Asian Journal of Psychiatry 114 (2025) 104759

Contents lists available at ScienceDirect

Asian Journal of Psychiatry

journal homepage: www.elsevier.com/locate/ajp

Comparative diagnostic performance of artificial intelligence models in structural MRI for schizophrenia: A systematic review and meta-analysis

Martin Kotochinsky [a], Pandora Eloa Oliveira Fonseca [b], Veronica Ramirez Lopera [c], Laura Mora [d], Wellgner Fernandes Oliveira Amador [e], Eduardo Cesar Teixeira Sirena [f], Felipe Bandeira de Melo Guimarães [g], Delfina Lahitou Herlyn [h,i], Nima Norbu Sherpa [j,k,*], Andrea Gonzalez Lezana [l], Thales Pardini Fagundes [m]

[a] Department of Medicine, National University of Cuyo, Argentina
[b] Department of Medicine, Federal University of Campina Grande, Brazil
[c] Department of Medicine, Metropolitan University of Barranquilla, Colombia
[d] Department of Medicine, Metropolitan University of Barranquilla, Colombia
[e] Department of Medicine, Federal University of Paraíba, Brazil
[f] Department of Medicine, University of Fortaleza, Brazil
[g] Institute of Psychiatry, Federal University of Rio de Janeiro, Brazil
[h] Research Group on Neurosciences Applied to Behavioral Disorders (INAAC Group), Institute of Neurosciences FLENI–CONICET (INEU), Argentina
[i] National Scientific and Technical Research Council (CONICET), Argentina
[j] Institute of Psychiatry, Psychology & Neuroscience, King's College London, United Kingdom
[k] Brain Research & Imaging Centre, University of Plymouth, United Kingdom
[l] Institute of Clinical Research Mar del Plata, Argentina
[m] Department of Head and Neck Surgery, Barretos Cancer Hospital, Brazil

ARTICLE INFO

Keywords:
Artificial Intelligence
Schizophrenia
Neuroimaging
Diagnostic Accuracy

ABSTRACT

Introduction: Timely diagnosis of schizophrenia is essential to ensure prompt treatment initiation and adherence. Structural magnetic resonance imaging (sMRI), when combined with artificial intelligence (AI), offers a promising avenue to enhance diagnostic accuracy. However, its performance and clinical use is a matter of debate.
Methods: PubMed, Embase, and Cochrane databases were searched for studies using AI models with sMRI to diagnose schizophrenia in adults. Eligible models encompass traditional machine learning methods and deep learning (DL) architectures, utilizing diverse neuroanatomical inputs, including gray matter (GM) features and whole-brain (WB) structural data. The outcomes of interest were diagnostic performance metrics as: sensitivity (SE), specificity (SP), area under the curve (AUC).
Results: A total of 16 studies were included, comprising 3601 participants. Overall pooled SE and SP were 0.76 (95 % CI: 0.71–0.80) and 0.78 (95 % CI: 0.73–0.82), respectively. When compared, DL models outperformed Support Vector Machine (SVM), achieving higher SP of 0.83 (95 % CI: 0.80–0.86) vs. 0.78 (95 % CI: 0.72–0.83), and AUC of 0.892 (95 % CI: 0.81–0.90) vs. 0.782 (95 % CI: 0.70–0.82). WB input models also outperformed GM performance, with SP of 0.86 (95 % CI: 0.78–0.92) vs. 0.80 (95 % CI: 0.73–0.85), and AUC of 0.89 (95 % CI: 0.70–0.93) vs. 0.816 (95 % CI: 0.71–0.84).
Conclusion: AI models using sMRI show promising but provisional diagnostic performance for schizophrenia. Across studies, DL architectures and WB inputs generally achieved higher specificity and AUC than SVM and GM features. Prospective, multi-site external validation cohorts are needed before routine clinical implementation.

Abbreviations: AI, Artificial Intelligence; AUC, Area Under the Curve; CI, Confidence Interval; DL, Deep Learning; FMRI, Functional Magnetic Resonance Imaging; GM, Gray Matter; ML, Machine Learning; MRI, Magnetic Resonance Imaging; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses; QUADAS-2, Quality Assessment of Diagnostic Accuracy Studies, version 2; SE, Sensitivity; SP, Specificity; SROC, Summary Receiver Operating Characteristic; SMRI, Structural Magnetic Resonance Imaging; SVM, Support Vector Machine; WB, Whole Brain.

* Corresponding author at: Institute of Psychiatry, Psychology & Neuroscience, King's College London, United Kingdom.
E-mail address: nima.sherpa@kcl.ac.uk (N.N. Sherpa).

https://doi.org/10.1016/j.ajp.2025.104759
Received 30 July 2025; Received in revised form 26 October 2025; Accepted 27 October 2025
Available online 4 November 2025

1. Introduction

Schizophrenia is a complex neuropsychiatric disorder that remains challenging to diagnose in some instances, particularly during early psychosis, due to its heterogeneous clinical features and symptom overlap with other psychiatric or neurodegenerative conditions (Kahn et al., 2015). This diagnostic uncertainty has driven the need for tools that offer objective, reproducible assessments, especially during the early stages of illness. Schizophrenia is a major contributor to global psychiatric disability, with high disability-adjusted life year rates and substantial societal costs from chronic disability and healthcare dependence (Charlson et al., 2018). Improving diagnostic accuracy is essential to support early intervention and reduce both personal and economic burden.

Structural magnetic resonance imaging (sMRI) has emerged as a critical tool that leads to the identification of biomarkers, and machine learning (ML) and deep learning (DL) models offer promising avenues for assistance on clinical diagnosis (Saha et al., 2024, Brugger & Howes, 2017). Compared to other imaging modalities, sMRI offers superior anatomical resolution, lower acquisition cost, and greater feasibility for routine implementation in clinical psychiatry, making it particularly suitable for diagnostic decision support and translational research.

Currently, up to 30 % of individuals with schizophrenia experience diagnostic delays or misclassification (Bradford et al., 2024). Long durations of untreated psychosis, often exceeding one year, are associated with poorer long-term outcomes, including increased relapse risk and reduced functional recovery (Lieberman & Fenton, 2000). Artificial Intelligence-driven (AI) sMRI offers an objective solution to address this limitation by detecting subtle neuroanatomical signatures that may appear with early clinical symptoms (Omlor et al., 2025, van Erp et al., 2018). Recent studies suggest traditional ML models can distinguish schizophrenia from bipolar disorder with 80 % accuracy using cortical thickness patterns, even outperforming clinical assessments in ambiguous cases (Madre et al., 2020). This capability positions sMRI-based AI tools as one potential adjuvant for diagnostic uncertainty, reducing reliance on symptomatic criteria alone.

Prior meta-analyses addressing this topic have focused on multimodal neuroimaging approaches and tried to pool together heterogeneous methodologies as sMRI, functional MRI (fMRI), resting-state functional MRI, task functional MRI, and multimodal MRI. Such strategy reduces the clinical utility of findings. It obstructs efforts to isolate sMRI-specific indicators of enduring neuroanatomical alterations that are minimally influenced by changes in clinical status or external factors (Di Camillo et al., 2024, Kambeitz et al., 2015).

Therefore this systematic review and diagnostic meta-analysis aims to assess the diagnostic performance of AI Models in sMRI for Schizophrenia. By delineating optimal practices for sMRI-based AI diagnostics, our goal is to advance translational applications in precision psychiatry while addressing reproducibility challenges in neuroimaging research.

2. Methods

This systematic review and diagnostic meta-analysis was conducted according to the Cochrane Handbook for Systematic Reviews and Interventions and is reported following the updated Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guideline (Page et al., 2021). The study protocol was registered in the International Prospective Register of Systematic Reviews in July 2025 (PROSPERO 2025 CRD420251087015).

2.1. Study selection and eligibility criteria

A systematic search was conducted in different databases, including PubMed, Embase, and Cochrane electronic databases, through April 2025. The primary outcomes of analysis included diagnostic performance parameters such as sensitivity (SE), specificity (SP), and area under the curve (AUC).

The search strategy included the following terms: (schizophrenia OR schizophrenic OR psychosis) AND (“machine learning" OR "deep learning" OR "artificial intelligence" OR AI OR ML OR DL) AND ("structural MRI" OR "sMRI" OR "magnetic resonance imaging") AND ("diagnosis" OR "diagnostic accuracy" OR "sensitivity" OR "specificity" OR "area under the curve" OR "AUC" OR "classification"). The Boolean search operators "AND" and "OR" were employed to link search terms. Additionally, we manually screened the reference lists of all included studies, prior systematic reviews, and meta-analyses to identify potentially relevant additional publications.

To be included, studies had to meet the following criteria: (1) include adult patients diagnosed with schizophrenia; (2) applied sMRI data as input for AI-based diagnosis; (3) reported metrics of diagnostic performance. The exclusion criteria were: (1) studies involving only functional or diffusion imaging without sMRI; (2) different study designs such as reviews, editorials, case reports, or methodological papers without model evaluation; and (3) incomplete preprocessing and model analysis description.

2.2. Study triage and data extraction

Two researchers (L.M. and P.E.O.F.) independently screened articles for inclusion criteria and extracted data from full texts and published appendices of included studies. Data extraction accuracy was independently cross-verified by each investigator. Any disagreements were resolved through consensus or, if necessary, by a third author (T.P.F.). The authors extracted baseline characteristics of studies and the prespecified outcomes.

2.3. Data analysis and synthesis

Pooled SE and SP for included studies with a 95 % confidence interval (CI) were obtained using a random-effects model. Forest plots were constructed to visually represent individual study estimates and the overall pooled effects. Heterogeneity was assessed using the inconsistency index (I^2) and the between-study variance (tau^2). In this study, traditional machine learning refers mainly to algorithms such as support vector machines (SVM), which are compared against deep learning architectures.

Summary receiver-operating characteristic (SROC) curves were generated using a bivariate random-effects model to assess overall diagnostic performance across varying thresholds. The area under the curve (AUC) and its 95 % confidence interval (CI) were calculated using a non-parametric bootstrap approach with 2000 resamples, implemented via the AUC_boot function from the dmetatools package in R 2025.05.1 (Noma, 2023). The 2.5th and 97.5th percentiles of the bootstrap distribution were used to define the lower and upper bounds of the 95 % CI, respectively.

Given expected heterogeneity between diagnostic accuracy studies, subgroup analyses were conducted based on algorithm type (DL and SVM) and input features (GM and WB). Leave-one-out analyses were also performed to assess the influence of individual studies on the pooled effect estimates.

2.4. Quality assessment

Risk of bias and quality assessment of individual studies were analyzed based on a modified Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS2) (Whiting et al., 2011). Two researchers (P.E.O.F. and F.B.M.G.) and all disagreements were resolved by the senior author’s decision (T.P.F.). The quality control instrument used consists of four domains: patient selection, index test, reference standard, and flow/timing. Each domain is evaluated for risk of bias, and the first three are also assessed for concerns regarding applicability to the review question. The overall risk of bias will be defined as: high, low or unclear.

3. Results

3.1. Study selection and characteristics

As detailed in Fig. 1, a total of 1266 studies were identified. After the removal of duplicate reports and non-relevant studies by title and abstract review, 35 studies were reviewed for full-text assessment. These were thoroughly evaluated to satisfy the inclusion criteria. A total of 16 studies, comprising 3601 patients, were included. A summary of the characteristics of the included studies is found in Table 1.

3.2. Pooled analysis of all studies

Figs. 2 and 3 present, respectively, the forest plots for the SE and SP with the appropriate 95 % CI. The pooled SE and SP were 0.76 (95 % CI: 0.71–0.80) and 0.78 (95 % CI: 0.73–0.82), respectively. There was statistically significant heterogeneity for both SE ($Tau^2 = 0.23$, $I^2 = 80.9$ %, $p < 0.01$) and SP ($Tau^2 = 0.25$, $I^2 = 81.0$ %, $p < 0.01$). The pooled SROC curve with the bivariate approach yielded an AUC of 0.82 (95 % CI: 0.77–0.85) (Fig. 4). A summary of all results can be found in Supplement 1.

3.3. Subgroup analysis: diagnostic performance by processing algorithms

The subgroup analysis for the SVM processing model yielded a pooled SE and SP of 0.71 (95 % CI: 0.67–0.75) and 0.78 (95 % CI: 0.72–0.83), respectively. There was statistically significant heterogeneity for both SE ($Tau^2 = 0.09$, $I^2 = 61.6$ %, $p < 0.01$) and SP ($Tau^2 = 0.33$, $I^2 = 79.6$ %, $p < 0.01$). The pooled SROC curve using the bivariate model showed an AUC of 0.78 (95 % CI: 0.70–0.82)(Supplement 2).

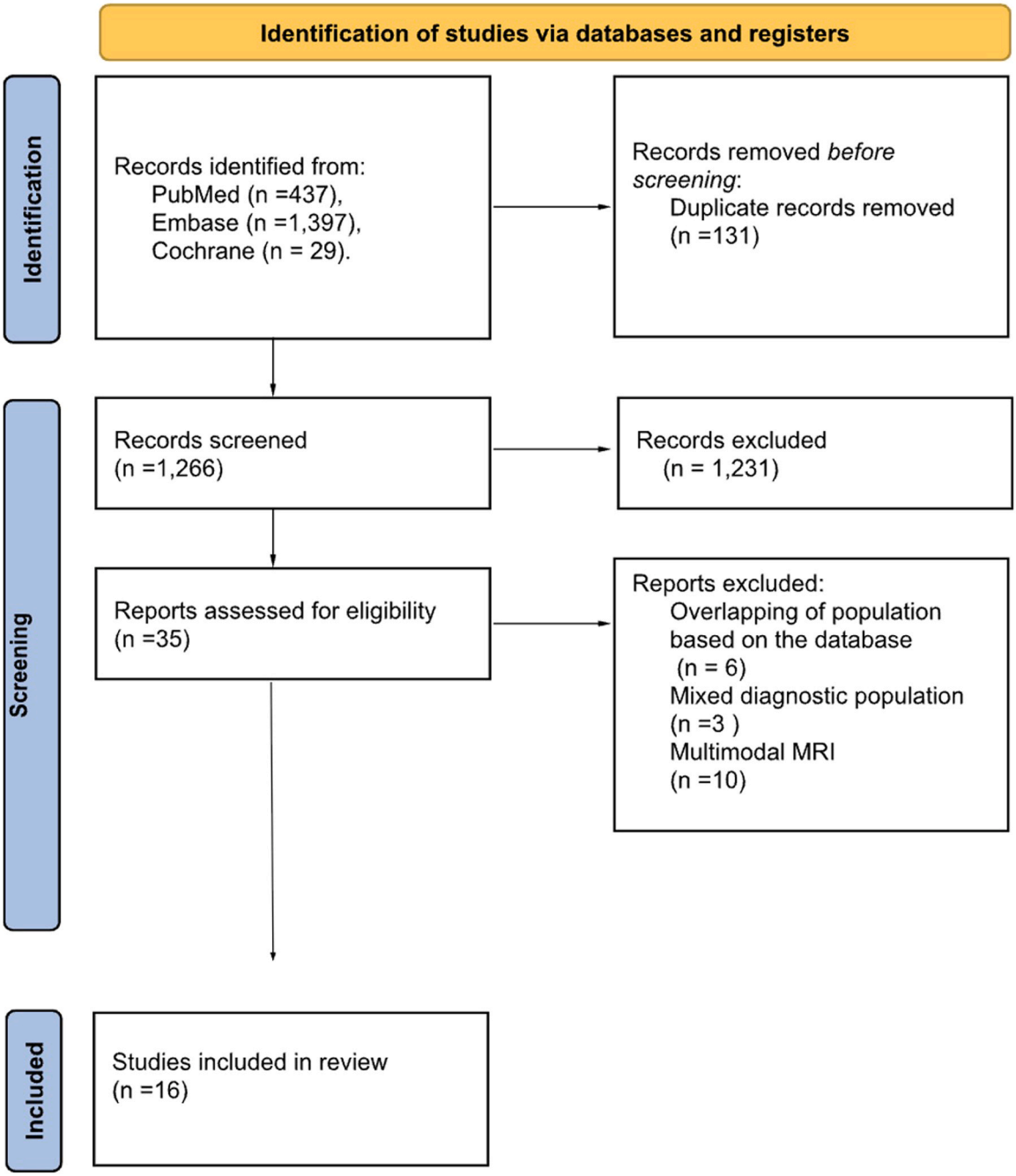

Fig. 1. PRISMA Flow diagram for the identification of studies via databases and registers.

Table 1
Baseline characteristics of included studies.

Author	Year	Country	Sample size		Female (%)		Age, Mean (SD)		MRI Dataset Name	Processing Method	Model Input
			SCZ	HC	SCZ	HC	SCZ	HC			
Bang et al.	2024	Republic of Korea	174	162	64.36 %	53.70 %	35.6 (11.5)	36.7 (10.2)	Bundang Medical Center	Light Gradient Boosting Machine (LightGBM) classifier with mutual information feature selection and SMOTE-ENN for imbalance handling, applied to radiomic features extracted via PyRadiomics, and interpreted the model using SHAP.	T1-weighted sMRI scans of the 12 cerebellar subregions (both hemispheres)
Chen et al. (GM)	2020	China	34	34	79.41 %	67.64 %	36.8 (10.9)	39.5 (10.6)	COBRE	2 T + PCA + SVM (Two-Tier Principal Component Analysis with Support Vector Machine)	MRI was segmented into three tissue probability maps (TPMs), including GM, WM, and cerebrospinal fluid (CSF). Third, the tissue volume was obtained by modulating the segmented tissue maps.
Chen et al. (WM)	2020	China	34	34	82.35 %	70.58 %	36.8 (10.9)	39.5 (10.6)	COBRE	2 T + PCA + SVM (Two-Tier Principal Component Analysis with Support Vector Machine)	MRI was segmented into three tissue probability maps (TPMs), including GM, WM, and cerebrospinal fluid (CSF). Third, the tissue volume was obtained by modulating the segmented tissue maps.
Chilla et al.	2022	Singapore	53	110	n/a	n/a	32.9 (9.4)		Local	Function kernel, Linear SVC, Nu-SVC)	Cortical and subcortical volume, cortical surface area, cortical mean curvature and cortical thickness
Cui et al.	2022	China	662	613	45.31 %	48.45 %	27.5 (6.5)	27.5 (6.1)	Peking University Six Hospital	DNN	Voxel-based volumetric features extracted from the normalised and modulated grey matter,
Davatziko et al.	2005	USA	69	79	33.33 %	48.10 %	29.9 (8.4)	28.2 (7.5)	Schizophrenia Research Center, University of Pennsylvania Medical Center (OWN)	They used a nonlinear multivariate classifier with leave-one-out validation on whole-brain RAVENS maps to distinguish schizophrenia from controls.	Voxel-based representations of brain sMRI
de Pierrefeu et al. (GM)	2018	France	43	90	55.81 %	54.44 %	34.5 (12.0)	32.4 (12.5)	PRAGUE	SVM	Grey matter VBM
de Pierrefeu et al. (GM)	2018	France	43	90	55.81 %	54.44 %	34.5 (12.0)	32.4 (12.5)	PRAGUE	SVM	Regions of interest based analysis (the volume of subcortical regions and the average thickness of cortical parcels.)
Dluhoš et al.	2017	Czech Republic	258	222	23.64 %	28.82 %	26.0 (6.4)	26.5 (6.3)	5 local sites	Geometric average of the SVM weights of local models	Graymatter (GM) and whitematter (WM) tissue segments were modulated with the determinant of Jacobian matrices of the deformations (DETJ)
Dluhoš et al. (GM)	2017	Czech Republic	258	222	23.64 %	28.82 %	26.0 (6.4)	26.5 (6.3)	5 local sites	Geometric average of the SVM weights of local models	Graymatter (GM) and whitematter (WM) tissue segments were modulated with the determinant of Jacobian

(*continued on next page*)

Table 1 (*continued*)

Author	Year	Country	Sample size		Female (%)		Age, Mean (SD)		MRI Dataset Name	Processing Method	Model Input
			SCZ	HC	SCZ	HC	SCZ	HC			
											matrices of the deformations (DETJ)
Dluhoš et al. (WM)	2017	Czech Republic	258	222	23.64 %	28.82 %	26.0 (6.4)	26.5 (6.3)	6 local sites	Geometric average of the SVM weights of local models	Graymatter (GM) and whitematter (WM) tissue segments were modulated with the determinant of Jacobian matrices of the deformations (DETJ)
Gould et al.	2016	China	98	83	18.36 %	15.66 %	24.3 (5.3)	23.8 (4.3)	Local Site	SVM	Cortical thickness
Kawasaki et al.	2007	Japan	30	30	0 %	0 %	24.7 (4.4)	25.4 (4.4)	Own Hospitals Records	Multilinear method (MLM) with discriminant function analysis (via eigenimage)	Gray matter maps
Lieslehto et al. (ROI)	2021	Finland	29	61	51.72 %	62.29 %	33.7 (0.75)	34.5 (0.71)	NFBC (1996)	SVM	Full image + demographic features
Liu et al.	2017	China	38	38	34.21 %	34.21 %	24.8 (4.6)	25.0 (4.9)	Xiangya Hospital of Central South University (Changsha, Hunan, China	SVM	Cortical thickness and pial surface
Oh et al.	2020	Republic of Korea	30	30	50.00 %	50.00 %	31.9 (7.2)		Single center in South Korea (Uijeongbu St. Mary's)	Deep Learning and 3DCNN. Validation used Linear Regression as well	Video of a subject's structural MR images. The input dimensions were 256 × 256 × 180. This architecture has four 3D convolutional layers, with max-pooling-based downsampling in each convolutional layer.
Wang et al.	2011	China	32	32	21.87 %	34.37 %	24.0 (5.7)	22.5 (4.2)	Own Hospitals Records	SVM	Gray matter density map
Winterburn et al.	2017	Canada	91	67	32.96 %	41.79 %	32.1 (12.2)	25.8 (9.8)	NUDAST	SVM (NON -LINEAR)	Cortical thickness
Yamamoto et al. (Nagoya)	2020	Japan	50	51	48 %	43.13 %	38.8 (6.9)	36.5 (7.1)	Nagoya University	SVM	Grey matter density
Yamamoto et al. (Toyama) [	2020	Japan	49	48	53.06 %	52.08 %	28.1 (5.0)	26.9 (3.3)	Toyama University	SVM	Grey matter density
Zhang et al.	2023	USA	16	16	18.75 %	18.75 %	37.6 (13.3)		BrainGluSchi	Modified 3D VGG model with batch normalization and squeeze-and-excitation block (SE-VGG−11BN)	T1 weighted sMRI, whole head imaging

Footnotes: SCZ: Schizophrenia; HC: Healthy Controls; USA: United States of America; helAI: Artificial Intelligence; CI: Confidence Interval; CNN: Convolutional Neural Network; CSF: Cerebrospinal Fluid; DETJ: Determinant of Jacobian; DL: Deep Learning; DNN: Deep Neural Network; GM: Gray Matter; ML: Machine Learning; MLM: Multilinear Method; MRI: Magnetic Resonance Imaging; PCA: Principal Component Analysis; ROI: Region of Interest; SE: Sensitivity; SE-VGG-11BN: Squeeze-and-Excitation VGG-11 with Batch Normalization; SHAP: SHapley Additive exPlanations; SMOTE-ENN: Synthetic Minority Oversampling Technique - Edited Nearest Neighbors; SP: Specificity; SVC: Support Vector Classifier; SVM: Support Vector Machine; TPM: Tissue Probability Map; VBM: Voxel-Based Morphometry; VGG: Visual Geometry Group; WM: White Matter.

Meanwhile, the subgroup analysis focusing on DL-based processing models demonstrated a pooled SE of 0.82 (95 % CI: 0.80–0.85) and a SP of 0.83 (95 % CI: 0.80–0.86). Statistical heterogeneity was observed for both SE ($Tau^2 = 0$, $I^2 = 24.3$ %, $p = 0.26$) and SP ($Tau^2 = 0$, $I^2 = 0$ %, $p = 0.97$), indicating variability across studies.The SROC curve for DL models (Fig. 5) yielded an AUC of 0.89 (95 % CI: 0.81–0.90).

3.4. Subgroup analysis: diagnostic performance by neuroanatomical feature input

In the subgroup analysis restricted to studies evaluating GM features, the pooled SE was 0.72 (95 % CI: 0.65–0.79), while the SP reached 0.80 (95 % CI: 0.73–0.85). Moderate heterogeneity was found for SE ($Tau^2 = 0.10$, $I^2 = 57.9$ %, $p = 0.02$) and SP ($Tau^2 = 0.10$, $I^2 = 51$ %, $p = 0.05$). The bivariate SROC analysis yielded an AUC of 0.81 (95 % CI: 0.71–0.84) (Supplement 3).

Expanding beyond GM, WB input features demonstrated a pooled SE of 0.75 (95 % CI: 0.65–0.83), while the SP was 0.86 (95 % CI: 0.78–0.92). Heterogeneity was moderate to substantial for SE ($Tau^2 = 0$, $I^2 = 32.7$ %, $p = 0.22$) and SP ($Tau^2 = 0$, $I^2 = 0$ %, $p = 0.52$). The bivariate SROC analysis produced an 0.88 (95 % CI: 0.70–0.93 (Supplement 4)).

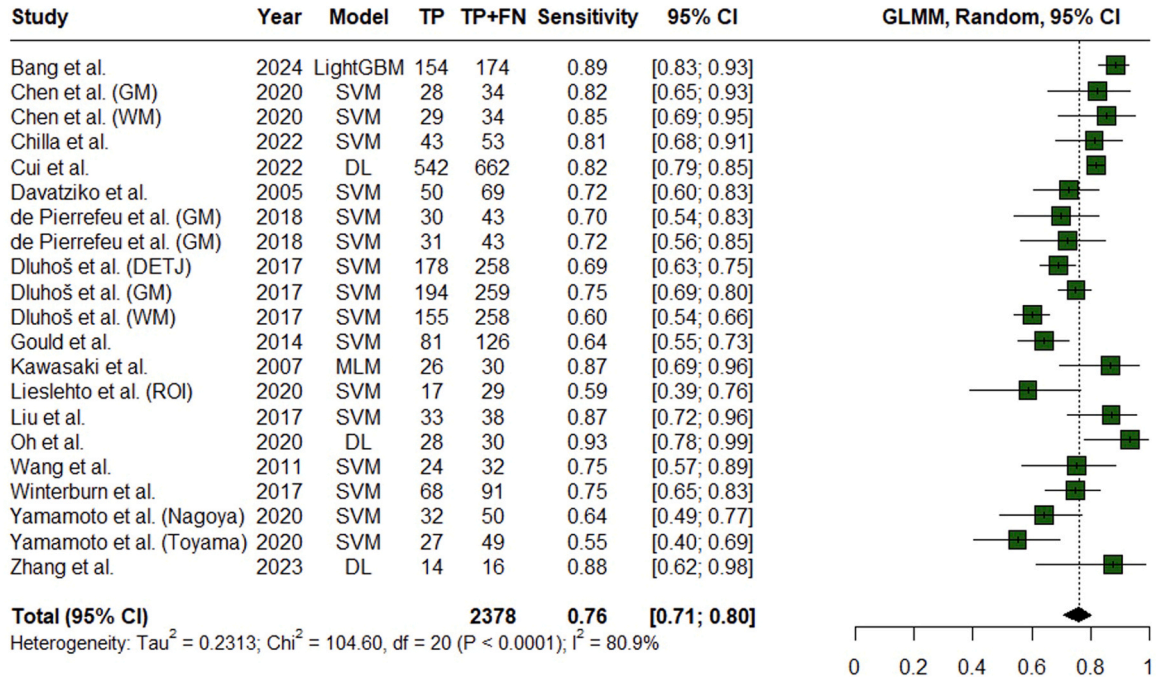

Fig. 2. Forest plot of the pooled sensitivity of AI models in diagnoing Schizophrenia.

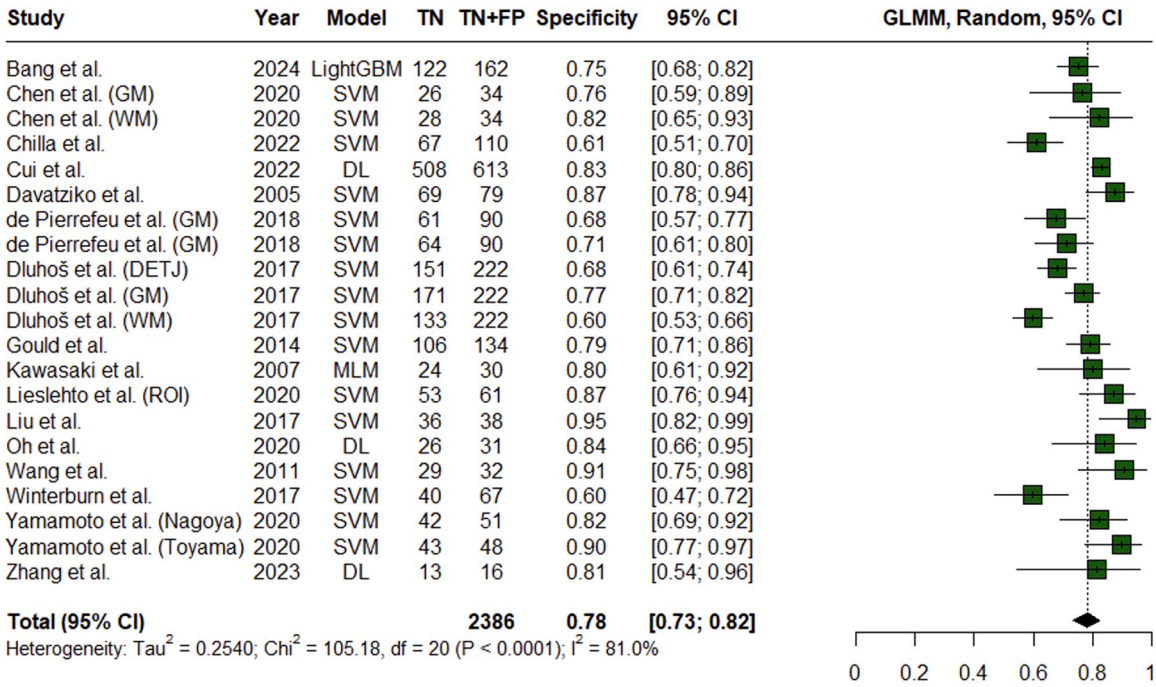

Fig. 3. Forest plot of the pooled specificity of AI models in diagnosing Schizophrenia.

3.5. Leave-one-out analysis

The leave-one-out analysis revealed that no single study exerted disproportionate influence on the overall pooled estimates of SE, SP, or AUC. Across all iterations, SE values ranged from 0.75 to 0.77, SP from 0.77 to 0.79, and AUC from 0.82 to 0.84 as seen in Supplement 5.

3.6. Risk of bias assessment

Risk of bias was generally low across studies on QUADAS-2. Nevertheless, concerns were noted in the index test domain in studies that did not report an independent validation set or a prespecified decision threshold (Wang et al.), which may affect accuracy estimates. Additional concerns arose in the reference standard domain where diagnostic ascertainment procedures were incompletely described (Zhang et al.), limiting appraisal of verification and applicability. Full, study-level judgments are presented in the graphical summary in Supplement 6.

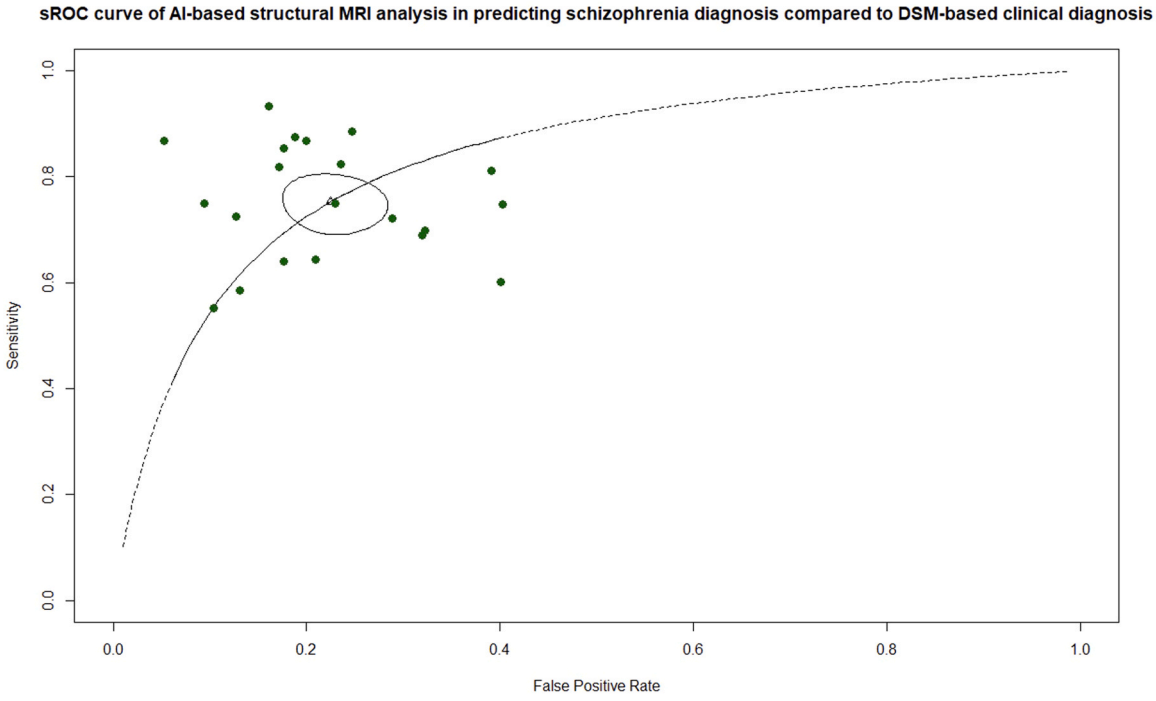

Fig. 4. Sumary Receiver Operating Characteristic (SROC) curve for all AI models.

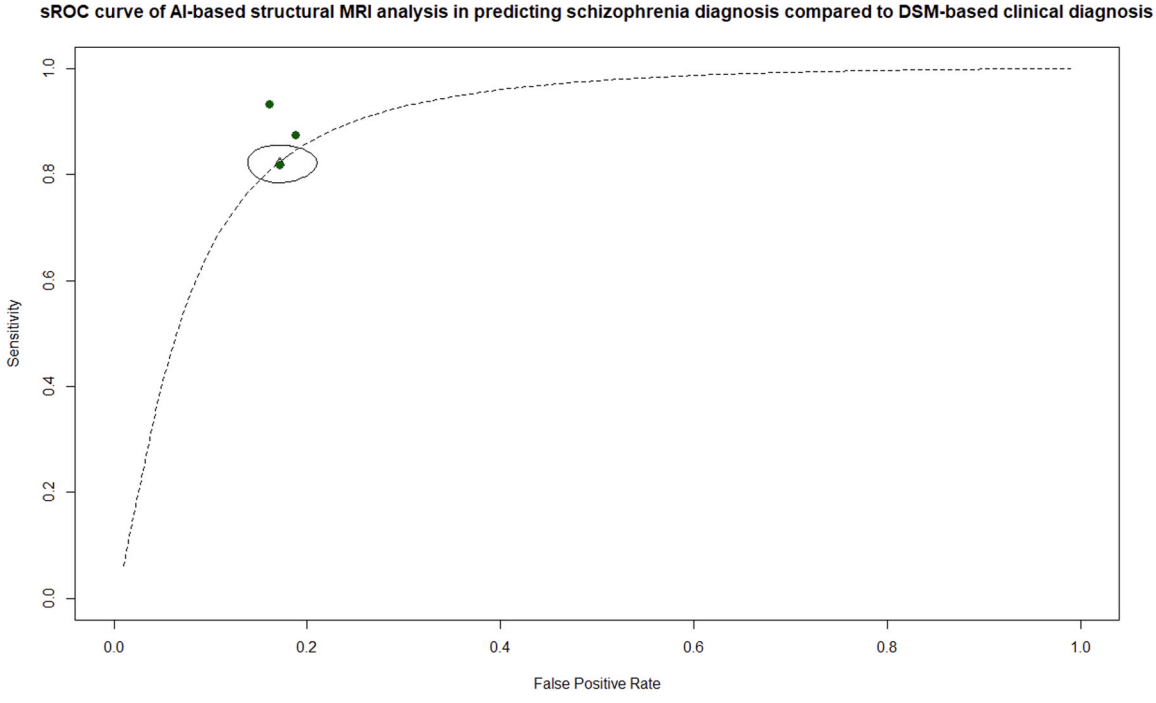

Fig. 5. Summary Receiver Operating Chracteristics (SROC) curve for Deep Learning models.

4. Discussion

This systematic review and meta-analysis of the diagnostic performance of AI models in schizophrenia analyzed 16 studies comprising 3601 participants. The overall pooled SE and SP were 0.76 (95 % CI: 0.71–0.80) and 0.78 (95 % CI: 0.73–0.82), respectively. Subgroup analyses revealed notable variation based on processing approach and neuroanatomical focus. Models using DL algorithms and WB input features generally outperformed those relying on traditional ML algorithms or GM features alone, suggesting potential gains in both SP and discriminative capacity when leveraging more comprehensive input data and advanced model architectures.

These results have strong clinical relevance by demonstrating that AI models using sMRI can support diagnostic decisions in patients with complex presentations, particularly when symptom profiles overlap with other psychiatric or neurological conditions (Cao et al., 2025). Thus, the use of AI-powered sMRI may assist psychiatrists in refining their assessments, particularly in such challenging cases, while

complementing, rather than replacing, their clinical expertise and judgment. Additionally, it should be noted that the currently reported high performance mainly derives from binary classification between patients and healthy controls. In real-world clinical diagnostic scenarios, the performance may decline, warranting further validation in future studies.

Previous meta-analyses, such as Di Camillo et al. (2024) pooled diagnostic accuracy metrics across multiple MRI modalities, condensing data on structural, functional, and diffusion imaging (Di Camillo et al., 2024). While comprehensive, the inclusion of state-dependent modalities, such as fMRI and DTI, may introduce additional heterogeneity and variability tied to symptom fluctuations and task-related factors. In contrast, our approach focused on restricting analysis to sMRI only, which allows for a more uniform methodological framework and focuses on stable neuroanatomical alterations that are less influenced by clinical state or external stimuli (Percie du Sert et al., 2023).

Stable neuroanatomical alterations in schizophrenia offer a trait-based signature detectable even in early stages, supporting reproducible diagnostic modeling. Recent evidence of uniform cortical folding in regions like the right caudal anterior cingulate cortex, shaped in early childhood, suggests constrained neurodevelopmental flexibility preceding symptom onset (de Vareilles et al., 2023, Sasabayashi et al., 2021). As such, the diagnostic patterns captured by AI models may be translatable to early psychosis, supporting their application in identifying individuals at risk before full clinical manifestation (Stevens et al., 2011).

The potential clinical applications of this technologies include aiding differential diagnosis in first-episode psychosis and supportive biological characterization in difficult-to-treat or diagnostically ambiguous cases. The stable-neuroanatomical changes may offer a pathway to reduce diagnosis time in young patients with first-episode psychosis, therefore enabling a timely treatment and better functioning outcomes (de Vareilles et al., 2023, Sasabayashi et al., 2021). Also, diagnostically ambiguous cases could benefit from a biological characterization and assist clinicians patient-specific treatments (Lieberman & Fenton, 2000).

The superior performance of DL models observed in this meta-analysis aligns with findings from previous studies, which consistently report higher diagnostic accuracy compared to traditional machine learning approaches (Abrol et al., 2021, Avberšek & Repovš, 2022). Unlike conventional SVMs, which often rely on manually engineered features and may require kernel functions to capture non-linear relationships, DL architectures can automatically learn complex, non-linear representations directly from high-dimensional imaging data (Chand et al., 2020). This enables them to capture subtle, distributed structural patterns associated with schizophrenia that simpler models may miss. Notably, because SVM is a specific traditional ML method, we refer to it by name to distinguish it from other traditional ML approaches in comparison with DL.

The present analysis reinforces this advantage by showing that DL-based models not only achieve higher specificity but also do so with lower heterogeneity, suggesting more consistent generalizability across diverse study populations and imaging protocols (Gengeç Benli & Andaç, 2023). Overall, these results suggest that DL architectures, capable of learning complex representations directly from high-dimensional data, may be inherently more robust to variations in acquisition protocols and preprocessing pipelines than traditional ML methods relying on handcrafted features.

The superior performance of WB models compared to both the overall pooled analysis and gray matter–focused approaches has important clinical implications. This aligns with recent evidence indicating that structural deviations in schizophrenia are highly individualized and span multiple brain systems, suggesting that whole-brain models may better accommodate such heterogeneity (Di Camillo et al., 2024). By incorporating structural information from across the entire brain, these models likely capture a broader range of disease-relevant alterations beyond localized GM changes, including abnormalities in white matter, ventricular size, and subcortical regions (Di Camillo et al., 2024, Si et al., 2024). This comprehensive input enhances diagnostic accuracy and may better reflect the diffuse and heterogeneous nature of schizophrenia-related brain changes (Mikolas et al., 2018, Mamah, 2023). Clinically, WB models offer a more robust and inclusive diagnostic approach, particularly practical in cases where structural abnormalities are subtle or distributed, supporting more accurate assessments across a broader spectrum of presentations.

In a multi dataset analysis, Oh et al. (2020) reported AUC 0.96 in sample and 0.71–0.90 on previously unseen datasets, exceeding semi trained clinician performance with AUC 0.61. Performance declined to 0.71 in younger, earlier-illness cohorts, with specific region-level contributions, underscoring disease stage and patient age dependence influence (Oh et al., 2020). Taken together, these findings indicate that sMRI models can outperform semi-trained clinicians yet display variable accuracy on new data. They align with evidence of structural brain changes preceding overt schizophrenia, while highlighting the need for targeted model training on early and prodromal populations and rigorous external validation before routine clinical use.

Recent studies have demonstrated the strong performance of AI models in diagnosing other neuropsychiatric conditions. For example, AI-based differential diagnosis of dementia subtypes using multimodal imaging has achieved AUCs up to 0.94 (Xue et al., 2024). In stroke populations, predictive models for progression to depression or Alzheimer's disease have reported AUCs ranging from 0.97 to 1.0 (Syaifullah et al., 2021). While these results stem from distinct clinical contexts and disease mechanisms, they underscore the relative underperformance of AI-based schizophrenia diagnostics and highlight the need for further methodological advances.

Clinical assessment of schizophrenia shows limited accuracy and reliability, underscoring the need for decision support. In standardized vignettes, Urkin et al. (2024) found that only one-in-three expert psychiatrists correctly identified schizophrenia-spectrum disorders, with very low inter-rater reliability ($\kappa = 0.08$) (Urkin et al., 2024). Likewise, a prior meta-analysis of 7912 patients reported low inter-rater and test–retest reliability (Santelmann et al., 2016). In this context, AI-assisted sMRI could help improve diagnostic consistency and provide structured support for physicians.

Substantial heterogeneity was observed across studies, likely driven by differences in AI model architecture, neuroanatomical input features, preprocessing pipelines, and validation strategies as described in Table 1. Methodological variabilities and divergent sample characteristics, such as sex distribution, contributed to variability. Kawasaki et al. (2007), Gould et al. (2016), and Zhang et al. (2023) reported unbalanced sex distributions of 18 % females, 18 %, and 0 % respectively. Moreover, the risk of bias assessment identified concerns about lack of external data validation and prespecified thresholds in Wang et al., potentially affecting the accuracy estimates. Thirdly, differences in patients' clinical characteristics (e.g., first-episode vs. chronic cases) may influence brain structure and thereby affect the discriminative features learned by the models. Fourthly, the results of the leave-one-out analysis on SE, SP, and AUC demonstrated sustained stability of the presented results. Also, there was no single study that was identified as a substantial source of heterogeneity, which further supports the hypothesis of methodological variabilities and demographic diversity. Notably, heterogeneity was substantially lower in the DL subgroup ($I^2 = 24.3$ % for SE, 0 % for SP) compared to the SVM subgroup. This finding, visually supported by the tighter clustering of study points in the DL-specific SROC curve (Fig. 5), suggests that DL architectures may be more robust to variations in preprocessing pipelines and imaging parameters, leading to more consistent performance across different settings. Finally, variations in model input from T1-weighted scans in Bang et al., (2024) to segmented tissue maps in Chen et al. (2020) might have influenced the observed heterogeneity in the overall pooled results.

International, multi-site studies with harmonized acquisition and analysis are essential for robust generalizability and interpretability.

ENIGMA-scale work and Omlor et al. (2025), reported reproducible structural variability, also in early-stage cohorts, that can serve as biologically grounded priors, reducing "black-box" opacity and improving stage-specific AI-model calibration. Prospective, preregistered, cross-site external validations built on these frameworks are critical to translate sMRI-based AI from research to routine care.

Future work should standardize model input, evaluation protocols, and diagnostic criteria, using diverse, multi-site cohorts to improve generalizability. Moreover, they should conduct head-to-head comparisons on harmonized datasets to identify clinically robust, scalable models for schizophrenia diagnosis.

This study has several notable strengths. Focusing exclusively on structural MRI avoids the methodological heterogeneity introduced by multimodal imaging and highlights the diagnostic potential of a clinically accessible modality. The inclusion of subgroup analyses by model type and neuroanatomical input provides valuable insight into the factors driving performance variability, revealing the superior accuracy of deep learning algorithms and whole-brain features. The inclusion criteria was tailored to minimize population overlap, particularly in studies using open-access MRI datasets, thereby improving the independence and reliability of pooled estimates.

This study also has some limitations. Clinical heterogeneity, including variable duration of illness, antipsychotic exposure, first-episode versus recurrent status, and treatment resistance, was inconsistently reported, precluding sensitivity analysis and motivating our sMRI-only focus on more trait-like structural features. Significant heterogeneity was observed across included studies, due to the factors explored above. Secondly, many studies relied on small, single-center samples with limited demographic diversity, reducing the generalizability of findings. Additionally, differences in diagnostic reference standards and a lack of transparency in reporting model training procedures may have influenced performance estimates. (Sun, 2023) Fourthly, the absence of head-to-head model comparisons within the same datasets further limits direct evaluation of algorithm superiority. These factors highlight the need for more standardized and collaborative research efforts in this field. Moreover, the intelligibility of ML studies in psychiatry and its likely applicability in patient-care poses questions about the ability of non-specialized personal to understand the limitations of these models. (Tandon, 2023, Tandon, 2019) Due to the technical complexity of these novel computational methods in psychiatric care, careful interpretation is recommended. (Thornton, 2023) Future studies should standardize and report more detailed clinical and imaging metadata to support deeper analyses of heterogeneity.

5. Conclusion

This systematic review and meta-analysis demonstrated that AI models using sMRI can achieve promising diagnostic accuracy for schizophrenia, with DL algorithms and WB inputs offering the highest performance. The overall pooled SE and SP were 0.76 (95 % CI: 0.71–0.80) and 0.78 (95 % CI: 0.73–0.82), respectively. Further evidence is needed to support the integration of AI-assisted sMRI tools into clinical practice as complementary aids in diagnostically challenging cases, particularly in early or ambiguous presentations.

Ethical Approval

This study was a systematic review and diagnostic meta-analysis of data from previously published studies. Because no new human participants were directly involved, institutional ethics approval and participant consent were not required. All data were derived from studies published in accordance with the Declaration of Helsinki and local regulations.

Informed Consent

Not applicable. Individual patient consent was obtained in the original studies; no new data collection was performed.

Disclosure

None of the authors have any conflicts of interest to declare related to this project. All authors affirm that there are no financial or personal relationships that could have influenced the work reported in this manuscript.

CRediT authorship contribution statement

Nima Norbu Sherpa: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis. **Delfina Lahitou Herlyn:** Writing – original draft, Resources, Project administration, Methodology, Investigation, Formal analysis. **Felipe Bandeira de Melo Guimarães:** Writing – review & editing, Writing – original draft, Validation, Resources, Project administration, Methodology. **Eduardo Cesar Teixeira Sirena:** Validation, Software, Resources, Project administration, Investigation, Formal analysis, Conceptualization. **Wellgner Fernandes Oliveira Amador:** Writing – original draft, Visualization, Validation, Project administration, Formal analysis, Data curation. **Laura Mora:** Software, Project administration, Methodology, Formal analysis. **Veronica Ramirez Lopera:** Visualization, Software, Resources, Formal analysis, Conceptualization. **Pandora Eloa Oliveira Fonseca:** Writing – original draft, Project administration, Methodology, Funding acquisition, Formal analysis, Data curation. **Martin Kotochinsky:** Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Thales Pardini Fagundes:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Andrea Gonzalez Lezana:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no competing financial interests or personal relationships that could be perceived to have influenced the work reported in this paper.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ajp.2025.104759.

Data Availability

All data analyzed in this work are publicly available in the cited publications. We conducted our review in accordance with the PRISMA 2020 guidelines and registered the protocol in PROSPERO (CRD420251087015).

References

Abrol, A., Fu, Z., Salman, M., 2021. Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. Nat. Commun. 12 (1), 353.

Avberšek, L.K., Repovš, G., 2022. Deep learning in neuroimaging data analysis: Applications, challenges, and solutions. Front Neuroimag. 1, 981642.

Bradford, A., Meyer, A.N.D., Khan, S., 2024. Diagnostic error in mental health: a review. BMJ Qual. Saf. 33 (10), 663–672.
Brugger, S.P., Howes, O.D., 2017. Heterogeneity and homogeneity of regional brain structure in schizophrenia: a meta-analysis. JAMA Psychiatry 74 (11), 1104–1111.
Cao, P., Li, R., Li, Y., 2025. Machine learning based differential diagnosis of schizophrenia, major depression disorder and bipolar disorder using structural magnetic resonance imaging. J. Affect Disord. 383, 20–31.
Chand, G.B., Dwyer, D.B., Erus, G., 2020. Two distinct neuroanatomical subtypes of schizophrenia revealed using machine learning. Brain 143 (3), 1027–1038.
Charlson, F.J., Ferrari, A.J., Santomauro, D.F., 2018. Global epidemiology and burden of schizophrenia: findings from the Global Burden of Disease Study 2016. Lancet Psychiatry 5 (12), 969–978.
Di Camillo, F., Grimaldi, D.A., Cattarinussi, G., 2024. Magnetic resonance imaging-based machine learning classification of schizophrenia spectrum disorders: a meta-analysis. Psychiatry Clin. Neurosci. 78 (12), 732–743.
van Erp, T.G.M., Walton, E., Hibar, D.P., 2018. Cortical brain abnormalities in 4474 individuals with schizophrenia and 5098 control subjects via the ENIGMA Consortium. Biol. Psychiatry 84 (9), 644–654.
Gengeç Benli Ş, Andaç, M., 2023. Constructing the schizophrenia recognition method employing GLCM features from multiple brain regions and machine learning techniques. Diagn. (Basel) 13 (13), 2140.
Kahn, R.S., Sommer, I.E., Murray, R.M., 2015. Schizophrenia. Nat. Rev. Dis. Prim. 1, 15067.
Kambeitz, J., Kambeitz-Ilankovic, L., Leucht, S., 2015. Detecting neuroimaging biomarkers for schizophrenia: a meta-analysis of multivariate pattern recognition studies. Neuropsychopharmacology 40 (7), 1742–1751. https://doi.org/10.1038/npp.2015.22.
Lieberman, J.A., Fenton, W.S., 2000. Delayed detection of psychosis: causes, consequences, and effect on public health. Am. J. Psychiatry 157 (11), 1727–1730.
Madre, M., Canales-Rodríguez, E.J., Fuentes-Claramonte, P., 2020. Structural abnormality in schizophrenia versus bipolar disorder: a whole brain cortical thickness, surface area, volume and gyrification analyses. Neuroimage Clin. 25, 102131.
Mamah, D., 2023. A review of potential neuroimaging biomarkers of schizophrenia-risk. J. Psychiatr. Brain Sci. 8 (2), e230005.
Mikolas, P., Hlinka, J., Skoch, A., 2018. Machine learning classification of first-episode schizophrenia spectrum disorders and controls using whole brain white matter fractional anisotropy. BMC Psychiatry 18 (1), 97.
Noma H. dmetatools: Diagnostic Meta-Analysis Tools. R package version 1.1.1. 2023. Available from: https://cran.r-project.org/package=dmetatools.
Oh, J., Oh, B.-L., Lee, K.-U., Chae, J.-H., Yun, K., 2020. Identifying Schizophrenia Using Structural MRI With a Deep Learning Algorithm. Front. Psychiatry 11, 16.
Omlor, W., Rabe, F., Fuchs, S., 2025. Estimating multimodal structural brain variability in schizophrenia spectrum disorders: a worldwide ENIGMA study. Am. J. Psychiatry 182 (4), 373–388.
Page, M.J., McKenzie, J.E., Bossuyt, P.M., 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. J. Clin. Epidemiol. 134, 178–189.
Percie du Sert, O., Unrau, J., Gauthier, C.J., 2023. Cerebral blood flow in schizophrenia: a systematic review and meta-analysis of MRI-based studies. Prog. Neuropsychopharmacol. Biol. Psychiatry 121, 110669.
Saha, A., Park, S., Geem, Z.W., 2024. Schizophrenia detection and classification: a systematic review of the last decade. Diagn. (Basel) 14 (23), 2698.
Santelmann, Hanno, et al., 2016. Interrater reliability of schizoaffective disorder compared with schizophrenia, bipolar disorder, and unipolar depression - A systematic review and meta-analysis. Schizophr. Res. 176 (2-3), 357–363.
Sasabayashi, D., Takahashi, T., Takayanagi, Y., 2021. Anomalous brain gyrification patterns in major psychiatric disorders: a systematic review and transdiagnostic integration. Transl. Psychiatry 11, 176.
Si, S., Bi, A., Yu, Z., 2024. Mapping gray and white matter volume abnormalities in early-onset psychosis: an ENIGMA multicenter voxel-based morphometry study. Mol. Psychiatry 29, 496–504.
Stevens, F.L., Hurley, R.A., Taber, K.H., 2011. Anterior cingulate cortex: unique role in cognition and emotion. J. Neuropsychiatry Clin. Neurosci. 23 (2), 121–125.
Sun, Jie, et al., 2023. Artificial intelligence in psychiatry research, diagnosis, and therapy. Asian J. Psychiatry 87, 103705. https://doi.org/10.1016/j.ajp.2023.103705.
Syaifullah, A.H., Shiino, A., Kitahara, H., 2021. Machine learning for diagnosis of AD and prediction of MCI progression from brain MRI using brain anatomical analysis using diffeomorphic deformation. Front Neurol. 11, 576029.
Tandon, Neeraj, Tandon, Rajiv, 2019. Machine learning in psychiatry- standards and guidelines. Asian J. Psychiatry 44, A1–A4. https://doi.org/10.1016/j.ajp.2019.09.009.
Tandon, Rajiv, 2023. Application of computational methods to the study of schizophrenia an exciting but treacherous frontier. Asian J. Psychiatry 87, 103752. https://doi.org/10.1016/j.ajp.2023.103752.
Thornton, Joseph, et al., 2023. Artificial intelligence and psychiatry research and practice. Asian J. Psychiatry 81, 103509. https://doi.org/10.1016/j.ajp.2023.103509.
Urkin, Bar, et al., 2024. Schizophrenia Spectrum Disorders: An Empirical Benchmark Study of Real-world Diagnostic Accuracy and Reliability Among Leading International Psychiatrists, 3 May Schizophr. Bull. Open 5 (1), sgae012, 3 May.
de Vareilles, H., Rivière, D., Mangin, J.F., 2023. Development of cortical folds in the human brain: an attempt to review biological hypotheses, early neuroimaging investigations and functional correlates. Dev. Cogn. Neurosci. 61, 101249.
Whiting, P.F., Rutjes, A.W., Westwood, M.E., 2011. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann. Intern Med 155 (8), 529–536.
Xue, C., Kowshik, S.S., Lteif, D., 2024. AI-based differential diagnosis of dementia etiologies on multimodal data. Nat. Med 30, 2977–2989.