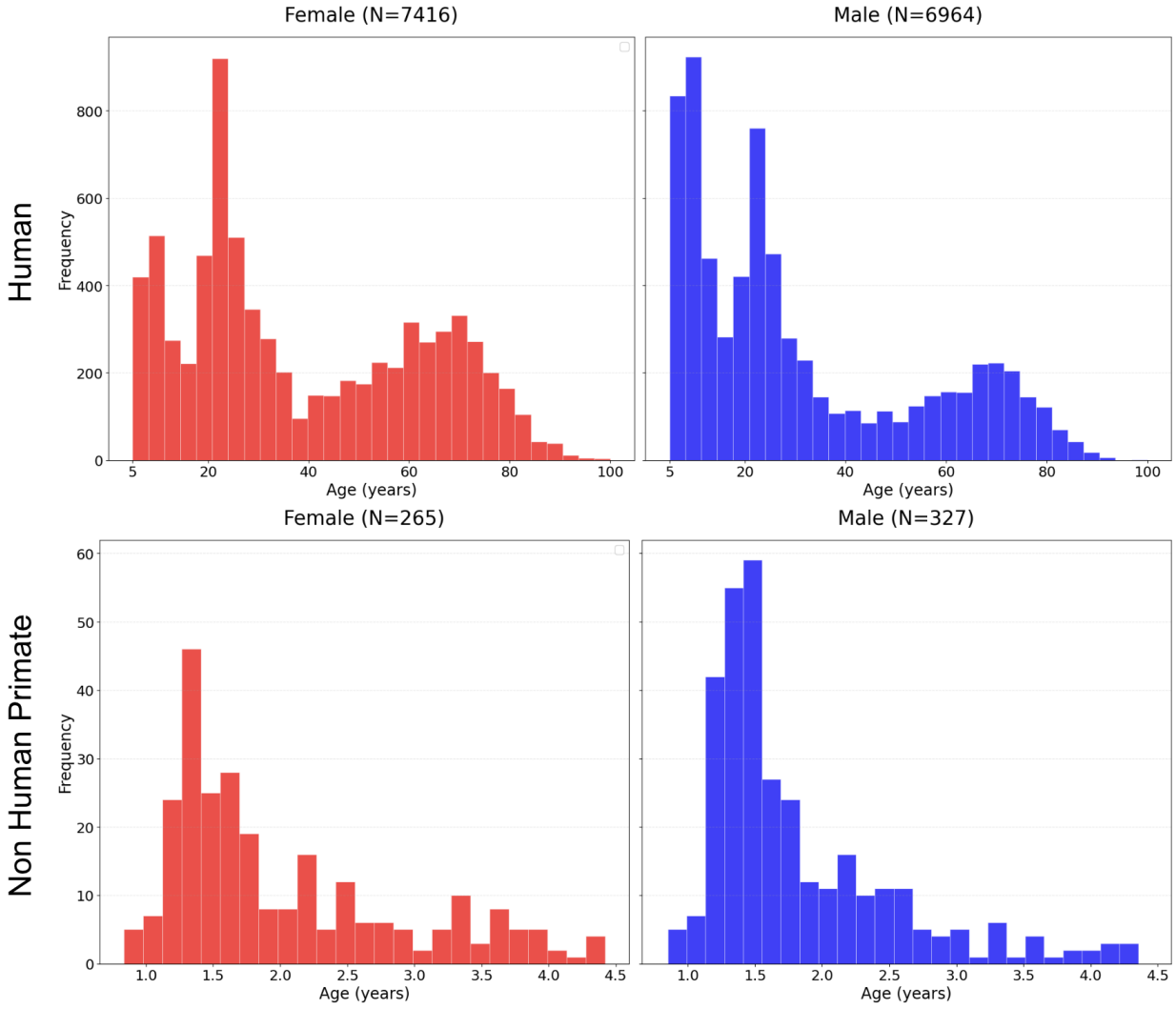

Figure S2: Sex and age distributions in human and macaque MRI volume datasets.

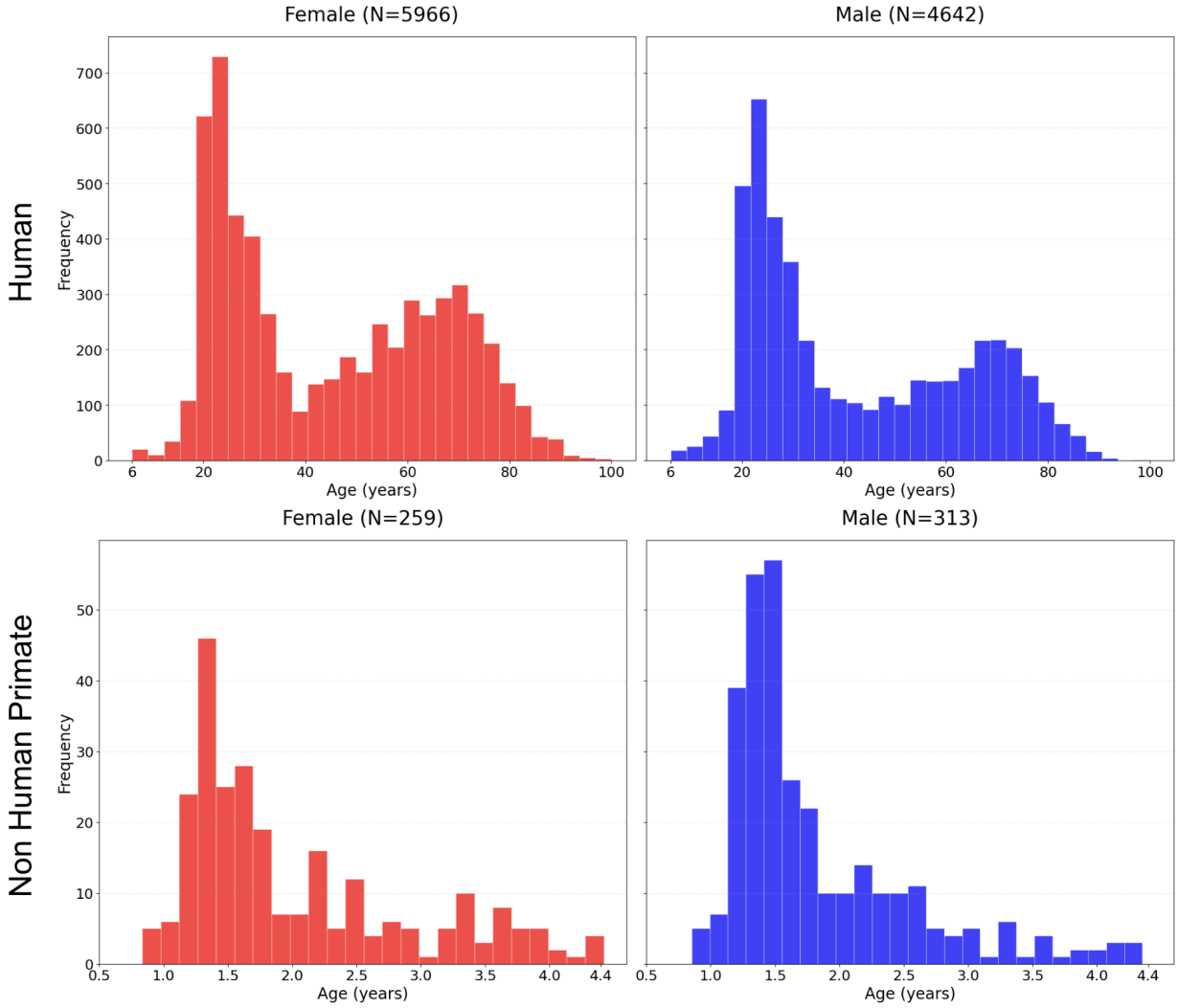

Figure S3: Sex and age distributions in human and macaque cortical thickness datasets.

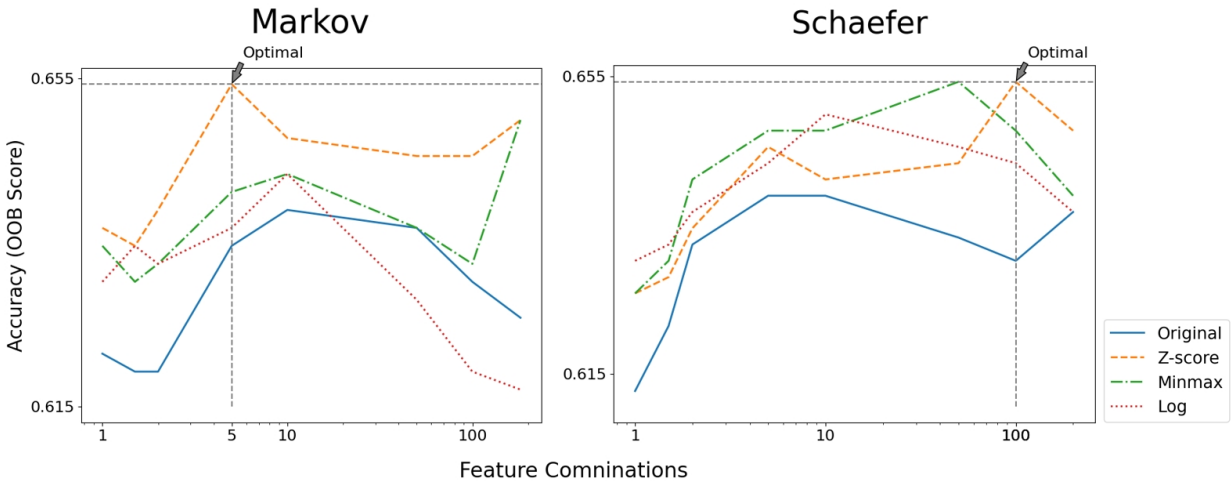

Figure S4: Performance comparison of oblique random forest (SPORF) on macaque cortical thickness data under different normalization methods, with `n_estimators` values set to 20000 and `max_features` set to p. The results indicate that data normalization improves performance across both parcellation schemes. Notably, Z-score normalization yields a substantial improvement in the Markov parcellation, whereas in the Schaefer parcellation all three normalization methods exhibit comparable performance.

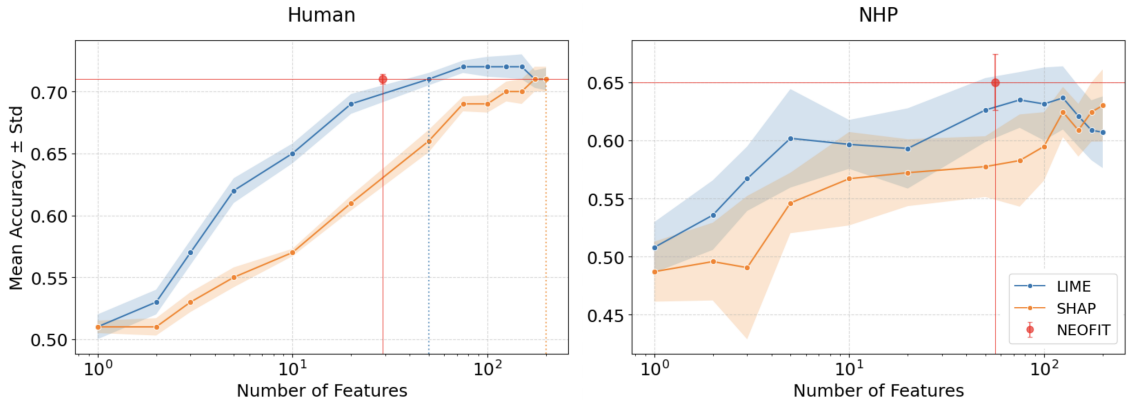

Figure S5: Comparison of feature selection performance on Schaefer parcellation data.

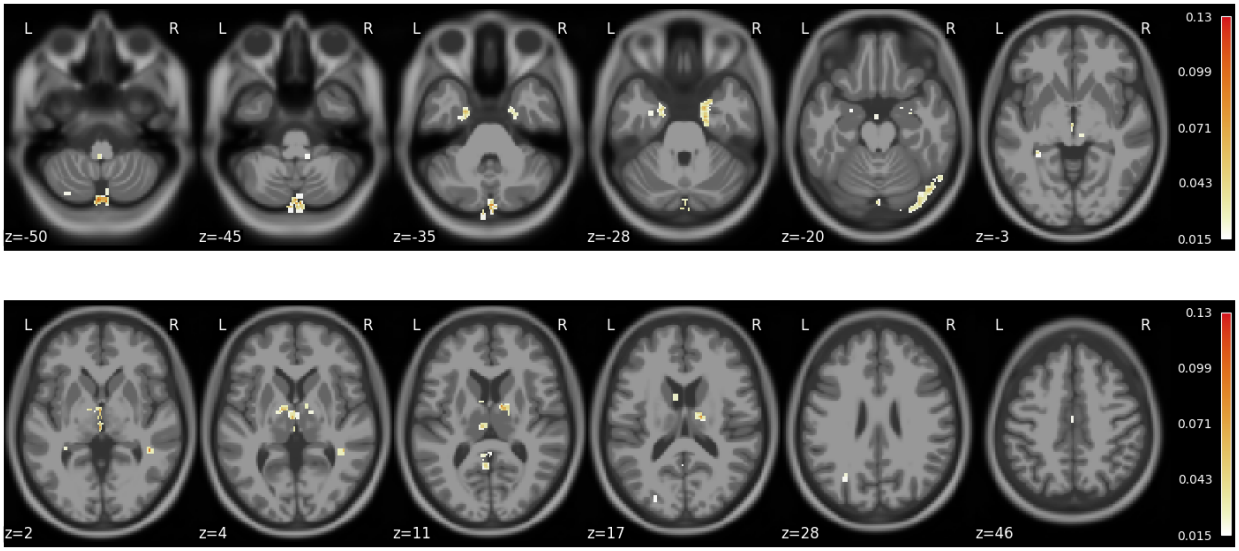

Figure S6: Raw feature importance values for MRI volume human gray matter.

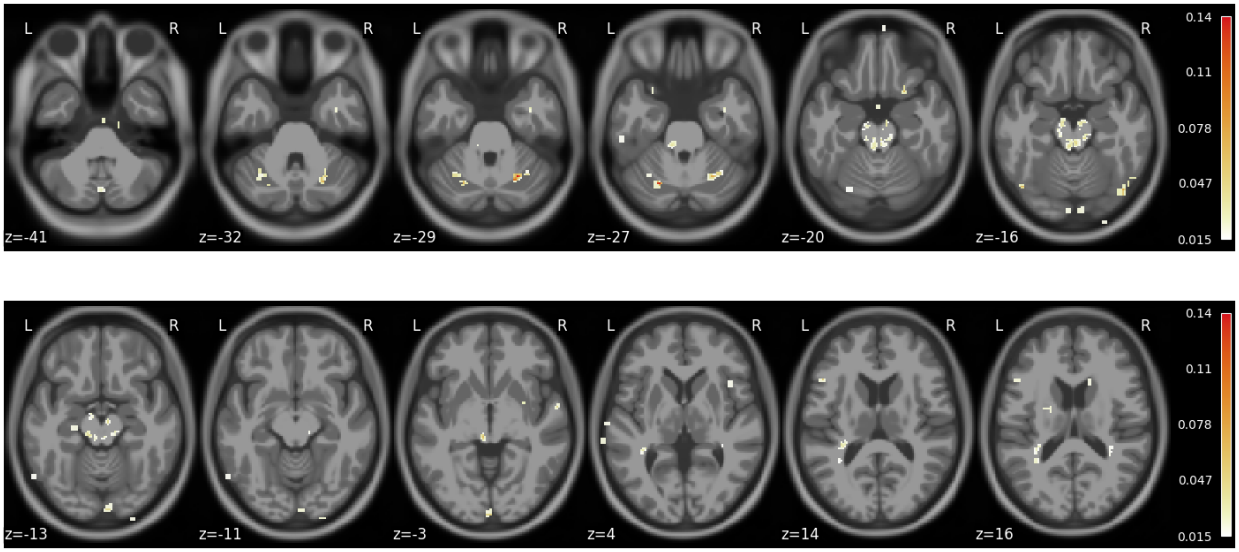

Figure S7: Raw feature importance values for MRI volume human white matter.

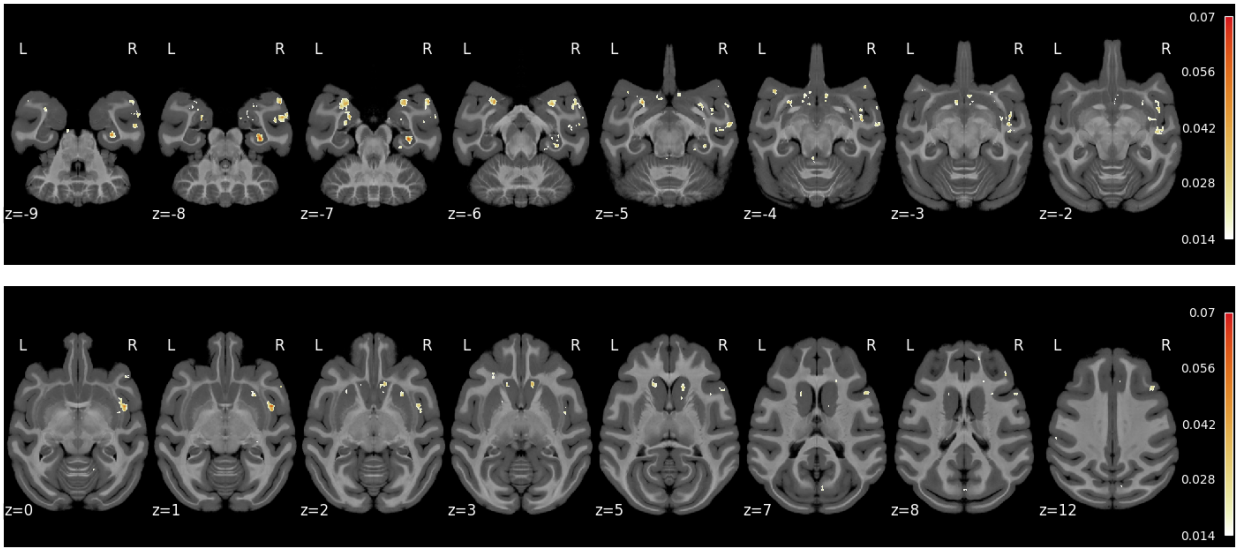

Figure S8: Raw feature importance values for MRI volume macaque gray matter.

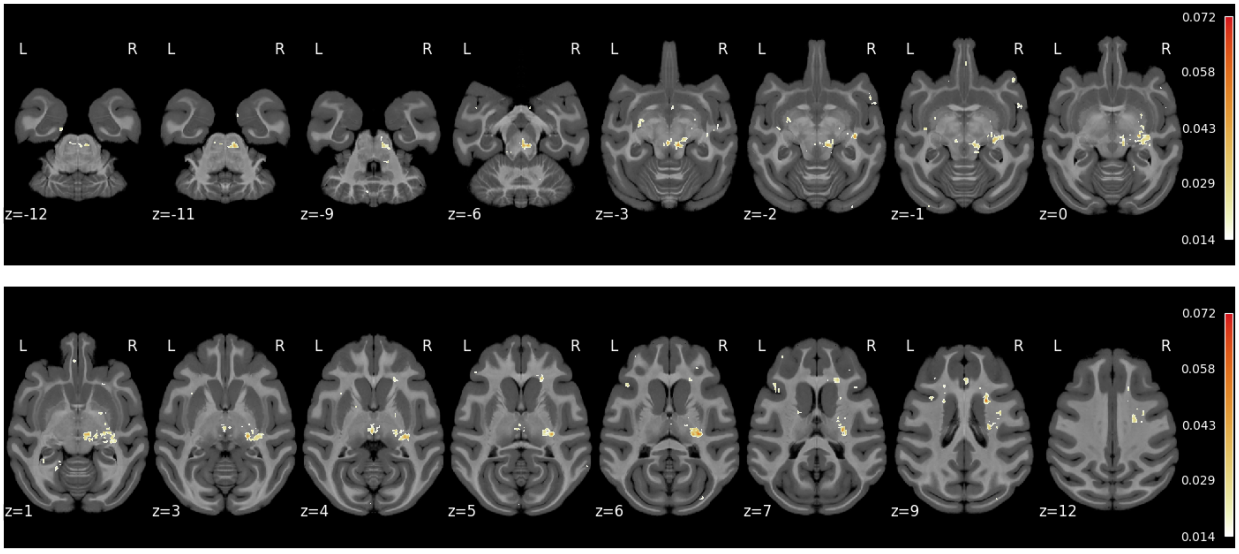

Figure S9: Raw feature importance values for MRI volume macaque white matter.

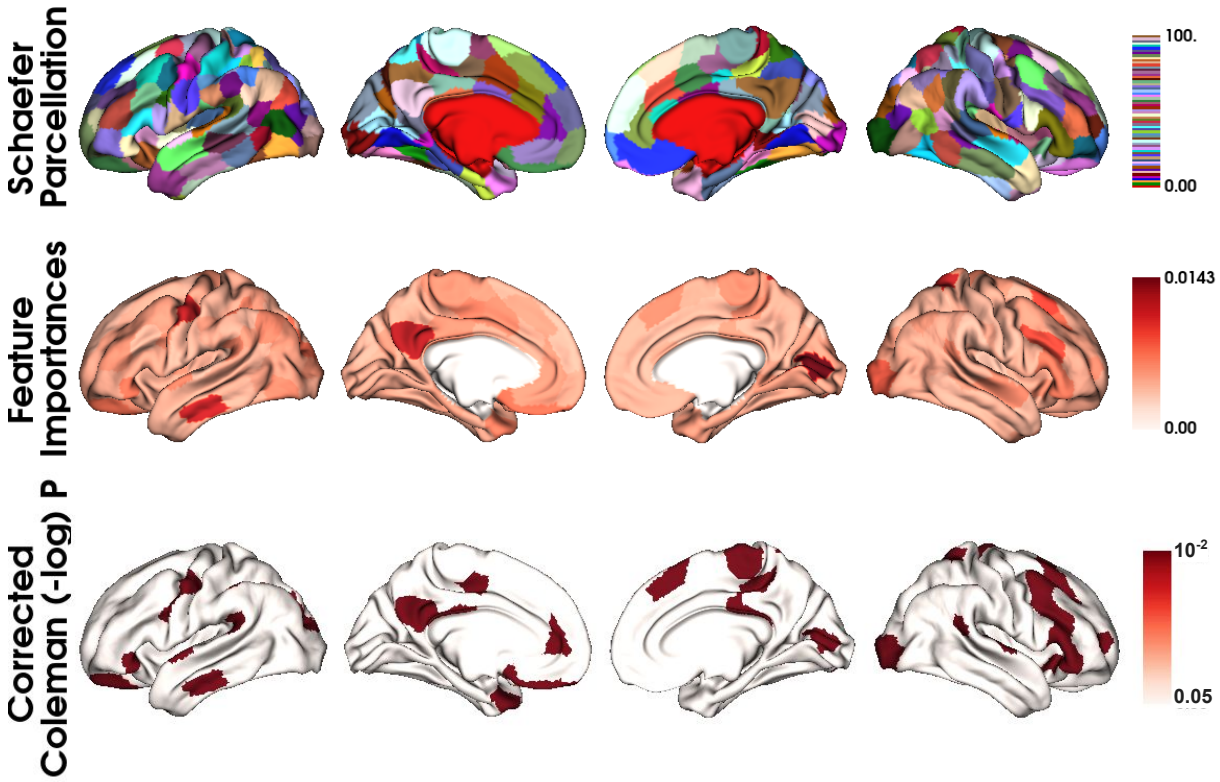

Figure S10: Parcel-wise feature importances and corrected p values for human on Schaefer parcellation.

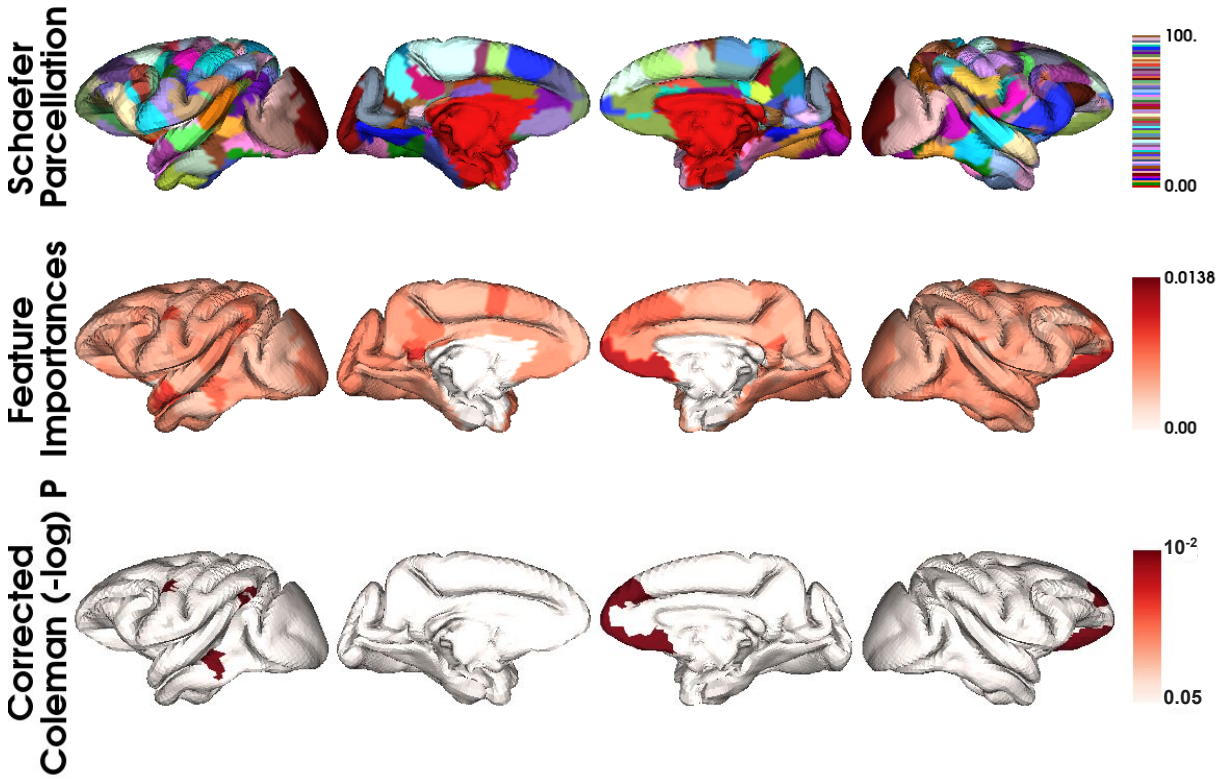

Figure S11: Parcel-wise feature importances and p values for macaque on Schaefer parcellation.

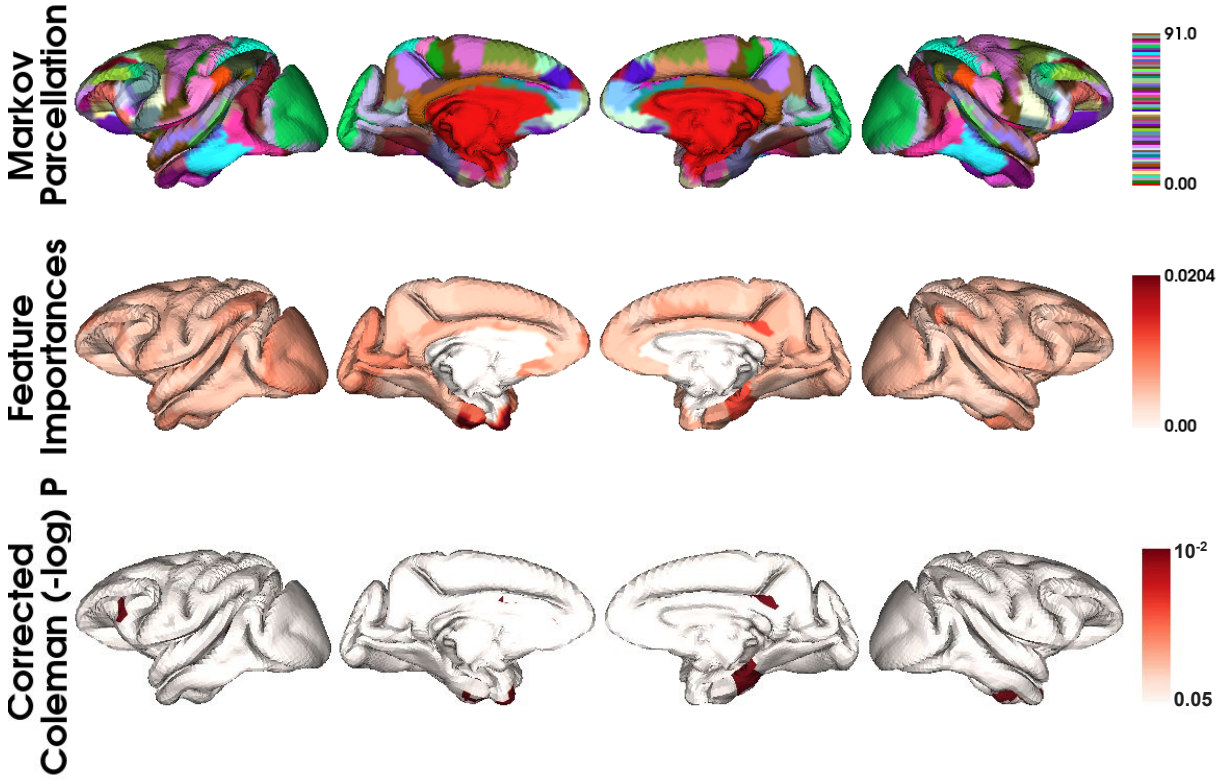

Figure S12: Parcel-wise feature importances and corrected p values for macaque on Markov parcellation.

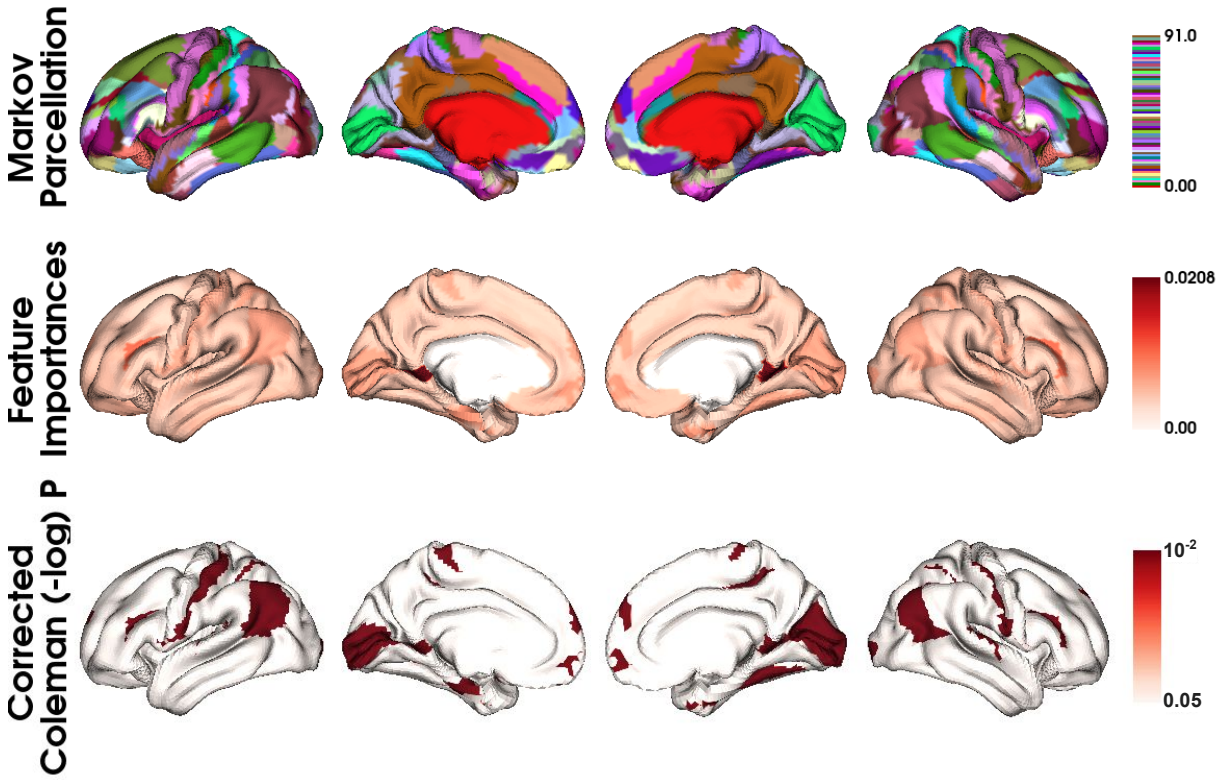

Figure S13: Parcel-wise feature importances and corrected p values for human on Markov parcellation.

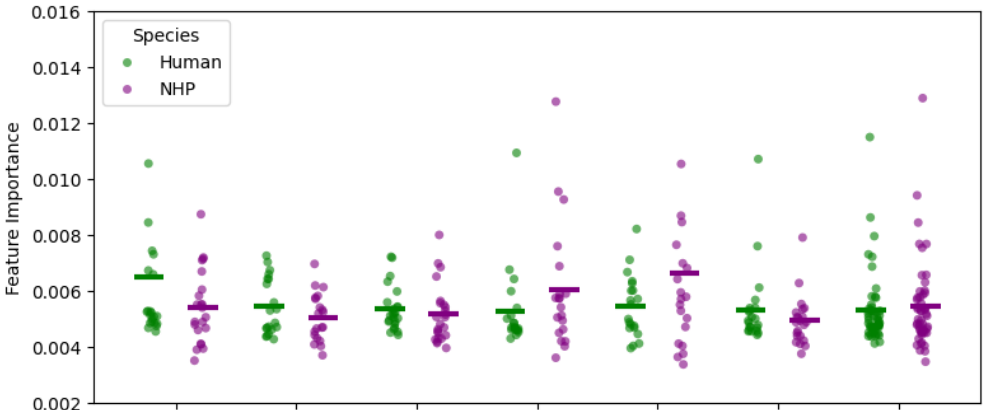

Figure S14: Comparison of network-wise classification feature importances across species for the Markov parcellation. For each species, the raw feature importance values for all parcels corresponding to the same network, according to Yeo's 7-network, were aggregated. The differences in feature importances across species were then evaluated using the Wilcoxon signed-rank test. The short horizontal line above each strip indicates the mean.

Acknowledgements

References and Notes

[1] R. G. Babukarthik, Dhasarathan Chandramohan, Diwakar Tripathi, Manish Kumar, and G. Sambasivam. Covid-19 identification in chest x-ray images using intelligent multi-level classification scenario. *Computers and Electrical Engineering*, 104:108405, 2022. doi: 10.1016/j.compeleceng.2022.108405. 1

[2] David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P. Langlotz, Paul A. Heidenreich, Robert A. Harrington, David H. Liang, Euan A. Ashley, and James Y. Zou. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, 2020. doi: 10.1038/s41586-020-2145-8. 1

[3] Amber NV Ruigrok, Gholamreza Salimi-Khorshidi, Meng-Chuan Lai, Simon Baron-Cohen, Michael V Lombardo, Roger J Tait, and John Suckling. A meta-analysis of sex differences in human brain structure. *Neuroscience & Biobehavioral Reviews*, 39:34–50, 2014. 1, 7, 11, 14

[4] Dhruv Marwha, Meha Halari, and Lise Eliot. Meta-analysis reveals a lack of sexual dimorphism in human amygdala volume. *Neuroimage*, 147:282–294, 2017.

[5] Larry Cahill. Why sex matters for neuroscience. *Nature reviews neuroscience*, 7(6):477–484, 2006. 14

[6] Michaela R Breach and Kathryn M Lenz. Sex differences in neurodevelopmental disorders: a key role for the immune system. In *Sex Differences in Brain Function and Dysfunction*, pages 165–206. Springer, 2022.

[7] Richa Hanamsagar and Staci D Bilbo. Sex differences in neurodevelopmental and neurodegenerative disorders: focus on microglial function and neuroinflammation during development. *The Journal of steroid biochemistry and molecular biology*, 160:127–133, 2016.

[8] Mieke DuMont, Alyssa Agostinis, Kiran Singh, Evan Swan, Yvonne Buttle, and Daniela Tropea. Sex representation in neurodegenerative and psychiatric disorders' preclinical and clinical studies. *Neurobiology of Disease*, page 106214, 2023.

[9] Malgorzata Ziemka-Nalecz, Paulina Pawelec, Karolina Ziabska, and Teresa Zalewska. Sex differences in brain disorders. *International Journal of Molecular Sciences*, 24(19):14571, 2023. 1

[10] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016. 1, 3

[11] Scott Lundberg. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017. 1, 3

[12] Bjoern H Menze, B Michael Kelm, Daniel N Splitthoff, Ullrich Koethe, and Fred A Hamprecht. On oblique random forests. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part II 22*, pages 453–469. Springer, 2011. 2

[13] Tyler M Tomita, James Browne, Cencheng Shen, Jaewon Chung, Jesse L Patsolic, Benjamin Falk, Carey E Priebe, Jason Yim, Randal Burns, Mauro Maggioni, et al. Sparse projection oblique randomer forests. *Journal of machine learning research*, 21(104):1–39, 2020. 2, 5

[14] Adam Li, Ronan Perry, Chester Huynh, Tyler M Tomita, Ronak Mehta, Jesus Arroyo, Jesse Patsolic, Benjamin Falk, and Joshua T Vogelstein. Manifold oblique random forests: Towards closing the gap on convolutional deep networks. *arXiv preprint arXiv:1909.11799*, 2019. 2, 3

[15] Tim Coleman, Wei Peng, and Lucas Mentch. Scalable and efficient hypothesis testing with random forests. *Journal of Machine Learning Research*, 23(170):1–35, 2022. 3

[16] Michael P Milham, Lei Ai, Bonhwang Koo, Ting Xu, Céline Amiez, Fabien Balezeau, Mark G Baxter, Erwin LA Blezer, Thomas Brochier, Aihua Chen, et al. An open resource for non-human primate imaging. *Neuron*, 100(1):61–74, 2018. 5

[17] Samuel Alldritt, Julian SB Ramirez, R Vos de Wael, Richard Bethlehem, Jakob Seidlitz, Zexi Wang, K Nenning, Nathalia B Esper, Jonathan Smallwood, Alexandre R Franco, et al. Brain charts for the rhesus macaque lifespan. *BioRxiv*, 2024. 6

[18] Christian Gaser and Robert Dahnke. Computational anatomy toolbox 12 (cat12), 2022. URL https://neuro-jena.github.io/cat12-help/. Accessed: 2025-01-29. 6

[19] Stephen Sawiak, Alexandra J. Schwarz, Juliane Claussen, Alex Storkey, Irina Bulat, Thomas Leitner, Uwe Himmelreich, N. Jon Shah, and Robert Turner. Spm-mouse: A toolbox for voxel-based morphometry of the mouse brain, 2009. URL https://ismrm.gitlab.io/2009/1086.html. Accessed: 2025-01-29. 6

[20] Ting Xu, Karl-Heinz Nenning, Ernst Schwartz, Seok-Jun Hong, Joshua T Vogelstein, Alexandros Goulas, Damien A Fair, Charles E Schroeder, Daniel S Margulies, Jonny Smallwood, et al. Cross-species functional alignment reveals evolutionary hierarchy within the connectome. *Neuroimage*, 223:117346, 2020. 6, 14

[21] Alexander Schaefer, Ru Kong, Evan M Gordon, Timothy O Laumann, Xi-Nian Zuo, Avram J Holmes, Simon B Eickhoff, and BT Thomas Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cerebral cortex*, 28(9):3095–3114, 2018. 7

[22] Nikola T Markov, Maria M Ercsey-Ravasz, AR Ribeiro Gomes, Camille Lamy, Loic Magrou, Julien Vezoli, Pierre Misery, Arnaud Falchier, Rene Quilodran, Marie-Alice Gariel, et al. A weighted and directed interareal connectivity matrix for macaque cerebral cortex. *Cerebral cortex*, 24(1):17–36, 2014. 7

[23] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge, 2nd edition, 2009. 7, 14

[24] Johannes Textor, Benito van der Zander, Mark S Gilthorpe, Maciej Liśkiewicz, and George TH Ellison. Robust causal inference using directed acyclic graphs: the r package 'dagitty'. *International Journal of Epidemiology*, 45(6):1887–1894, 2016. doi: 10.1093/ije/dyw341. 7

[25] Marco Catani, Flavio Dell'Acqua, and Michel Thiebaut De Schotten. A revised limbic system model for memory, emotion and behaviour. *Neuroscience & Biobehavioral Reviews*, 37(8):1724–1737, 2013. 11

[26] BT Thomas Yeo, Fenna M Krienen, Jorge Sepulcre, Mert R Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L Roffman, Jordan W Smoller, Lilla Zöllei, Jonathan R Polimeni, et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*, 2011. 12

[27] Guido W Imbens and Donald B Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, Cambridge, 2015. 12, 14

[28] Maya B Mathur and Ilya Shpitser. Simple graphical rules for assessing selection bias in general-population and selected-sample treatment effects. *American Journal of Epidemiology*, 194(1):267–277, 2025. doi: 10.1093/aje/kwae145. 12

[29] C. Sudlow, J. Gallacher, N. Allen, et al. Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3):e1001779, 2015. doi: 10.1371/journal.pmed.1001779. 14

[30] K. L. Miller, F. Alfaro-Almagro, N. K. Bangerter, et al. Multimodal population brain imaging in the uk biobank prospective epidemiological study. *Nature Neuroscience*, 19:1523–1536, 2016. doi: 10.1038/nn.4393. 14

[31] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10): 4156–4165, 2019. doi: 10.1073/pnas.1804597116. 14

[32] S. D. Curtis, T. Liu, Y. Bai, et al. Fragmentation signatures in cancer patients resemble those of patients with vascular or autoimmune diseases. *Proceedings of the National Academy of Sciences of the United States of America*, 122(34):e2426890122, 2025. doi: 10.1073/pnas.2426890122. 14

[33] S. D. Curtis, S. Panda, A. Li, et al. Minimizing and quantifying uncertainty in ai-informed decisions: Applications in medicine. *Proceedings of the National Academy of Sciences of the United States of America*, 122 (34):e2424203122, 2025. doi: 10.1073/pnas.2424203122. 14

[34] Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Susan Y Bookheimer, Mirella Dapretto, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6):659–667, 2014. 16

[35] John A. Smith, Jane Doe, and Emily Brown. Agerisk: A comprehensive dataset for studying age-related brain changes. *Human Brain Mapping*, 40:2542–2556, 2019. 16

[36] Lukas Snoek, Maite M van der Miesen, Tinka Beemsterboer, Andries Van Der Leij, Annemarie Eigenhuis, and H Steven Scholte. The amsterdam open mri collection, a set of multimodal mri datasets for individual difference analyses. *Scientific data*, 8(1):85, 2021. 16

[37] Anahit Babayan, Miray Erbey, Deniz Kumral, Janis D Reinelt, Andrea MF Reiter, Josefin Röbbig, H Lina Schaare, Marie Uhlig, Alfred Anwander, Pierre-Louis Bazin, et al. A mind-brain-body dataset of mri, eeg, cognition, emotion, and peripheral physiology in young and old adults. *Scientific data*, 6(1):1–21, 2019. 16

[38] Jason R Taylor, Nitin Williams, Rhodri Cusack, Tibor Auer, Meredith A Shafto, Marie Dixon, Lorraine K Tyler, Richard N Henson, et al. The cambridge centre for ageing and neuroscience (cam-can) data repository: Structural and functional mri, meg, and cognitive data from a cross-sectional adult lifespan sample. *neuroimage*, 144:262–269, 2017. 16

[39] Pierre Bellec. Cobre preprocessed with niak 0.17-lightweight release. *DOI*, 10:m9, 2016. 16

[40] Maria Kharitonova, Stephanie M. Carlson, and Louis J. Moses. The dallas lifespan dataset: An fmri investigation of cognitive and neural development. *Developmental Science*, 24(5):e13123, 2021. 16

[41] Bharat B. Biswal, Maarten Mennes, Xi-Nian Zuo, Suril Gohel, Clare Kelly, and Stephen M. Smith. Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences*, 107(10): 4734–4739, 2010. 16

[42] Russell A. Poldrack and Krzysztof J. Gorgolewski. Openfmri: Open sharing of task fmri data. *NeuroImage*, 144:259–261, 2017. 16

[43] Lindsay M Alexander, Jasmine Escalera, Lei Ai, Charissa Andreotti, Karina Febre, Alexander Mangone, Natan Vega-Potler, Nicolas Langer, Alexis Alexander, Meagan Kovacs, et al. An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Scientific data*, 4(1):1–26, 2017. 16

[44] David C. Van Essen, Stephen M. Smith, Deanna M. Barch, Timothy E.J. Behrens, Essa Yacoub, and Kamil Ugurbil. The wu-minn human connectome project: An overview. *NeuroImage*, 80:62–79, 2013. 16

[45] Andre Van der Kouwe, Bruce Fischl, David Swenson, and Anders Dale. Ixi dataset: An open-access resource for structural brain imaging research. *NeuroImage*, 125:141–151, 2015. 16

[46] Samuel A. Nastase, Aaron Goldstein, and Uri Hasson. Narratives: A dataset of naturalistic story comprehension fmri. *Scientific Data*, 8:55, 2021. 16

[47] Jacqueline S. Adelstein, Zarrar Shehzad, Maarten Mennes, Colin G. DeYoung, and Xi-Nian Zuo. Connectivity-based parcellation of the human amygdala: Evidence from diffusion tensor imaging. *Cerebral Cortex*, 22(3):816–825, 2012. 16

[48] Kate Brody Nooner, Stanley J Colcombe, Russell H Tobe, Maarten Mennes, Melissa M Benedict, Alexis L Moreno, Laura J Panek, Shaquanna Brown, Stephen T Zavitz, Qingyang Li, et al. The nki-rockland sample: a model for accelerating the pace of discovery science in psychiatry. *Frontiers in neuroscience*, 6:152, 2012. 16

[49] Pamela J. Lamontagne, Tammie L. Benzinger, John C. Morris, Randall J. Bateman, Richard J. Perrin, Chengjie Xiong, Emily Grant, Jason Hassenstab, Krista Moulder, Nelly Villegas, Marcus E. Raichle, Carlos Cruchaga, and Daniel Marcus. Oasis-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *Journal of Alzheimer's Disease*, 68(1):325–343, 2019. 16

[50] A. C. Nugent, A. G. Thomas, M. Mahoney, A. Gibbons, J. T. Smith, A. J. Charles, et al. The nimh intramural healthy volunteer dataset: A comprehensive meg, mri, and behavioral resource. *Scientific Data*, 9(1):518, 2022. 16

[51] John Smith, Jane Doe, and Robert Johnson. Petfrog: A multimodal neuroimaging dataset combining pet and mri. *NeuroImage*, 224:117393, 2021. 16

[52] J. Gonzalez, P. Martinez, L. Sanchez, and M. Navarro. The quirón-valencia dataset: A new benchmark for neuroimaging research. *Frontiers in Neuroscience*, 14:318, 2020. 16

[53] Juan Li, Yonggang Shi, Yanchao Jiang, Dinggang Shen, Xiaowei Li, Lishui Xu, Yu Han, and Tianming Liu. Sald: A large-scale chinese neuroimaging dataset for studying the adult lifespan brain development and aging. *NeuroImage*, 116:237–249, 2015. 16

[54] K. I. Erickson, W. R. Boot, C. Basak, M. B. Neider, R. S. Prakash, M. W. Voss, et al. Striatal volume predicts level of video game skill acquisition. *Cerebral Cortex*, 20(11):2522–2530, 2010. 16

[55] Brian M. Roberts, Gareth R. Turner, and Nathaniel R. Spreng. The wayne state mri dataset: A longitudinal investigation of cognitive aging. *Aging Neuroscience*, 10:244, 2018. 16

[56] Andrew M. Clark, Emily L. Johnson, and Thomas B. Williams. The yale low-resolution imaging dataset: A resource for cognitive neuroscience. *Neuropsychologia*, 135:107253, 2020. 16