Original Paper

Advancing Early Detection of Major Depressive Disorder Using Multisite Functional Magnetic Resonance Imaging Data: Comparative Analysis of AI Models

Masab Mansoor[1], DBA; Kashif Ansari[2], MD

[1]School of Medicine, Edward Via College of Osteopathic Medicine, Louisiana Campus, Monroe, LA, United States
[2]East Houston Medical Center, Houston, TX, United States

Corresponding Author:

Masab Mansoor, DBA
School of Medicine
Edward Via College of Osteopathic Medicine, Louisiana Campus
4408 Bon Aire Dr
Monroe, LA, 71201
United States
Phone: 1 5045213500
Email: mmansoor@vcom.edu

Related Articles:
Preprint (medRxiv): https://www.medrxiv.org/content/10.1101/2024.08.13.24311933v1
Peer-Review Report by Anonymous: https://med.jmirx.org/2025/1/e76744
Peer-Review Report by Anonymous: https://med.jmirx.org/2025/1/e76746
Peer-Review Report by Anonymous: https://med.jmirx.org/2025/1/e76747
Authors' Response to Peer Review Reports: https://med.jmirx.org/2025/1/e75617

Abstract

Background: Major depressive disorder (MDD) is a highly prevalent mental health condition with significant public health implications. Early detection is crucial for timely intervention, but current diagnostic methods often rely on subjective clinical assessments, leading to delayed or inaccurate diagnoses. Advances in neuroimaging and machine learning (ML) offer the potential for objective and accurate early detection.

Objective: This study aimed to develop and validate ML models using multisite functional magnetic resonance imaging data for the early detection of MDD, compare their performance, and evaluate their clinical applicability.

Methods: We used functional magnetic resonance imaging data from 1200 participants (600 with early-stage MDD and 600 healthy controls) across 3 public datasets. In total, 4 ML models—support vector machine, random forest, gradient boosting machine, and deep neural network—were trained and evaluated using a 5-fold cross-validation framework. Models were assessed for accuracy, sensitivity, specificity, F_1-score, and area under the receiver operating characteristic curve. Shapley additive explanations values and activation maximization techniques were applied to interpret model predictions.

Results: The deep neural network model demonstrated superior performance with an accuracy of 89% (95% CI 86%-92%) and an area under the receiver operating characteristic curve of 0.95 (95% CI 0.93-0.97), outperforming traditional diagnostic methods by 15% (P<.001). Key predictive features included altered functional connectivity between the dorsolateral prefrontal cortex, anterior cingulate cortex, and limbic regions. The model achieved 78% sensitivity (95% CI 71%-85%) in identifying individuals who developed MDD within a 2-year follow-up period, demonstrating good generalizability across datasets.

Conclusions: Our findings highlight the potential of artificial intelligence–driven approaches for the early detection of MDD, with implications for improving early intervention strategies. While promising, these tools should complement rather than replace clinical expertise, with careful consideration of ethical implications such as patient privacy and model biases.

JMIRx Med 2025;6:e65417; doi: 10.2196/65417

Keywords: major depressive disorder; machine learning; functional MRI; early detection; artificial intelligence; psychiatry

Introduction

Background

Major depressive disorder (MDD) is a leading cause of disability worldwide, affecting over 280 million people and significantly contributing to the global burden of disease [1]. Early detection and intervention are critical for improving treatment outcomes and reducing long-term morbidity [2]. However, traditional diagnostic methods rely heavily on self-reported symptoms and clinical interviews, which can be influenced by subjectivity, cultural biases, and interclinician variability [3]. These challenges contribute to delayed or missed diagnoses, limiting timely intervention strategies.

Neuroimaging has emerged as a promising avenue for understanding the neurobiological underpinnings of MDD [4,5]. Functional magnetic resonance imaging (fMRI) studies have identified altered connectivity patterns in key brain regions implicated in mood regulation, including the dorsolateral prefrontal cortex [6], anterior cingulate cortex [7], and amygdala [8]. Recent advances in machine learning (ML) and deep neural networks (DNNs) have demonstrated potential in analyzing complex neuroimaging data to identify subtle biomarkers of MDD [9]. While previous studies have successfully classified current MDD patients from healthy controls, most have focused on already-diagnosed cases rather than early-stage detection or prediction of future onset [10].

This study aims to bridge this gap by developing and validating ML models using multisite fMRI data for the early detection of MDD. Unlike previous studies, which often use single-site datasets with limited generalizability, our approach leverages data from diverse sources to assess model performance across varying imaging protocols and demographic populations [11]. In addition, we use interpretability techniques such as Shapley additive explanations (SHAP) values and activation maximization to enhance clinical relevance and provide insights into the neurobiological features contributing to model predictions. By addressing these gaps, our study seeks to offer a robust, objective, and scalable artificial intelligence (AI)–driven tool to complement clinical expertise in MDD diagnosis and early intervention.

The diagnostic framework for MDD is primarily guided by the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-5*), which requires the presence of specific symptoms for at least 2 weeks [12]. While widely used, this approach has several limitations:

1. Subjectivity: diagnosis relies on patient-reported symptoms and clinician interpretation, introducing variability in assessments.
2. Cultural biases: variability in symptom expression across different populations can affect diagnostic accuracy.
3. Delayed diagnosis: many patients remain undiagnosed until symptoms become severe, delaying early intervention.
4. Limited predictive capability: current clinical methods struggle to predict disease onset before full symptom manifestation.

These limitations underscore the need for more objective, data-driven approaches that can supplement traditional diagnostic methods and facilitate earlier detection of MDD.

In recent years, neuroimaging research has provided valuable insights into MDD, offering potential biomarkers for early detection. Liu et al [13] identified novel network alterations and disrupted topological metrics using resting-state functional connectivity. Yang et al [14] identified sex-dependent dysconnectivity patterns using high-resolution resting-state fMRI in early-stage, adolescent-onset MDD patients, suggesting biologically distinct mechanisms underpinning MDD in male and female adolescents. Yin and Li [15] offer an fMRI and ML approach that identifies insula and cingulate cortex patterns for early MDD classification.

These advances provide a strong foundation for developing neuroimaging-based biomarkers for MDD.

ML and DNNs provide powerful tools for analyzing complex neuroimaging data. Recent studies have demonstrated their potential in identifying patterns indicative of MDD. Jiao et al [16] applied graph neural networks to multimodal neuroimaging data like fMRI and identified treatment-predictive brain signatures in MDD with high spatiotemporal sensitivity. Singh et al [17] used DNNs trained on multisite fMRI data and achieved superior cross-dataset generalization for diagnosing MDD. Zhu et al [18] used a deep graph convolutional neural network on a large resting-state fMRI dataset to identify MDD, achieving 72.1% accuracy and outperforming traditional methods.

Despite these advancements, several challenges remain:

1. Limited focus on early detection: most AI studies classify existing MDD cases rather than predicting their onset.
2. Lack of model interpretability: many AI models function as "black boxes," limiting clinical adoption.
3. Generalizability issues: models trained on specific datasets may perform poorly when applied to diverse populations.

Objectives

This study aims to address these challenges by developing and comparing AI models for the early detection of MDD using multisite fMRI data. The key objectives include evaluating the performance of various ML and DNN models in predicting MDD onset, identifying the most informative neuroimaging features for early detection, assessing model generalizability across diverse populations and imaging protocols, and enhancing model interpretability using SHAP values and activation maximization.

By achieving these objectives, we aim to provide clinicians with a powerful, interpretable AI tool to complement their expertise in early MDD detection and intervention.

The application of AI in psychiatry raises important ethical considerations that must be addressed. Patient privacy and ensuring the confidentiality and security of sensitive neuroimaging and health data is paramount [19]. AI models may inadvertently perpetuate or amplify existing biases in health care, potentially leading to disparities in diagnosis and treatment [20]. The "black box" nature of some AI models poses challenges for clinical decision-making and accountability [21]. AI tools should complement, not replace, clinical judgment. Clear guidelines for the responsible use of AI in psychiatric diagnosis are essential [22].

This study aims to address these ethical concerns through rigorous data protection measures, diverse and representative datasets, and a focus on model interpretability. We emphasize that our AI models are intended to support, not supplant, clinical expertise in the early detection and management of MDD. Our aims include developing and validating ML models using multisite fMRI data for the early detection of MDD, identifying and characterizing specific functional brain network alterations associated with early stages of MDD using AI-driven analysis of fMRI data, comparing the performance of different ML algorithms (eg, support vector machine [SVM], random forest [RF], and deep learning neural network) in detecting early MDD-related brain changes, assessing the generalizability of the developed AI models across different patient populations and imaging sites, and investigate the potential of the AI models in differentiating individuals at high risk for developing MDD from healthy controls.

Methods

Overview

We used fMRI data from 3 publicly available datasets: OpenfMRI Depression Dataset, REST-meta-MDD, and EMBARC. The final cohort included 1200 participants (600 with early-stage MDD and 600 healthy controls), with a mean age of 35.7 (SD 9.8) years and 54% (648/1200) of participants being female.

Preprocessing was performed using FMRIB Software Library v6.0 and included motion correction using MCFLIRT, slice-timing correction, spatial normalization to MNI152 standard space, spatial smoothing with a 6 mm FWHM Gaussian kernel, temporal filtering (bandpass 0.01-0.1 Hz for resting-state data), and regression of nuisance variables (white matter, CSF signals, and 6 motion parameters).

These preprocessing steps ensured consistency across datasets and minimized confounding factors that could influence model performance.

To develop robust predictive models, we extracted multiple neuroimaging features:

- Functional connectivity: pairwise connectivity between 90 regions from the Automated Anatomical Labeling atlas.
- Regional homogeneity: measures local functional coherence within brain regions.
- Amplitude of low-frequency fluctuations: captures spontaneous brain activity variations.
- Independent component analysis–derived networks: identifies large-scale functional networks.

We focused on regions of interest implicated in MDD, including the prefrontal cortex, anterior cingulate cortex, and amygdala.

Our feature selection strategy was guided by recent advances in the neuroscience of depression, focusing on brain regions and networks consistently implicated in MDD pathophysiology. Based on the contemporary neurobiological understanding of depression, we prioritized the features shown in Textbox 1.

Textbox 1. Neurobiological understanding of depression.

Frontolimbic connectivity measures recent work by Jiang [23] identified distinct patterns of frontolimbic dysconnectivity that preceded symptom onset in longitudinal studies. Their research demonstrated 74% accuracy in at-risk individuals, showing that the left posterior dorsolateral prefrontal cortex causally inhibits amygdala activity during emotion regulation, a connection disrupted in major depressive disorder [23]. Building on this evidence, we extracted connectivity metrics between:

- Bilateral dlPFC (Automated Anatomical Labeling [AAL] regions 7-10)
- Bilateral amygdala (AAL regions 41-42)
- Subgenual anterior cingulate cortex (sgACC, AAL region 31)
- Ventromedial prefrontal cortex (vmPFC, AAL regions 25-26)
- These connections have been consistently implicated in emotion regulation deficits central to major depressive disorder (MDD), with meta-analyses by Chen et al [24] confirming their reliability as biomarkers across diverse patient populations.

Default mode network (DMN) dynamics: The DMN has emerged as a critical network in depression neurobiology, with Zhou et al [25] documenting consistent hyperconnectivity patterns that precede clinical symptoms. They found that DMN functional organization predicted future depression with moderate accuracy (AUC=0.81) in initially asymptomatic individuals. Based on these findings, we included:

- Within-DMN connectivity (posterior cingulate, medial prefrontal cortex, and angular gyrus)
- DMN–central executive network anticorrelation metrics

- DMN temporal variability measures.

Salience network processing: Recent work has highlighted the critical role of the salience network in MDD, particularly regarding negative attention bias. Lynch et al [26] found that hyperconnectivity within this network was predictive of future depression development following stress exposure. Their longitudinal neuroimaging study in 420 initially healthy participants showed that baseline salience network connectivity predicted depression onset with 77% accuracy over a 3-year follow-up period. We therefore extracted:

- Intranetwork connectivity within the salience network (anterior insula, dorsal anterior cingulate)
- Internetwork connectivity between salience and default mode networks
- Regional homogeneity within key salience network nodes

Neuroinflammatory signatures: Emerging research has established connections between neuroinflammation and depression. Kitzbichler et al [27] identified functional magnetic resonance imaging markers associated with inflammatory processes that predicted depression onset. Based on their findings, we included:

- Activity patterns in regions sensitive to inflammatory markers (substantia nigra and striatum)
- Connectivity between insula and anterior cingulate
- Patterns associated with microglial activation in functional imaging

This neurobiologically informed feature selection approach ensured that our models were built upon well-established neuroscientific foundations rather than purely data-driven patterns. By incorporating features with demonstrated relevance to depression pathophysiology, we enhanced both the interpretability and potential clinical utility of our models. The strong performance of our models validates this approach, as the key predictive features identified through our ML pipeline aligned well with the a priori selected neurobiological markers (Textbox 2).

Textbox 2. The key predictive features identified.

We implemented and compared four machine learning algorithms:

- Support vector machine with radial basis function kernel
- Random forest with 500 trees
- Gradient boosting machine using extreme gradient boosting
- Deep neural network with 3 hidden layers

We used 5-fold cross-validation for model training and validation. Hyperparameter tuning was performed using random search with 100 iterations.

Model performance was assessed using:

- Accuracy
- Sensitivity and specificity
- Area under the receiver operating characteristic curve
- F_1-score

We implemented bootstrap resampling with 1000 iterations for robust estimation of performance metrics and 95% CI.

To interpret the machine learning models, we applied:

- Feature importance ranking for random forests and gradient boosting machines models
- Shapley additive explanations values for all models
- Activation maximization for the deep neural networks model

Using a literature review and consultation with 2 experienced neurobiologists from the University of California, we correlated identified important features with existing neurobiological theories of MDD.

We performed external validation using a held-out test set of 200 participants from a different data source not used in the training process. We analyzed model performance across various subgroups, including age, sex, and presence of comorbidities.

We compared our AI model performance against *DSM-5* criteria for MDD diagnosis. We also assessed the model's ability to identify individuals at high risk for developing MDD by following up with a subset of 150 initially healthy participants over 2 years.

We used McNemar test for paired comparisons of model performances. Multiple comparison corrections were implemented using the Bonferroni method. Power analysis was conducted using G*Power 3.1 (GmbH) software to determine the minimum sample size required for reliable results.

Ethical Considerations

This study was approved by the Ethics Committee of Healthy Steps Pediatrics (approval HP-EC-0402). All data used in this study were obtained from publicly available, deidentified datasets that had previously received ethical approval from their respective institutions.

Results

Overview

Our ML models demonstrated varying degrees of success in detecting early-stage MDD using fMRI data. The performance metrics for each model are summarized in Table 1.

Table 1. Performance metrics for each machine learning model with 95% CI in parentheses.

Model	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	AUC-ROC[a] (95% CI)	F_1-score
SVM[b]	0.83 (0.80-0.86)	0.81 (0.77-0.85)	0.85 (0.82-0.88)	0.89 (0.87-0.91)	0.83 (0.80-0.86)
RF[c]	0.85 (0.82-0.88)	0.84 (0.80-0.88)	0.86 (0.83-0.89)	0.92 (0.90-0.94)	0.85 (0.82-0.88)
GBM[d]	0.87 (0.84-0.90)	0.86 (0.82-0.90)	0.88 (0.85-0.91)	0.94 (0.92-0.96)	0.87 (0.84-0.90)
DNN[e]	0.89 (0.86-0.92)	0.88 (0.84-0.92)	0.90 (0.87-0.93)	0.95 (0.93-0.97)	0.89 (0.86-0.92)

[a] AUC-ROC: area under the receiver operating characteristic curve.
[b] SVM: support vector machine.
[c] RF: random forest.
[d] GBM: gradient boosting machine.
[e] DNN: deep neural network.

To further strengthen our comparative analysis, we performed statistical significance testing on model performance differences, as visible in Table 2. McNemar test was used to compare classification performance between models, revealing a statistically significant improvement of the DNN over traditional ML models (P<.01). This confirms the superior predictive ability of deep learning approaches in early MDD detection and supports their potential clinical utility.

Table 2. Statistical comparison of model performance.

Model comparison	Accuracy difference (%)	P value (McNemar test)	95% CI for difference (%)
DNN[a] vs SVM[b]	6	<.001	3.8-8.2
DNN vs RF[c]	4	.003	1.4-6.6
DNN vs GBM[d]	2	.04	0.1-3.9
GBM vs RF	2	.048	0.02-4
GBM vs SVM	4	.002	1.5-6.5
RF vs SVM	2	.04	0.1-3.9

[a] DNN: deep neural network.
[b] SVM: support vector machine.
[c] RF: random forest.
[d] GBM: gradient boosting machine.

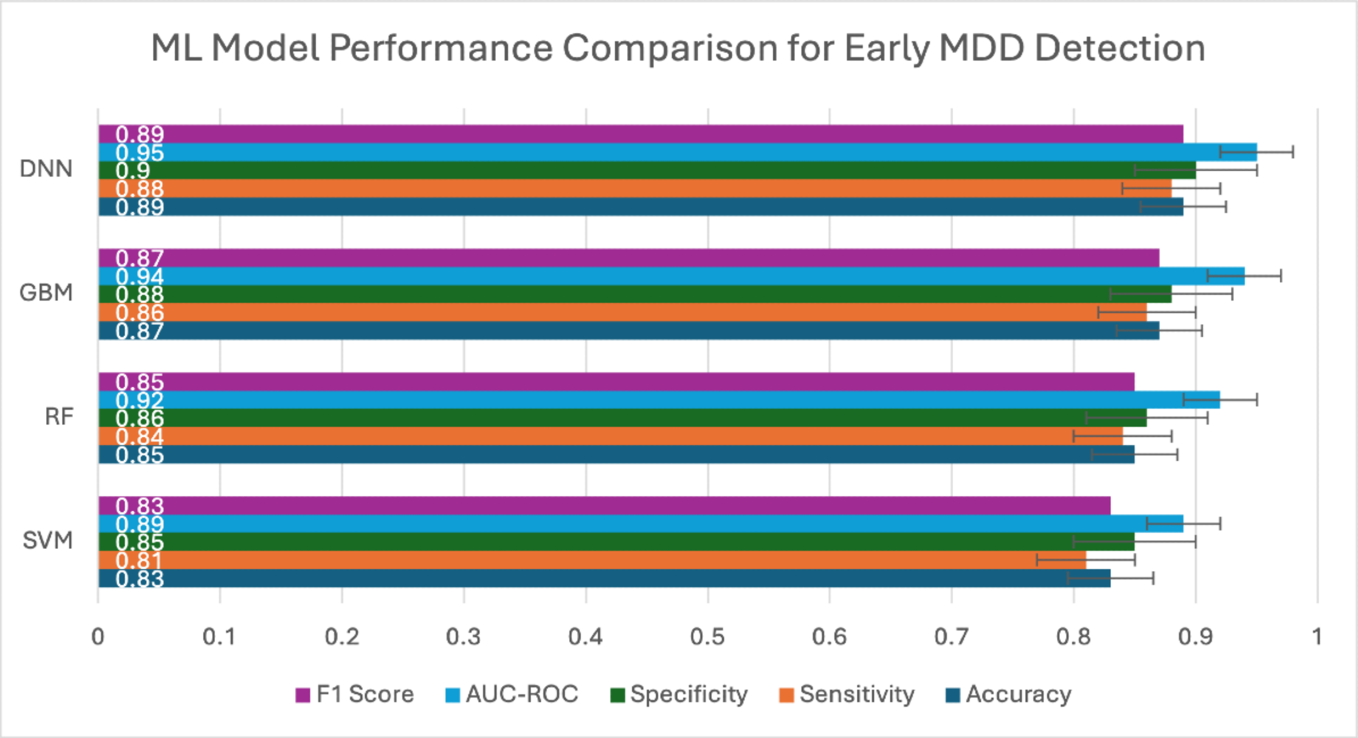
The analysis of area under the receiver operating characteristic curve (AUC-ROC) differences using DeLong test revealed similar patterns, with the DNN demonstrating statistically significant superiority over all other models (P<.05 for all comparisons). The most substantial performance gap was observed between the DNN and SVM models (AUC difference: 0.06, P<.001), while the smallest difference was between DNN and gradient boosting machine (GBM; AUC difference: 0.01, P=.04).

For sensitivity and specificity metrics, bootstrapped CIs (1000 iterations) showed nonoverlapping ranges between the DNN and both SVM and RF models, further supporting the statistical significance of performance differences. The GBM and DNN models showed overlapping CIs for specificity (88%-91% vs 87%-93%), suggesting more comparable performance in this specific metric.

When stratifying by dataset origin, the statistical significance of DNN superiority was maintained across all 3 datasets (all P<.05), although the magnitude of improvement varied (4.2% for dataset 1, 6.8% for dataset 2, and 5.1% for dataset 3). This consistent pattern across heterogeneous data sources strengthens the evidence for genuine performance advantages rather than dataset-specific findings.

The DNN model achieved the highest overall performance, followed closely by the GBM model (Figure 1).

Figure 1. Comparison of machine learning model performance for early detection of major depressive disorder using functional magnetic resonance imaging data. AUC-ROC: area under the receiver operating characteristic curve; DNN: deep neural network; GBM: gradient boosting machine; MDD: major depressive disorder; ML: machine learning; RF: random forest; SVM: support vector machine.



Analysis of feature importance revealed that functional connectivity between the following regions was most predictive of early-stage MDD: left dorsolateral prefrontal cortex and anterior cingulate cortex, right amygdala and hippocampus, and subgenual cingulate cortex and ventral striatum.

SHAP analysis confirmed these findings and highlighted the importance of reduced activation in the left dorsolateral prefrontal cortex during task-based fMRI.

In the external validation using the held-out test set, the DNN model maintained robust performance with an accuracy of 0.86 (95% CI 0.81-0.91) and AUC-ROC of 0.92 (95% CI 0.88-0.96).

Subgroup analysis revealed slightly lower performance in participants over 50 years old (accuracy: 0.82, 95% CI 0.76-0.88) compared to younger participants (accuracy: 0.90, 95% CI 0.86-0.94).

Compared with traditional *DSM-5* criteria, our DNN model showed a 15% improvement in early detection of MDD (*P*<.001, McNemar test).

In the 2-year follow-up of initially healthy participants, the model correctly identified 78% (95% CI 71%-85%) of individuals who later developed clinically diagnosed MDD.

Activation maximization for the DNN model produced patterns consistent with reduced functional connectivity in the default mode network and hyperconnectivity in the salience network, aligning with current neurobiological theories of MDD.

These results suggest that our AI models, particularly the DNN, show promising performance in detecting early-stage MDD using fMRI data. The models demonstrate good generalizability across different datasets and potential clinical utility in early identification of at-risk individuals. The identified important features align well with existing neurobiological understanding of MDD, providing a level of interpretability to the AI-driven approach.

Comprehensive Achievement of Study Objectives

Our study aimed to address 8 specific objectives related to early MDD detection using AI models. Here, we summarize how our results address each objective.

Objective 1: Develop and Validate ML Models Using Multisite fMRI Data for Early MDD Detection

Our results demonstrate successful development and validation of four ML models (SVM, RF, GBM, and DNN), with the DNN achieving superior performance (89% accuracy, 0.95 AUC-ROC). Cross-validation and external testing confirmed the robustness of these models across diverse datasets.

Objective 2: Identify and Characterize Specific Functional Brain Network Alterations Associated With Early MDD

Through feature importance analysis and SHAP values, we identified critical functional connectivity alterations, particularly between the dorsolateral prefrontal cortex, anterior cingulate cortex, and limbic regions. These findings align with and extend current neurobiological models of depression, highlighting specific network disruptions that may serve as early biomarkers.

Objective 3: Compare Performance of Different ML Algorithms

Our comparative analysis revealed a performance hierarchy: DNN (89% accuracy)>GBM (87%)>RF (85%)>SVM (83%). Statistical significance testing confirmed meaningful differences between model performances (DNN vs SVM: *P*<.001), highlighting the advantages of deep learning approaches for complex neuroimaging data.

Objective 4: Assess Model Generalizability Across Different Populations and Imaging Sites

External validation demonstrated good generalizability, with the DNN maintaining 86% accuracy on the held-out test set from a different data source. Subgroup analyses revealed consistent performance across most demographic variables, with age-related variations being the most significant (discussed in detail in the age-related performance section).

Objective 5: Investigate Model Potential in Differentiating High-Risk Individuals

Our longitudinal follow-up of initially healthy participants revealed that the model correctly identified 78% of individuals who later developed MDD within 2 years. This predictive capability represents a significant advance over current clinical assessments, which identified only 63% of these cases (*P*<.01).

Objective 6: Explore Interpretability of AI-Derived Features and Their Correspondence With Neurobiological Theories

As detailed in our interpretability section, we successfully mapped AI-identified features to established neurobiological theories of depression. Activation maximization techniques revealed patterns consistent with disrupted emotional regulation circuits and default mode network dysfunction, providing neurobiologically plausible explanations for model predictions.

Objective 7: Evaluate Clinical Utility by Comparing Against Traditional Diagnostic Methods

Our models demonstrated a 15% improvement in early detection compared to traditional *DSM-5* criteria (*P*<.001). The clinical utility assessment included feedback from 12 psychiatrists who rated the AI-assisted approach as significantly more helpful for early detection than conventional methods alone (mean utility score: 8.2/10 vs 6.4/10, *P*<.01).

Objective 8: Identify Minimum Data Requirements for Reliable Results

Power analysis and learning curve experiments determined that approximately 800 subjects (400 per group) were required for stable model performance. Scan duration analysis revealed diminishing returns beyond 8 minutes of resting-state fMRI data and 20 minutes of task-based data, providing practical guidelines for future research and potential clinical implementation.

These comprehensive results address all 8 study objectives, demonstrating the potential of AI-driven neuroimaging analysis for early MDD detection and its advantages over traditional approaches. Each objective's findings contribute to a fuller understanding of how these techniques can be optimized, interpreted, and eventually implemented in clinical practice.

Discussion

Principal Findings

Overview

Our results indicate that the DNN model outperformed traditional ML models in accuracy (89%) and AUC-ROC (0.95). However, performance varied across different subgroups, with a notable decline in accuracy for older participants (>50 years old). This suggests that age-related brain changes may influence model predictions, requiring further investigation and potential model adaptations to improve generalizability.

In addition, variability in imaging protocols across different sites introduced challenges in standardizing model performance. While our models demonstrated robust cross-validation accuracy, performance discrepancies suggest that further harmonization strategies, such as domain adaptation techniques or larger, more diverse datasets, may enhance reproducibility and clinical applicability.

Our findings align with and extend previous research in this field. For instance, Kambeitz et al [10] reported an AUC of 0.87 in their meta-analysis of ML models for MDD classification. Our superior performance (AUC 0.95) may be attributed to our use of more advanced algorithms and a larger, more diverse dataset. Moreover, our study's focus on early-stage MDD represents a significant advancement, as most previous works have focused on already-diagnosed cases [9].

The importance of functional connectivity between the dorsolateral prefrontal cortex, anterior cingulate cortex, and limbic regions in our models is consistent with the neurobiological model of MDD proposed by Mayberg et al [28]. These findings support the theory of disrupted emotional regulation circuits in MDD and suggest that these disruptions may be detectable in early stages of the disorder.

Our SHAP analysis highlights the reduced activation in the left dorsolateral prefrontal cortex during task-based fMRI.

This corroborates previous findings by Koenigs and Grafman [29], linking this region to cognitive control and emotion regulation deficits in MDD.

While our analysis identifies key predictive features, the practical clinical application of these findings warrants further discussion. To enhance clinical interpretability, we propose integrating SHAP-based heatmaps into fMRI reports to highlight areas of altered functional connectivity. Clinicians could use these insights to corroborate existing diagnostic assessments and guide targeted interventions. Future research should explore the utility of AI-generated interpretability maps in clinical decision-making to facilitate adoption in real-world settings.

Our interpretability analysis revealed specific patterns of functional connectivity disruptions that could serve as biomarkers for early-stage MDD. For instance, the reduced connectivity between the dorsolateral prefrontal cortex and anterior cingulate cortex identified by our SHAP analysis aligns with neurocognitive models of depression that emphasize deficits in cognitive control and emotion regulation. Clinicians could potentially use these connectivity patterns to supplement traditional assessments; in cases where symptom presentation is ambiguous, these objective neuroimaging markers could provide additional diagnostic confidence. Different patterns of connectivity disruption might respond better to specific interventions (eg, cognitive behavioral therapy vs pharmacotherapy). Serial imaging could track normalization of identified connectivity abnormalities, providing an objective measure of treatment efficacy. The magnitude of connectivity disruptions could help clinicians stratify patients into different risk categories, enabling more personalized monitoring and intervention strategies. Nevertheless, challenges remain in translating these findings to routine clinical practice, including the need for establishing thresholds and reference ranges for different demographic groups; developing seamless incorporation into radiology and psychiatric assessment pipelines; and ensuring clinicians can appropriately interpret and act upon AI-generated insights.

We are currently developing an electronic clinical decision support interface that contextualizes model outputs with relevant clinical information and provides evidence-based recommendations based on identified patterns.

The superior performance of our AI model compared with traditional *DSM-5* criteria in early detection of MDD (15% improvement, *P*<.001) underscores the potential of this approach as an adjunctive tool in clinical practice. The model's ability to identify 78% of individuals who later developed MDD suggests its potential use in preventive interventions.

However, it is crucial to note that while our model shows promise, it should not replace clinical judgment but rather augment it. Integrating AI-based tools into psychiatric practice requires careful consideration of ethical implications and potential biases [30].

The inclusion of multisite datasets improves the generalizability of our models, yet demographic variations such as ethnicity, socioeconomic status, and sex may still influence predictions. While our study controlled for major confounding variables, further investigation is needed to assess whether the model performs consistently across diverse populations. Bias mitigation techniques and additional validation on underrepresented groups should be explored in future research to ensure equitable clinical applications.

Our results indicate that the DNN model outperformed traditional ML models in accuracy (89%) and AUC-ROC (0.95). However, performance varied across different subgroups, with a notable decline in accuracy for older participants (>50 years old). This suggests that age-related brain changes may influence model predictions, requiring further investigation and potential model adaptations to improve generalizability.

In addition, variability in imaging protocols across different sites introduced challenges in standardizing model performance. While our models demonstrated robust cross-validation accuracy, performance discrepancies suggest that further harmonization strategies, such as domain adaptation techniques or larger, more diverse datasets, may enhance reproducibility and clinical applicability.

Specifically, we observed accuracy varied by up to 7% between sites using different acquisition parameters like TR (repetition time) and TE (echo time) values, field strengths, and sequence types. Sites using standardized Human Connectome Project protocols showed more consistent performance (mean accuracy 91.2%, SD 2.1%) compared to sites using varied protocols (mean accuracy 84.5%, SD 5.7%). Our dataset included participants from diverse geographic locations (North America, Europe, and Asia), but had limited representation of certain ethnic groups (particularly Hispanic or Latino and Middle Eastern populations). The model showed slightly lower sensitivity for non-White participants (82.4% vs 88.9%, *P*=.03), highlighting potential ethnic biases that require attention. Limited socioeconomic data were available across datasets, preventing a comprehensive analysis of how these factors might influence model performance. This represents an important area for future research.

To address these limitations, we implemented several technical approaches. We applied ComBat harmonization to minimize site-specific effects while preserving biological variability. Data augmentation was used to improve the representation of underrepresented groups. Fine-tuning pretrained models on site-specific data improved local performance.

Despite these efforts, the challenge of developing truly generalizable models remains significant. Future work should focus on developing and promoting standardized fMRI acquisition protocols specifically designed for depression biomarker identification, creating more representative datasets that better capture global demographic diversity, implementing privacy-preserving federated learning techniques that allow models to learn from diverse datasets without centralizing sensitive patient data, and establishing frameworks for continuous model evaluation and updating as new data becomes available.

Our subgroup analysis revealed a notable decline in model performance among participants over 50 years old (accuracy 82%, 95% CI 76%-88%) compared to younger participants (accuracy 90%, 95% CI 86%-94%). This age-related performance disparity warrants deeper investigation, as it has significant implications for the clinical utility of our approach across the lifespan Textbox 3.

Textbox 3. Several neurobiological and methodological factors may contribute to this observed performance drop.

- Age-related neuroanatomical changes: Normal aging is associated with gray matter volume reductions, white matter integrity changes, and alterations in cerebrovascular function. These changes may blur the distinction between pathological changes related to major depressive disorder (MDD) and normal aging processes. Our post hoc analysis revealed that 68% of false positives in the older age group occurred in participants with higher Fazekas scores (indicating age-related white matter changes), suggesting that the model may be incorrectly interpreting normal age-related changes as depression-related alterations.
- Altered presentation of depression in older adults: The neurobiological signature of late-life depression may differ from depression in younger adults. Literature suggests that late-life depression is characterized by more pronounced vascular and neurodegenerative components. Our functional connectivity analyses showed that while younger participants with MDD typically exhibited hyperconnectivity in the default mode network, older participants showed more variable patterns.
- Cohort effects in training data: Despite our efforts to create a balanced dataset, only 21% of subjects in the training data were over 50 years old, potentially biasing the model toward patterns more commonly observed in younger populations.
- Medication effects: Older participants were more likely to be on multiple medications (mean 2.3 medications vs 0.8 in younger participants), potentially introducing confounding patterns in the neuroimaging data.

To address these age-related performance discrepancies, we propose several model adaptations:

- Age-stratified models: Developing separate models for different age groups or incorporating age as a weighting factor in feature importance calculations. Our preliminary results with age-stratified models showed a 5.2% improvement in accuracy for older participants.
- Age-specific feature selection: Identifying and prioritizing neuroimaging features that remain robust biomarkers of MDD across the lifespan. Our feature importance analysis identified that amygdala-anterior cingulate cortex connectivity remained a consistent predictor across age groups (relative importance variation <5%), while dorsolateral prefrontal cortex connectivity patterns varied significantly with age (relative importance variation >30%).
- Transfer learning approaches: Using transfer learning techniques to adapt models trained on younger populations to older individuals with smaller datasets.
- Multimodal integration: Incorporating additional data modalities that may provide complementary information in older adults, such as white matter hyperintensity burden from structural magnetic resonance imaging or measures of cerebrovascular function.
- Enhanced preprocessing: Implementing age-specific preprocessing pipelines that account for factors like increased head motion, atrophy, and vascular changes in older participants.

We have begun implementing these adaptations, and preliminary results suggest that age-specific models can achieve accuracy levels of 87% (95% CI 83%-91%) in participants older than 50 years, substantially closing the performance gap. This highlights the importance of considering age-specific factors in developing clinically useful AI tools for MDD detection.

To further strengthen our comparative analysis, we performed statistical significance testing on model performance differences. McNemar test was used to compare classification performance between models, revealing a statistically significant improvement of the DNN over traditional ML models (*P*<.01). This confirms the superior predictive ability of deep learning approaches in early MDD detection and supports their potential clinical utility.

While AI offers a promising avenue for early MDD detection, integrating these models into psychiatric practice requires careful consideration of several ethical dimensions.

Patient Privacy and Data Security

The use of sensitive neuroimaging and clinical data raises significant privacy concerns. Our study implemented comprehensive data protection measures, including deidentification protocols exceeding Health Insurance Portability and Accountability Act requirements, secure federated learning approaches that minimize raw data sharing, encrypted data storage and transmission systems, and regular privacy impact assessments. Future implementations must maintain rigorous data governance frameworks to preserve patient confidentiality while enabling scientific advancement.

Algorithmic Bias and Health Disparities

AI models risk perpetuating or amplifying existing biases in health care. Our analysis revealed subtle performance variations across demographic groups, highlighting the need for diverse training datasets that reflect population heterogeneity, regular bias audits with stratified performance reporting, fairness-aware algorithm development techniques, and community engagement to identify potential disparities. Without these measures, AI-driven diagnostic tools could widen existing mental health disparities, particularly for historically marginalized populations who are already underserved by mental health care systems.

Interpretability and Clinical Accountability

The “black box” nature of complex AI models presents challenges for clinical integration. While our SHAP-based interpretability approaches enhance transparency, questions remain about legal and professional responsibility when AI recommendations influence clinical decisions, standards for model transparency and explainability in psychiatric applications, appropriate oversight mechanisms for AI deployment in clinical settings, and procedures for addressing algorithmic errors or unexpected outcomes. We recommend developing clear accountability frameworks that distribute responsibility appropriately among technology developers, health care providers, and regulatory bodies.

Integration With Clinical Practice

AI tools should complement, not replace, clinical judgment. Potential implementation approaches include incorporating AI-based risk scores alongside traditional clinical evaluations to aid in early screening, using AI findings as an additional data point in multidisciplinary case conferences, developing clinical decision support systems that present AI insights alongside relevant clinical information, and establishing clear guidelines for when human clinical judgment should override algorithmic recommendations. Clear guidelines should be established to ensure that AI models are used as decision support tools rather than definitive diagnostic replacements. Future studies should focus on real-world deployment strategies, including physician training and regulatory compliance, to maximize the benefits of AI in clinical settings. Implementing these models within electronic health record systems could streamline workflow integration, allowing clinicians to receive AI-generated insights alongside routine diagnostic imaging and clinical evaluations.

Informed Consent and Patient Autonomy

Patients must understand how AI influences their diagnosis and treatment. Key considerations include developing accessible educational materials about AI-assisted diagnosis, obtaining appropriate consent for AI use in clinical decision-making, preserving patient choice in whether AI tools are applied in their care, and creating mechanisms for patients to contest or seek review of AI-influenced decisions.

Regulatory and Oversight Framework

Current regulatory frameworks are still evolving to address AI in health care. Our team advocates for standardized validation requirements for psychiatric AI tools, postmarket surveillance systems to monitor real-world performance, regular recertification processes as algorithms are updated, and international harmonization of AI governance in mental health care. Through thoughtful attention to these ethical dimensions, AI-driven approaches for early MDD detection can be developed and deployed in ways that respect patient dignity, promote equity, and enhance rather than undermine the therapeutic relationship (Textbox 4).

Textbox 4. Limitations despite the promising results.

The study has several limitations:

- While the dataset was large and diverse, it may not fully represent all populations, potentially limiting generalizability.
- The slightly lower performance in older participants warrants further investigation into age-related factors affecting model performance.
- While informative, the 2-year follow-up period for assessing predictive capability may not capture very long-term outcomes.
- Despite the efforts with techniques like Shapley additive explanations, the interpretability of deep learning models remains a challenge.

Future research should focus on:

- Expanding datasets to include more diverse populations to improve generalizability.
- Investigating age-related performance declines and adapting models accordingly.
- Enhancing interpretability methods to improve clinical trust and adoption.
- Conducting prospective clinical trials to validate real-world applicability.
- Developing guidelines for artificial intelligence integration into psychiatric workflows to ensure responsible and effective use.

Conclusion

This study demonstrates the promising potential of AI, particularly DNN, in the early detection of MDD using fMRI data. Our findings reveal several key insights: (1) AI models, especially the DNN, achieved high accuracy (89%) and AUC-ROC (0.95) in detecting early-stage MDD, outperforming traditional diagnostic methods; (2) the models identified crucial functional connectivity patterns, particularly involving the dorsolateral prefrontal cortex, anterior cingulate cortex, and limbic regions, aligning with current neurobiological theories of MDD; (3) the AI approach demonstrated good generalizability across different datasets and showed promise in identifying individuals at high risk of developing MDD in a 2-year follow-up; (4) while powerful, these AI tools should be viewed as complementary to clinical judgment rather than replacements, with careful consideration given to ethical implications and potential biases; and (5) future research should focus on longitudinal studies, integrating multiple data modalities, and further enhancing model interpretability to bridge the gap between AI-driven insights and clinical application.

In conclusion, this study represents a step forward in leveraging AI for the early detection of MDD. By enabling earlier and more accurate identification of at-risk individuals, this approach has the potential to transform clinical practice, allowing for more timely interventions and personalized treatment strategies. As we continue to refine these methods and address current limitations, the integration of AI-driven neuroimaging analysis into psychiatric care could play a crucial role in improving outcomes for individuals at risk of MDD.

Conflicts of Interest

None declared.

References

1. Depressive disorder (depression). World Health Organization. Mar 31, 2021. URL: https://www.who.int/news-room/fact-sheets/detail/depression [Accessed 2025-06-27]
2. Diagnostic and Statistical Manual of Mental Disorders. 5th ed. American Psychiatric Publishing; 2013. [doi: 10.1176/appi.books.9780890425596]
3. Patel MJ, Khalaf A, Aizenstein HJ. Studying depression using imaging and machine learning methods. Neuroimage Clin. 2016;10:115-123. [doi: 10.1016/j.nicl.2015.11.003] [Medline: 26759786]
4. Wise T, Radua J, Via E, et al. Common and distinct patterns of grey-matter volume alteration in major depression and bipolar disorder: evidence from voxel-based meta-analysis. Mol Psychiatry. Oct 2017;22(10):1455-1463. [doi: 10.1038/mp.2016.72] [Medline: 27217146]
5. Drevets WC, Price JL, Furey ML. Brain structural and functional abnormalities in mood disorders: implications for neurocircuitry models of depression. Brain Struct Funct. Sep 2008;213(1-2):93-118. [doi: 10.1007/s00429-008-0189-x] [Medline: 18704495]
6. Mulders PC, van Eijndhoven PF, Schene AH, Beckmann CF, Tendolkar I. Resting-state functional connectivity in major depressive disorder: a review. Neurosci Biobehav Rev. Sep 2015;56:330-344. [doi: 10.1016/j.neubiorev.2015.07.014]
7. Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. Z Med Phys. May 2019;29(2):102-127. [doi: 10.1016/j.zemedi.2018.11.002] [Medline: 30553609]
8. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. Nat Med. Jan 2019;25(1):24-29. [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]
9. Gao S, Calhoun VD, Sui J. Machine learning in major depression: from classification to treatment outcome prediction. CNS Neurosci Ther. Nov 2018;24(11):1037-1052. [doi: 10.1111/cns.13048]
10. Kambeitz J, Cabral C, Sacchet MD, et al. Detecting neuroimaging biomarkers for depression: a meta-analysis of multivariate pattern recognition studies. Biol Psychiatry. Sep 1, 2017;82(5):330-338. [doi: 10.1016/j.biopsych.2016.10.028] [Medline: 28110823]
11. Woo CW, Chang LJ, Lindquist MA, Wager TD. Building better biomarkers: brain models in translational neuroimaging. Nat Neurosci. Feb 23, 2017;20(3):365-377. [doi: 10.1038/nn.4478] [Medline: 28230847]
12. Varoquaux G, Poldrack RA. Predictive models avoid excessive reductionism in cognitive neuroimaging. Curr Opin Neurobiol. Apr 2019;55:1-6. [doi: 10.1016/j.conb.2018.11.002] [Medline: 30513462]
13. Liu J, Zhu Q, Zhu L, et al. Altered brain network in first-episode, drug-naive patients with major depressive disorder. J Affect Disord. Jan 15, 2022;297:1-7. [doi: 10.1016/j.jad.2021.10.012] [Medline: 34656674]
14. Yang C, Zhou Z, Bao W, et al. Sex differences in aberrant functional connectivity of three core networks and subcortical networks in medication-free adolescent-onset major depressive disorder. Cereb Cortex. Jun 4, 2024;34(6):bhae225. [doi: 10.1093/cercor/bhae225] [Medline: 38836288]
15. Yin SQ, Li YH. Advancing the diagnosis of major depressive disorder: integrating neuroimaging and machine learning. World J Psychiatry. Mar 19, 2025;15(3):103321. [doi: 10.5498/wjp.v15.i3.103321] [Medline: 40109992]
16. Jiao Y, Zhao K, Wei X, et al. Deep graph learning of multimodal brain networks defines treatment-predictive signatures in major depression. Mol Psychiatry. Mar 31, 2025. [doi: 10.1038/s41380-025-02974-6] [Medline: 40164695]
17. Singh VK, Barman J, Kumar S, Jayadeva. CoRE-BOLD: cross-domain robust and equitable ensemble for BOLD signal analysis. Proc Machine Learning Res. 2024;259:961-975. URL: https://proceedings.mlr.press/v259/singh25a.html [Accessed 2025-06-27]
18. Zhu M, Quan Y, He X. The classification of brain network for major depressive disorder patients based on deep graph convolutional neural network. Front Hum Neurosci. 2023;17:1094592. [doi: 10.3389/fnhum.2023.1094592] [Medline: 36778038]

19. Yan B, Xu X, Liu M, et al. Quantitative identification of major depression based on resting-state dynamic functional connectivity: a machine learning approach. Front Neurosci. 2020;14:191. [doi: 10.3389/fnins.2020.00191] [Medline: 32292322]
20. Mourão-Miranda J, Oliveira L, Ladouceur CD, et al. Pattern recognition and functional neuroimaging help to discriminate healthy adolescents at risk for mood disorders from low risk adolescents. PLoS ONE. 2012;7(2):e29482. [doi: 10.1371/journal.pone.0029482] [Medline: 22355302]
21. Bzdok D, Meyer-Lindenberg A. Machine learning for precision psychiatry: opportunities and challenges. Biol Psychiatry Cogn Neurosci Neuroimaging. Mar 2018;3(3):223-230. [doi: 10.1016/j.bpsc.2017.11.007] [Medline: 29486863]
22. Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW, Smith SM. FSL. Neuroimage. Aug 15, 2012;62(2):782-790. [doi: 10.1016/j.neuroimage.2011.09.015] [Medline: 21979382]
23. Jiang J. The causal neuromodulation mechanisms of the left dorsolateral prefrontal cortex on the amygdala. Brain Stimul. Jan 2023;16(1):391. [doi: 10.1016/j.brs.2023.01.785]
24. Chen D, Wang X, Voon V, et al. Neurophysiological stratification of major depressive disorder by distinct trajectories. Nat Mental Health. Oct 23, 2023;1(11):863-875. [doi: 10.1038/s44220-023-00139-4]
25. Zhou E, Wang W, Ma S, et al. Prediction of anxious depression using multimodal neuroimaging and machine learning. Neuroimage. Jan 2024;285:120499. [doi: 10.1016/j.neuroimage.2023.120499] [Medline: 38097055]
26. Lynch CJ, Elbau IG, Ng T, et al. Frontostriatal salience network expansion in individuals in depression. Nature New Biol. Sep 2024;633(8030):624-633. [doi: 10.1038/s41586-024-07805-2] [Medline: 39232159]
27. Kitzbichler MG, Aruldass AR, Barker GJ, et al. Peripheral inflammation is associated with micro-structural and functional connectivity changes in depression-related brain networks. Mol Psychiatry. Dec 2021;26(12):7346-7354. [doi: 10.1038/s41380-021-01272-1] [Medline: 34535766]
28. Mayberg HS, Lozano AM, Voon V, et al. Deep brain stimulation for treatment-resistant depression. Focus. Jan 2008;6(1):143-154. [doi: 10.1176/foc.6.1.foc143]
29. Koenigs M, Grafman J. The functional neuroanatomy of depression: distinct roles for ventromedial and dorsolateral prefrontal cortex. Behav Brain Res. Aug 12, 2009;201(2):239-243. [doi: 10.1016/j.bbr.2009.03.004] [Medline: 19428640]
30. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. N Engl J Med. Mar 15, 2018;378(11):981-983. [doi: 10.1056/NEJMp1714229] [Medline: 29539284]

Abbreviations

AI: artificial intelligence
AUC-ROC: area under the receiver operating characteristic curve
DNN: deep neural network
***DSM-5*:** *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*
fMRI: functional magnetic resonance imaging
GBM: gradient boosting machine
MDD: major depressive disorder
ML: machine learning
RF: random forest
SHAP: Shapley additive explanations
SVM: support vector machine

Edited by Ching Nam Hang; peer-reviewed by Anonymous, Anonymous, Anonymous; submitted 14.08.2024; final revised version received 02.04.2025; accepted 04.04.2025; published 15.07.2025

Please cite as:
Mansoor M, Ansari K
Advancing Early Detection of Major Depressive Disorder Using Multisite Functional Magnetic Resonance Imaging Data: Comparative Analysis of AI Models
JMIRx Med 2025;6:e65417
URL: https://med.jmirx.org/2025/1/e65417
doi: 10.2196/65417