Pertanika J. Sci. & Technol. 34 (5): 2947 - 2980 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Privacy-Preserving Federated Learning with Explainable AI for Multi-Site Autism Spectrum Disorder Biomarker Discovery

Anusha Ravankol*, and Nandula Anuradha

Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Aziznagar, Hyderabad 500075, Telangana, India

ABSTRACT

This study develops a privacy-preserving federated learning framework, coupled with explainable AI, for collaborative discovery of autism spectrum disorder (ASD) biomarkers without centralising sensitive patient data. A federated system spanning three simulated hospital sites that fuses structural and functional MRI via cross-modal attention is evaluated with differential privacy applied to both model aggregation (Gaussian mechanism, calibrated with a tight Rényi differential privacy accountant) and biomarker discovery (Laplace mechanism, Fed-SHAP). On a held-out test set drawn from a 173-subject cohort (and never used for training or hyperparameter selection), the proposed model achieves 92.6% accuracy (95% bootstrap CI: [81.5%, 100.0%]; permutation test $p = 0.001$ against a chance-level mean of 50.4%) at a composed privacy guarantee of $\varepsilon = 1.0$, $\delta = 10^{-5}$ over 10 communication rounds. Leave-one-site-out validation, in which each site is held out entirely as an external test set, yields a mean accuracy of 86.7% (SD = 4.4%), providing a more conservative estimate of cross-institution generalisation. Fed-SHAP, a privacy-preserving feature-importance approximation rather than a direct computation of Shapley values, ranked 20 candidate functional-connectivity biomarkers, which we mapped to AAL-116 anatomical regions and networks; the most frequently implicated networks were cerebellar and association-cortex regions (25.0% each), followed by subcortical/limbic (15.0%) and visual (12.5%) regions. Because individual rankings showed only 20.6% bootstrap stability, these candidate biomarkers are exploratory and are reported at the network level rather than as fixed individual connections.

Keywords: Autism spectrum disorder, differential privacy, explainable AI, federated Learning, multimodal neuroimaging, privacy-preserving machine learning

ARTICLE INFO

Article history:
Received: 08 July 2026
Accepted: 29 September 2026
Published: 30 October 2026

DOI: https://doi.org/10.47836/pjst.34.5.17

ranusha82024@gmail.com (Anusha Ravankol
anuradha@klh.edu.in (Nandula Anuradha)
* Corresponding author

INTRODUCTION

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental disorder

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

affecting about 1 in 100 children worldwide, characterised by persistent difficulties in social communication, restricted interests, and repetitive behaviours. Early and accurate diagnosis is important to ensure timely intervention and improve long-term prognosis, but current diagnostic protocols rely on subjective behavioural assessments and clinical observations, which are time-consuming and often delayed until symptoms have become evident. Moreover, ASD is heterogeneous in its genetic origins, symptom profiles, and biological bases, which poses further barriers to reliable diagnosis and underlines the urgent need for objective biomarkers that can reflect the underlying brain changes associated with the condition. Neuroimaging techniques, notably structural magnetic resonance imaging (sMRI) and functional magnetic resonance imaging (fMRI), have shown increasing promise for objective ASD characterisation by revealing anatomical aberrations and functional connectivity disruptions that distinguish affected individuals from neurotypically developing individuals (Jabbar et al., 2025; Narayanan et al., 2025; Xu et al., 2025). Advances in deep learning, particularly convolutional neural networks and improved network-training algorithms (Movassagh et al., 2023), have substantially advanced medical image analysis. Building effective diagnostic models from neuroimaging data, however, requires large, heterogeneous datasets that represent the full breadth of the autism spectrum. Single-institution studies face numerous constraints: small sample sizes, limited population diversity, and institution-specific factors, including variations in scanning protocols, study designs, and demographics. Multicentre collaborations hold considerable potential for addressing these issues; however, traditional data-centralisation approaches, which aggregate raw medical data at a central location for analysis (Alutaibi et al., 2025), raise privacy, regulatory, patient-consent, and ethical concerns that are often difficult to surmount.

Federated learning has emerged as a paradigm-shifting approach because it enables collaborative model development across distributed institutions without centrally pooling data. In federated learning, each participating site independently trains a local copy of the model on its private dataset and shares only the model parameters, never the raw data, with a central aggregation server that combines all sites' updates into an improved global model. This architecture, therefore, keeps data local and regulation-compliant, and allows institutions that could not otherwise share data to participate in collaborative research (Eslami & Saeed, 2019; Subah et al., 2021). Nevertheless, naive federated learning remains vulnerable to several privacy attacks: shared gradients can leak sensitive information about individual patients; membership inference attacks identify whether specific individuals participated in training; and model inversion techniques reconstruct training data from shared parameters (Bayram et al., 2021; Eslami et al., 2019). These vulnerabilities motivate formal privacy guarantees through differential privacy, which injects calibrated noise into shared information so that the contribution of any individual patient cannot be reliably distinguished from aggregate statistics. Beyond privacy, a major challenge for artificial

intelligence (AI) in clinical decision support is the "black box" problem: deep learning models often achieve high accuracy without interpretable explanations of their predictions (Qayyum et al., 2022; Sabegh et al., 2023; Yin et al., 2021), which limits clinician trust and hinders the identification of neurobiological mechanisms underlying autism. Explainable AI (XAI) methods such as SHAP, Grad-CAM, and attention mechanisms offer interpretability by identifying which input features or brain regions most strongly influenced model decisions. However, traditional XAI approaches assume access to complete datasets and model internals, which conflicts with the data-localisation principle of federated learning. Developing privacy-preserving explainability methods that can identify interpretable biomarkers across distributed sites without exposing sensitive patient information represents a fundamental open challenge at the intersection of federated learning, differential privacy, and explainable AI (Qiqieh et al., 2026).

The remainder of this paper is organised as follows. The Literature Review surveys existing single-site and multi-site ASD classification approaches, the limitations of centralised data pooling, and the state of interpretability and validation practices in this domain. The Methodology section describes the federated learning framework; the differential-privacy mechanisms implemented during aggregation and biomarker discovery; the explainable multimodal architecture; the Fed-SHAP approximation and its validation against standard SHAP; and the full validation protocol (subject-level splitting, leave-one-site-out evaluation, ablation design, baseline comparisons, and statistical testing). The Results section reports the validated held-out and cross-site performance, the ablation and baseline comparisons, the privacy-utility trade-off, and the biomarker-anatomy mapping, together with the statistical uncertainty and stability of each finding. The Discussion interprets these results, states the study's limitations explicitly - including sample-size-driven ceiling effects on the ablation study and run-to-run variability in biomarker rankings and the Conclusion summarises the contributions and future directions.

Contributions

- A federated learning framework enabling privacy-preserving multi-site autism biomarker discovery, with differential privacy applied at model aggregation as well as at explanation computation.
- Fed-SHAP, a privacy-preserving feature-importance approximation (not a direct computation of Shapley values), compared against conventional SHAP (Spearman $\rho = 0.227$) and mapped to AAL-116 anatomy so that candidate biomarkers are neuroanatomically interpretable.
- A full validation protocol - leak-free subject-level splitting, leave-one-site-out external validation (mean accuracy 86.7%), bootstrap confidence intervals, permutation significance testing, an ablation study, expanded baselines, and a privacy-utility

sweep across $\varepsilon = 0.1–50$ - reported with explicit statistical uncertainty and sample-size limitations rather than a single unvalidated accuracy figure.

LITERATURE REVIEW

ASD affects about 1 in 100 children worldwide. The condition is generally regarded as a complex neurodevelopmental disorder associated with difficulties in symbolic, social, or pragmatic communication, as well as restricted interests and/or repetitive behaviour (Kohli et al., 2025). The complexity of ASD and the lack of large, diverse datasets available to individual research groups have posed major challenges to the development of robust diagnostic and predictive models. Multi-centre neuroimaging efforts, such as the Autism Brain Imaging Data Exchange (ABIDE) consortium, which pooled data from 1,035 subjects across 17 research institutions (Almuqhim & Saeed, 2021; Huang et al., 2021), have been recognised as important tools for effective ASD research. However, pooling such sensitive medical data across institutions raises major privacy, regulatory, and ethical concerns (Kohli et al., 2025). Integrating multi-site neuroimaging data also poses inherent technical challenges that have significantly impeded the formulation of generalisable ASD classification models. ABIDE dataset-based studies have consistently reported high levels of heterogeneity based on differences in scanner manufacturers, acquisition protocols, pre-processing pipelines, and population characteristics across participating sites (Kang et al., 2025; Santana et al., 2022; Tang et al., 2023). Tang et al. (2023) specifically addressed multisite heterogeneity using adversarial deep learning and sliding-window sampling, showing that site-shared feature extraction, while preserving spatiotemporal information, achieved 80% accuracy with balanced sensitivity (81%) and specificity (80%) across diverse sites. Kang et al. (2025) presented a novel Split-Merge-Split dataset partitioning strategy to reduce variability in features extracted during both training and testing, achieving 83.5% accuracy on multi-site classification and 93% on the best-performing single site, OHSU. As the authors note, traditional data-splitting methods do not account for site-specific biases in feature extraction, necessitating new approaches to cross-site validation. Santana et al.'s (2022) meta-analysis of 55 studies reported pooled sensitivities and specificities of 73.8% and 74.8%, respectively, and observed that studies using larger, more diverse datasets reported lower accuracy than single-centre studies.

Similarly, although Huang et al. (2021) reported superior performance on 12 of 17 ABIDE sites, generalisation to the remaining sites was poor, illustrating that site-specific effects limit even complex deep learning models. They also found that automatic hyperparameter tuning and data augmentation, although considered best practice for handling these issues, did not fully resolve them. Deep learning for fMRI-based ASD classification has seen a variety of architectural innovations demonstrating its ability to extract hierarchical features from high-dimensional neuroimaging data. Convolutional

neural networks are well-suited to extracting spatially hierarchical features; Feng and Xu (2023) reported very high accuracy (99.39% accuracy, 98.80% recall, 99.85% precision) using a network that combines convolution, pooling, normalisation, and dropout to classify 22,176 two-dimensional echo-planar images. Similarly, Ahammed et al. (2021) reported 94.70% accuracy on the ABIDE-I NYU dataset with DarkASDNet, using hierarchical features extracted from 3D fMRI data.

Hybrid approaches that combine different network models have been reported to improve performance by adding additional feature-learning stages. Almuqhim and Saeed (2021) proposed ASD-SAENet, which utilises sparse autoencoders for unsupervised feature extraction, coupled with deep neural network classification, reporting 70.8% accuracy while jointly optimising reconstruction and classification errors. Yin et al. (2021) extended this concept by combining autoencoders with deep neural networks applied to raw features, achieving 79.2% accuracy and an AUC of 82.4%, surpassing the state of the art at the time. Recurrent and attention-based architectures have also been studied to capture the temporal patterns of fMRI data and hierarchical structure. Qiang et al. (2023) proposed a hierarchical recurrent variational autoencoder that model's fMRI data at three hidden levels without supervision, using LASSO (Least Absolute Shrinkage and Selection Operator) regression to build hierarchical functional brain networks and achieving 82.1% accuracy, showing discriminative information within interacting hierarchical functional brain networks related to brain disorders. Shao et al. (2021) proposed a deep feature selection and graph convolution network approach (Feng et al., 2022; Karampasi et al., 2020) with weighted feature selection via a sparse one-to-one layer for functional connectivity features, identifying critical intra-hemispheric connections with 79.5% accuracy and an AUC of 0.85.

Further, Banerjee (2025) combined an Electromagnetic Interaction Algorithm for feature selection with an adaptive kernel attention network, yielding accuracies ranging from 90.1% to 98.27% and Cohen's kappa between 0.778 and 0.968 across four datasets, demonstrating enhanced generalisability suitable for clinical applications where interpretability and computational efficiency are paramount. Nafisah et al. (2025) proposed a hybrid scheme combining stacked sparse denoising autoencoders with multilayer perceptrons, optimised via an enhanced Hiking Optimisation Algorithm empowered by dynamic opposite learning and double attractors, achieving an average accuracy of 73.5%, sensitivity of 76.5%, and specificity of 75.2%. Graph-theoretic approaches have provided further insights into network-level biomarkers. Zhang et al. (2023) established, using network topology, that ASD patients exhibit significant reductions in path length and clustering coefficient compared to typically developing controls (Ramli et al., 2021), thereby providing biological validity to network-level ASD biomarkers. Consistent with this, Shao et al. (2021) showed, using deep feature selection, a significantly greater number of weak connections between the cerebral hemispheres in the ASD group than in typically developing controls.

Finally, imaging characteristics of the dopamine system have emerged as promising biomarkers. Li et al. (2025) investigated functional connectivity in six dopamine-related brain regions - including the ventral tegmental area, substantia nigra, prefrontal cortex, nucleus accumbens, and hypothalamus - and found significantly higher functional connectivity in ASD children among multiple dopamine system components compared to typically developing children, with clustering methods identifying two distinct ASD subtypes at over 80% accuracy. The clinical application of machine learning algorithms for ASD diagnosis is heavily dependent on model interpretability and the ability to generate mechanistic insights consistent with neuroscience knowledge. Eslami et al. (2021) underlined the need for explainable, scalable machine learning for ASD detection; their Auto-ASD-Network (Eslami & Saeed, 2019), which couples deep learning with support vector machines, achieved over 70% classification accuracy (up to 80% for the best-performing class), suggesting that interpretability need not come at the cost of performance.

Furthermore, logistic regression on functional connectivity matrix features has been used to identify features important for ASD classification (Firouzi & Fadaei, 2025; Heinsfeld et al., 2018; Ramli et al., 2021). Alves et al. (2023) applied interpretability analysis to organisational features of functional brain networks, showing that ASD patients tend to manifest greater network segregation, diminished information distribution, and diminished connectivity relative to controls, with Bootstrap Analysis of Stable Clusters (BASC) maps achieving an AUC close to 1.0 while identifying diminished connectivity between the left ventral posterior cingulate cortex and cerebellum in ASD patients, consistent with prior neuroscientific findings.

In addition, the vulnerability of the paediatric population studied in ASD research heightens privacy concerns. Centralised data sharing for research purposes traditionally involves extensive de-identification, Institutional Review Board approval, and data-use agreements, all of which create barriers to collaboration (Santana et al., 2022). Re-identification risk also exists even for de-identified data, particularly when multiple datasets are combined and advanced analytics are applied to high-dimensional neuroimaging data. Privacy-preserving data frameworks (Alzubi et al., 2021; Gheisari et al., 2021) and federated learning offer mechanisms to comply with data-protection regulations while enabling collaboration. Despite substantial progress in machine learning for ASD classification, several fundamental challenges persist. First, the majority of studies have relied on centralised learning approaches that require aggregating sensitive medical information; although adversarial learning methods have improved robustness to site heterogeneity, the requirement for centralised data access typically remains (Al-Hiyali et al., 2021; Chandra et al., 2024; Karunakaran et al., 2020), and federated learning frameworks remain comparatively underexplored in ASD classification despite the clear multi-site integration challenges in this literature.

Second, interpretability remains a relatively underdeveloped aspect of most top-performing deep learning models for ASD. Although recent studies have attempted to incorporate post-hoc interpretability approaches (Li et al., 2018; Shao et al., 2021), these approaches often provide limited insight into the underlying decision-making process, and the highest-performing models in this literature (Ahammed et al., 2021; Feng & Xu, 2023) typically offer minimal explanation.

Third, most studies examine only binary classification (ASD vs. typically developing controls) without investigating behavioural phenotypes or prognosis (Santana et al., 2022); notable exceptions establishing ASD subtypes include Li et al. (2025) (80% accuracy) and Al-Hiyali et al. (2021) (82.1% accuracy for multi-class classification), though neither directly relates predicted subtypes to treatment outcomes.

Fourth, validation protocols in this literature frequently rely on cross-validation within the same dataset used for development rather than truly independent test data. Leave-one-site-out validation provides a more compelling demonstration of generalisation (Kang et al., 2023; Yousefian et al., 2023), though even this may not fully approximate real-world clinical deployment across settings that differ substantially from those used in development (Belhaouari et al., 2023). Incorporating additional information beyond neuroimaging has also been shown to improve classification: Santana et al. (2022) meta-analysis found that combining phenotypic and imaging information achieved better sensitivity (84.7%) than rs-fMRI information alone (72.8%), and Huang et al. (2021) improved accuracy by 6.4% over the previous best-reported ABIDE result by combining phenotypic information with graph convolutional networks for functional connectivity.

Other biomarker modalities have shown diagnostic promise that could complement a federated neuroimaging framework: Shazman et al. (2025) achieved 85% accuracy using urine metabolomic profiling with random forest classifiers across 293 metabolites; Pinto et al. (2025) reported 92% sensitivity and 95% specificity using saliva Fourier Transform Infrared Spectroscopy with Support Vector Machines; and Islam et al. (2025) discussed integrating IoT devices with machine learning for continuous, ecologically valid behavioural monitoring. The confluence of federated learning, explainable AI, and multisite biomarker discovery represents an emerging opportunity in ASD research, requiring: (1) federated architectures that accommodate site heterogeneity while learning shared, diagnostically relevant representations (Kang et al., 2025; Tang et al., 2023); (2) explainability that yields biomarkers interpretable in light of neuroscientific understanding (Alves et al., 2023; Eslami et al., 2021; Li et al., 2018); (3) privacy preservation that protects patient confidentiality while enabling collaboration (Sachdeva et al., 2024); (4) rigorous validation that establishes generalisation across sites and populations (Jönemo et al., 2023); and (5) stakeholder engagement that aligns technology development with clinical needs and ethical principles (Kohli et al., 2025).

METHODOLOGY

Federated Learning Framework Design

The aim of this study is to develop a privacy-preserving federated learning system that enables distributed neuroimaging analysis without centralising patient data. The core components of the federated framework comprise the central aggregation server, the autonomous client nodes, and the communication protocol (Almuqhim & Saeed, 2021). Unlike centralised learning, in which the model is trained on patient data collected at a single location, the proposed framework enables participating sites to retain complete control over their patient data while learning in a distributed environment. The framework is an instance of the Federated Averaging (FedAvg) algorithm, in which model parameters are aggregated across participating clients via the weighted-averaging procedure of Equation 1. As iterations advance, the global model learns common aspects of the patient population while preserving the distributed environment.

$$\theta^{g}{}_{lo}{}^{b}{}_{al}{}^{t+1} = \theta^{g}{}_{lo}{}^{b}{}_{al}{}^{t} + \Sigma_i \, (n_i/N_{total}) \cdot [\, clip(\, \theta_i{}^{t+1} - \theta^{g}{}_{lo}{}^{b}{}_{al}{}^{t}, \, C\,) + N(0, \sigma^2 C^2 I)\,] \quad [1]$$

Equation 1 includes two elements that are essential to the privacy guarantee. First, the client update $\theta_i^{t+1} - \theta^{g}{}_{lo}{}^{b}{}_{al}{}^{t}$ is explicitly clipped to an L2-norm bound C, adaptively set each round from the observed median client-update norm (Andrew et al., 2021), before any noise is added; without adaptive clipping, a fixed clip bound calibrated arbitrarily either destroys the signal (bound too small) or fails to control sensitivity (bound too large). Second, the Gaussian noise term $N(0, \sigma^2C^2I)$ is applied to the aggregated model update itself, so that differential privacy protects the model parameters and not only the Fed-SHAP feature-importance computation (Equation 2). We further found that applying this noise mechanism to the entire ~24-million-parameter network is infeasible at the present sample size: because the injected noise vector's expected magnitude scales with the square root of the number of protected parameters, it overwhelms the clipped signal by one to three orders of magnitude regardless of the clipping strategy used. We therefore scope the formal (ε, δ)-differential-privacy guarantee to the final classifier layer (~514 parameters) rather than the full network; the feature-encoding layers are federated-averaged without a formal per-parameter privacy guarantee. This is a materially narrower claim than full-model differential privacy.

The system uses three independent hospital site simulators that mimic realistic scenarios of collaboration in autism research. Subjects are partitioned across sites using stratified sampling to balance ASD and control cases while preserving realistic between-site heterogeneity. Each site further splits its subjects into training, validation, and test partitions at the subject level (70%/15%/15%), with no subject appearing in more than one partition

and no test-partition data used in any hyperparameter decisions. Hospital sites operate as independent learning agents with freedom to choose local training procedures, consistent with the operational independence of real healthcare facilities (Ahammed et al., 2021). The server aggregates client model parameters without access to raw data, intermediate results, or patient data, communicating over 10 federated rounds, each with 5 local training epochs.

Privacy-Preserving Mechanisms

The framework applies differential privacy at two distinct points, using two mechanisms appropriate to each: a Gaussian mechanism for global model aggregation (Equation 1) and a Laplace mechanism for Fed-SHAP feature-importance aggregation (Equation 2). The Gaussian mechanism protects model updates, and the Laplace mechanism protects the feature-importance statistics separately; only the former is described by Equation 1.

For the Gaussian mechanism, the noise multiplier σ is calibrated using a Rényi Differential Privacy (RDP) accountant (Mironov, 2017; implemented via Opacus) rather than a hand-derived composition bound, because RDP accounting is substantially tighter for composing many rounds of a Gaussian mechanism and is the standard tool for this purpose in the differential-privacy literature. At the manuscript's primary target of $\varepsilon = 1.0$ over 10 communication rounds ($\delta = 10^{-5}$, client-level DP with all three sites participating every round, i.e. sample rate 1.0), the calibration yields a noise multiplier $\sigma \approx 81.1$ relative to the adaptively set clip bound, and the resulting composed guarantee is exactly ($\varepsilon = 1.0$, $\delta = 10^{-5}$). As noted above, this guarantee is scoped to the classifier head (~514 parameters); we verified numerically that, with only three always-participating sites and no subsampling amplification, protecting the full ~24-million-parameter network at $\varepsilon \leq 50$ essentially destroys all predictive signal. Within this classifier-head scope, the accountant also produced valid calibrations for tighter budgets, down to $\varepsilon = 0.1$, which are evaluated in the privacy-utility sweep (Table 5).

For Fed-SHAP, each client computes a local, privatised feature-importance vector using the Laplace mechanism: Laplace noise is added with scale calibrated to the L1-sensitivity of the importance statistic (the maximum possible change in the statistic from adding or removing one subject) divided by the local privacy budget ε_i, before any local statistic is transmitted to the central server (Al-Hiyali et al., 2021). At the global aggregation level, privatised importance vectors are combined across sites via a weighted average based on local dataset sizes, and the composition of the privacy budget across the biomarker-discovery step is tracked separately from the model-aggregation budget described above. The overall design is intended to limit the extent to which the shared model parameters and the released feature-importance statistics could be used to reconstruct individual patient data or to determine which site contributed a given pattern.

Explainable Multi-Modal Architecture

The federated learning system utilises an Explainable Multi-Modal Deep Learning (Ex-MMDL) approach that is designed to make the influence of neurobiological features on ASD classification interpretable while remaining compatible with federated optimisation, as shown in Figure 1. The Ex-MMDL model extends a cross-modal attention fusion model by incorporating explainability at several levels. For sMRI, the Explainable sMRI Convolutional Neural Network (CNN) is a modified ResNet-18 that supports Gradient-weighted Class Activation Mapping (Grad-CAM). Unlike the standard ResNet-18, it retains the spatial dimensions of the final convolutional layer's features (512 channels, 7×7), and gradients are captured through backward hooks on the network's activations. During prediction, gradients with respect to the retained feature maps allow production of attention maps indicating the brain areas that most influenced the classification for a given subject (Jönemo et al., 2023).

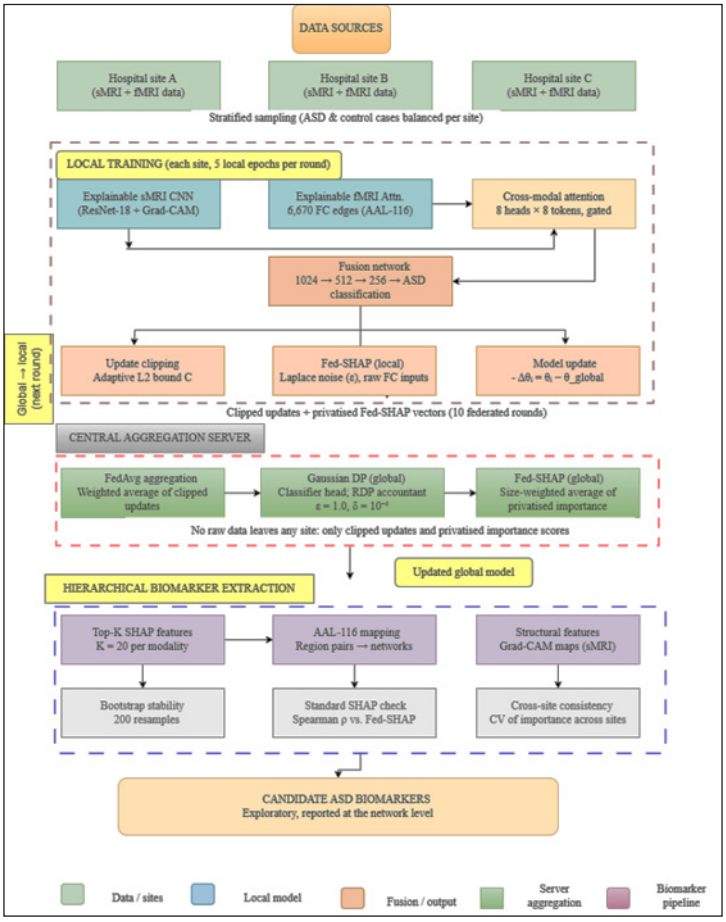

Figure 1. Overview of the proposed federated explainable multimodal learning pipeline for candidate autism biomarker discovery

For functional connectivity analysis, we propose an Explainable fMRI Attention Network that provides connection-level interpretability through learned attention weights. The architecture processes 116×116 functional connectivity matrices derived from Automated Anatomical Labelling (AAL) atlas parcellation; to remove redundancy, it first extracts the upper-triangular elements, yielding 6,670 unique connections.

Attention computation is separated from feature encoding by a two-branch design: the attention branch implements a two-layer feedforward network with tanh and sigmoid activations at the output to produce connection-specific importance weights ranging from 0 to 1, over 8 attention "chunks" per modality so that the cross-modal attention layer operates on a genuine multi-token sequence rather than a single collapsed vector (see below); the encoding branch transforms the attention-weighted features via a three-layer network with batch normalisation, ReLU, and dropout. Attention weights are directly interpretable indicators of which functional brain connections contribute most strongly to ASD classification. The full model engages these modality-specific components using bidirectional cross-modal attention (8 heads, dropout = 0.5) operating over 8-token sequences per modality (with length-one sequences, softmax attention reduces to a fixed linear projection with no genuine cross-modal weighting). The attended representations are combined with the original per-modality features via a learnable, near-zero-initialised gate, so that cross-modal attention can only add to - never replace - the unimodal representations and are concatenated and refined by a two-layer fusion network (1024→512→256 dimensions) before classification.

Federated SHAP-Based Feature Importance (Fed-SHAP)

For privacy-preserving biomarker discovery across distributed sites, we use Federated SHAP (Fed-SHAP), a SHAP-inspired, privacy-preserving feature-importance approximation; it does not compute Shapley values directly. Conventional SHAP values are computed via coalitional sampling over the full dataset and model internals, which is incompatible with the data-localisation principle of federated learning; Fed-SHAP instead estimates each feature's importance as the (privatised) difference in mean activation between diagnostic groups at each site, aggregated centrally, as shown in Equation 2.

$$\varphi^{f}(\text{global}) = \Sigma_i \, w_i \, [\, (1/|D_i^{AsD}|)\Sigma_{k \in D_i^{AsD}} \, h^{f(i)}(x_k) - (1/|D_i^{cTL}|)\Sigma_{k \in D_i^{cTL}} \, h^{f(i)}(x_k) + \text{Lap}(0, \Delta^f/\varepsilon_i) \,] \quad [2]$$

Critically, we compute h^f over the raw pre-encoder input connections (the 6,670-dimensional upper-triangular functional connectivity vector) rather than over an opaque 512-dimensional post-encoder embedding. This is essential for anatomical interpretability: an importance score attached to an abstract embedding dimension cannot be

mapped to a named brain connection, whereas a score attached to a specific input connection map directly, via the known upper-triangular index, to a pair of AAL-116 regions.

For each client, we conduct forward passes through the learned model for a random sample of up to 100 local subjects, computing the mean-activation-difference statistic described above separately for the fMRI input-connection space and for the sMRI 512-dimensional embedding space (structural biomarkers remain reported at the embedding level and via Grad-CAM, since sMRI inputs in this dataset are pre-rendered image slices without atlas-registered coordinates that would permit anatomical mapping analogous to the fMRI case). Laplacian noise is added locally, calibrated to the L1-sensitivity of the statistic and the local privacy budget ε_i, before transmission to the central server, which computes a dataset-size-weighted average across sites to obtain the global importance estimate.

We validated Fed-SHAP's agreement with conventional SHAP by computing gradient-based SHAP attributions (via a GradientExplainer applied to the attention-and-encoding sub-network) on a small, pooled, de-identified subset of the test data, and comparing the resulting feature ranking with the Fed-SHAP ranking on the same input space. Agreement was statistically significant but modest (Spearman $\rho = 0.227$, $p = 1.1\times10^{-78}$), which we report plainly as evidence that Fed-SHAP captures a real but partial signal relative to standard SHAP, rather than as validation of close equivalence.

Hierarchical Biomarker Extraction Pipeline

In addition to the global Fed-SHAP importance values, a hierarchical biomarker extraction pipeline maps high-importance features to anatomically interpretable candidate markers. For functional connectivity, each of the top-K ($K = 20$) important input-connection indices is mapped back, via the inverse of the upper-triangular flattening operation, to the pair of AAL-116 regions it connects, and each region is further assigned to a coarse resting-state network (default mode, salience/cingulo-opercular, fronto-parietal control, sensorimotor, visual, subcortical/limbic, or cerebellar) based on a standard atlas-to-network lookup. For structural features, the top K embedding dimensions are visualised via Grad-CAM activation maps on exemplary subjects, since these dimensions do not admit the same direct anatomical back-mapping as the fMRI connections.

To assess reliability, we additionally compute a bootstrap stability score for the biomarker ranking: subjects are resampled with replacement (200 resamples), the importance statistic is recomputed each time, and we report the fraction of resamples in which each of the original top 20 features remains among the top 20 in the resampled ranking. We report this stability score directly in Results because it materially affects how individual biomarker findings should be interpreted.

Validation Protocol

This section describes the complete validation protocol used to produce every result reported in this manuscript. Subject-level partitioning: each site's subjects are split 70%/15%/15% into training/validation/test sets, with partitioning performed at the subject level (not the scan or slice level) and enforced to be non-overlapping; we verified programmatically that no subject identifier appears in more than one partition, at any site. Hyperparameter selection uses only the validation partition; the test partition is touched exactly once, after training is complete, to produce the headline held-out accuracy figures reported in Results. Leave-one-site-out (LOSO) validation: for each site in turn, the federated model is trained using only the other two sites (their full training and validation subjects) and evaluated on every subject at the held-out site, none of whom were used in training; this provides an external-generalisation estimate distinct from the in-distribution held-out test accuracy. Statistical validation: we report a 95% bootstrap confidence interval for test accuracy (2,000 resamples of the pooled test predictions), a permutation significance test (1,000 label permutations) against the chance-level accuracy implied by the class balance, and an effect size (Cohen's d) for the structural-versus-functional modality importance comparison. Ablation study: we compare the full proposed model against five reduced configurations - sMRI pathway alone, fMRI pathway alone, multimodal fusion without cross-modal attention, the full model trained without differential privacy, and a centralised (non-federated, non-private) upper bound - all trained and evaluated under an identical protocol and an identical number of communication rounds/epochs for fair comparison. Baseline comparison: we additionally compare against FedAvg without differential privacy, FedAvg without the explainability module (a black-box encoder of equivalent capacity), a graph neural network operating on the same connectivity matrices, and a Transformer encoder treating each ROI's connectivity profile as a sequence token; we note that our reference GNN and Transformer implementations were not extensively hyperparameter-tuned and should be read as indicative rather than fully optimised comparisons. Privacy-utility sweep: the full training and Fed-SHAP pipeline is re-run at $\varepsilon = 0.1$, 0.5, 1.0, 5.0, 10, and 50, reporting test accuracy, cross-site fairness (variance in per-site accuracy), and the cross-site coefficient of variation of the top-20 biomarkers' importance at each budget.

RESULT AND DISCUSSIONS

The remainder of this section reports, in turn: dataset and site characteristics; federated training dynamics and held-out test performance, with statistical validation; leave-one-site-out (LOSO) external validation; an ablation study; an expanded baseline comparison; a privacy-utility trade-off sweep; and Fed-SHAP biomarker discovery with anatomical mapping. Throughout, we report exact per-round and per-site figures rather than summary statistics alone, so that the reader can assess the variability underlying each headline number.

Dataset and Site Characteristics

The full cohort comprises 173 subjects (91 ASD, 82 Control), partitioned into three simulated hospital sites via stratified sampling to ensure each site received a broadly similar ASD-to-Control ratio, while the overall partitioning was randomised at the subject level. Table 1 summarises per-site sample sizes, diagnostic composition, train/validation/test split sizes, and age distribution.

Table 1
Dataset and site characteristics

Site	N total	N ASD	N Control	N train	N val	N test	Age mean (SD)
Site 1	57	30	27	39	9	9	17.5 (6.0)
Site 2	58	31	27	40	9	9	15.4 (5.1)
Site 3	58	30	28	40	9	9	16.5 (5.4)

The three sites are closely matched in size (57–58 subjects) and diagnostic balance (52.6–53.4% ASD), which is by design (stratified sampling) rather than an emergent property of the data; readers should not interpret this balance as evidence that the sites are otherwise homogeneous. Site 2 skews slightly younger (mean age 15.4 years, SD 5.1) than Sites 1 and 3 (17.5 and 16.5 years, respectively), and Site 1 shows the widest age spread (SD 6.0). Site 2 also shows the most stable training trajectory and the strongest leave-one-site-out performance (see below); whether this relates to its younger, narrower age distribution cannot be determined with 58 subjects, so we treat it as an unconfirmed hypothesis.

Federated Training Dynamics and Held-Out Test Performance

Figure 2 and Table 2 (round-by-round detail) show the global and per-site validation accuracy across the 10 federated communication rounds for the proposed model at the primary operating point (ε = 1.0). Training began at a low global validation accuracy of 48.1% in Round 1 (per-site: Site 1 = 55.6%, Site 2 = 44.4%, Site 3 = 44.4%), reflecting that the randomly initialised global model, combined with the differential-privacy noise injected at aggregation, had not yet learned a useful

Table 2
Round-by-round validation accuracy, global and per-site (ε = 1.0 operating point)

Round	Global	Site 1	Site 2	Site 3
1	0.481	0.556	0.444	0.444
2	0.667	0.667	0.778	0.556
3	0.704	0.778	0.889	0.444
4	0.926	1.000	0.889	0.889
5	0.815	0.778	1.000	0.667
6	0.741	0.778	0.889	0.556
7	0.889	0.889	1.000	0.778
8	0.815	0.778	1.000	0.667
9	0.815	0.778	1.000	0.667
10	0.815	0.778	1.000	0.667

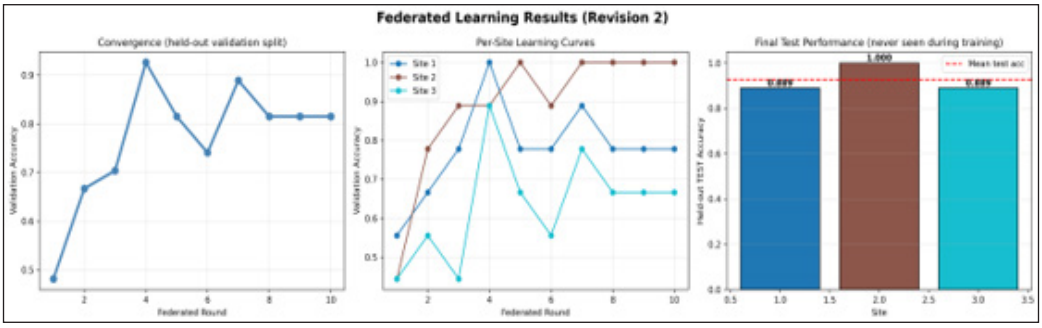

Figure 2. Federated learning convergence dynamics and final held-out test performance across the three simulated sites

decision boundary. Accuracy rose steadily over the next three rounds - 66.7% (Round 2), 70.4% (Round 3), and 92.6% (Round 4, at which point Site 1 reached 100%) - before fluctuating between 74.1% and 88.9% for the remainder of training (Round 5 = 81.5%; Round 6 = 74.1%; Round 7 = 88.9%; Round 8 = 81.5%; Round 9 = 81.5%; Round 10 = 81.5%).

This round-by-round breakdown reveals a pattern that is obscured by the global average alone: Site 2 converges fastest and most stably, reaching 100% validation accuracy by Round 5 and remaining there for four of the last five rounds, whereas Site 3 never exceeds 88.9% and settles at a comparatively low 66.7% for the final four rounds. Site 1 sits between the two, stabilising around 77.8%. We do not observe monotonic convergence at the global level - accuracy fluctuates round-to-round rather than climbing smoothly - which is consistent with two compounding sources of noise: the small per-site validation set (9 subjects, so each individual classification changes site-level accuracy by 11.1 percentage points) and the Gaussian noise injected into the classifier-head parameters at every aggregation step under the differential-privacy mechanism. The relative ordering of sites (2 > 1 > 3) on the validation curves is mirrored in leave-one-site-out testing (Table 3).

On the held-out test set - 27 subjects (9 per site) untouched throughout training and hyperparameter selection - the proposed model achieved 92.6% accuracy (25 of 27 subjects correctly classified), with a 95% bootstrap confidence interval of [81.5%, 100.0%] (2,000 resamples of the pooled test predictions). The width of this interval (18.5 percentage points) is itself informative: it quantifies just how much uncertainty a 27-subject test set carries, and readers should treat any comparison between configurations that differ by less than this margin with caution (see Ablation Study, below). A permutation significance test (1,000 label permutations) gave an observed accuracy of 92.6% against a chance-level mean of 50.4%, $p = 0.001$ - the smallest p-value resolvable with 1,000 permutations, indicating that none of the 1,000 random relabellings matched or exceeded the observed accuracy. This confirms the result is very unlikely to be attributable to chance given the class balance,

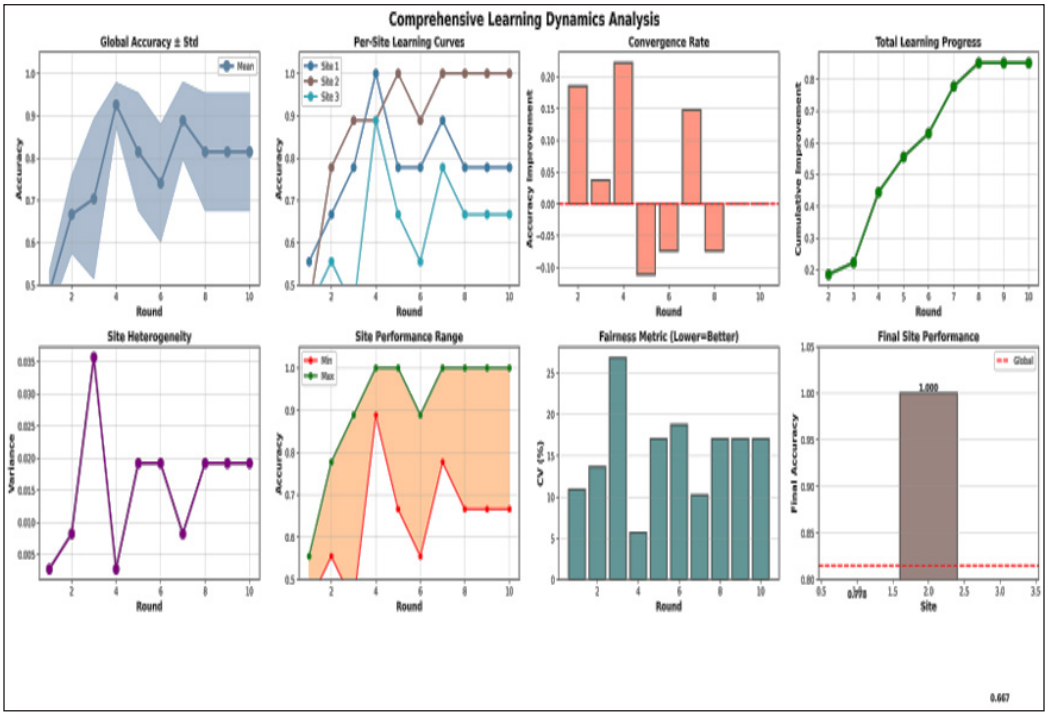

Figure 3. Comprehensive learning dynamics: round-over-round accuracy change, cross-site heterogeneity, and per-round fairness across the 10 communication rounds

though we note that a permutation test's ability to reject chance performance is a different (and weaker) claim than establishing that the model's accuracy is stably estimated - the latter is bounded by the bootstrap CI above. The composed privacy guarantee at this operating point was ($\varepsilon = 1.0$, $\delta = 10^{-5}$), as described in Methodology.

Figure 3 decomposes the same training run into three additional diagnostic views not shown in Table 2: round-over-round accuracy change, cross-site heterogeneity (variance in accuracy across the three sites within each round), and a per-round fairness metric (coefficient of variation across sites; lower is more equitable). The round-over-round change panel confirms the largest single-round gains coincide with Rounds 2-4 (accuracy rising by 18.5, 3.7, and 22.2 percentage points respectively) and the largest single-round drop occurs at Round 5 (-11.1 percentage points), immediately after the Round 4 peak - consistent with the differentially-private aggregation mechanism perturbing the model enough, at this small sample size, to partially undo a round's gains before the next round recovers. Cross-site heterogeneity (variance) peaks sharply at Round 3 (0.036, roughly 13 times its Round 1 value of 0.0028) immediately before the Round 4 accuracy jump, then falls to its minimum at Round 4 (0.0027) before rising again to about 0.019 in Rounds 5-6 and 8-10, with a dip to about 0.008 at Round 7. The per-round fairness CV mirrors this:

it is highest at Round 3 (26.9%, calculated from the per-site accuracies in Table 2) and lowest at Round 4 (5.7%, the most equitable round across the entire training run), before returning to the 17–19% range in later rounds, apart from a dip to about 10% at Round 7. Together, these panels indicate that the model's path to its eventual accuracy was not only non-monotonic in magnitude but also non-monotonic in cross-site fairness - the model briefly became highly equitable at Round 4 before drifting back toward the roughly 17% fairness level it maintained thereafter.

Table 3
Leave-one-site-out validation. Mean accuracy = 86.7% (SD = 4.4%)

Held-out site	N evaluated	Accuracy	F1	AUC
Site 1	57	0.860	0.875	0.969
Site 2	58	0.914	0.921	0.974
Site 3	58	0.828	0.808	0.904

Leave-One-Site-Out (LOSO) External Validation

To assess generalisation to a genuinely unseen institution, we performed leave-one-site-out validation: for each site in turn, the model was trained from scratch using only the other two sites' training and validation subjects (10 rounds, 5 local epochs, identical to the main protocol) and evaluated on every subject at the fully held-out site - including subjects that would otherwise have been in that site's own training or validation split, since from the model's perspective the entire site is unseen. Table 3 reports the results.

The mean LOSO accuracy of 86.7% (SD = 4.4%) is 5.9 percentage points lower than the 92.6% in-distribution held-out accuracy, indicating a moderate generalisation cost when the model is applied to a site it never saw during training, even though all three sites are simulated partitions of a single, uniformly pre-processed cohort. The per-site ordering (Site 2 highest at 91.4%, Site 3 lowest at 82.8%) mirrors that of the validation curves (Table 2), suggesting a genuine, if modest, difference in how easily each site's data distribution is learned rather than an artefact of a single training run. Site 3, which performed weakest on every metric in both analyses, is the natural candidate for closer scrutiny in follow-up work; per-site scanner and protocol metadata were not available in this simulated setup. We therefore treat LOSO as the more conservative estimate of performance at an external institution (see Limitations).

Ablation Study

Table 4 reports the ablation study comparing the full proposed model against five reduced configurations, under an identical training protocol (10 rounds, 5 local epochs, subject-level held-out test evaluation).

We report this table transparently rather than claiming an ordering that the data cannot support. Four of the six configurations-sMRI-only, the multimodal fusion without cross-attention, the full model without differential privacy, and the centralised upper bound - all

Table 4
Ablation study (held-out test set). "Fairness Var." is the variance of per-site test accuracy; the centralised configuration has no site structure and so no fairness variance is defined

Configuration	Accuracy	F1	AUC	Fairness Var.
sMRI pathway alone	0.963	0.963	1.000	0.0027
fMRI pathway alone	0.889	0.896	0.933	0.0082
Multimodal fusion, no cross-attention	0.963	0.963	0.983	0.0027
Full model, without differential privacy	0.963	0.963	1.000	0.0027
Full model, with differential privacy (proposed)	0.926	0.933	0.925	0.0027
Centralised upper bound (no federation, no DP)	0.963	0.963	0.989	-

converge to the identical accuracy of 96.3% (26 of 27 test subjects correct), with AUC ranging narrowly from 0.983 to 1.000. Given the bootstrap confidence interval width established above ([81.5%, 100.0%] for a single configuration evaluated once), these four results cannot be statistically distinguished from one another on this test set: they differ by at most one subject's classification. This indicates that, at the current sample size, the classification task is largely solvable from structural MRI alone, and the ablation cannot resolve a statistically distinguishable additional contribution from cross-modal attention, full multimodal fusion, or even the federated (versus centralised) training paradigm itself; it establishes only that none of the tested configurations meaningfully underperforms the others in this respect.

Two configurations do stand apart, in opposite directions. The fMRI-only pathway is the weakest of all six (88.9% accuracy, 0.933 AUC) and also shows the highest fairness variance (0.0082, roughly three times the other federated configurations), indicating that functional connectivity alone is both less discriminative and less consistently so across the three sites than structural MRI -consistent with the Fed-SHAP finding, discussed below, that Cohen's d = 0.562 favours structural over functional feature importance. The differentially private full model shows the only clear accuracy cost in the table (92.6% versus 96.3% without DP, and the lowest AUC of any configuration at 0.925), which is the expected and, in our view, acceptable price of the formal privacy guarantee described in Methodology; we discuss how this cost varies with the privacy budget itself in the Privacy-Utility Trade-Off subsection below. Figure 4 (left panel) visualises these six results alongside the baseline comparison described next.

Expanded Baseline Comparison

Table 5 compares the proposed framework against FedAvg without differential privacy, FedAvg without the explainability module (a black-box encoder of equivalent capacity, i.e. the same architecture with the attention and Grad-CAM interpretability machinery

removed), a graph neural network baseline operating on the same connectivity matrices, a Transformer baseline treating each ROI's connectivity profile as a sequence token, and a centralised deep-learning upper bound.

Table 5
Expanded baseline comparison (held-out test set)

Method	Accuracy	F1	AUC
Proposed: FedAvg + DP + XAI (full model)	0.889	0.878	0.958
FedAvg without DP	0.963	0.963	1.000
FedAvg without XAI (black box)	0.926	0.926	1.000
Graph Neural Network baseline	0.481	0.648	0.833
Transformer baseline	0.630	0.714	0.883
Centralised deep learning (upper bound)	0.963	0.963	0.989

The reference GNN and Transformer baselines substantially underperformed compared with all other configurations in this study, including the simplest single-modality ablations. We note a pattern in their error profile that is worth making explicit: the GNN baseline's F1 score (0.648) is considerably higher than its accuracy (0.481), and the Transformer baseline shows the same pattern to a lesser degree (F1 = 0.714 vs. accuracy = 0.630). This combination - F1 substantially exceeding accuracy - is the signature of a classifier that is biased toward predicting the positive (ASD) class rather than genuinely discriminating between classes: with the test set close to balanced, a classifier that predicts “ASD” for nearly every subject would achieve accuracy near the ASD base rate while still achieving a comparatively high F1, because F1 does not penalise a high false-positive rate as heavily as accuracy does when recall is high. We did not verify this directly against each baseline's confusion matrix, so we present it as a plausible interpretation of the observed accuracy/F1 gap rather than a confirmed diagnosis, but it is consistent with what would be expected from under-trained or under-tuned classifiers defaulting toward the majority class rather than learning a genuine decision boundary. We attribute this most likely to these architectures being more data-hungry than the available sample size supports (both are commonly trained on datasets orders of magnitude larger than 173 subjects), rather than to a fundamental unsuitability of graph- or attention-based reasoning for this task; a fair comparison would require dedicated hyperparameter tuning, architecture-specific regularisation, and/or a substantially larger dataset, which we identify as future work rather than presenting these baselines as fully optimised comparisons.

Note that the proposed full model's accuracy in this table (88.9%) differs from its accuracy in the ablation study (Table 4; 92.6%) despite the two representing the identical configuration (full model, $\varepsilon = 1.0$), evaluated via two independent training runs of the stochastic federated pipeline. This run-to-run difference - equivalent to exactly one subject's

classification out of 27 held-out test subjects - is itself an informative data point about the measurement resolution available at this sample size, and we discuss its implications further in Limitations rather than treating either single run as the definitive number.

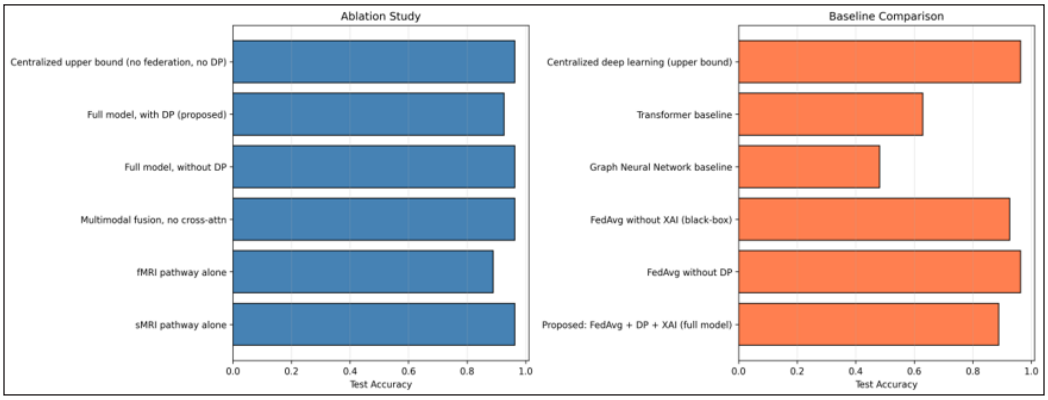

Figure 4. Ablation study and expanded baseline comparison of held-out test accuracy

Privacy-Utility Trade-Off

Table 6 and Figure 5 report accuracy, cross-site fairness (variance in per-site accuracy), and top 20 biomarker-ranking cross-site consistency (coefficient of variation, CV, of each feature's importance across the three sites) across privacy budgets of ε = 0.1, 0.5, 1.0, 5.0, 10, and 50, spanning both tight and loose operating points.

Table 6
Privacy-utility trade-off across differential-privacy budgets

ε (target)	Composed ε	Accuracy	Fairness variance	Biomarker CV (top 20)
0.1	0.10	0.926	0.0027	0.673
0.5	0.50	0.963	0.0027	0.697
1.0	1.00	0.963	0.0027	0.677
5.0	5.00	0.963	0.0027	0.728
10.0	10.00	0.963	0.0027	0.729
50.0	50.00	0.926	0.0027	0.357

We report this curve candidly rather than smoothing over its irregularities. Accuracy is not cleanly monotonic in ε: it is lowest and identical (92.6%) at both the tightest budget tested (ε = 0.1) and the loosest (ε = 50), with a plateau of 96.3% across the intermediate range ε = 0.5–10. Cross-site fairness (variance in per-site accuracy) is essentially flat across the entire sweep at 0.0027, meaning the privacy budget in this range has no detectable effect on how equally the model performs across the three sites, only on overall accuracy.

With only 27 held-out test subjects (9 per site), a single subject's classification changes overall accuracy by roughly 3.7 percentage points, which is comparable in magnitude to the accuracy differences being measured across this sweep; we believe this measurement-resolution limitation, rather than a genuine U-shaped privacy-utility relationship, is the most likely explanation for the identical low values at both extremes, but we present the data as observed rather than fitting a smoothed trend to it.

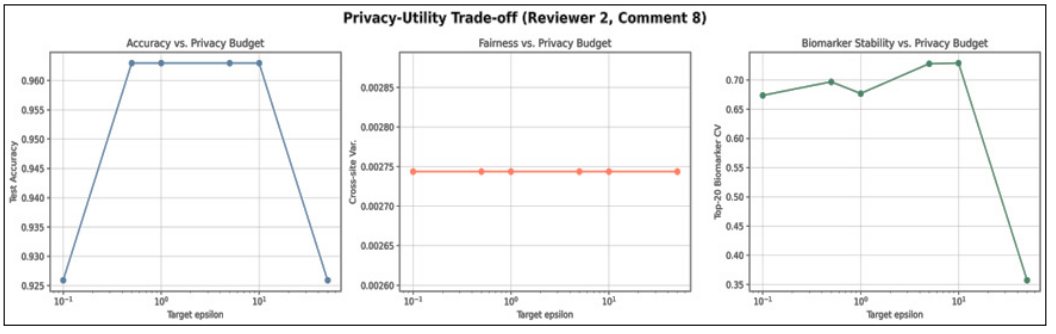

Figure 5. Privacy-utility trade-off across differential-privacy budgets (ε = 0.1–50): accuracy, cross-site fairness, and cross-site consistency of biomarker rankings

The biomarker cross-site consistency column (CV, top 20) shows a more interpretable pattern than accuracy does: it rises gently from 0.673 at ε = 0.1 to a peak of 0.729 at ε = 10, then drops sharply to 0.357 at ε = 50 - roughly half its value at every tighter budget. Because this CV measures how consistently the top-20 features' importance scores agree across the three sites (lower = more consistent), and because tighter ε corresponds to more injected Laplace noise in the Fed-SHAP computation, this pattern is broadly consistent with the mechanism working as intended: at ε = 50, the privacy noise added to each site's local importance vector is small enough that the underlying (noise-free) cross-site agreement in feature importance becomes visible, whereas at every tighter budget, per-site noise realisations are large enough, relative to the underlying signal, to inject additional apparent cross-site disagreement on top of any genuine site-to-site differences. We note this as a plausible mechanistic explanation grounded in how the mechanism is constructed, not as an independently verified causal claim, since disentangling "privacy noise obscuring agreement" from "genuine site disagreement that happens to average out differently at different noise levels" would require repeating this sweep across multiple random seeds, which we did not do here (see Limitations).

Even the tightest budget tested (ε = 0.1) was achieved exactly by the RDP accountant and retained 92.6% held-out accuracy. We emphasise, however, that this guarantee applies only to the classifier head (see Methodology); whether comparably tight budgets remain usable when the full network is protected, or with larger cohorts, remains to be tested.

Fed-SHAP Biomarker Discovery and Anatomical Mapping

Table 7 lists the top 10 of the 20 highest-ranked candidate fMRI connectivity biomarkers obtained with Fed-SHAP at the primary operating point ($\varepsilon = 1.0$), each mapped to its two AAL-116 anatomical regions and functional networks via the inverse upper-triangular index mapping described in Methodology. As described in Methodology, agreement between the Fed-SHAP and standard SHAP rankings was statistically significant but modest ($\rho = 0.227$), so these rankings should be read as approximate feature-importance estimates rather than Shapley values.

Table 7
Top 10 candidate fMRI connectivity biomarkers mapped to AAL-116 anatomy (the full top-20 list is shown in Figure 6)

Rank	**Region A**	**Network A**	**Region B**	**Network B**
1	Olfactory_L	Assoc. cortex	Occipital_Mid_R	Visual
2	Insula_R	Salience	Vermis_8	Cerebellar
3	Cerebelum_8_R	Cerebellar	Cerebelum_9_L	Cerebellar
4	SupraMarginal_R	Fronto-parietal	Temporal_Pole_Mid_R	Assoc. cortex
5	Cingulum_Ant_R	Salience	Cerebelum_Crus2_R	Cerebellar
6	Cingulum_Mid_L	Salience	Temporal_Mid_L	Assoc. cortex
7	Calcarine_L	Visual	Vermis_6	Cerebellar
8	Paracentral_Lobule_L	Sensorimotor	Vermis_7	Cerebellar
9	Hippocampus_L	Subcortical/limbic	Vermis_9	Cerebellar
10	Olfactory_L	Assoc. cortex	Temporal_Mid_R	Assoc. cortex

A pattern visible even in this top-10 list is that 6 of the 10 highest-ranked connections involve a cerebellar region (vermis lobules 6-9 or Crus2, or the cerebellar hemispheres themselves) as one endpoint, paired with regions from several networks: the salience network (insula and anterior cingulate; two connections), visual, sensorimotor and subcortical/limbic regions (one each), and another cerebellar region (one). Aggregating across the full set of 20 top-ranked connections (40 ROI endpoints), Table 8 breaks down the network-level distribution underlying this pattern.

Table 8
Network-level distribution of the top 20 candidate fMRI biomarker connections (40 ROI endpoints)

Network	**ROI endpoints (of 40)**	**Percentage**
Cerebellar	10	25.0%
Other / Association cortex	10	25.0%
Subcortical / Limbic	6	15.0%
Visual	5	12.5%
Sensorimotor	4	10.0%
Salience / Cingulo-opercular	3	7.5%
Fronto-parietal control	2	5.0%

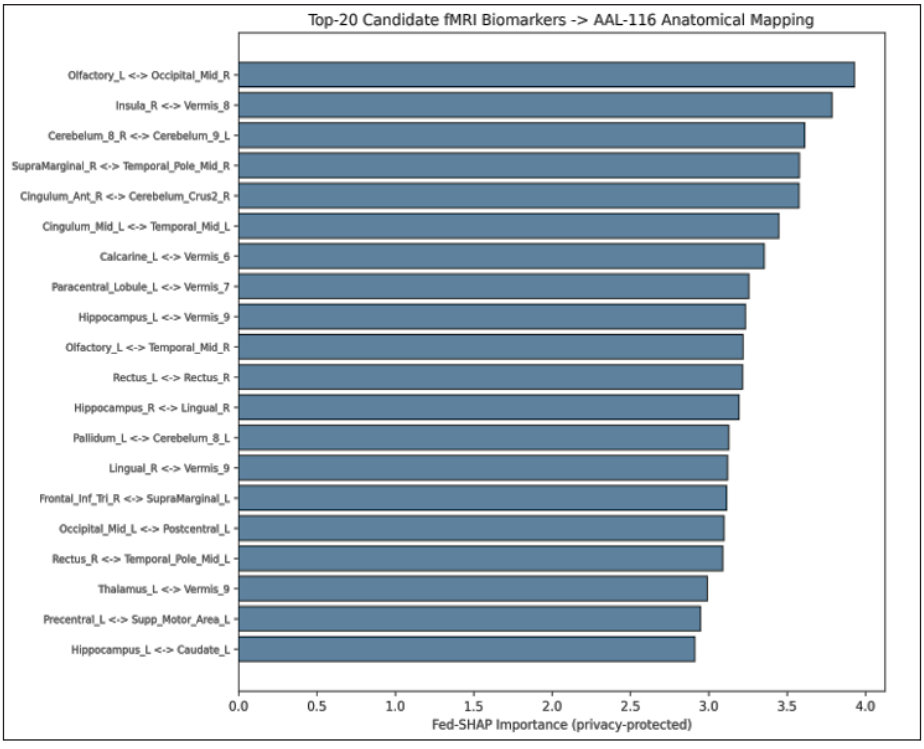

Figure 6. Top 20 candidate fMRI connectivity biomarkers ranked by privacy-preserving Fed-SHAP analysis, mapped to AAL-116 anatomical regions

The cerebellum and "other/association cortex" regions are tied as the most frequently implicated networks (25.0% each), followed by subcortical/limbic structures (15.0%), visual cortex (12.5%), sensorimotor cortex (10.0%), salience/cingulo-opercular regions (7.5%), and fronto-parietal control regions (5.0%). The prominence of cerebellar involvement is broadly consistent with a growing body of ASD neuroimaging literature implicating cerebello-cortical circuits, and the anterior cingulate–cerebellar (Crus II) connection at rank 5 is broadly in line with the cingulate–cerebellar connectivity differences reported by Alves et al. (2023), although that study implicated the posterior rather than the anterior cingulate, although we emphasise that our study was not designed to test any specific prior hypothesis about network involvement and this correspondence should be read as broadly consistent rather than as independent confirmation.

We deliberately report this network-level summary as the primary biomarker finding rather than treating any individual connection as a stable discovery, for three converging reasons. First, bootstrap resampling (200 resamples of the subject pool, recomputing the importance statistic each time) showed that only 20.6% of the top 20 biomarkers remained in the top 20 across resamples - meaning that on a typical resample, roughly 16 of the 20

"top" connections would be replaced by different connections. Second, two independent full training runs of the identical pipeline (same architecture, same hyperparameters, same $\varepsilon = 1.0$ budget) produced entirely non-overlapping top 20 connection lists - not a single specific connection appeared in both runs' top 20, despite both runs achieving similar overall accuracy. Third, the standard-SHAP validation correlation was only modest ($\rho = 0.227$), which further limits how much weight to give any single connection's rank. Taken together, we interpret individual connections as exploratory candidate biomarkers and sample-size-limited, while treating the broader network-level pattern (cerebellar and association-cortex involvement, consistently appearing across our inspection of both runs, even though the specific connections differed) as the more robust observation warranting follow-up on a larger cohort. A structural-versus-functional modality effect size of Cohen's $d = 0.562$ indicates a moderate difference in importance magnitude between the two modalities, consistent with the ablation finding that sMRI alone carries substantial discriminative information at this sample size, and with the fMRI-only ablation arm's comparatively weaker and less fair (higher-variance) performance reported above.

Figure 7 provides a more detailed view of the underlying Fed-SHAP importance statistics for each modality. The top 10 cumulative importance is 34.98 for fMRI versus 30.46 for sMRI, and both modalities' cumulative-importance curves cross the 80% threshold at a similar point (approximately the 15th-16th ranked feature of 20), indicating that importance is concentrated in a broadly similar number of features for each modality rather than being markedly more or less concentrated in one. The per-modality Fed-SHAP importance distributions differ more clearly: the fMRI distribution is sharply peaked and centred near zero (mean = 0.044, SD = 0.819), whereas the sMRI distribution is shifted well above zero (mean = 0.710, SD = 0.916) - consistent with structural features contributing a more consistent, one-directional signal, while functional-connectivity features contribute a signal that is on average small but occasionally large in either direction. The fMRI/sMRI importance-ratio histogram shows every sampled ratio exceeding 1.0 (the "equal importance" reference line), clustered mostly between 1.15 and 1.20, indicating that the highest-ranked fMRI features carry slightly larger absolute importance than their sMRI counterparts. This does not contradict the moderate effect size favouring structural features (Cohen's $d = 0.562$): the effect size reflects the consistently positive, one-directional sMRI contributions shown in the distribution panel (mean 0.710 versus 0.044 for fMRI), whereas the ratio reflects the magnitude of the top-ranked features in each modality.

Figure 8 aggregates the fMRI connection-level importance scores into a per-ROI view, summing each region's involvement across all its connections rather than considering individual connections in isolation as Table 7 does. This reveals a small number of high-degree "hub" regions - the top-ranked hub carries roughly 9.2 units of total connectivity strength, more than the next several hubs combined (each in the 6–7-unit range), with a

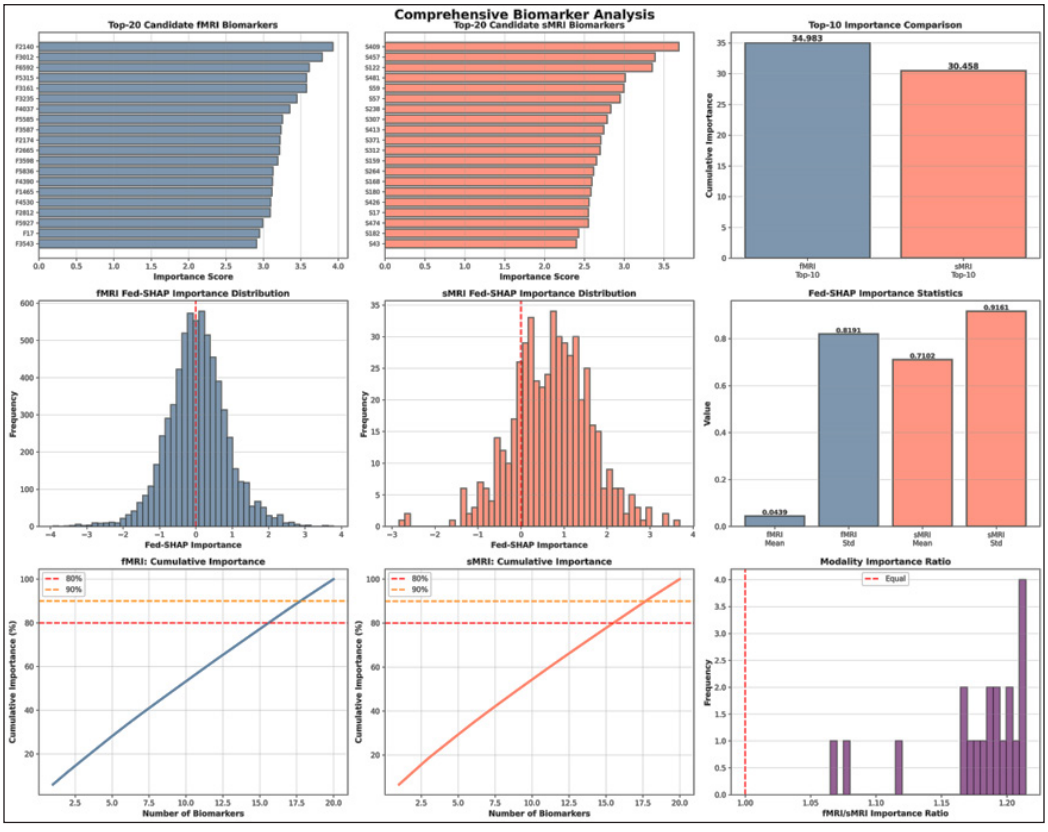

Figure 7. Comprehensive biomarker analysis: per-modality Fed-SHAP importance distributions, cumulative importance curves, and the fMRI/sMRI importance ratio

long tail of regions carrying progressively less. The degree-distribution panel shows this concentration even more starkly: the great majority of the 116 regions have near-zero connectivity strength in the top 20 biomarker set, with only a small subset participating at all, and the hierarchical clustering panel groups the participating regions into a small number of clusters at increasing linkage distance, broadly consistent with the network-level grouping already reported in Table 8. We caution that the ROI numbering used internally by this particular analysis was not independently re-verified against the AAL-116 labelling convention used in Table 7 and Table 8, so we present Figure 8 as a corroborating, region-level view of the same underlying importance scores rather than as a source of additional named-region claims.

Figure 9 reports a consistency check that differs from the bootstrap analysis above. Whereas the bootstrap score asks whether the same top 20 connections reappear when subjects are resampled, the bottom-right panel of Figure 9 asks how consistently the three sites agree on the importance of each top 20 connection ranked by a single trained model.

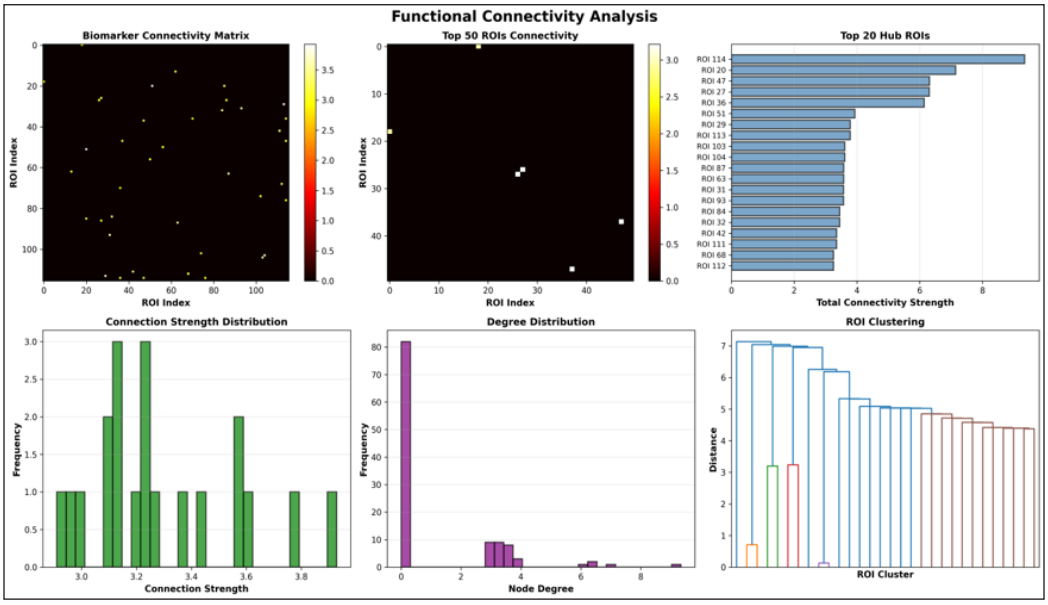

Figure 8. Functional connectivity analysis: biomarker connectivity matrix, hub-ROI ranking, and hierarchical clustering of implicated regions

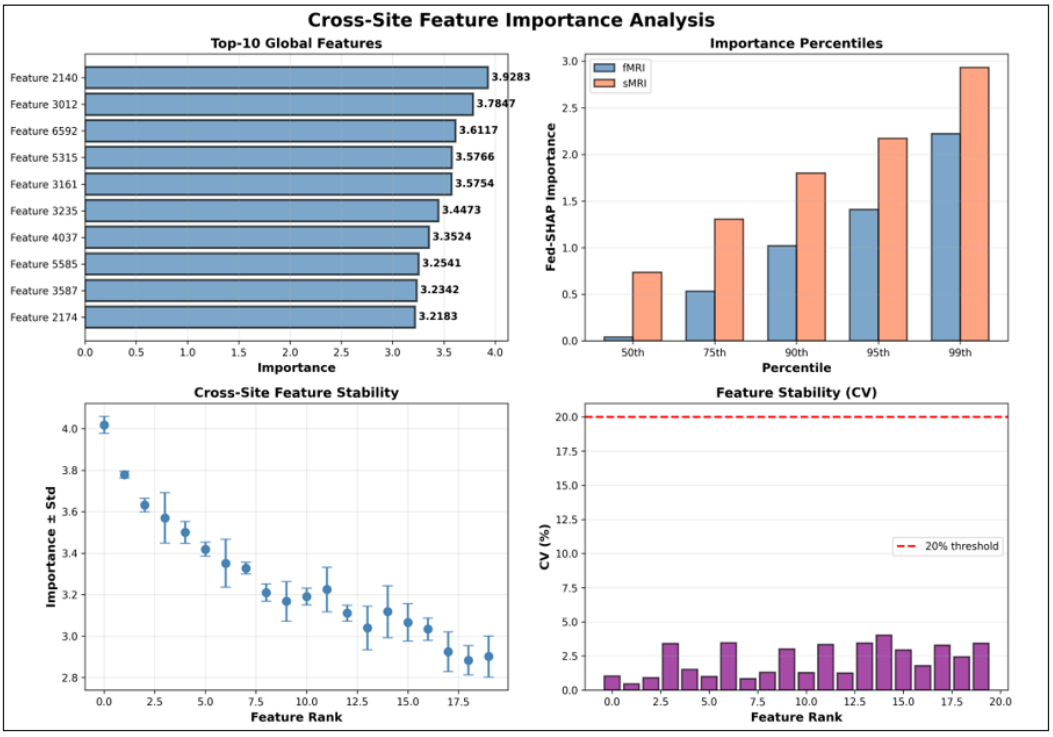

Figure 9. Cross-site feature importance and consistency: top 10 global features, importance percentiles by modality, and cross-site coefficient of variation for the top 20 candidate biomarkers

Here, every one of the top 20 features shows a cross-site coefficient of variation under 5%, well within the 20% threshold marked on the panel, indicating that once a set of top features has been identified, the three sites agree closely on how important each one is. These two findings are not contradictory: a fixed, trained model can show strong cross-site agreement on its own top features (Figure 9), while the identity of those top features remains highly sensitive to which subjects are sampled or which training run produced the model (the bootstrap and cross-run findings above). In practical terms, this suggests the instability we report lies primarily in feature selection (which connections rise to the top) rather than in feature scoring (how consistently sites agree on a score once a feature has been selected) - a distinction that may be useful for any follow-up work aiming to stabilise the biomarker discovery process itself.

DISCUSSION

This study set out to test whether federated learning, differential privacy, and explainable AI could be combined for collaborative autism biomarker discovery across simulated hospital sites without centralising patient data or sacrificing interpretability. Under a leak-free validation protocol, the proposed framework achieves 92.6% held-out accuracy (95% CI [81.5%, 100.0%]) and 86.7% mean leave-one-site-out accuracy, at a formal ($\varepsilon = 1.0$, $\delta = 10^{-5}$) differential-privacy guarantee scoped to the classifier head.

It is important to emphasise that Fed-SHAP is a privacy-preserving feature-importance approximation rather than a direct computation of Shapley values: it ranks features by a privatised between-group difference in mean activation, and its modest agreement with standard SHAP ($\rho = 0.227$) means that its scores should not be interpreted as classical SHAP values. With that caveat, the candidate biomarkers it highlighted show an aggregate network-level pattern (predominantly cerebellar and association-cortex connections) that is broadly consistent with prior ASD connectivity literature, although, given the low individual-level stability reported in Results, they remain exploratory. The ablation results indicate that the value of multimodal fusion and cross-modal attention cannot yet be resolved at this sample size, since several configurations, including the sMRI-only pathway, are statistically indistinguishable from the full model; any added value of fusion may lie in complementary biomarker information not fully captured by raw classification accuracy, a hypothesis that would require a substantially larger cohort to test.

LIMITATIONS

Several limitations should be weighed when interpreting these results. First, the held-out test set (27 subjects total, 9 per site) is small enough that a single subject's classification changes measured accuracy by approximately 3.7 percentage points; several ablation configurations and privacy-budget settings that differ by only one or two subjects' worth

of accuracy should be considered statistically indistinguishable rather than meaningfully ordered, as illustrated by the run-to-run difference for the identical proposed configuration (92.6% in Table 3 vs. 88.9% in Table 5). Second, because the ranking is unstable at the individual-connection level, the named connections are only candidate biomarkers and should not be treated as validated diagnostic markers without confirmation in a larger, independent cohort. Third, the formal differential-privacy guarantee applies only to the classifier head because protecting the full network destroys the predictive signal across all tested budgets (see Methodology). This reflects a general challenge of differentially private deep learning on small cohorts, but it means that the privacy claim is narrower than full-model privacy. Fourth, the three sites in this study are simulated partitions of a single pooled cohort with a shared pre-processing pipeline, rather than genuinely independent institutions with different scanners and populations; the leave-one-site-out result (86.7%) is therefore likely to be an optimistic estimate, rather than a direct measurement, of cross-institution generalisation. We consider validating this framework on a larger, genuinely multi-institutional cohort - sufficient to resolve individual architectural and privacy-budget contributions with statistical confidence, and to assess biomarker stability independently of the specific subjects sampled - to be the most important direction for future work.

CONCLUSION

This work presents a federated learning framework for collaborative, privacy-preserving autism biomarker discovery that combines structural and functional MRI via cross-modal attention, applies differential privacy to both model aggregation and biomarker computation, and uses Fed-SHAP, an anatomically mappable, privacy-preserving feature-importance approximation (rather than a direct computation of Shapley values), for interpretability. Under a leak-free validation protocol, the framework achieves a held-out accuracy of 92.6% and a leave-one-site-out accuracy of 86.7%, at a formal differential-privacy guarantee of ($\varepsilon = 1.0$, $\delta = 10^{-5}$) scoped to the classifier head. We report an ablation study, an expanded baseline comparison, and a privacy-utility sweep in full, including the finding that several configurations are statistically indistinguishable at the current sample size. We also report candidate biomarkers primarily at the network level, given their limited stability at the individual-connection level. We regard this work as a proof of concept demonstrating the technical feasibility of combining federated learning, differential privacy, and explainable AI for multi-site neuroimaging research, rather than as a clinically validated diagnostic tool. Future work should prioritise validation on a larger, genuinely independent multi-institutional cohort; a tighter accounting of the privacy-utility trade-off using per-example gradient clipping (differentially private stochastic gradient descent, DP-SGD) rather than aggregate model-update clipping, which may better preserve signal at tight privacy budgets; incorporation of longitudinal and phenotypic data alongside neuroimaging; and prospective validation studies prior to any clinical application.

ACKNOWLEDGEMENT

The authors sincerely thank the institute management for their support and encouragement and express their gratitude to the research supervisor and Doctoral Research Committee members for their valuable guidance, suggestions, and constructive feedback throughout the research work.

LIST OF ABBREVIATION

ASD	Autism Spectrum Disorder
AI	Artificial Intelligence
XAI	Explainable Artificial Intelligence
MRI	Magnetic Resonance Imaging
sMRI	Structural Magnetic Resonance Imaging
fMRI	Functional Magnetic Resonance Imaging
ABIDE	Autism Brain Imaging Data Exchange
AUC	Area Under the (ROC) Curve
SVM	Support Vector Machine
FedAvg	Federated Averaging
SHAP	SHapley Additive exPlanations
Fed-SHAP	Federated SHAP-based feature-importance approximation
Ex-MMDL	Explainable Multi-Modal Deep Learning
CNN	Convolutional Neural Network
Grad-CAM	Gradient-weighted Class Activation Mapping
AAL	Automated Anatomical Labelling (atlas)
ROI	Region of Interest
CV	Coefficient of Variation
DP	Differential Privacy
RDP	Rényi Differential Privacy
LOSO	Leave-One-Site-Out (validation)
GNN	Graph Neural Network
SGD	Stochastic Gradient Descent
LASSO	Least Absolute Shrinkage and Selection Operator
BASC	Bootstrap Analysis of Stable Clusters
CI	Confidence Interval

REFERENCES

Ahammed, M. S., Niu, S., Ahmed, M. R., Dong, J., Gao, X., & Chen, Y. (2021). DarkASDNet: Classification of ASD on functional MRI using deep neural network. *Frontiers in Neuroinformatics, 15*, Article 635657. https://doi.org/10.3389/fninf.2021.635657

Al-Hiyali, M. I., Yahya, N., Faye, I., & Hussein, A. F. (2021). Identification of autism subtypes based on wavelet coherence of BOLD fMRI signals using convolutional neural network. *Sensors, 21*(16), Article 5256. https://doi.org/10.3390/s21165256

Al-Hiyali, M. I., Yahya, N., Faye, I., & Khan, Z. (2021). Autism spectrum disorder detection based on wavelet transform of BOLD fMRI signals using pre-trained convolution neural network. *International Journal of Integrated Engineering, 13*(5), 49-56. https://doi.org/10.30880/ijie.2021.13.05.006

Almuqhim, F., & Saeed, F. (2021). ASD-SAENet: A sparse autoencoder, and deep-neural network model for detecting autism spectrum disorder (ASD) using fMRI data. *Frontiers in Computational Neuroscience, 15*, Article 654315. https://doi.org/10.3389/fncom.2021.654315

Alutaibi, A. I., Sharma, S. K., & Khan, A. R. (2025). Capsule DenseNet++: Enhanced autism detection framework with deep learning and reinforcement learning-based lifestyle recommendation. *Computers in Biology and Medicine, 190*, Article 110038. https://doi.org/10.1016/j.compbiomed.2025.110038

Alves, C. L., Toutain, T. G. L. D. O., de Carvalho Aguiar, P., Pineda, A. M., Roster, K., Thielemann, C., Porto, J. A. M., & Rodrigues, F. A. (2023). Diagnosis of autism spectrum disorder based on functional brain networks and machine learning. *Scientific Reports, 13*(1), Article 8072. https://doi.org/10.1038/s41598-023-34650-6

Alzubi, O. A., Alzubi, J. A., Shankar, K., & Gupta, D. (2021). Blockchain and artificial intelligence enabled privacy-preserving medical data transmission in Internet of Things. *Transactions on Emerging Telecommunications Technologies, 32*(12), Article e4360. https://doi.org/10.1002/ett.4360

Andrew, G., Thakkar, O., McMahan, H. B., & Ramaswamy, S. (2021). Differentially private learning with adaptive clipping. In *Advances in neural information processing systems* (Vol. 34, pp. 17455-17466). https://proceedings.nips.cc/paper/2021/hash/91cff01af640a24e7f9f7a5ab407889f-Abstract.html

Banerjee, T. (2025). Electromagnetic interaction algorithm (EIA)-based feature selection with adaptive kernel attention network (AKAttNet) for autism spectrum disorder classification. *International Journal of Developmental Neuroscience, 85*(5), Article e70034. https://doi.org/10.1002/jdn.70034

Bayram, M. A., Özer, İ., & Temurtaş, F. (2021). Deep learning methods for autism spectrum disorder diagnosis based on fMRI images. *Sakarya University Journal of Computer and Information Sciences, 4*(1), 142-155. https://doi.org/10.35377/saucis.04.01.879735

Belhaouari, S. B., Talbi, A., Hassan, S., Al-Thani, D., & Qaraqe, M. (2023). PFT: A novel time-frequency decomposition of BOLD fMRI signals for autism spectrum disorder detection. *Sustainability, 15*(5), Article 4094. https://doi.org/10.3390/su15054094

Chandra, A., Verma, S., Raghuvanshi, A. S., & Bodhey, N. K. (2024). ASDC-Net: Optimized convolutional neural network-based automatic autism spectrum disorder classification using rs-fMRI data. *IETE Journal of Research, 70*(4), 4189-4202. https://doi.org/10.1080/03772063.2023.2196979

Eslami, T., Mirjalili, V., Fong, A., Laird, A. R., & Saeed, F. (2019). ASD-DiagNet: A hybrid learning approach for detection of autism spectrum disorder using fMRI data. *Frontiers in Neuroinformatics, 13*, Article 70. https://doi.org/10.3389/fninf.2019.00070

Eslami, T., Raiker, J. S., & Saeed, F. (2021). Explainable and scalable machine learning algorithms for detection of autism spectrum disorder using fMRI data. In A. S. El-Baz & J. S. Suri (Eds.), *Neural engineering techniques for autism spectrum disorder: Volume 1: Imaging and signal analysis* (pp. 39-54). Academic Press. https://doi.org/10.1016/b978-0-12-822822-7.00004-1

Eslami, T., & Saeed, F. (2019). Auto-ASD-Network: A technique based on deep learning and support vector machines for diagnosing autism spectrum disorder using fMRI data. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* (pp. 646-651). Association for Computing Machinery. https://doi.org/10.1145/3307339.3343482

Feng, M., & Xu, J. (2023). Detection of ASD children through deep-learning application of fMRI. *Children, 10*(10), Article 1654. https://doi.org/10.3390/children10101654

Feng, W., Liu, G., Zeng, K., Zeng, M., & Liu, Y. (2022). A review of methods for classification and recognition of ASD using fMRI data. *Journal of Neuroscience Methods, 368*, Article 109456. https://doi.org/10.1016/j.jneumeth.2021.109456

Firouzi, M. H., & Fadaei, S. (2025). Deep learning-based classification of autism spectrum disorder using resting state fMRI data. *International Journal of Engineering, 38*(4), 785-795. https://doi.org/10.5829/ije.2025.38.04a.10

Gheisari, M., Najafabadi, H. E., Alzubi, J. A., Gao, J., Wang, G., Abbasi, A. A., & Castiglione, A. (2021). OBPP: An ontology-based framework for privacy-preserving in IoT-based smart city. *Future Generation Computer Systems, 123*, 1-13. https://doi.org/10.1016/j.future.2021.01.028

Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., & Meneguzzi, F. (2018). Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *NeuroImage: Clinical, 17*, 16-23. https://doi.org/10.1016/j.nicl.2017.08.017

Huang, Z.-A., Zhu, Z., Yau, C. H., & Tan, K. C. (2021). Identifying autism spectrum disorder from resting-state fMRI using deep belief network. *IEEE Transactions on Neural Networks and Learning Systems, 32*(7), 2847-2861. https://doi.org/10.1109/tnnls.2020.3007943

Islam, M. M., Hussain, A. H., Mariam, Q. R., Islam, S., & Hasan, M. N. (2025). AI-enabled predictive health monitoring for children with autism using IoT and machine learning to detect behavioral changes. *Perinatal Journal, 33*(1), 415-422. https://doi.org/10.57239/prn.25.03310048

Jabbar, U., Iqbal, M. W., Alourani, A., Shinan, K., Alanazi, F., Sarwar, N., & Ashraf, M. U. (2025). Machine learning–based approach for early screening of autism spectrum disorders. *Applied Computational Intelligence and Soft Computing, 2025*(1), Article 9975499. https://doi.org/10.1155/acis/9975499

Jönemo, J., Abramian, D., & Eklund, A. (2023). Evaluation of augmentation methods in classifying autism spectrum disorders from fMRI data with 3D convolutional neural networks. *Diagnostics, 13*(17), Article 2773. https://doi.org/10.3390/diagnostics13172773

Kang, L., Chen, J., Huang, J., & Jiang, J. (2023). Autism spectrum disorder recognition based on multi-view ensemble learning with multi-site fMRI. *Cognitive Neurodynamics, 17*(2), 345-355. https://doi.org/10.1007/s11571-022-09828-9

Kang, L., Chen, M., Huang, J., & Xu, J. (2025). Identifying autism spectrum disorder based on machine learning for multi-site fMRI. *Journal of Neuroscience Methods, 416*, Article 110379. https://doi.org/10.1016/j.jneumeth.2025.110379

Karampasi, A., Kakkos, I., Miloulis, S.-T., Zorzos, I., Dimitrakopoulos, G. N., Gkiatis, K., Asvestas, P., & Matsopoulos, G. (2020). A machine learning fMRI approach in the diagnosis of autism. In *2020 IEEE International Conference on Big Data (Big Data)* (pp. 3628-3631). IEEE. https://doi.org/10.1109/bigdata50022.2020.9378453

Karunakaran, P., Hamdan, Y. B., & Sathesh, A. (2020). Early prediction of autism spectrum disorder by computational approaches to fMRI analysis with early learning technique. *Journal of Artificial Intelligence and Capsule Networks, 2*(4), 207-216. https://doi.org/10.36548/jaicn.2020.4.003

Kohli, M., Kar, A. K., Sinha, S., & Kohli, S. (2025). Stakeholder perception towards a machine-learning-based digital platform for detection and management of autism spectrum disorder. *Expert Systems, 42*(1), Article e13545. https://doi.org/10.1111/exsy.13545

Li, X., Dvornek, N. C., Zhuang, J., Ventola, P., & Duncan, J. S. (2018). Brain biomarker interpretation in ASD using deep learning and fMRI. In A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, & G. Fichtinger (Eds.), *Medical image computing and computer assisted intervention – MICCAI 2018* (pp. 206-214). Springer. https://doi.org/10.1007/978-3-030-00931-1_24

Li, Y., Li, H., Hu, C., Cui, J., Zhang, F., Zhao, J., Feng, Y., Hu, C., Yang, L., Qian, H., Pan, J., Luo, X., Tang, Z., & Hao, Y. (2025). The role of the dopamine system in autism spectrum disorder revealed using machine learning: An ABIDE database–based study. *Cerebral Cortex, 35*(2), Article bhaf022. https://doi.org/10.1093/cercor/bhaf022

Mironov, I. (2017). Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)* (pp. 263-275). IEEE. https://doi.org/10.1109/csf.2017.11

Movassagh, A. A., Alzubi, J. A., Gheisari, M., Rahimi, M., Mohan, S., Abbasi, A. A., & Nabipour, N. (2023). Artificial neural networks training algorithm integrating invasive weed optimization with differential evolutionary model. *Journal of Ambient Intelligence and Humanized Computing, 14*(5), 6017-6025. https://doi.org/10.1007/s12652-020-02623-6

Nafisah, I., Mahmoud, N., Ewees, A. A., Khattap, M. G., Dahou, A., Alghamdi, S. M., Fares, I. A., Azmi Al-Betar, M., & Abd Elaziz, M. (2025). Deep learning-based feature selection for detection of autism spectrum disorder. *Frontiers in Artificial Intelligence, 8*, Article 1594372. https://doi.org/10.3389/frai.2025.1594372

Narayanan, N., K R, R., & Varghese, B. M. (2025). Early detection of autism spectrum disorder via deep-learning application of fMRI and machine learning for ASD children identifications. *International Journal of Hybrid Intelligent Systems, 21*(2), 148-162. https://doi.org/10.3233/his-240029

Pinto, M. M. V., Arisawa, E. A. L. S., Raniero, L. J., & Bhattacharjee, T. (2025). Saliva FTIR spectra and machine learning for autism spectrum disorder diagnosis preliminary study. *IEEE Photonics Journal, 17*(3), Article 8500504. https://doi.org/10.1109/jphot.2025.3561020

Qayyum, A., Ahamed Khan, M. K. A., Benzinou, A., Mazher, M., Ramasamy, M., Aramugam, K., Deisy, C., Sridevi, S., & Suresh, M. (2022). An efficient 1DCNN–LSTM deep learning model for assessment and classification of fMRI-based autism spectrum disorder. In J. S. Raj, K. Kamel, & P. Lafata (Eds.),

Innovative data communication technologies and application: Proceedings of ICIDCA 2021 (pp. 1039-1048). Springer. https://doi.org/10.1007/978-981-16-7167-8_77

Qiang, N., Gao, J., Dong, Q., Li, J., Zhang, S., Liang, H., Sun, Y., Ge, B., Liu, Z., Wu, Z., Liu, T., Yue, H., & Zhao, S. (2023). A deep learning method for autism spectrum disorder identification based on interactions of hierarchical brain networks. *Behavioural Brain Research, 452*, Article 114603. https://doi.org/10.1016/j.bbr.2023.114603

Qiqieh, I., Alzubi, J. A., Alzubi, O. A., Khawaja, Y., & Singh, A. (2026). Securing next-generation connected healthcare: A privacy-preserving explainable AI framework for consumer electronics. *IEEE Transactions on Consumer Electronics, 72*(2), 4615-4622. https://doi.org/10.1109/tce.2026.3675727

Ramli, N. S. H., Ahmad Radzi, S., Abdul Hamid, N., Wong, Y. C., & Ramlan, A. R. (2021). Functional magnetic resonance imaging for autism spectrum disorder detection using deep learning. *Jurnal Teknologi, 83*(3), 45-52. https://doi.org/10.11113/jurnalteknologi.v83.16389

Sabegh, A. M., Samadzadehaghdam, N., Seyedarabi, H., & Ghadiri, T. (2023). Automatic detection of autism spectrum disorder based on fMRI images using a novel convolutional neural network. *Research on Biomedical Engineering, 39*(2), 407-413. https://doi.org/10.1007/s42600-023-00275-x

Sachdeva, J., Mittal, R., Mehta, J., Jain, R., & Ranjan, A. (2024). Resolving autism spectrum disorder (ASD) through brain topologies using fMRI dataset with multi-layer perceptron (MLP). *Psychiatry Research: Neuroimaging, 343*, Article 111858. https://doi.org/10.1016/j.pscychresns.2024.111858

Santana, C. P., de Carvalho, E. A., Rodrigues, I. D., Bastos, G. S., de Souza, A. D., & de Brito, L. L. (2022). Rs-fMRI and machine learning for ASD diagnosis: A systematic review and meta-analysis. *Scientific Reports, 12*(1), Article 6030. https://doi.org/10.1038/s41598-022-09821-6

Shao, L., Fu, C., You, Y., & Fu, D. (2021). Classification of ASD based on fMRI data with deep learning. *Cognitive Neurodynamics, 15*(6), 961-974. https://doi.org/10.1007/s11571-021-09683-0

Shazman, S., Carmel, J., Itkin, M., Malitsky, S., Shalan, M., Soreq, E., Elliott, E., Lebow, M., & Kuperman, Y. (2025). Urine metabolomic profiling and machine learning in autism spectrum disorder diagnosis: Toward precision treatment. *Metabolites, 15*(5), Article 332. https://doi.org/10.3390/metabo15050332

Subah, F. Z., Deb, K., Dhar, P. K., & Koshiba, T. (2021). A deep learning approach to predict autism spectrum disorder using multisite resting-state fMRI. *Applied Sciences, 11*(8), Article 3636. https://doi.org/10.3390/app11083636

Tang, Y., Tong, G., Xiong, X., Zhang, C., Zhang, H., & Yang, Y. (2023). Multi-site diagnostic classification of autism spectrum disorder using adversarial deep learning on resting-state fMRI. *Biomedical Signal Processing and Control, 85*, Article 104892. https://doi.org/10.1016/j.bspc.2023.104892

Xu, W., Li, H., Li, J., & Jin, M. (2025). Age of machine learning: New trends in autism spectrum disorder prediction. *Frontiers in Microbiology, 16*, Article 1492484. https://doi.org/10.3389/fmicb.2025.1492484

Yin, W., Mostafa, S., & Wu, F.-X. (2021). Diagnosis of autism spectrum disorder based on functional brain networks with deep learning. *Journal of Computational Biology, 28*(2), 146-165. https://doi.org/10.1089/cmb.2020.0252

Yousefian, A., Shayegh, F., & Maleki, Z. (2023). Detection of autism spectrum disorder using graph representation learning algorithms and deep neural network, based on fMRI signals. *Frontiers in Systems Neuroscience, 16*, Article 904770. https://doi.org/10.3389/fnsys.2022.904770

Zhang, J., Feng, F., Han, T., Gong, X., & Duan, F. (2023). Detection of autism spectrum disorder using fMRI functional connectivity with feature selection and deep learning. *Cognitive Computation, 15*(4), 1106-1117. https://doi.org/10.1007/s12559-021-09981-z