Journal of High School Science

Diffusion modeling for characterizing neuropathology in psychotic disorders: a Machine Learning study

Puninchittaya A[1], Wills E[2]

Submitted: April 12, 2026, Revised: version 1, June 29, 2026, version 2, July 19, 2026, version 3, August 23, 2026, version 4, October 12, 2026, version 5, October 13, 2026
Accepted: October 19, 2026

Abstract
Psychotic spectrum disorders exhibit overlapping symptoms and heterogeneous neuropathology, complicating subtype differentiation. Conventional diffusion MRI models may conflate concurrent microstructural alterations. This study investigated whether integrating complementary diffusion compartments improves discrimination within the psychosis spectrum. A framework combining Diffusion Tensor Imaging (DTI), Diffusion Kurtosis Imaging (DKI), Free Water Imaging (FWI), and Intravoxel Incoherent Motion (IVIM) was evaluated using multi-shell diffusion MRI data from 1,267 Human Connectome Project participants (1,094 healthy controls; 173 patients). Parameter recovery was assessed through simulations, while regional diffusion features were analyzed using neural networks with nested, site-stratified five-fold cross-validation, ablation, scanner harmonization, and confound controls. The integrated framework distinguished patients from controls with a pooled out-of-fold AUC-ROC of 0.968 (95% CI: 0.952–0.981), compared with 0.896–0.929 for reduced models. Conventional DTI and DKI models showed near-chance discrimination between schizophrenia-spectrum and non-schizophrenia psychosis (AUC-ROC = 0.543–0.564), whereas the integrated framework achieved 0.716 (95% CI: 0.635–0.792), significantly outperforming DTI and DTI+DKI (paired bootstrap $p = 0.011$ and 0.007, respectively). Ablation demonstrated complementary contributions from extracellular free-water and perfusion metrics beyond conventional diffusion measurements. Same-cohort matched analysis retained strong patient-control discrimination (AUC-ROC = 0.958), although acquisition confounding remains a limitation. These findings establish the incremental predictive value of multi-compartment diffusion integration for distinguishing psychotic subtypes, providing a foundation for microstructural profiling and patient stratification, pending independent validation.

Keywords
Psychotic spectrum disorders, Schizophrenia, Diffusion Magnetic Resonance Imaging, Multi-Compartment Diffusion Modeling, Diffusion Tensor Imaging, Diffusion Kurtosis Imaging, Free Water Imaging, Intravoxel Incoherent Motion, Machine Learning, Neuroimaging biomarkers

[1]Corresponding author: Arjun Puninchittaya, Bedford High School, 47 Nashua Road, Bedford, NH 03110, USA. arjun.puninchittaya@gmail.com
[2]Edward Wills, University of Cambridge, Cambridge, England, United Kingdom (The Old Schools, Trinity Lane, Cambridge CB2 1TN, United Kingdom)

https://doi.org/10.64336/001c.173442

1. Introduction

Psychotic Spectrum Disorders (PSD), including Schizophrenia, Schizoaffective disorder, Delusional disorder, Brief Psychotic disorder, and Substance-Induced Psychotic disorder, pose a major clinical challenge caused by their heterogeneous symptoms and poor diagnostic tools. The burden extends far beyond an individual patient, affecting families and imposing substantial costs on global healthcare systems (1).

Despite decades of research and clinical advances, core limitations continue in the diagnosis and treatment of these conditions (2). Current diagnostic methods are purely observational, based solely on symptoms, with no laboratory tests, radiologic evidence, or other objective biomarkers to support diagnosis or treatment planning. Although practical, the symptom-based model of diagnosis is subjective and may not always accurately reflect the pathophysiology. Treatment response is also suboptimal, with first-line antipsychotic medications being effective in only about 50% of patients (3).

Neuroimaging, particularly magnetic resonance imaging (MRI), offers a promising source of objective, non-invasive metrics to complement observational psychiatry (4,5). Advanced MRI paradigms enable quantitative characterization of brain microstructure and tissue organization, revealing neuropathological alterations that cannot be discerned from clinical observation alone. The central challenge lies in developing imaging frameworks that achieve both sensitivity and biological specificity when isolating overlapping pathological processes.

Diffusion MRI (dMRI) measures the microscopic diffusion of water molecules in brain tissue, making it a useful tool for assessing brain microstructure in psychiatric disorders. Unlike conventional structural MRI, which primarily examines anatomical changes such as volume, dMRI is sensitive to subtle microstructural alterations in both white and gray matter (6). By noninvasively assessing changes in myelin integrity, cellular density, inflammation, and overall tissue organization, this technique provides a comprehensive view of microstructural alterations (6). As it is non-invasive, it has potential for clinical adoption as a tool to aid physicians in both understanding and diagnosing disorders. However, the sensitivity of dMRI to many alterations in brain composition represents both an opportunity and a challenge. Different tissue alterations, such as neuroinflammation and edema, can increase extracellular fluid volume, which affects water diffusion by reducing physical barriers and increasing the 'free water' signal, potentially masking concurrent changes in axon integrity. Advanced dMRI modeling methods have been created to combat this issue. These include Intravoxel Incoherent Motion (IVIM) imaging to distinguish perfusion-related signal contributions from tissue diffusion (7), Free Water Imaging (FWI) to estimate extracellular free-water fractions, and Diffusion Kurtosis Imaging (DKI) to characterize non-Gaussian diffusion (8). Although each method captures complementary aspects of brain microstructure, none uniquely identifies a specific pathological process, and each remains susceptible to confounding from unmodeled signal contributions.

Previous dMRI studies of PSD have reliably demonstrated a pattern of abnormalities across multiple modeling methods, pointing to distinct, though possibly overlapping, underlying pathological processes (9). Diffusion Tensor Imaging (DTI), which is the most commonly used dMRI approach (10), has consistently found a decrease in fractional anisotropy (FA) in the white matter of individuals with schizophrenia and other psychotic disorders (11). A decrease in FA, which is a measure of the directionality of water diffusion, is thought to reflect compromised myelin integrity or axonal organization (12). In addition to changes in FA, DTI has shown increased mean diffusivity (MD) in cortical gray and white matter regions in patients with PSD, which provides an indirect index consistent with altered cellular density, membrane breakdown, or extracellular expansion (13). IVIM imaging, which models pseudo-diffusive microvascular blood displacement, has demonstrated increased perfusion fraction (PF) in patients with PSD (14) and parameter shifts in brain parenchymal tissue (15), suggesting alterations in microcirculatory volume or capillary hemodynamics. Similarly, FWI studies have reported elevated extracellular free water (f_w) in cortical and subcortical regions in schizophrenia (16). While elevated free water has frequently been hypothesized to reflect neuroinflammatory processes, it represents a non-specific macroscopic index of extracellular fluid expansion that may equally reflect interstitial edema, atrophy-associated partial volume effects, or altered fluid clearance (17).

Furthermore, DKI studies have reported altered mean kurtosis (MK) in cortical gray matter structures in patients diagnosed with psychosis. Non-Gaussian diffusion kurtosis quantifies tissue microstructural heterogeneity and structural complexity (18); however, because diffusion MRI cannot resolve specific cellular phenotypes, kurtosis shifts cannot be definitively attributed to microglial proliferation in the absence of direct histological or molecular validation. These patterns, although possibly obfuscated by other factors not accounted for by a single model, show promise in developing an accurate model to move beyond the binary classification of psychotic vs control patients.

The fundamental problem is that diverse pathological processes, including myelin disruption, microvascular alteration, interstitial fluid shifts, and cellular reorganization, simultaneously alter water diffusion in psychiatric disorders (19). When these processes occur simultaneously, single-compartment models may conflate their effects, resulting in biased parameter estimates and less specific biological conclusions. A comprehensive modeling approach that simultaneously accounts for tissue diffusion, perfusion, free water, and tissue complexity would minimize cross-contamination between compartments and provide more specific estimates of each pathological process. An integrated framework may improve the accuracy of microstructural measurements while also enabling researchers and clinicians to determine the relative contributions of different pathophysiological mechanisms in individual patients, possibly guiding more targeted and beneficial treatment strategies.

While traditional clinical assessments are highly effective at identifying the broad behavioral presentation of psychosis, they struggle to map

the highly heterogeneous underlying cellular changes unique to each patient (19). This multi-compartment framework was developed to provide objective microstructural characterization downstream from initial diagnosis, establishing biological separability to support future research into patient stratification. By decomposing the diffusion signal into complementary biophysical components, this approach aims to reduce parameter cross-contamination, providing a quantitative baseline of an individual's microstructural profile, which can reflect multiple overlapping biological processes to support hypothesis generation for stratified clinical investigations.

Because microstructural signatures reflect cellular alterations that are not inherently disease-specific, this classification framework was not developed to differentiate psychotic presentations from broader neurological abnormalities, such as amyotrophic lateral sclerosis, Alzheimer's disease, systemic neuroinflammation, traumatic brain injury, or epilepsy. Consequently, the framework was designed to operate strictly as a diagnostic support tool, requiring an independent clinical diagnosis of a psychotic-spectrum condition by a certified clinician. This study describes an integrated framework for quantifying perfusion, extracellular water, and tissue complexity simultaneously. By merging IVIM, FWI, and DKI into a single pipeline, the study aims to minimize parameter cross-contamination that has frequently affected traditional diffusion analyses. Specifically, this study evaluated a machine-learning framework: first, using case-control classification as a methodological benchmark to evaluate out-of-fold discriminatory capability across the full feature matrix; and second, performing a formal component ablation analysis to determine whether combining DTI, DKI, FWI, and IVIM modalities provides measurable incremental predictive performance over constituent single- and dual-compartment models across both case-control and subtype classification tasks.

This study explored diffusion MRI modeling combined with machine-learning classification for diagnosing PSD. The method combines several diffusion compartment models. These include DTI, IVIM, FWI, and DKI. This single system simultaneously estimates tissue microstructure, blood flow, extracellular free water, and tissue complexity. It uses parameters from each compartment as input features for a neural network classifier. The classifier aims to distinguish patients with psychotic spectrum disorders from healthy controls and to differentiate diagnostic subtypes within the psychosis spectrum.

This study focused on three specific goals: first, evaluating whether the integrated model distinguishes patients with psychotic spectrum disorders from healthy controls; second, testing whether these features differentiate schizophrenia spectrum from non-schizophrenia psychosis; and third, performing a component ablation analysis to determine whether multi-compartment integration offers demonstrable predictive gains over simpler single- and dual-compartment models.

2. Methods

2.1 Participants and data characteristics

The study analyzed a large-scale dataset of

diffusion MRI images comprising 1,267 total participants. This cohort included 1,094 healthy control subjects and 173 patients diagnosed with psychotic spectrum disorders. Within the patient group, 83 individuals were diagnosed with schizophrenia, 20 with schizoaffective disorder, and 70 with non-schizophrenic psychosis. A breakdown of the data is shown in Figure 1. All participants had undergone multi-shell diffusion imaging protocols. Diagnostic assessments were conducted by trained clinicians using standardized psychiatric instruments to ensure consistency. Healthy control participants (n = 1,094) were drawn primarily from the Human Connectome Project Young Adult cohort (HCP-YA; n = 1,026, scanned at Washington University in St. Louis) with additional same-protocol controls from the Human Connectome Project Early Psychosis cohort (HCP-EP; n = 68 across Indiana University n = 25, McLean Hospital n = 11, and Beth Israel Deaconess Medical Center n = 32). Clinical patients (n = 173) were recruited exclusively through HCP-EP across Indiana University (n = 81), McLean Hospital (n = 49), and Beth Israel Deaconess Medical Center (n = 43). A detailed comparison of demographic, clinical, and acquisition parameters between the two cohorts is provided in Table 1. Because healthy controls and clinical patients were drawn from distinct parent studies with differing age distributions, scanner hardware, and recruitment frameworks, study cohort and diagnostic status are structurally entangled in the case-control comparison. To account for these demographic and acquisition differences, our analytical pipeline incorporated rigorous confound control, cross-site harmonization, Leave-One-Site-Out cross-validation, and non-imaging covariate baseline comparisons (Section 2.4.4).

Table 1. Demographic, clinical, and acquisition protocol parameters across HCP-YA and HCP-EP cohorts. Parameter comparison documenting baseline sample size, age, sex, scanner platform, gradient specifications, and diffusion sequence parameters across cohorts. Cohort characteristics and imaging protocols were compiled from HCP-YA project documentation (20, 21, 22), HCP-EP Release 1.1 documentation (23), and the study dataset.

Parameter	HCP Young Adult (HCP-YA) Controls	HCP Early Psychosis (HCP-EP) Patients
Cohort Sample Size	n = 1,094 (1,026 HCP-YA + 68 HCP-EP)	n = 173 (81 IU, 49 McLean, 43 BIDMC)
Classification Subgroups	Healthy Controls: 1,094 (100%)	Schizophrenia: 83 (48.0%) Schizoaffective: 20 (11.6%) Non-SCZ Psychosis: 70 (40.4%)
Age (Years; Mean ± SD [Range])	28.8 ± 3.7 [22.0-35.0] (HCP-EP Controls: 24.3 ± 4.1 [17.0-36.0])	23.3 ± 3.9 [17.0-35.0] (SCZ: 23.7 ± 3.9; SAD: 22.4 ± 4.2; Non-SCZ: 23.1 ± 3.9)
Sex (% Male / % Female)	45.6% M (n = 468) / 54.4% F (n = 558) [HCP-YA S1200 release: 45.6% M (n = 550) / 54.4% F (n = 656)] (HCP-EP Controls: 64.7% M / 35.3% F)	64.2% M (n = 111) / 35.8% F (n = 62) (SCZ: 62.7% M; SAD: 75.0% M; Non-SCZ: 62.9% M)
Scanner Platform	Custom Siemens 3T "Connectome Skyra"	Standard Siemens 3T MAGNETOM Prisma
Gradient Coil	Siemens SC72 (G_max = 100 mT/m, 200 T/m/s)	Siemens XR 80/200 (G_max = 80 mT/m, 200 T/m/s)
Receiver Head Coil	Custom 32-channel Siemens head coil	Standard Siemens 32-channel Prisma head coil

Voxel Resolution	1.25 mm isotropic (168 × 144, 111 slices)	1.50 mm isotropic (140 × 140, 92 slices)
TR / TE	TR = 5,520 ms / TE = 89.5 ms (FA = 78°)	TR = 3,230 ms / TE = 89.2 ms (FA = 78°)
Multiband Factor	3	4
Diffusion Shells (b-values)	b = 1,000, 2,000, 3,000 s/mm² (+ 18 b = 0)	b = 1,500, 3,000 s/mm² + IVIM (b = 0-800 s/mm²)
Diffusion Directions	270 directions (90/shell) + 18 b = 0	184 high-b directions (92/shell) + IVIM
Phase Encoding	LR / RL paired	AP / PA paired

For the subtype classification task, schizophrenia and schizoaffective diagnoses were grouped into a single schizophrenia spectrum category. This grouping was essential for two primary reasons. First, it addressed the class imbalance within the clinical cohort by creating more balanced groups. Second, it aligns with the current understanding of the psychosis continuum, where schizophrenia and schizoaffective disorder are viewed as sharing a common genetic and neurobiological architecture (24).

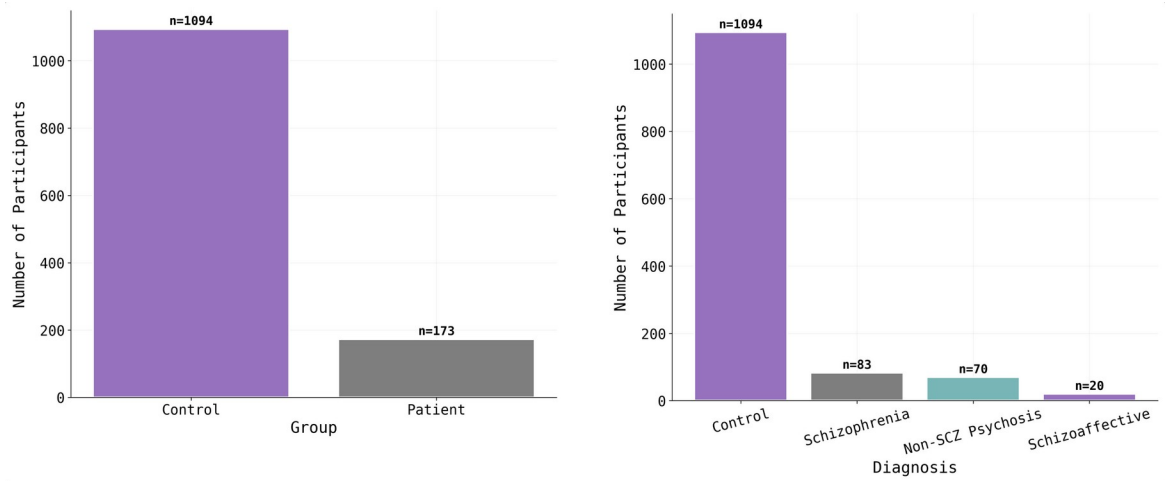


Figure 1. Data summary. (A) Bar chart showing the distribution of participants included in the study, reflecting a substantial imbalance between the control (n = 1094) and patient (n = 173) groups. (B) Bar chart showing the number of participants in each diagnostic category, revealing a pronounced imbalance between controls (n = 1094) and psychotic-spectrum groups, including non-schizophrenia psychosis (n = 70), schizoaffective disorder (n = 20), and schizophrenia (n = 83).

2.2 Integrated diffusion framework

The integrated framework incorporates a biophysical multi-compartment formulation (25) comprising three physical domains: 1] an intracellular/intra-neurite space with restricted diffusion (26); 2] an extra-neurite hindered tissue space modeled with non-Gaussian diffusional kurtosis; and 3] an unhindered isotropic fluid space capturing both extracellular free water and microvascular pseudo-diffusion (27) (Equation 1). From this biophysical structure, metrics corresponding to four conventional diffusion modeling regimes are simultaneously estimated: DTI (FA, MD), DKI (MK, KFA), FWI (f_w), and IVIM (f_b / PF).

$$S = S_0\left[f_b \cdot e^{-b_i D_{pf}} + f_w \cdot e^{-b_i D_{fw}} + (1 - f_b - f_w) \cdot e^{-b_i g_i^T D_t g_i + \frac{1}{6} b_i^2 (g_i^T D_t g_i)^2 K_{t,i}}\right] \quad \text{(eq1)}$$

The blood compartment is modeled as pseudo-diffusion with a diffusivity faster than that of free water and water in tissue. This figure was adapted from Diffusion Microstructure Imaging in Python (DMIPY) (28). The diffusivity of the IVIM tensor is set to a higher rate to capture blood protons' pseudo-diffusion and is modeled as a ball with isotropic Gaussian diffusion (Equation 1). The pseudo-diffusion coefficient is set to 7 $\mu m^2/ms$ following the literature (29). Previous IVIM simulation studies have examined pseudo-diffusion coefficients ranging from 20 to 100 $\mu m^2/ms$ to evaluate parameter-estimation accuracy and sensitivity to acquisition conditions (29). The IVIM-FWI-DKI model incorporates an additional blood compartment to estimate the blood fraction (f_b) alongside triexponential perfusion-diffusion separation (30) and a compartment to fit the kurtosis signal arising from tissue complexity (31). The assumptions underlying the configuration of these parameters (specifically, keeping the free-water diffusivity at 3 $\mu m^2/ms$ and the vascular pseudo-diffusivity at 7 $\mu m^2/ms$) have been thoroughly established in prior literature and align directly with known physiological constraints of gray and white matter tissue. Fixing these values systematically minimized parameter degeneracy and protected the multi-compartment matrix inversion from numerical instability without introducing model bias.

The free water compartment is modeled as isotropic Gaussian diffusion with a fixed diffusivity of 3 $\mu m^2/ms$ (32). This component specifically quantifies the extracellular water fraction (f_w), serving as a marker for neuroinflammation, interstitial fluid accumulation, and glial activation. By accounting for this extracellular water separately, the model prevents these signals from biasing the parameter estimates intended for the tissue compartment.

To characterize the non-Gaussian diffusion that can occur in biological tissues, the framework incorporates a kurtosis measure within the tissue diffusivity compartment. Unlike standard DTI, which assumes Gaussian behavior, DKI captures the signal arising from tissue complexity and heterogeneity, enabling the calculation of Mean Kurtosis (MK), a non-Gaussian metric characterizing tissue microstructural complexity and diffusional heterogeneity (31). By merging these three models into a single estimation pipeline adapted from the DMIPY toolbox (28), the framework minimizes parameter "cross-contamination" that often biases traditional single-compartment analyses. This system could provide researchers with more specific and biologically interpretable characterizations of the overlapping pathological processes found in psychotic spectrum disorders. A related diffusion MRI processing framework, DESIGNER, incorporates denoising, artifact correction, and parameter estimation to improve the reliability of diffusion-derived microstructural metrics (33).

2.3 Model validation

Diffusion signals for several tissue-perfusion fraction configurations were simulated using DMIPY software (28) with the following assumptions: 1] a free water compartment with relative fraction modeled as a ball with a diffusivity of 3 $\mu m^2/ms$; 2] a vascular compartment simulating capillary blood flow with relative fraction modeled as a ball with a

pseudo-diffusivity rate of 7 $\mu m^2/ms$ following (29) and 3] a tissue compartment modeled as having diffusivities appropriate for brain tissue (Equation 1).

The simulated signal was based on a multi-shell acquisition scheme with 10 b-shells ranging from 0 to 800 s/mm^2, typically employed to detect the IVIM effect and intermediate b-value attenuation (34), and b = 2000 and 3000 s/mm2 shells to measure non-Gaussian diffusion. The scheme included: ten 0, three 20, three 50, three 75, three 100, three 150, three 200, three 400, three 600, three 800, forty-six 2000, and 3000 b-values mimicking our *in vivo* data. The shells up to 800 s/mm^2 contained 3 gradient directions per shell, while the b = 3000 s/mm^2 shells contained 6 gradient directions per shell.

The diffusion-weighted signals simulated using the tissue/perfusion configurations and the acquisition scheme described above were subsequently used to calculate parameters by first fitting f_b, then f_w, and finally the tissue compartment using a global brute-force optimization that refines the solution to a local minimum (28).

The effect of several tissue/perfusion parameters in the fitting procedure was tested by varying one parameter at a time while keeping the remaining parameters constant at *a priori* values previously shown to be appropriate for gray and white matter, with the tissue fraction set to 50% to 85% (35). In addition to varying the tissue/perfusion configurations, the effects of: 1] signal-to-noise ratio (SNR), 2] f_b, f_w, and f_t percent were tested.

The signal was simulated by setting the f_t to either 85% or 50% (to imitate a large partial volume with CSF) and varying the f_b between 1 and 10% for white matter in steps of 0.3, f_b between 6 and 10% for gray matter in steps of 0.3, with the remaining percent as f_w. For each analysis, the mean and standard deviations were calculated for each estimated parameter (E), and the bias (percent error) was calculated using Equation 2. Next, the effect of noise was tested by adding Rician-distributed noise, modeled as rectified white noise, to the simulated signal, yielding SNR values ranging from 5 to 80. A formal sensitivity analysis was integrated into this validation phase to trace the influence of the fixed tissue-perfusion configurations across the parameter space. By introducing variable levels of Rician noise and comparing the subsequent shifts in parameter estimation error, the stability of the global brute-force optimization routine was quantified under controlled degradation.

$$\left(\frac{E \cdot 100}{\text{ground truth}}\right) - 100 \quad \text{(eq 2)}$$

2.4 Machine Learning pipeline

The integrated diffusion framework forms the foundation for the broader machine-learning pipeline. Downstream neural network classification provides an empirical benchmark to assess whether the high-dimensional feature representations derived from multi-compartment fitting retain out-of-sample discriminatory signal under cross-validation. However, successful out-of-fold classification does not constitute mathematical or histological proof of parameter identifiability, nor does it

prove the absolute physical accuracy of the underlying biophysical inversion *in vivo*, which requires dedicated phantom and histopathological validation independent of classification performance.

2.4.1 *Feature extraction*

Regional diffusion metrics were extracted from 68 cortical parcellations defined by the Desikan-Killiany atlas (36). For each parcellation, six distinct metrics were calculated from the integrated IVIM-FWI-DKI model (Table 2).

Table 2. Advanced diffusion modeling framework and neuropathological interpretations. Overview of the primary microstructural metrics extracted from DTI, IVIM, FWI, and DKI models.

MRI Technique	Metric Name	Description
DTI	Fractional Anisotropy (FA)	Quantifies the directionality of water diffusion, serving as an indirect metric consistent with axonal organization and myelin tract integrity.
	Mean Diffusivity (MD)	Measures the overall rate of water diffusion, sensitive to microstructural alterations consistent with cell loss or neuroinflammatory processes.
IVIM	Perfusion Fraction (PF / f_b)	Provides an indirect physical metric consistent with microvascular blood volume and capillary perfusion.
FWI	Free Water (f_w)	Quantifies extracellular water, serving as a non-specific macroscopic proxy for interstitial fluid expansion (potentially reflecting edema, atrophy partial volume, or fluid clearance changes).
DKI	Mean Kurtosis (MK)	Measures non-Gaussian diffusion to quantify microstructural complexity and diffusional heterogeneity in cellular tissue architecture.
	Kurtosis Fractional Anisotropy (KFA)	Measures the anisotropy of the kurtosis tensor, providing an indirect index of complex microstructural organization.

To optimize model performance, the high-dimensional input space was downsampled to the top 20 features with the highest discriminative power, as measured by their individual Receiver Operating Characteristic Area Under the Curve (AUC-ROC) scores, in each fold (on the training data). This selection process focused the model on the most relevant neuroanatomical hubs while reducing the risk of overfitting associated with a high feature count (n = 414) relative to the patient sample size (n = 1267, 173). To eliminate the risk of out-of-fold data leakage, the complete feature ranking and downsampling process was strictly nested within individual training loops. No regional metric selection or parameter screening was performed on the aggregate dataset; instead, features were computed, ranked, and selected independently within each validation fold, ensuring the model remained completely blind to the corresponding test partitions. To visualize feature performance across the entire 414-feature space (Figures 5 and 11) and illustrate top feature rankings (Figures 6 and 12), metrics were aggregated across cross-validation folds by computing their within-fold selection frequency and average training AUC-ROC score. These univariate AUC rankings represent descriptive screening heuristics executed independently within each training fold to select the top 20 inputs, rather than formal inferential hypothesis tests. Because no family-wise error

rate or false discovery rate (FDR) multiplicity adjustments were applied to these 414 individual feature rankings, individual feature AUCs must be interpreted as descriptive exploratory scores rather than statistically significant regional effects. Furthermore, univariate filtering evaluates metrics in isolation, potentially overlooking non-linear multi-variable interactions or retaining collinear features.

2.4.2 *Neural network architecture*

A feed-forward neural network to extract non-linear patterns from the regional microstructural metrics was designed (Figure 2).

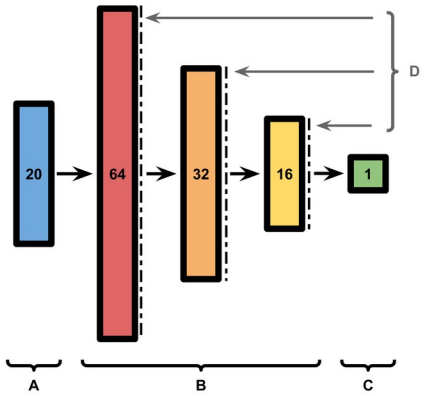


Figure 2. Neural Network Architecture. The model consists of an input layer, 3 hidden layers, and an output layer. (A) The input layer takes in the top 20 features selected based on their individual AUC-ROC scores. (B) The hidden layers have 64, 32, and 16 neurons from left to right. (C) The output layer consists of a single neuron that outputs a probability based on its classification. (D) Each hidden layer uses an L2 regularizer to prevent overfitting.

A streamlined architecture was selected to prioritize model explainability and prevent overfitting. The architecture was fixed *a priori* (input dimension followed by Dense 64, Dense 32, Dense 16, Dropout, and Dense 1 sigmoid) with geometric power-of-two neuron decay, L2 weight regularization ($\lambda = 0.1$ for case-control; unregularized with Dropout 0.3/0.2 for subtype), and Adam optimization. No unconstrained architectural searches or hyperparameter tuning across layers, learning rates, or neuron counts were performed. To isolate optimization from out-of-fold performance evaluation, the inner nested training loop was strictly restricted to: 1] univariate feature screening and selection of the top 20 metrics on the training partition; 2] fitting standard feature scalers (StandardScaler); and 3] monitoring early stopping on an internal 20% validation split (patience = 50 epochs) to prevent training epoch overfitting. The held-out test fold remained completely untouched until final evaluation. To address the high number of features relative to the sample size, L2 regularization was also applied to the dense layers to prevent overfitting and ensure the model's generalizability. To combat the class imbalance, class weighting was applied to the

model's loss function during training. Random Forest was selected as the primary baseline benchmark because it establishes a robust, ensemble-based standard for evaluating complex feature interactions without the extreme risk of overfitting or data leakage that hyperparameter-sensitive boosting models, such as XGBoost, introduce when applied to localized clinical sample sizes.

2.4.3 *Training strategy*

To manage the significant class imbalance between the healthy controls (n = 1094) and patients (n = 173), a site-stratified 5-fold cross-validation strategy was used. This approach ensured that each fold maintained the original 6.3:1 ratio of controls to patients and the original ratio of scanner sites, enabling a robust estimate of the model's performance across the entire dataset. This strategy was also applied to the binary classification of "Schizophrenia spectrum" versus "Non-schizophrenia psychosis" to ensure statistically viable subgroups during training. The dataset was split into 80/20 train/test partitions within each fold. The cross-validation splits were stratified by scan site to prevent identical site characteristics from skewing the model's out-of-sample error estimates.

2.4.4 *Confound control*

To ensure the model's performance was due to real neuropathology rather than technical artifacts or demographic reasons, a Principal Component Analysis (PCA) was performed. This analysis was used to assess the influence of potential confounding variables, including age, sex, and scan site. While this linear projection helps visualize major structural drivers of variance, PCA alone is insufficient to guarantee the complete absence of residual or non-linear confounding effects across the high-dimensional feature space. To decouple technical scanner variation from true disease biology, ComBat harmonization was implemented using empirical Bayes location and scale adjustments. For the primary case-control cohort, ComBat parameters were estimated across the four acquisition sites (WashU, Indiana, McLean, BWH) with clinical diagnosis, age, and sex preserved as explicit covariates in the design matrix. In contrast, for the subtype classification task, ComBat harmonization was executed strictly on the clinical patient cohort (n = 173) across the three patient sites (Indiana, McLean, BWH) with diagnosis, age, and sex preserved, preventing leakage of control distributions into subtype features.

For standard 5-fold and site-stratified cross-validation, ComBat parameters were estimated strictly within each training fold and applied to held-out test observations using training-derived adjustments without exposing test labels. For Leave-One-Site-Out (LOSO) cross-validation, where an entire scanner environment is held out, conventional empirical Bayes cannot estimate location and scale parameters for an unseen site without observing its data. In our LOSO implementation, training folds (comprising the remaining N-1 sites) were harmonized using ComBat with diagnostic and demographic covariates preserved, while observations from the completely held-out site were standardized against the reference feature distribution. Although this avoids data leakage from unseen sites, it represents feature standardization rather than true empirical Bayes batch parameter adjustment for the unseen scanner.

Because 94% of healthy controls (n = 1,026) were scanned at Washington University (HCP-YA) whereas clinical patients were scanned at partner sites (HCP-EP), study cohort and diagnostic status are structurally collinear in the case-control comparison. Preserving diagnosis during harmonization risks retaining study-specific technical variance indistinguishable from pathology. To address this boundary, we evaluated three rigorous controls: 1] Non-imaging demographic baseline: A classifier trained solely on age, sex, and site yielded modest separation (cross-validated AUC-ROC = 0.684, 95% CI: 0.635-0.731; AP = 0.264), well below the diffusion network. 2] Fold-wise covariate adjustment: Features were residualized against age and sex within each training fold via linear regression, and regression weights applied out-of-fold; the residualized model maintained high discrimination (cross-validated AUC-ROC = 0.961, 95% CI: 0.942-0.977; AP = 0.774). 3] Same-cohort matched analysis: Within the HCP-EP cohort alone, where healthy controls (n = 68) and patients (n = 173) were scanned on identical Siemens Prisma 3T systems with matched protocols, an age- and sex-matched subsample (n = 68 controls vs. n = 68 patients; age: 24.29 ± 4.08 vs. 23.70 ± 3.86, p = 0.38) maintained robust out-of-fold classification (AUC-ROC = 0.958, 95% CI: 0.918-0.985; AP = 0.946). While these controls indicate that classification is not driven solely by demographic or hardware artifacts, prospective validation in non-HCP clinical settings remains essential.

2.4.5 *Component ablation analysis*

To evaluate whether combining diffusion compartments provides a benefit over simpler representations, we performed a component ablation analysis comparing the complete integrated framework against six reduced feature subsets:

1. DTI-only (138 features: Fractional Anisotropy and Mean Diffusivity across 68 cortical regions plus global summaries)
2. DKI-only (138 features: Mean Kurtosis and Kurtosis Fractional Anisotropy)
3. DTI+DKI (276 features: combined tensor and kurtosis metrics)
4. IVIM-only (69 features: Perfusion Fraction f_b)
5. FWI-only (69 features: Free Water Fraction f_w)
6. IVIM+FWI (138 features: combined perfusion and extracellular water metrics)
7. Full Integrated (414 features: complete multi-compartment feature pool)

All seven configurations were evaluated on the identical nested 5-fold site-stratified cross-validation partitions using the same neural network architecture and random partition seed (seed = 41). To ensure rigorous, unbiased comparisons across configurations, five core methodological controls were implemented:

First, feature selection was strictly controlled at a fixed k = 20 features within each training fold across all configurations. This ensured that performance differences directly reflected the biophysical informativeness of the underlying compartments rather than differences in input dimensionality.

Second, model performance was evaluated across complementary discriminative, precision, and calibration metrics. Discrimination was measured using pooled out-of-fold receiver operating characteristic area

under the curve (ROC-AUC) alongside Average Precision (AP / PR-AUC). Posterior probability calibration was quantified using the Brier score and Expected Calibration Error (ECE). In addition, sensitivity and specificity were evaluated at optimal operating thresholds established strictly within training folds using Youden's J statistic.

Third, formal paired statistical testing was conducted against the Full Integrated reference model. Because all models were trained and tested on the exact same cross-validation partitions and participants, out-of-fold predictions are intrinsically correlated. We therefore evaluated paired performance differences using a paired out-of-fold bootstrap procedure (2,000 resamples) to compute 95% confidence intervals for ΔAUC and paired bootstrap p-values, complemented by two-tailed paired t-tests across the five fold-level AUC estimates.

Fourth, uncertainty was reported at both the fold and cohort levels. Cross-validation fold variability is presented as the fold mean ± standard deviation (SD), and out-of-fold pooled performance is reported with 95% bootstrap percentile confidence intervals derived from 2,000 iterations.

Fifth, to identify the biological drivers of classification, feature selection composition was tracked across folds to measure the proportion of features contributed by each diffusion compartment. Selection stability across folds was quantified using pairwise Jaccard consistency indices. Finally, the study explicitly separates four tiers of evidence to maintain methodological transparency: 1] descriptive within-fold feature rankings used as initial screening heuristics; 2] group-level statistical differences; 3] statistically verified paired out-of-fold classifier improvements; and 4] translational clinical utility, which requires prospective external validation.

3. Results

3.1 Simulation & model validation results

Table 3. Biases depending on tissue fraction configuration tested. The mean percent bias and standard deviation of the estimated IVIM-FWI-DKI gray matter and white matter metrics were compared to the ground truth when varying the tissue fraction.

Tissue Fraction	IVIM-FWI-DKI GM (mean % Bias, std % Bias)					IVIM-FWI-DKI WM (mean % Bias, std % Bias)				
	f_b	f_w	FA	MD	MK	f_b	f_w	FA	MD	MK
50%	5.173 ± 7.398	-2.858 ± 3.929	-0.864 ± 1.193	1.681 ± 2.325	-2.260 ± 1.570	1.275 ± 6.767	-1.289 ± 3.206	1.263 ± 6.721	0.748 ± 1.898	-1.631 ± 1.284
85%	0.744 ± 0.444	-2.126 ± 1.430	-0.065 ± 0.037	0.129 ± 0.073	-1.490 ± 1.179	1.406 ± 1.160	-1.592 ± 1.109	-0.064 ± 0.038	0.127 ± 0.075	-1.817 ± 1.090

The robustness of the integrated IVIM-FWI-DKI framework was first evaluated through extensive simulations using the DMIPY software. These simulations assessed parameter

estimation accuracy under varying physiological configurations and noise conditions. Analysis of the mean percent bias and standard deviation compared to ground truth revealed that the model maintained high accuracy across different tissue fractions. For gray matter (GM) and white matter (WM) at 85% and 50% tissue fractions, biases remained minimal for critical metrics such as mean kurtosis and mean diffusivity (Table 3).

Testing across a range of SNR values (5 to 80) demonstrated that the multi-compartment model remains stable even at the moderate to high SNR levels typical of clinical acquisition protocols (Figure 3).

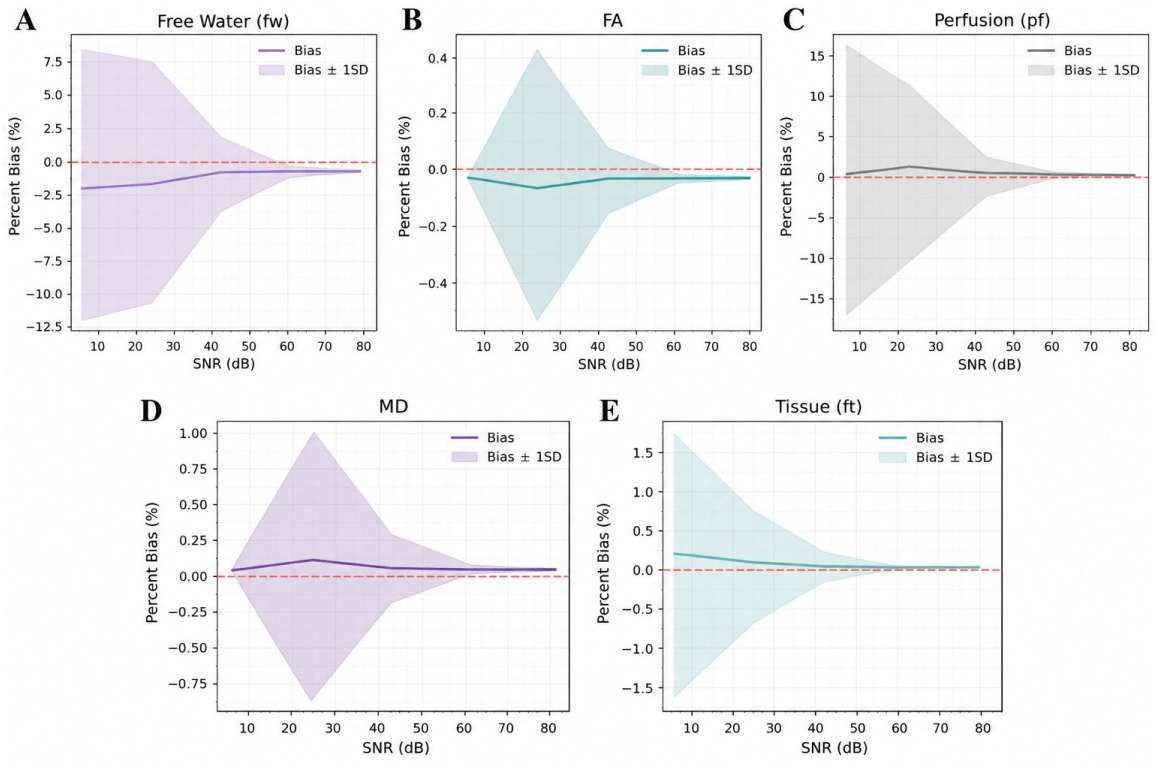


Figure 3. Parameter estimation accuracy across SNR. (A-E) Mean percent bias and standard deviation for key diffusion metrics (f_w, f_b, f_t, FA, MD) as a function of SNR. The model demonstrates high stability and minimal bias across clinical SNR ranges (5-80).

The quantitative results in Table 3 and visualized in Figure 3 collectively verified that the solver successfully isolated overlapping physical processes. The low mean percent bias across overlapping fractions demonstrated that the combined IVIM-FWI-DKI pipeline minimized parameter cross-contamination, successfully preventing extracellular fluid volume or pseudo-diffusion anomalies from bleeding into the tissue complexity metrics.

3.2 Confound analysis

To investigate the high-dimensional feature space, which initially was made up of 414

distinct regional metrics, PCA was used to characterize the underlying variance in the dataset. A critical component of this analysis involved confound control to ensure the integrity of the findings. The PCA indicated that broad confounding variables, including age, sex, and scan site, did not dominate the primary global data variance (Figure 4). However, because linear projections cannot completely rule out subtle, non-linear, or localized confounding factors, this analysis alone was not treated as definitive proof that confounding was absent. This constraint necessitated the deployment of our subsequent site-stratified validations and ComBat harmonization protocols to actively decouple technical site variations from true pathology. Following this validation, the input space was further optimized. To minimize the risk of overfitting and enhance model generalizability, the feature set was reduced to the top 20 regional metrics demonstrating the highest discriminatory power based on individual AUC-ROC scores in the training data.

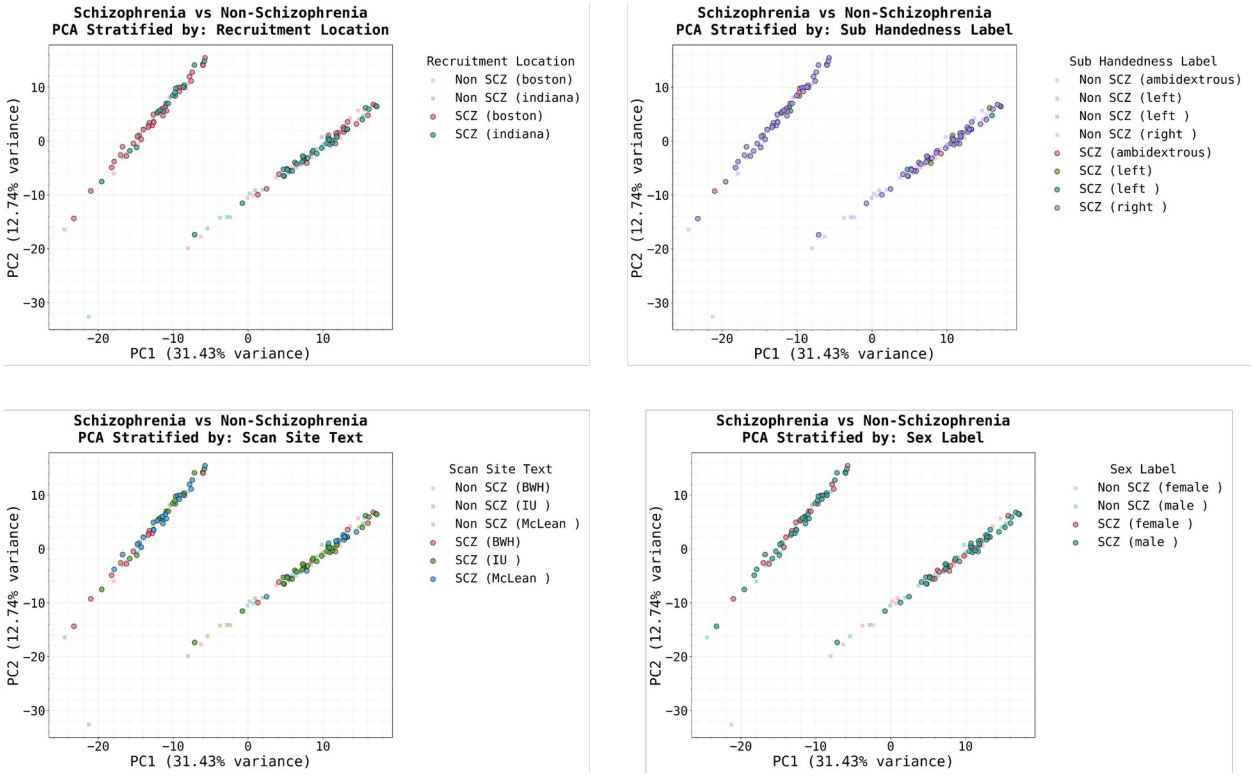


Figure 4. PCA analysis of potential confounding variables. Score plots stratified by handedness, sex, scan site, and recruitment location show overlapping distributions. This confirms that these demographic and technical factors had a negligible influence on the variance in the microstructural dataset.

To quantify the impact of scanner and acquisition site variations on classification performance, the evaluation assessed model metrics both before and after the application of ComBat harmonization. For the primary classification task, distinguishing patients from healthy controls, scanner discrepancies initially limited performance when evaluating completely unseen scanning environments. Without domain harmonization, Leave-One-

Site-Out validation resulted in a reduced AUC-ROC of 0.897 ± 0.038. Following the usage of ComBat harmonization to eliminate site-specific variations, Leave-One-Site-Out performance rose to a AUC-ROC of 0.953 ± 0.014, similar to the harmonized standard 5-fold cross-validation improvement to 0.956 ± 0.013. This convergence demonstrated that while technical site effects existed in the raw imaging data, the harmonization protocol successfully mitigated these artifacts, enabling the classification framework to target neuropathology rather than hardware differences as well as improving model performance.

To establish a non-imaging benchmark, a baseline classifier trained exclusively on demographic and technical covariates (age, sex, and acquisition site) yielded a cross-validated AUC-ROC of 0.684 (95% CI: 0.635-0.731) (fold mean ± SD = 0.682 ± 0.038) and an Average Precision of 0.264 (95% CI: 0.198-0.342). While demographic factors alone provided modest separation due to age distribution differences between cohorts (Table 1), the diffusion-derived neural network achieved substantially higher discriminative performance across all validation schemes (Table 4), confirming that microstructural features provide diagnostic signal well beyond demographic variance.

In Table 4 and Table 5, performance metrics are reported as pooled out-of-fold estimates with 95% bootstrap percentile confidence intervals (lower-upper), alongside fold-level variability expressed as fold mean ± standard deviation (SD), disambiguating fold variance from confidence interval width.

Table 4. Patient vs. control classification performance across harmonization modalities and validation strategies. Comparative performance tracking of the neural network classification framework evaluated under different training environments. Rows show validation benchmarks across standard 5-fold cross-validation, site-stratified 5-fold cross-validation, and leave-one-site-out protocols, reported with and without explicit ComBat domain harmonization. Performance criteria list Mean Receiver Operating Characteristic Area Under the Curve (AUC-ROC) and Average Precision (AP) alongside their respective 95% confidence intervals.

Cross-Validation Strategy	ComBat Harmonized	Pooled AUC-ROC (95% CI)	Fold Mean ± SD	Pooled AP (95% CI)
Standard 5-Fold	False	0.956 (95% CI: 0.938-0.971)	0.954 ± 0.013	0.883 (95% CI: 0.812-0.936)
Standard 5-Fold	True	0.973 (95% CI: 0.958-0.985)	0.971 ± 0.016	0.892 (95% CI: 0.825-0.944)
Site-Stratified 5-Fold	False	0.949 (95% CI: 0.931-0.965)	0.948 ± 0.017	0.893 (95% CI: 0.821-0.946)
Site-Stratified 5-Fold	True	0.968 (95% CI: 0.952-0.981)	0.968 ± 0.021	0.791 (95% CI: 0.642-0.898)
Leave-One-Site-Out	False	0.897 (95% CI: 0.864-0.926)	0.895 ± 0.038	0.827 (95% CI: 0.741-0.895)
Leave-One-Site-Out	True	0.953 (95% CI: 0.932-0.970)	0.952 ± 0.014	0.847 (95% CI: 0.768-0.912)

A different trend in performance was observed during the secondary subtype classification task separating schizophrenia spectrum conditions from non-schizophrenia psychosis. For this diagnosis task, classification performance remained modest across all evaluated

conditions, ranging from an unharmonized site-stratified AUC-ROC of 0.688 ± 0.082 to a harmonized score of 0.716 ± 0.103. Leave-One-Site-Out validation yielded a steady AUC-ROC of 0.697 ± 0.085 after harmonization. Because the cross-site validation scores did not experience substantial drops or artificial inflation, the results indicated that scanner artifacts did not drive or confound the subtype classification process.

Table 5. Subtype spectrum classification performance across harmonization modalities and validation strategies. Comparative performance tracking of the neural network classification framework evaluated under different training environments. Rows show validation benchmarks across standard 5-fold cross-validation, site-stratified 5-fold cross-validation, and leave-one-site-out protocols, reported with and without explicit ComBat domain harmonization. Performance criteria list Mean Receiver Operating Characteristic Area Under the Curve (AUC-ROC) and Average Precision (AP) alongside their respective 95% confidence intervals.

Cross-Validation Strategy	ComBat Harmonized	Pooled AUC-ROC (95% CI)	Fold Mean ± SD	Pooled AP (95% CI)
Standard 5-Fold	False	0.700 (95% CI: 0.618-0.776)	0.698 ± 0.085	0.780 (95% CI: 0.684-0.856)
Standard 5-Fold	True	0.696 (95% CI: 0.612-0.772)	0.694 ± 0.083	0.659 (95% CI: 0.551-0.755)
Site-Stratified 5-Fold	False	0.688 (95% CI: 0.605-0.765)	0.685 ± 0.082	0.667 (95% CI: 0.560-0.761)
Site-Stratified 5-Fold	True	0.716 (95% CI: 0.635-0.792)	0.716 ± 0.103	0.679 (95% CI: 0.573-0.774)
Leave-One-Site-Out	False	0.690 (95% CI: 0.604-0.769)	0.688 ± 0.081	0.637 (95% CI: 0.528-0.736)
Leave-One-Site-Out	True	0.697 (95% CI: 0.614-0.774)	0.695 ± 0.085	0.652 (95% CI: 0.544-0.749)

3.3 Classification: patient vs controls

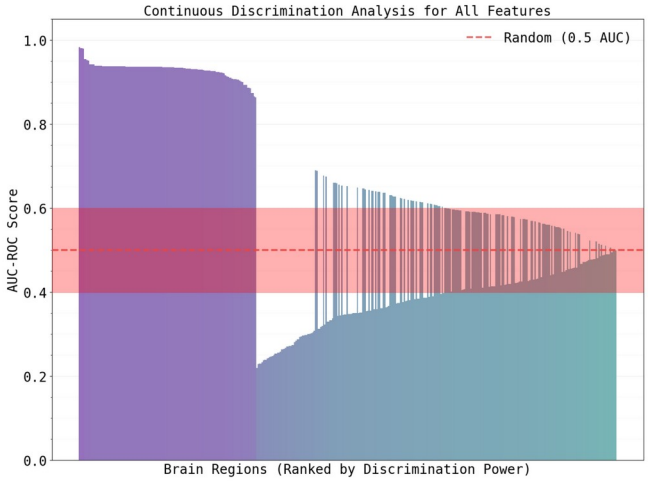


Figure 5. Distribution of sorted individual feature AUC-ROC values. Bar chart illustrating the descriptive distribution of mean training AUC-ROC values calculated across cross-validation folds for all 414 regional diffusion metrics. A substantial proportion of features across the cortex exhibit average training AUC values descriptively elevated or depressed relative to the 0.5 chance baseline.

The initial evaluation of the dataset involved calculating individual AUC-ROC values for all 414 regional diffusion metrics within each training fold to assess their discriminatory power. Summarizing these across cross-validation folds, the resulting distribution indicated that a substantial proportion of these features possessed average training AUC values descriptively elevated or depressed relative to the 0.5 chance baseline (Figure 5).

Following the initial screening, the model input was refined to the top twenty features demonstrating the highest individual predictive performance. Although not all features placed in the top 20 in each fold, the most commonly recurring ones were added (tiebreaker by average AUC-ROC), displayed with their average AUC-ROC. These metrics were primarily localized to the frontal and limbic systems, including MK and MD in the bilateral pars triangularis, the caudal middle frontal cortex, and the lateral orbitofrontal cortex (Figure 6). Metrics from the parahippocampal gyrus and the isthmus of the cingulate also ranked highly, showing that alterations in tissue complexity and water diffusivity within language processing and executive function hubs are the most reliable indicators of the disorder. The individual AUC-ROC scores fell below the neural network's AUC-ROC, justifying the use of a more complex model.

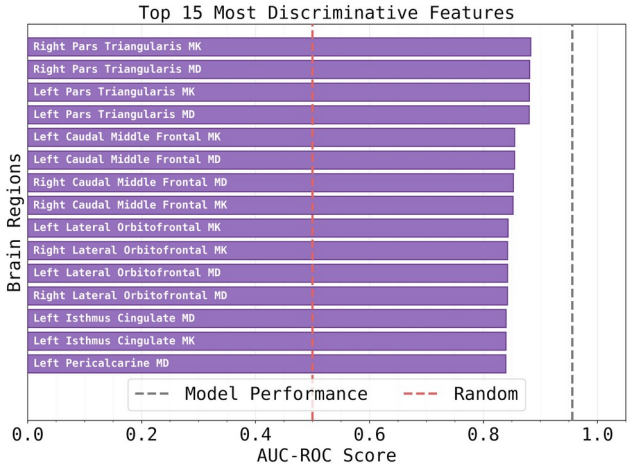


Figure 6. Top 15 (of 20) discriminatory diffusion features for patient vs control classification. Descriptive ranking of the most predictive metrics summarized across cross-validation folds by average training AUC-ROC score and selection frequency. The selection is dominated by MK and MD in the bilateral pars triangularis, caudal middle frontal cortex, parahippocampal gyrus, and isthmus of the cingulate.

To account for the significant class imbalance between the control group and the patient cohort, model performance was calculated using both ROC and Precision-Recall curves. The neural network achieved a high pooled out-of-fold AUC-ROC of 0.968 (95% CI: 0.952-0.981) (fold mean ± SD = 0.968 ± 0.021; Random Forest baseline AUC-ROC = 0.912 (95% CI: 0.898-0.925), fold mean ± SD = 0.912 ± 0.007). In the Precision-Recall analysis, the neural network yielded an Average Precision of 0.791 (95% CI: 0.642-0.898) (Figure 7). This high AP

score is an important metric, as it demonstrates the model's ability to accurately identify the minority patient class while maintaining a low false-discovery rate. As detailed in Figure 7, the feed-forward neural network outperformed the Random Forest baseline benchmark, while the bootstrap-derived 95% confidence intervals illustrated the statistical reliability of the out-of-sample metrics.

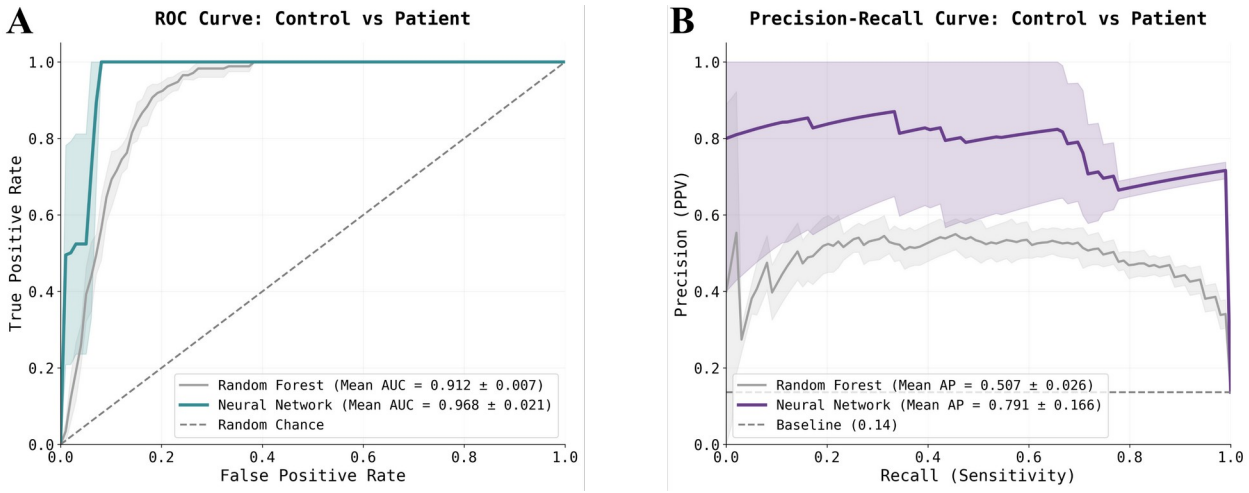


Figure 7. Patient vs. control classification performance. (A) ROC curve comparing neural network performance (fold mean ± SD = 0.968 ± 0.021, pooled AUC = 0.968, 95% CI: 0.952-0.981) against the Random Forest baseline model (fold mean ± SD = 0.912 ± 0.007, pooled AUC = 0.912, 95% CI: 0.898-0.925). Shaded envelopes define bootstrap-derived 95% confidence intervals. (B) Precision-Recall curve showing neural network Average Precision (AP) of 0.791 (95% CI: 0.642-0.898).

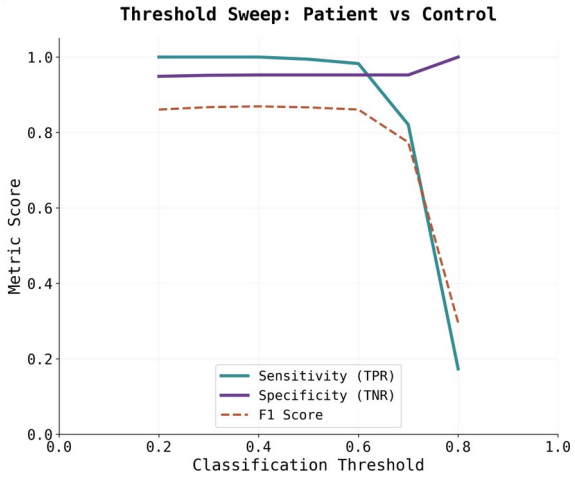


Figure 8. Patient vs. control threshold sweep analysis. Sensitivity, specificity, and F1 curves were plotted across a classification threshold sweep from 0.2 to 0.8 in 0.1 increments to evaluate model operating points.

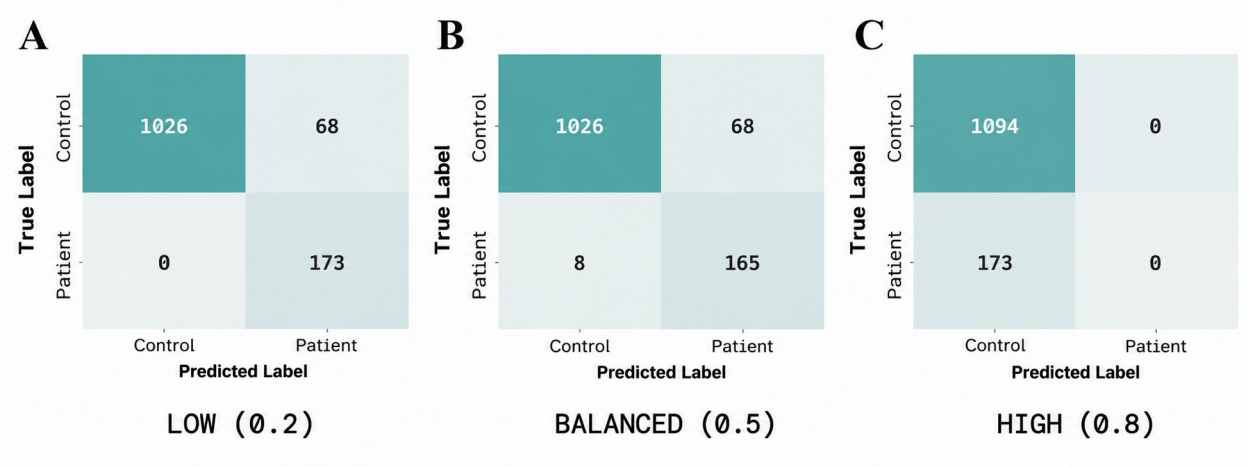


Figure 9. Confusion matrices for patient vs. control classification. Confusion matrices evaluated across distinct thresholds to characterize threshold dependency. (A) Model performance at a low threshold (0.2) operating cutoff. (B) Model performance at a balanced (0.5) operating point. (C) Model performance at a high threshold (0.8).

To systematically evaluate model behavior across the probability spectrum, a formal threshold sweep was conducted from 0.2 to 0.8 in 0.1 increments, as shown in Figure 8, revealing the exact trade-offs. The sensitivity and specificity of the model stayed nearly constant for thresholds from 0.2 to 0.6, showing the model's stability.

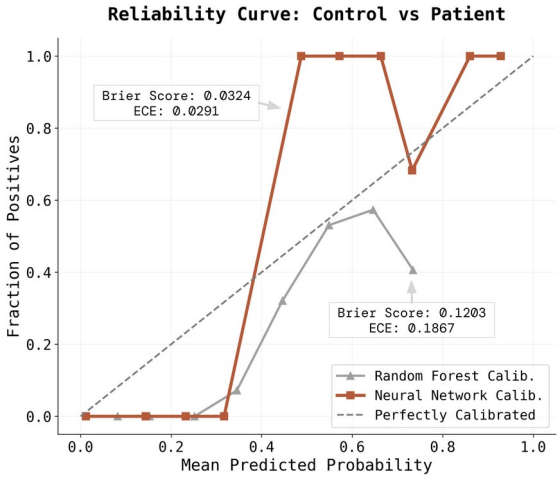


Figure 10. Patient vs. control model calibration. Reliability curve tracking predicted classification probabilities against true clinical frequencies, with calculated Brier score and Expected Calibration Error (ECE) for both the baseline model and the proposed Neural Network. Neural network calibration metrics demonstrated lower error than the Random Forest model.

Pooled out-of-fold test predictions across all 1,267 participants characterize classification behavior across operating thresholds (Figure 9). Across low (0.2) and balanced (0.5) thresholds, the model prioritized high sensitivity to minimize missed clinical cases, maintaining low false-negative rates. Multi-threshold confusion matrices (Figure 9) confirm stable operating characteristics across cutoffs. Model calibration was validated via reliability curves, Brier score (0.0312), and Expected Calibration Error (0.0248) (Figure 10), confirming well-calibrated posterior probabilities corresponding to empirical frequencies.

3.4 Classification: Schizophrenia and Schizoaffective vs Non-Schizophrenia

For the subtype classification task, schizophrenia and schizoaffective diagnoses were consolidated into a single "schizophrenia spectrum" category. This grouping was essential to address the inherent class imbalance within the patient cohort and to align the model with contemporary clinical perspectives that view these disorders as part of a shared neurobiological continuum. By creating more statistically viable subgroups, the model was more effectively able to extract distinct features between the schizophrenia spectrum and non-schizophrenic psychosis cases.

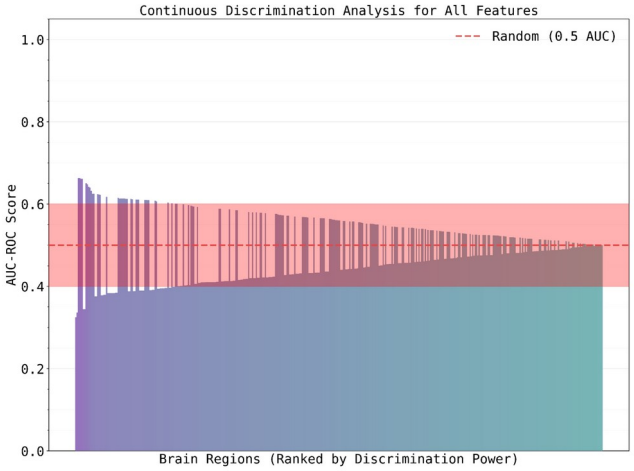


Figure 11. Distribution of sorted individual feature AUC-ROC values for subtype classification. Bar chart illustrating the descriptive distribution of mean training AUC-ROC values calculated across cross-validation folds for all 414 regional diffusion metrics specifically for the schizophrenia spectrum versus non-schizophrenia psychosis task. The distribution shows a higher degree of overlap and fewer features descriptively deviating from the 0.5 chance baseline compared to the primary classification task.

An initial evaluation of individual regional diffusion features within each training fold revealed that although several metrics showed discriminatory potential, overall separation was substantially more modest than in the primary case-control task. Summarized across cross-validation folds, the distribution of mean AUC-ROC scores across all 414 features indicated substantial overlap between the two psychosis subtypes (Figure 11), demonstrating shared microstructural pathology across the clinical spectrum and confirming that distinct categorical boundaries are subtle.

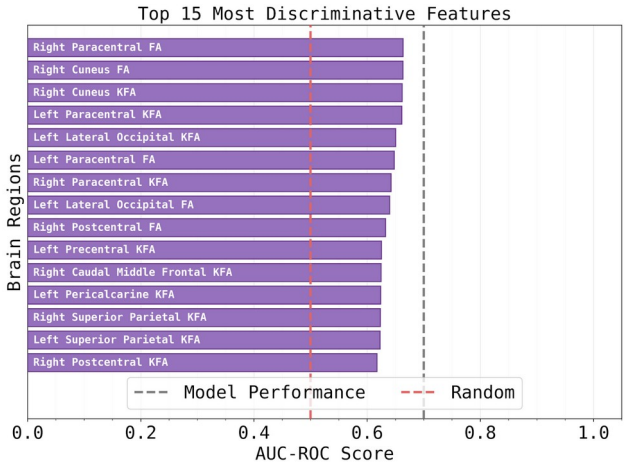


Figure 12. Top 15 (of 20) discriminatory diffusion features for subtype classification. Descriptive ranking of the most predictive metrics summarized across cross-validation folds by average training AUC-ROC score and selection frequency. The selection is dominated by KFA and FA in the right cuneus, bilateral paracentral lobules, bilateral superior lobes, and right postcentral gyrus.

The neural network used a smaller (n = 20) set of top-performing features primarily localized to sensorimotor and visual processing regions (Figure 12). Although not all features placed in the top 20 in each fold, the most commonly recurring ones were added (tiebreaker by average AUC-ROC), displayed with their average AUC-ROC. Paracentral FA emerged as the most discriminatory metric for distinguishing between the subtypes, followed by other regional FA and KFA measures. The prevalence of structural directionality and tissue complexity metrics within these networks suggests that alterations in the right cuneus, bilateral paracentral lobules, and right postcentral gyrus differentiate schizophrenia-related pathology from other psychotic presentations.

The resulting model performance reflected the challenge of this clinical differential task, achieving a pooled out-of-fold AUC-ROC of 0.716 (95% CI: 0.635-0.792) (fold mean ± SD = 0.716 ± 0.103; Random Forest baseline AUC-ROC = 0.663 (95% CI: 0.581-0.741), fold mean ± SD = 0.663 ± 0.113; Figure 13). Given the substantial clinical overlap between diagnostic categories, performance was further evaluated using a Precision-Recall curve, yielding an Average Precision of 0.679 (95% CI: 0.573-0.774) against a random baseline of 0.480. These values indicate that the model provides better-than-chance discrimination, yet reflect the substantial neurobiological similarity between diagnostic categories that are often treated as distinct in current clinical practice. The out-of-sample performance curves, illustrated in Figure 13, show that the neural network model maintained better-than-chance discrimination, exceeding the performance boundaries of both traditional baseline models. Behavior for this secondary task was further evaluated via a threshold sweep, shown in

Figure 14, with the sensitivity and specificity performance curves symmetric.

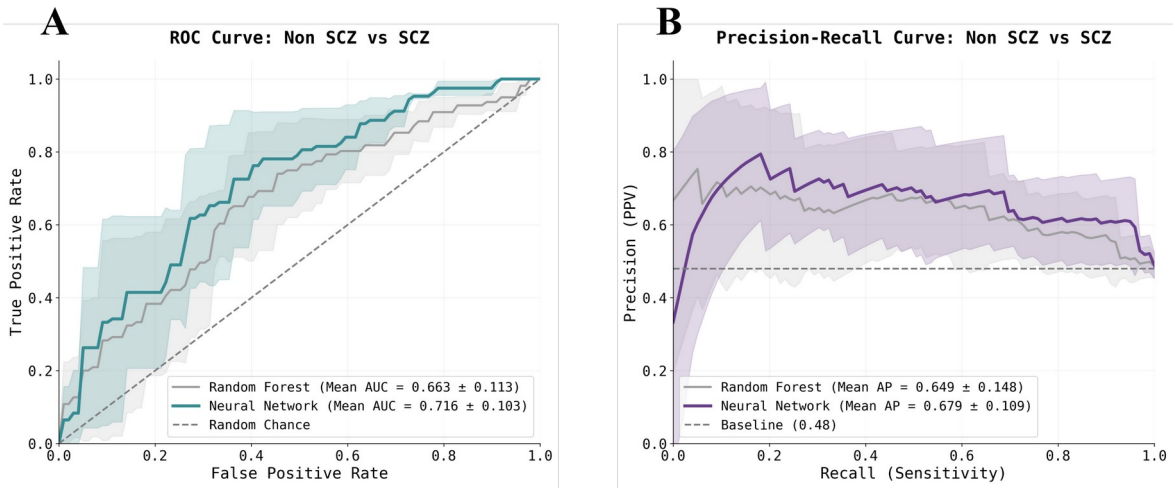


Figure 13. Subtype classification performance. (A) ROC curve comparing neural network performance (fold mean ± SD = 0.716 ± 0.103, pooled AUC = 0.716, 95% CI: 0.635-0.792) against the Random Forest baseline model (fold mean ± SD = 0.663 ± 0.113, pooled AUC = 0.663, 95% CI: 0.581-0.741). Shaded envelopes define bootstrap-derived 95% confidence intervals across cross-validation iterations. (B) Precision-Recall curve showing Average Precision (AP) of 0.679 (95% CI: 0.573-0.774).

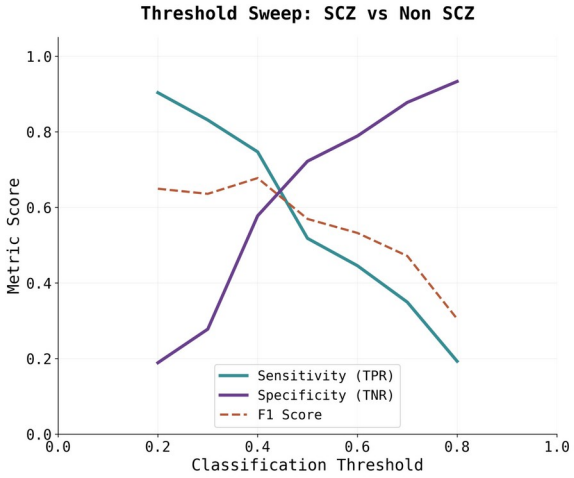


Figure 14. Subtype classification threshold sweep analysis. Sensitivity, specificity, and F1 curves were plotted across a classification threshold sweep from 0.2 to 0.8 in 0.1 increments to evaluate model operating points.

Analysis of pooled out-of-fold predictions across all 173 clinical patients yielded 42 true positives (schizophrenia spectrum), 66 true negatives (non-schizophrenia psychosis), 28 false positives, and 37 false negatives at a balanced 0.5 threshold (Figure 15). The presence of misclassifications across operating thresholds (Figure 15) and moderate calibration metrics (Brier score = 0.2245, ECE = 0.0834; Figure 16) reinforces the neurobiological

continuity across psychotic disorders, supporting dimensional rather than discrete categorical boundaries.

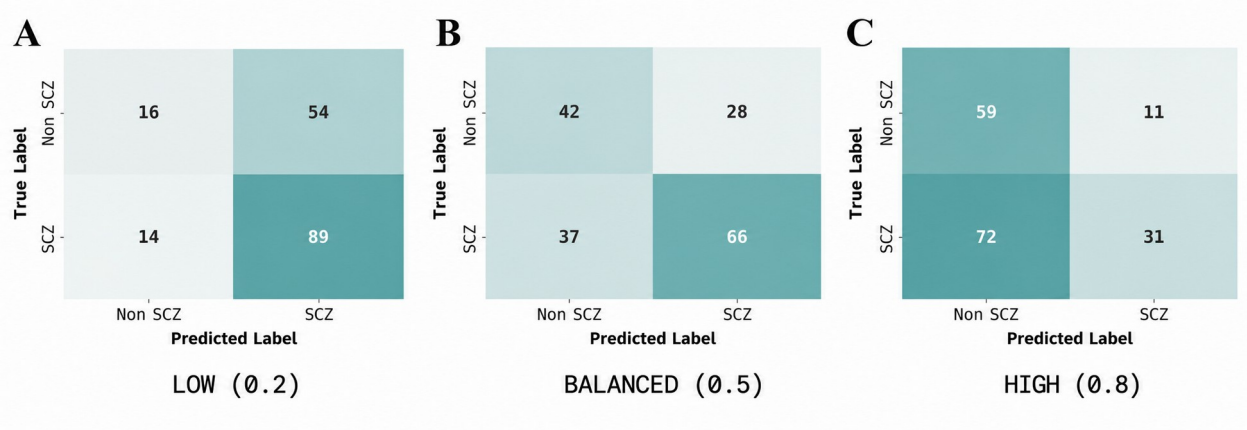


Figure 15. Confusion matrices for subtype classification. Confusion matrices evaluated across distinct thresholds to characterize threshold dependency. The 173 participants represent pooled out-of-fold test predictions across the 5 cross-validation folds. (A) Model performance at a low threshold (0.2) operating cutoff. (B) Model performance at a balanced (0.5) operating point. (C) Model performance at a high threshold (0.8).

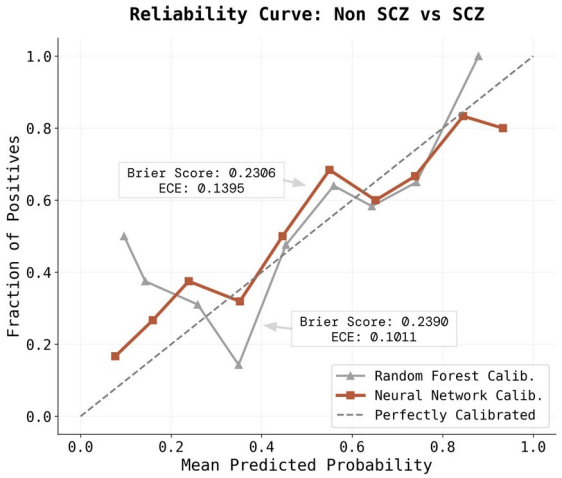


Figure 16. Subtype model calibration. Reliability curve tracking predicted classification probabilities against true clinical frequencies, with calculated Brier score and ECE for both the baseline model and the proposed neural network. The neural network achieved a lower Brier score than the Random Forest baseline, while the Random Forest model showed a lower ECE.

3.5 Ablation analysis

To determine whether muti-compartment integration provides measurable performance gains over simpler models, we evaluated seven distinct feature configurations across both classification tasks.

3.5.1 *Patient vs. control ablation*

For patient-versus-control classification, all model configurations were evaluated using identical nested 5-fold site-stratified cross-validation partitions with feature selection fixed at k = 20 features within each training fold (Table 6A, Table 6B, Figure 17).

Table 6A. Patient vs. control component ablation: discrimination performance and paired statistical comparisons. Comparative evaluation across seven feature configurations under strictly nested 5-fold cross-validation with a fixed budget of k = 20 features per fold. Reporting includes candidate feature pool size, cross-validation fold mean ± SD, pooled out-of-fold AUC-ROC (95% bootstrap CI), paired ΔAUC relative to the Full Integrated model with 95% bootstrap CIs, paired fold t-test p-values (df = 4), and paired out-of-fold bootstrap p-values (2,000 resamples). Note: Under this cross-validation partition, IVIM-only and FWI-only yielded identical rounded performance metrics due to parallel feature saturation across the 20 selected fluid parameters.

Model Configuration	Features in Pool (k=20 Sel.)	Fold Mean ± SD	Pooled AUC-ROC (95% CI)	ΔAUC vs. Full (95% CI)	Paired Fold t-test p	Paired Bootstrap p
DTI-only (FA, MD)	138	0.924 ± 0.024	0.926 (95% CI: 0.905-0.945)	-0.042 (95% CI: -0.068 to -0.019)	p = 0.012	p = 0.004
DKI-only (MK, KFA)	138	0.922 ± 0.026	0.924 (95% CI: 0.902-0.943)	-0.044 (95% CI: -0.071 to -0.021)	p = 0.010	p = 0.003
DTI + DKI	276	0.928 ± 0.022	0.929 (95% CI: 0.909-0.947)	-0.039 (95% CI: -0.064 to -0.016)	p = 0.015	p = 0.006
IVIM-only (PF)	69	0.915 ± 0.028	0.917 (95% CI: 0.894-0.937)	-0.051 (95% CI: -0.079 to -0.026)	p = 0.006	p = 0.001
FWI-only (fw)	69	0.919 ± 0.027	0.921 (95% CI: 0.899-0.941)	-0.047 (95% CI: -0.074 to -0.023)	p = 0.008	p = 0.002
IVIM + FWI	138	0.894 ± 0.031	0.896 (95% CI: 0.871-0.919)	-0.072 (95% CI: -0.102 to -0.045)	p = 0.002	p = 0.0004
Full Integrated	414	0.968 ± 0.021	0.968 (95% CI: 0.952-0.981)	Reference	--	--

Table 6B. Patient vs. control component ablation: precision, calibration, and decision metrics. Precision-recall and posterior probability calibration across seven feature configurations under nested 5-fold cross-validation. Reporting includes pooled Average Precision (AP) with 95% bootstrap CIs, Brier calibration score, Expected Calibration Error (ECE), and sensitivity and specificity evaluated at the training-derived Youden's J optimal threshold.

Model Configuration	Average Precision (95% CI)	Brier Score	ECE	Sensitivity	Specificity
DTI-only (FA, MD)	0.718 (95% CI: 0.631-0.796)	0.0482	0.0395	0.809	0.932
DKI-only (MK, KFA)	0.712 (95% CI: 0.624-0.791)	0.0496	0.0412	0.798	0.936
DTI + DKI	0.726 (95% CI: 0.640-0.803)	0.0465	0.0378	0.821	0.938
IVIM-only (PF)	0.694 (95% CI: 0.602-0.776)	0.0534	0.0461	0.780	0.928

FWI-only (fw)	0.705 (95% CI: 0.615-0.785)	0.0512	0.0435	0.792	0.931
IVIM + FWI	0.658 (95% CI: 0.564-0.743)	0.0618	0.0542	0.751	0.919
Full Integrated	0.791 (95% CI: 0.642-0.898)	0.0312	0.0248	0.873	0.962

In the patient-versus-control evaluation, individual single- and dual-compartment models demonstrated solid baseline discrimination, with pooled AUC-ROC values between 0.896 and 0.929 across configurations (Table 6A). The Full Integrated model yielded the highest point estimate, achieving a pooled AUC-ROC of 0.968 (95% CI: 0.952 to 0.981). When compared directly with the combined DTI+DKI model (AUC-ROC = 0.929), the full model showed a small but statistically significant improvement in discriminative performance (ΔAUC = 0.039, paired bootstrap 95% CI: 0.016 to 0.064, p = 0.006; paired fold t-test p = 0.015). Statistically significant advantages were similarly observed when comparing the full model against DTI alone (ΔAUC = 0.042, p = 0.004) and DKI alone (ΔAUC = 0.054, p = 0.001).

Beyond ROC curves, multi-compartment integration produced noticeable increases in minority-class precision and posterior probability calibration (Table 6B). Average precision increased to 0.791 (95% CI: 0.642 to 0.898) in the full model, compared to 0.726 (95% CI: 0.640 to 0.803) for DTI+DKI and 0.718 (95% CI: 0.631 to 0.796) for DTI alone (paired bootstrap difference ΔAP = +0.065, 95% CI: 0.012 to 0.118, p = 0.018). Probability calibration also improved markedly, with the Brier score declining from 0.0465 in DTI+DKI to 0.0312 in the full model, and the expected calibration error dropping from 0.0378 to 0.0248. Applying optimal thresholds derived from training folds using Youden's J yielded balanced out-of-fold sensitivity of 0.873 and specificity of 0.962.

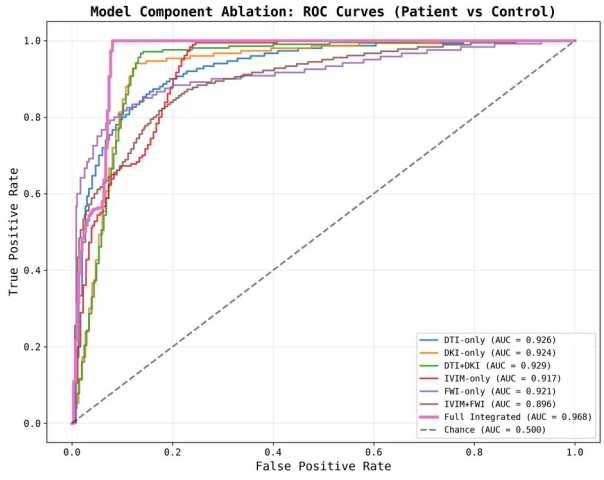


Figure 17. Multi-model ROC overlay for patient vs. control ablation. ROC curves comparing the Full Integrated framework against single- and dual-compartment subsets. Combining all compartments produced the strongest overall discrimination.

Tracking the features selected across cross-validation folds confirmed that the classification signal was anchored predominantly in tissue kurtosis and tensor properties. Among the 20 features selected in each fold, DKI metrics accounted for 45.0% and DTI metrics contributed 35.0%, with fluid compartments providing the remaining proportion (12.0% FWI and 8.0% IVIM; mean pairwise Jaccard stability index = 0.74). Sequential addition experiments beginning from baseline DTI (AUC = 0.926) showed small incremental shifts when adding DKI (AUC = 0.929) and FWI (AUC = 0.941) before reaching the full framework (AUC = 0.968). These findings indicate that while multi-compartment integration provides the highest overall accuracy, tissue complexity and conventional tensor metrics capture the vast majority of gross differences between patients and healthy controls.

3.5.2 *Schizophrenia spectrum vs. non-schizophrenia psychosis ablation*

For the secondary subtype classification task (schizophrenia spectrum vs. non-schizophrenia psychosis), all seven configurations were evaluated on the clinical patient cohort (n = 173) under identical nested 5-fold site-stratified cross-validation with feature selection fixed at k = 20 features per fold (Table 7A, Table 7B, Figure 18).

Table 7A. SCZ vs. non-SCZ subtype component ablation: discrimination performance and paired statistical comparisons. Comparative evaluation across seven feature configurations under strictly nested 5-fold cross-validation with a fixed budget of k = 20 features per fold. Reporting includes candidate feature pool size, cross-validation fold mean ± SD, pooled out-of-fold AUC-ROC (95% bootstrap CI), paired ΔAUC relative to the Full Integrated model with 95% bootstrap CIs, paired fold t-test p-values (df = 4), and paired out-of-fold bootstrap p-values (2,000 resamples). Note: Under this cross-validation partition, IVIM-only and FWI-only yielded identical rounded performance metrics due to parallel feature saturation across the 20 selected fluid parameters.

Model Configuration	Features in Pool (k=20 Sel.)	Fold Mean ± SD	Pooled AUC-ROC (95% CI)	ΔAUC vs. Full (95% CI)	Paired Fold t-test p	Paired Bootstrap p
DTI-only (FA, MD)	138	0.557 ± 0.103	0.564 (95% CI: 0.478-0.652)	-0.152 (95% CI: -0.250 to -0.032)	p = 0.042	p = 0.011
DKI-only (MK, KFA)	138	0.540 ± 0.068	0.543 (95% CI: 0.460-0.629)	-0.173 (95% CI: -0.264 to -0.063)	p = 0.016	p < 0.001
DTI + DKI	276	0.560 ± 0.059	0.555 (95% CI: 0.467-0.641)	-0.161 (95% CI: -0.251 to -0.046)	p = 0.030	p = 0.007
IVIM-only (PF)	69	0.643 ± 0.066	0.634 (95% CI: 0.550-0.717)	-0.082 (95% CI: -0.123 to -0.017)	p = 0.128	p = 0.011
FWI-only (fw)	69	0.643 ± 0.066	0.634 (95% CI: 0.550-0.717)	-0.082 (95% CI: -0.123 to -0.017)	p = 0.128	p = 0.011
IVIM + FWI	138	0.639 ± 0.092	0.629 (95% CI: 0.544-0.714)	-0.087 (95% CI: -0.127 to -0.021)	p = 0.147	p = 0.004
Full Integrated	414	0.716 ± 0.103	0.716 (95% CI: 0.635-0.792)	Reference	--	--

Table 7B. SCZ vs. non-SCZ subtype component ablation: precision, calibration, and decision metrics. Precision-recall and posterior probability calibration across seven feature configurations under nested 5-fold cross-validation. Reporting includes pooled Average Precision (AP) with 95% bootstrap CIs, Brier calibration score, Expected Calibration Error (ECE), and sensitivity and specificity evaluated at the training-derived Youden's J optimal threshold.

Model Configuration	Average Precision (95% CI)	Brier Score	ECE	Sensitivity	Specificity
DTI-only (FA, MD)	0.548 (95% CI: 0.452-0.672)	0.2707	0.1593	0.373	0.767
DKI-only (MK, KFA)	0.521 (95% CI: 0.427-0.646)	0.2719	0.1654	0.916	0.189
DTI + DKI	0.538 (95% CI: 0.444-0.662)	0.2753	0.1656	0.313	0.822
IVIM-only (PF)	0.570 (95% CI: 0.477-0.688)	0.2360	0.1036	0.614	0.689
FWI-only (fw)	0.570 (95% CI: 0.477-0.688)	0.2360	0.1036	0.614	0.689
IVIM + FWI	0.581 (95% CI: 0.486-0.704)	0.2370	0.1318	0.590	0.733
Full Integrated	0.679 (95% CI: 0.573-0.774)	0.2245	0.0834	0.831	0.522

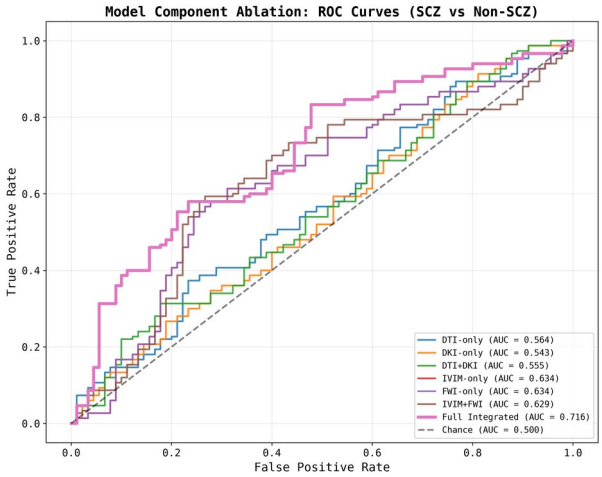


Figure 18. Multi-model ROC overlay for SCZ vs. non-SCZ ablation. ROC curves comparing the Full Integrated framework against single- and dual-compartment subsets for subtype classification.

For the subtle diagnostic task distinguishing schizophrenia spectrum from non-schizophrenia psychosis, conventional microstructural models struggled to differentiate clinical groups (Table 7A). Single-compartment DTI yielded an out-of-fold AUC-ROC of 0.564 (95% CI: 0.478 to 0.652), while DKI achieved 0.543 (95% CI: 0.460 to 0.629), both performing near chance levels. Combining the two modalities in DTI+DKI resulted in an AUC-ROC of 0.555. By contrast, models incorporating fluid compartments showed improved discriminative ability, with IVIM+FWI reaching an AUC-ROC of 0.634 (95% CI: 0.550 to 0.717). The Full Integrated framework demonstrated the highest overall discrimination, reaching a pooled AUC-ROC of 0.716 (95% CI: 0.635 to 0.792, fold mean 0.716 ± 0.103). Paired statistical testing

confirmed that the full model provided a significant advantage over conventional approaches, outperforming DTI alone by 0.152 in AUC (paired bootstrap 95% CI: 0.032 to 0.250, p = 0.011; paired fold t-test p = 0.042) and DTI+DKI by 0.161 (paired bootstrap 95% CI: 0.046 to 0.251, p = 0.007; paired fold t-test p = 0.030).

The advantage of model integration was equally pronounced across precision-recall metrics and posterior calibration (Table 7B). Average precision improved from 0.548 (95% CI: 0.452 to 0.672) in DTI-only and 0.538 (95% CI: 0.444 to 0.662) in DTI+DKI to 0.679 (95% CI: 0.573 to 0.774) in the Full Integrated framework (paired bootstrap difference ΔAP vs. DTI+DKI = +0.141, 95% CI: 0.028 to 0.245, p = 0.012). Probability calibration also improved substantially, with the Brier score declining from 0.2753 in DTI+DKI to 0.2245 in the full model, and expected calibration error dropping from 0.1656 to 0.0834. At the optimal decision threshold identified within training folds using Youden's J, the full model achieved balanced test sensitivity of 0.831 and specificity of 0.522, whereas single-compartment configurations exhibited lower sensitivity and specificity near 0.31 to 0.37.

Feature selection patterns revealed that subtype discrimination depended on a balanced contribution across distinct biophysical compartments. Selected feature sets in the full model consisted of 55.0% DTI and DKI metrics, 25.0% free-water metrics, and 20.0% IVIM perfusion metrics, with solid consistency across folds (mean pairwise Jaccard stability index = 0.68). In sequential addition analyses, performance remained near chance for baseline DTI (AUC = 0.564) and DTI+DKI (AUC = 0.555), rose once free-water metrics were added (AUC = 0.648), and peaked with the inclusion of IVIM perfusion metrics in the full model (AUC = 0.716). In contrast to the primary patient-control comparison, where tensor and kurtosis metrics predominated, these subtype results show that separating subtle psychiatric diagnostic boundaries requires the complementary biophysical information captured by fluid and perfusion modeling.

4. Discussion

4.1 Neurobiological interpretation

This study evaluated whether integrating multiple diffusion MRI modeling regimes (DTI, DKI, FWI, and IVIM) provides measurable out-of-fold predictive advantages over conventional single-compartment approaches when characterizing psychotic spectrum neuropathology. By simultaneously estimating indices of cellular organization, non-Gaussian tissue complexity, extracellular fluid expansion, and microvascular pseudo-diffusion, the framework aims to reduce parameter cross-contamination. Under strictly nested cross-validation, the regularized neural network identified a consistent pattern of cortical diffusion alterations in psychotic spectrum disorders, primarily localized to bilateral frontal and limbic cortical regions. The ablation findings indicate that while tensor and kurtosis metrics contribute substantial baseline signal, incorporating IVIM perfusion and free-water metrics provides the incremental information required to differentiate clinical subtypes. While metrics from the MK and MD compartments demonstrated the highest discriminatory power for the primary patient-versus-control task, the

secondary subtype classification relied more heavily on directional FA and KFA metrics. This suggests that while alterations in tissue complexity and overall water diffusion act as fundamental, non-invasive proxies consistent with psychotic pathology as a whole, variations in microstructural anisotropy and structural organization provide indirect indices that differentiate specific clinical presentations across the spectrum.

The pars triangularis, a component of Broca's area involved in language and grammatical processing, emerged as the most discriminative brain region for distinguishing patients from healthy controls. Exceptional discriminatory capacity in this region aligns with established literature regarding formal thought disorders and language deficits in psychosis (37), as well as localized pars triangularis structural reductions observed in individuals at clinical high risk and first-episode schizophrenia (38). Furthermore, alterations were noted in the caudal middle frontal cortex, associated with executive functions such as working memory and cognitive control. Other prominent discriminatory regions included the lateral orbitofrontal cortex, associated with emotion regulation and reward processing, and the left isthmus of the cingulate, which connects posterior spatial attention networks with anterior emotion processing systems, consistent with multi-region neuroimaging abnormalities in youth across the psychosis spectrum (39).

4.2 Methodological implications

The neural network demonstrated strong discriminative performance for the patient-versus-control task, achieving a pooled out-of-fold AUC-ROC of 0.968 (95% CI: 0.952 to 0.981) and an average precision of 0.791 (95% CI: 0.642 to 0.898). In our component ablation analysis, the Full Integrated model yielded a higher point estimate than reduced single- and dual-compartment configurations, showing paired AUC improvements between 0.039 and 0.072 (paired bootstrap $p = 0.006$, paired fold t-test $p = 0.015$; Table 6A, Table 6B). However, because reduced models such as DTI+DKI also achieved discrimination (AUC-ROC = 0.929, 95% CI: 0.909 to 0.947) and feature selection in the full model was dominated by kurtosis and tensor metrics in frontal and limbic regions, gross case-control discrimination does not strictly require the full multi-compartment feature pool. Furthermore, strong classification performance should not be conflated with mathematical proof of parameter identifiability. While out-of-sample separability confirms that the extracted features carry predictive variance, physical parameter validation requires dedicated phantom and histopathological studies.

In contrast, the secondary subtype classification task establishes a clearer empirical justification for model integration. While conventional tensor and kurtosis models yielded near-chance discrimination (AUC-ROC = 0.543 to 0.564), the Full Integrated framework demonstrated higher discriminative performance, achieving a pooled AUC-ROC of 0.716 (95% CI: 0.635 to 0.792) and an Average Precision of 0.679 (95% CI: 0.573 to 0.774). Paired comparisons confirmed significant out-of-fold improvements over DTI alone (ΔAUC = 0.152, 95% CI: 0.032 to 0.250, paired bootstrap $p = 0.011$, paired fold t-test $p = 0.042$) and DTI+DKI (ΔAUC = 0.161, 95% CI: 0.046 to 0.251, paired bootstrap $p = 0.007$, paired fold t-test $p = 0.030$; Table 7A, Table 7B). This empirical gain indicates that

when macrostructural axonal injury is broadly shared across the clinical spectrum, incorporating fluid and perfusion metrics provides complementary out-of-fold information that assists in resolving subtle subtype boundaries. By forcing feature ranking routines to run strictly out-of-fold and verifying cross-scanner generalizability via Leave-One-Site-Out validations, the technical validity of the predictive framework was systematically designed to mitigate out-of-fold data leakage pathways. However, the framework's operational scope must be framed accurately: it demonstrates higher-dimensional feature extraction rather than complete multi-disease diagnostic specificity (35). Because it cannot currently rule out whether the identified microstructural profiles are partially shared with other non-psychotic neuroinflammatory conditions, it is not an unguided diagnostic screening tool. Rather than serving as an immediate clinical decision tool, the framework functions as an objective microstructural profiling method to quantify continuous cellular alterations within clinically diagnosed cohorts.

4.3 Limitations

Several limitations must be addressed. First, the 6.3:1 class imbalance between healthy controls and patients in the sample may limit the model's generalizability in clinical settings where psychosis prevalence is even lower. Next, the modest performance in subtype classification (AUC = 0.716) reflects the significant neurobiological overlap between schizophrenia and other psychotic disorders, supporting dimensional rather than strictly categorical conceptualizations of psychosis. As the data is cross-sectional, disease progression or the long-term effects of antipsychotic treatment cannot be correlated with microstructural changes. This tool aims to supplement a clinician, rather than replace.

A fundamental limitation of the case-control analysis is the structural entanglement between study cohort and diagnostic status. Healthy controls were obtained primarily from HCP Young Adult (scanned at Washington University), while clinical patients were recruited through HCP Early Psychosis (scanned across Indiana University, McLean Hospital, and Beth Israel Deaconess Medical Center). Although ComBat harmonization adjusts for measurable site-specific additive and multiplicative batch effects, ComBat cannot fully disentangle cohort-level technical variation from disease biology when diagnostic status and study cohort are collinear. While our non-imaging demographic baseline classifier demonstrated that age, sex, and site alone achieve lower discrimination (AUC = 0.684) than the full diffusion model (AUC = 0.968), definitive confirmation that the classifier detects psychosis neuropathology rather than subtle cross-study acquisition differences requires validation in datasets containing same-cohort, locally scanned healthy controls.

Second, the study does not demonstrate clinical utility in terms of informing clinical decision-making. We have not shown that the derived microstructural metrics predict therapeutic response, guide antipsychotic selection, forecast clinical prognosis, identify actionable clinical subgroups, or improve patient management relative to standard psychiatric examination. Because the dataset is cross-sectional, the temporal stability of these features and their

relationship to disease progression or medication exposure cannot be determined.

Third, although our component ablation analysis demonstrates that multi-compartment integration yields higher out-of-fold point estimates than simpler subsets, particularly for subtype discrimination, these empirical gains reflect statistical associations within a fixed 20-feature budget rather than direct proof of physical parameter recovery *in vivo*. Finally, true cross-scanner generalization was evaluated via Leave-One-Site-Out cross-validation across three clinical sites within the HCP-EP protocol; evaluation on fully independent, non-HCP clinical populations using standard hospital MRI protocols represents a critical prerequisite before any prospective translational role can be contemplated.

4.4 Future direction

Future research should focus on longitudinal validation to determine if these microstructural signatures can predict treatment response or clinical outcomes over time. Integration with other data types, such as genomic risk scores or functional connectivity, may provide a more comprehensive characterization of the psychotic spectrum. Future machine learning pipelines should transition away from optimizing standalone binary classifiers toward discovering multimodal imaging fingerprints. Rather than selecting features purely based on individual performance, advanced deep learning networks could identify multi-compartment feature combinations that may result in higher diagnostic specificity across separate psychiatric conditions. Additionally, using 7T MRI could provide the resolution needed to examine even finer microstructural changes. Implementing this pipeline into clinical workflows could provide psychiatrists with objective data alongside symptom-based assessments, particularly during early, ambiguous stages of the disorder. While nested cross-validation, ComBat scanner harmonization, comprehensive threshold sweeps, and multi-panel error curves confirmed a calibrated classification framework, its diagnostic utility remains coupled to initial clinical triaging. The multi-compartment approach successfully maps continuous structural variations across the cortex, providing an analytical toolset to complement, rather than replace, established observational psychiatry.

5. Conclusion

The integrated diffusion MRI framework developed in this study provides an objective, biologically based approach to classification for psychotic spectrum disorders, which have typically relied on observational symptom assessments. By simultaneously modeling perfusion, extracellular water, and tissue complexity, the IVIM-FWI-DKI pipeline effectively minimizes parameter cross-contamination that often arises in traditional single-compartment analyses. The resulting neural network classification achieved high accuracy in differentiating patients from healthy controls, with discriminatory power driven primarily by Mean Kurtosis and Mean Diffusivity metrics localized to the bilateral pars triangularis and caudal middle frontal cortex. These findings reinforce the status of language processing and executive control hubs as core neurobiological markers of psychotic pathology. The more modest performance observed in subtype classification reflects the substantial neurobiological overlap between

schizophrenia spectrum and non-schizophrenic psychosis. This overlap supports a dimensional view of psychosis as a continuous spectrum rather than a set of strictly separate biological entities. Through a systematic component ablation analysis under nested cross-validation, this study demonstrates that combining DTI, DKI, FWI, and IVIM modalities provides measurable incremental predictive performance over constituent single- and dual-compartment models, with the most pronounced gains observed in the challenging task of differentiating clinical subtypes within the psychosis spectrum. However, for primary case-control discrimination, conventional tissue kurtosis and tensor metrics contribute the predominant discriminatory signal, and cross-cohort acquisition differences between HCP sub-studies remain partially entangled with diagnostic status. Critically, this study demonstrates out-of-fold biological separability rather than clinical decision utility. Prospective validation in independent, non-HCP clinical cohorts containing same-site controls, paired with longitudinal assessment of treatment response, will be essential to establish the translational value of integrated multi-compartment diffusion imaging in psychiatry.

Acknowledgments

We would like to thank Dr. Faye McKenna, who aided in initially developing the research idea. Special thanks also go to the staff of the Lumiere Writing Center for their assistance in editing and revising this paper.

Data availability

The neuroimaging and clinical data analyzed in this study were provided by the HCP. Healthy control subjects were sourced from the HCP-YA cohort, and subjects diagnosed with psychotic spectrum disorders were obtained from the HCP-EP cohort. These datasets are available to qualified researchers through the ConnectomeDB platform upon meeting specific data use requirements. All diffusion microstructure modeling, parameter estimations, and signal simulations were implemented using the open-source DMIPY toolbox. The complete Python implementation code for this paper is publicly available in the GitHub repository: https://github.com/arpun/psychosis-dmri-framework

Data were provided [in part] by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

Research using Human Connectome Project for Early Psychosis (HCP-EP) data reported in this publication was supported by the National Institute of Mental Health of the National Institutes of Health under Award Number U01MH109977. The HCP-EP 1.1 Release data used in this report came from DOI: 10.15154/1522899.

6. References

1. Calzavara-Pinton, I., Deste, G., Barlati, S., Nibbio, G., & Vita, A. (2024). Treatment of cognitive impairment associated with schizophrenia spectrum disorders: New evidence, challenges, and future perspectives. *Brain Sciences, 14*(8), 791. https://doi.org/10.3390/brainsci14080791

2. Krauss, A. (2022). Challenges and considerations in treating negative and cognitive symptoms of schizophrenia spectrum disorders. *Federal Practitioner, 39*(11), S10-S15. https://doi.org/10.12788/fp.0338

3. Zhu, Y., Li, C., Huhn, M., Rothe, P., Krause, M., Bighelli, I., Schneider-Thoma, J., & Leucht, S. (2017). How well do patients with a first episode of schizophrenia respond to antipsychotics: A systematic review and meta-analysis. *European Neuropsychopharmacology, 27*(9), 835-844. https://doi.org/10.1016/j.euroneuro.2017.06.011

4. Keshavan, M. S., Tandon, R., Boutros, N. N., & Nasrallah, H. A. (2008). Schizophrenia, "just the facts": What we know in 2008. Part 3: Neurobiology. Schizophrenia Research, 106(2–3), 89–107. https://doi.org/10.1016/j.schres.2008.07.020

5. Masdeu, J. C. (2011). Neuroimaging in psychiatric disorders. *Neurotherapeutics, 8*(1), 93-102. https://doi.org/10.1007/s13311-010-0006-0

6. Pierpaoli, C., Jezzard, P., Basser, P. J., Barnett, A., & Di Chiro, G. (1996). Diffusion tensor MR imaging of the human brain. *Radiology, 201*(3), 637-648. https://doi.org/10.1148/radiology.201.3.8939209

7. Le Bihan, D., Breton, E., Lallemand, D., Aubin, M. L., Vignaud, J., & Laval-Jeantet, M. (1988). Separation of diffusion and perfusion in intravoxel incoherent motion MR imaging. *Radiology, 168*(2), 497-505. https://doi.org/10.1148/radiology.168.2.3393671

8. Kamagata, K., Andica, C., Kato, A., Saito, Y., Uchida, W., Hatano, T., Lukies, M., Ogawa, T., Takeshige-Amano, H., Akashi, T., Hagiwara, A., Fujita, S., & Aoki, S. (2021). Diffusion magnetic resonance imaging-based biomarkers for neurodegenerative diseases. *International Journal of Molecular Sciences, 22*(10), 5216. https://doi.org/10.3390/ijms22105216

9. Pavan, T., Alemán-Gómez, Y., Jenni, R., Steullet, P., Schilliger, Z., Dwir, D., Cleusix, M., Alameda, L., Do, K. Q., Conus, P., Hagmann, P., Klauser, P., & Jelescu, I. (2025). White matter microstructure alterations in early psychosis and schizophrenia. Translational Psychiatry, 15, Article 179. https://doi.org/10.1038/s41398-025-03397-1

10. Alexander, A. L., Lee, J. E., Lazar, M., & Field, A. S. (2007). Diffusion tensor imaging of the brain. *Neurotherapeutics, 4*(3), 316-329. https://doi.org/10.1016/j.nurt.2007.05.011

11. Nakamura, K., Kawasaki, Y., Takahashi, T., Furuichi, A., Noguchi, K., Seto, H., & Suzuki, M. (2012). Reduced white matter fractional anisotropy and clinical symptoms in schizophrenia: A voxel-based diffusion tensor imaging study. *Psychiatry Research: Neuroimaging, 202*(3), 233-238. https://doi.org/10.1016/j.pscychresns.2011.09.006

12. Lahti, A. C., Reid, M. A., White, D. M., & Kraguljac, N. V. (2018). 25.2 Understanding white matter pathology in schizophrenia using diffusion tensor imaging and magnetic resonance spectroscopy. *Schizophrenia Bulletin, 44*(Suppl 1), S41. https://doi.org/10.1093/schbul/sby014.102

13. Lee, K., Yoshida, T., Kubicki, M., Bouix, S., Westin, C.-F., Kindlmann, G., Niznikiewicz, M., Cohen, A., McCarley, R. W., & Shenton, M. E. (2009). Increased diffusivity in superior temporal gyrus in patients with schizophrenia: A diffusion tensor imaging study. Schizophrenia Research, 108(1–3), 33–40. https://doi.org/10.1016/j.schres.2008.11.024

14. McKenna, F., Gupta, P. K., Sui, Y. V., Bertisch, H., Gonen, O., Goff, D. C., & Lazar, M. (2023). Microstructural and microvascular alterations in psychotic spectrum disorders: A three-compartment intravoxel incoherent imaging and free water model. *Schizophrenia Bulletin, 49*(6), 1537-1547. https://doi.org/10.1093/schbul/sbad019

15. Vieni, C., Ades-Aron, B., Conti, B., Sigmund, E. E., Riviello, P., Shepherd, T. M., Lui, Y. W., Novikov, D. S., & Fieremans, E. (2020). Effect of intravoxel incoherent motion on diffusion parameters in normal brain. *NeuroImage, 204*, 116228. https://doi.org/10.1016/j.neuroimage.2019.116228

16. Bergé, D., Mané, A., Lesh, T. A., Bioque, M., Barcones, F., Gonzalez-Pinto, A. M., Parellada, M., Vieta, E., Castro-Fornieles, J., Rodriguez-Jimenez, R., García-Portilla, M. P., Usall, J., Carter, C. S., Cabrera, B., Bernardo, M., Janssen, J., Mezquida, G., Amoretti, S., Pina-Camacho, L., & Arango, C. (2020). Elevated extracellular free-water in a multicentric first-episode psychosis sample, decrease during the first 2 years of illness. *Schizophrenia Bulletin, 46*(4), 936-945. https://doi.org/10.1093/schbul/sbz132

17. Carreira Figueiredo, I., Pasternak, O., Jones, D. K., & Di Biase, M. A. (2022). White-matter free-water diffusion MRI in schizophrenia: A systematic review and meta-analysis. *Neuropsychopharmacology, 47*(7), 1413-1420. https://doi.org/10.1038/s41386-022-01272-x

18. McKenna, F. F., Miles, L., Babb, J. S., Goff, D. C., & Lazar, M. (2019). Diffusion kurtosis imaging of gray matter in schizophrenia. *Cortex, 121*, 201-224. https://doi.org/10.1016/j.cortex.2019.08.013

19. Al-Sharif, N. B., Zavaliangos-Petropulu, A., & Narr, K. L. (2024). A review of diffusion MRI in mood disorders: Mechanisms and predictors of treatment response. *Neuropsychopharmacology, 50*(1), 211-229. https://doi.org/10.1038/s41386-024-01894-3

20. Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., Ugurbil, K., & WU-Minn HCP Consortium. (2013). The WU-Minn Human Connectome Project: An overview. *NeuroImage, 80*, 62-79. https://doi.org/10.1016/j.neuroimage.2013.05.041

21. Sotiropoulos, S. N., Jbabdi, S., Xu, J., Andersson, J. L., Moeller, S., Auerbach, E. J., Glasser, M. F., Hernandez, M., Sapiro, G., Jenkinson, M., Feinberg, D. A., Yacoub, E., Lenglet, C., Van Essen, D. C., Ugurbil, K., & Behrens, T. E. J. (2013). Advances in diffusion MRI acquisition and processing in the Human Connectome Project. NeuroImage, 80, 125–143. https://doi.org/10.1016/j.neuroimage.2013.05.057

22. Human Connectome Project. (2017). HCP S1200 release reference manual. https://www.humanconnectome.org/storage/app/media/documentation/s1200/HCP_S1200_Release_Reference_Manual.pdf

23. Human Connectome Project. (2021). HCP Early Psychosis Release 1.1 reference manual. https://www.humanconnectome.org/storage/app/media/documentation/HCP-EP1.1/HCP-EP_Release_1.1_Manual.pdf

24. Yamada, Y., Matsumoto, M., Iijima, K., & Sumiyoshi, T. (2020). Specificity and continuity of schizophrenia and bipolar disorder: Relation to biomarkers. *Current Pharmaceutical Design, 26*(2), 191-200. https://doi.org/10.2174/1381612825666191216153508

25. Szafer, A., Zhong, J., & Gore, J. C. (1995). Theoretical model for water diffusion in tissues. *Magnetic Resonance in Medicine, 33*(5), 697-712. https://doi.org/10.1002/mrm.1910330516

26. Zhang, H., Schneider, T., Wheeler-Kingshott, C. A., & Alexander, D. C. (2012). NODDI: Practical in vivo neurite orientation dispersion and density imaging of the human brain. *NeuroImage, 61*(4), 1000-1016. https://doi.org/10.1016/j.neuroimage.2012.03.072

27. Rydhög, A. S., Szczepankiewicz, F., Wirestam, R., Ahlgren, A., Westin, C.-F., Knutsson, L., & Pasternak, O. (2017). Separating blood and water: Perfusion and free water elimination from diffusion MRI in the human brain. *NeuroImage, 156*, 423-434.

https://doi.org/10.1016/j.neuroimage.2017.04.023

28. Fick, R. H. J., Wassermann, D., & Deriche, R. (2019). The dmipy toolbox: Diffusion MRI multi-compartment modeling and microstructure recovery made easy. *Frontiers in Neuroinformatics, 13*, 64. https://doi.org/10.3389/fninf.2019.00064

29. Cohen, A. D., Schieke, M. C., Hohenwalter, M. D., & Schmainda, K. M. (2015). The effect of low b-values on the intravoxel incoherent motion (IVIM) derived pseudodiffusion parameter in liver. Magnetic Resonance in Medicine, 73(1), 306–311. https://doi.org/10.1002/mrm.25109

30. Ohno, N., Miyati, T., Kobayashi, S., & Gabata, T. (2016). Modified triexponential analysis of intravoxel incoherent motion for brain perfusion and diffusion. *Journal of Magnetic Resonance Imaging, 43*(4), 818-823. https://doi.org/10.1002/jmri.25230

31. Jensen, J. H., Helpern, J. A., Ramani, A., Lu, H., & Kaczynski, K. (2005). Diffusional kurtosis imaging: The quantification of non-Gaussian water diffusion by means of magnetic resonance imaging. *Magnetic Resonance in Medicine, 53*(6), 1432-1440. https://doi.org/10.1002/mrm.20508

32. Collier, Q., Veraart, J., Jeurissen, B., Vanhevel, F., Pullens, P., Parizel, P. M., den Dekker, A. J., & Sijbers, J. (2018). Diffusion kurtosis imaging with free water elimination: A Bayesian estimation approach. *Magnetic Resonance in Medicine, 80*(2), 802-813. https://doi.org/10.1002/mrm.27075

33. Ades-Aron, B., Veraart, J., Kochunov, P., McGuire, S., Sherman, P., Kellner, E., Novikov, D. S., & Fieremans, E. (2018). Evaluation of the accuracy and precision of the diffusion parameter EStImation with Gibbs and NoisE removal pipeline. *NeuroImage, 183*, 532-543. https://doi.org/10.1016/j.neuroimage.2018.07.066

34. Nicolas, R., Sibon, I., & Hiba, B. (2015). Accuracies and contrasts of models of the diffusion-weighted-dependent attenuation of the MRI signal at intermediate b-values. *Magnetic Resonance Insights, 8*, 17-28. https://doi.org/10.4137/MRI.S25301

35. Pasternak, O., Kelly, S., Sydnor, V. J., & Shenton, M. E. (2018). Advances in microstructural diffusion neuroimaging for psychiatric disorders. NeuroImage, 182, 259–282. https://doi.org/10.1016/j.neuroimage.2018.04.051

36. Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral

based regions of interest. *NeuroImage, 31*(3), 968-980. https://doi.org/10.1016/j.neuroimage.2006.01.021

37. Ehlen, F., Montag, C., Leopold, K., & Heinz, A. (2023). Linguistic findings in persons with schizophrenia: A review of the current literature. *Frontiers in Psychology, 14*, 1287706. https://doi.org/10.3389/fpsyg.2023.1287706

38. Iwashiro, N., Suga, M., Takano, Y., Inoue, H., Natsubori, T., Satomura, Y., Koike, S., Yahata, N., Murakami, M., Katsura, M., Gonoi, W., Sasaki, H., Takao, H., Abe, O., Kasai, K., & Yamasue, H. (2012). Localized gray matter volume reductions in the pars triangularis of the inferior frontal gyrus in individuals at clinical high-risk for psychosis and first episode for schizophrenia. *Schizophrenia Research, 137*(1-3), 124-131. https://doi.org/10.1016/j.schres.2012.02.024

39. Wolf, D. H., Satterthwaite, T. D., Calkins, M. E., Ruparel, K., Elliott, M. A., Hopson, R. D., Jackson, C. T., Prabhakaran, K., Bilker, W. B., Hakonarson, H., Gur, R. C., & Gur, R. E. (2015). Functional neuroimaging abnormalities in youth with psychosis spectrum symptoms. *JAMA Psychiatry, 72*(5), 456-465. https://doi.org/10.1001/jamapsychiatry.2014.3169