Table S1. Characteristics of included studies investigating MRI-based machine learning models for sex classification

Author (year)	Population (Male/Female)	Age range (mean ± SD)	Morphometric Features	Classification model	Prediction accuracy/AUC
(Besson et al., 2021) [16]	6410 (2893/3517)[4,6,11]	6-89 (33.32±17.53)	Inner and outer cortical surfaces	Graph convolutional neural networks (gCNNs)	87.99 % on both surfaces 85.23% inner surface
(Bi et al., 2023) [17]	11700[1]	9-12	Grey matter data	3D Convolutional neural network (3D CNN)	97.3%
(Brennan et al., 2021) [18]	8325 (4308/4017)	9–10 (119.07±7.48 months)	Cortical Volume Cortical Thickness Cortical Surface Area Sulcal Depth Grey matter intensity White matter intensity Grey-white matter boundary intensity (GWC)	Support Vector Machine (SVM)	76.1% cortical area 75.20% cortical volume 73.21% SulcD 72.55% cortical thickness 71.37% grey matter intensity 71.01% GWC 69.38% white matter intensity 86.33% all features
(Chen et al., 2024) [19]	1031 (471/560) [4]	22-37	Fractional anisotropy mean diffusivity mean kurtosis	2D CNN 3D CNN Vision Transformer	91% highest on 2D CNN using fractional anisotropy 83% lowest on Vision Transformer using fractional anisotropy
(Dibaji et al., 2024) [20]	2110 (995/1115)	18-95	Grey matter data	Simple Fully Convolutional Network	89.5%
(Dhamala et al., 2020) [21]	1003 (469/534)[4]	22-37	Hurst exponent Volume	SVM	81.2% on Hurst exponent 85.38% on volume
(Dumitru, 2023) [4]	826 (396/430)[2]	19-26	Subtraction index (SI) Distance index (DI) Laterality index (LI)	Random Forest (RF)	63% highest for SI using grey matter volume compared to other target measure, 64% highest for DI using grey matter surface area compared to other target measure, 62% highest for LI using grey matter volume compared to other target measure.
(Ebel et al., 2023) [9]	3298[3] 399[4]	21-90 (53) 22-36 (29)	Z-score normalization of grey matter voxels	CNN Logistic regression	95.78% with LR on SHIP data and 91.8% on HCP data 91.4% with CNN on SHIP data and 85.17% on HCP data

(Fan et al., 2020) [22]	1206 (481/569)[4]	22-35	Dynamic functional connectivity	CNN + long short-term memory network (CNN-LSTM)	93.05%
(Feis et al., 2013) [23]	121 (54/67)	20-30 (25±2)	Grey matter segments on T1 weighted, T2 weighted and fractional anisotropy (FA)	SVM	88% T1 weighted grey matter segment 85% T2 weighted grey matter segment 83% FA segment 96% multimodal: T1 weighted, T2 weighted and diffusion weighted (FA)
(Ge et al., 2021) [24]	1063[4] 549 (236/313)[5]	22-37 17-27	Cortical Volume Cortical Thickness Cortical Folding (gyrification index (GI), fractal dimension (FD), sulcal depth (SulcD))	Linear Discriminant Analysis	80% on cortical thickness, 72% on GI, 71% on cortical volume, 70% on SulcD, 69% on FD. 85% on composite metrics.
(Hu et al., 2019) [25]	876 (385/491)[4]	22–37	Grey matter data	Multi-Layer 3D Convolution Extreme Learning Machine (MCN-ELM)	98.06%
(Jeon et al., 2024) [26]	1157 (607/550)	14-89	Subcortical and cortical volume	Linear Regression K-Nearest Neighbor Extreme Gradient Boosting Random Forest Support Vector Classifier Multilayer Perceptron Variational Quantum Circuit	81 % on linear regression, 77.9 % on k-nearest neighbor, 76.2 % on extreme gradient boosting, 77 % on random forest, 77.1 % on support vector classifier, 75.3 % on multilayer perceptron, 81.8 % on variational quantum circuit
(Leming & Suckling, 2021) [27]	16970		Resting-state data Emotion task functional data	Convolutional neural network	0.8459 AUC of combined data 0.8923 AUC of resting state data 0.7683 AUC of task fmri data
(Luo et al., 2019) [28]	1113 (507/606)[4]	22–37	Cortical 3-D Morphology (0.7mm, 1.4 mm and 2.8 mm)	Hierarchical Sparse Representation Classifier (HSRC) Direct sparsity Principal Component Analysis (PCA)	96.77% highest using HSRC on 0.7mm data 94.7% highest using direct sparsity on 1.4mm data 94.52% highest using PCA on 1.4mm data
(Matte Bon et al., 2024) [29]	992 (496/496)[4] 1017 (390/627)	22-38 18-30	Volume	Gradient Boosted Trees (GBedT)	82.3% on limbic volume 77.6% on non-limbic volume 79.7% on whole brain volume

Reference	Sample (M/F)	Age	Data type	Method	Results
(Mendes et al., 2021) [30]	580 (428/152)[6] 922 (582/ 340)[7]	6-20 (12.12 ± 3.16) 7.1–19.9 (11.72 ± 2.99)	Grey matter data	3D CNN	AUC= 0.85 for ADHD-200 AUC=0.82 for ABIDE-II
(Metoki et al., 2024) [48]	3129 (1542/1587)	9-11	Resting state functional connectivity Cortical thickness Cortical volume	SVM	85 % rsFC 76 % Cortical thickness 70 % Cortical volume
(Nebli & Rekik, 2020) [31]	698 (308/390)	21-23	Maximum Principal Curvature Cortical Thickness Network Sulcal Depth Network Average Curvature Network	SVM	73% sulcal depth network on both hemispheres 70% cortical thickness network on both hemispheres 64% maximum principal curvature on left hemisphere and 66% on right hemisphere 56% on average curvature network on both hemispheres
(Patel et al., 2024) [32]	622 (279/43)[8]	8-22	Functional connectivity matrix	SVM Graph isomorphism network Brain graph neural networks Multi-view graph convolutional network Multimodal graph convolutional network Multi-layer perceptron model Multi-modal graph isomorphism network	67.53% on support vector machine 77.18% on graph isomorphism network 68.6% on brain graph neural networks 80.06% on multi-view graph convolutional network 75.01% on multimodal graph convolutional network 80.97% on multi-layer perceptron model 81.67% on multi-modal graph isomorphism network
(Ryali et al., 2024) [33]	1088 (499/589) [4]	22-38	Resting state data	spatiotemporal deep neural network (stDNN)	94.72% highest over 4 sessions
(Sanchis-Segura et al., 2022) [34]	876 (438/438)[4]	(28.5±3.4)	Grey matter volume	Linear Discriminant Analysis (LDA) Logistic Regression (LR) Multiple Adaptive Regression Splines (MARS) Random Forest SVM	90% highest using SVM, 88.3% using LDA, 88% using MARS, 87.7% using RF, 86.3% using LR.

(Satterthwaite et al., 2015) [35]	674 (312/362)	9-22	Functional connectivity data	SVM	71%
(Sen & Parhi, 2021) [36]	475 (196/279)[4]	22-36	Dynamic functional connectivity (tensor parallel factor (PARAFAC))	Random Forest	94%
(Sepehrband et al., 2018) [37]	967 (469/498)[8] 682 (360/322)[9]	8-23 (14.64±3.44) 3-21 (12.29±5.03)	Cortical morphology	SVM	83% PNC 77% PING
(Shanmugan et al., 2022) [38]	693 (298/395)	8-22	Functional networks	SVM	82.9%
(Wang et al., 2012) [39]	140 (70/70)	18-26	Grey matter volume	SVM	86.1% grey matter volume 89.1% multimodal: grey matter volume and resting state
(Weber et al., 2022) [40]	1113 (507/606) [4]	(28.8±3.7)	Grey matter feature White matter feature Resting state feature	Linear Regression	89.3% highest with resting state pipeline, 84% on grey matter features, 85.3 on white matter feature, 93.3% grey matter, white matter, resting state and task evoked fMRI
(Weis et al., 2020) [41]	1685 (884/801) [4, 11]	18-88	Resting state connectivity parcels	SVM	75.1% highest on right middle cingulate cortex
(X. Zhang et al., 2020) [42]	290 (135/55)	18-29	Grey matter volume Regional homogeneity Functional connectivity	SVM	94.3% on grey matter volume 90.73% on regional homogeneity 83.89% on functional connectivity 96.6% on grey matter volume, regional homogeneity and functional connectivity
(Xin et al., 2019) [43]	1065 (490/575) [4]	22-36	Fractional anisotropy	3D CNN SVM	93.3% on 3D Convolutional neural network 78.2% on Support Vector Machine
(Xu et al., 2020) [44]	58 (28/30)	19-31	Language task fMRI	SVM	75.86%
(Y. Zhang et al., 2021)[45]	9620 (4495/5125)[4, 10]	19-62	Resting state data	SVM	74.55%
(Yang et al., 2021) [46]	188 (128/60)	0-15	Grey matter volume data Diffusion tensor imaging (DTI) data	SVM	66% on T1 weighted feature 55% on DTI feature 72% on both multimodal features

(Yeung et al., 2023) [47]	3152 (1485/1667)[10]	44.6-77.1	Fraction anisotropy Intracellular volume fraction Isotropic volume fraction Mean diffusivity Orientation dispersion Streamline count	BrainNetCNN	86.91% on streamline count 78.15% on mean diffusivity 79.74% on fractional anisotropy 82.88% on orientation dispersion 81.82% on isotropic volume fraction 78.34% on intracellular volume fraction

[1]Source from public data: Adolescent Brain Cognitive Development (ABCD)

[2]Source from public data: Amsterdam Open MRI collection

[3]Source from public data: Study of Health of Pomerania (SHIP)

[4] Source from public data: Human Connectome Project (HCP)

[5] Source from public data: Southwest University Longitudinal Imaging Multimodal (SHIM)

[6]Source from public data: Autism Brain Imaging Data Exchange II (ABIDE-II)

[7]Source from public data: Attention Deficit Hyperactivity Disorder (ADHD-200)

[8]Source from public data: Philadelphia Neurodevelopmental Cohort (PNC)

[9]Source from public data: Pediatric Imaging, Neurocognition and Genetics (PING)

[10]Source from public data: UK Biobank (UKB)

[11]Other Sources: Age-ility, Cambridge Centre for Ageing Neuroscience (CamCan), Consortium for Reliability and Reproducibility (CoRR), Dallas Lifespan Brain Study (DLBS), Brain Genomics Superstruct Project (GSP), Information eXtraction from Images (IXI), MPI-Leipzig Mind Brain Body (MPI-LMBB), Enhanced Nathan Kline Institute – Rockland Sample (NKI-RS), and Southwest University Adult Lifespan Dataset (SALD).