Mapping heterogeneity in the neuroanatomical correlates of depression

Devon Watts[1], Travis T Mallard[1,2], Lorenza Dall' Aglio[1], Evan Giangrande[3], Chris Kennedy[1-2], Na Cai[4], Karmel Choi[1,2], Tian Ge[2,3], Jordan W. Smoller, MD, ScD[1,2,3]

Affiliations

1. Center for Precision Psychiatry, Department of Psychiatry, Massachusetts General Hospital. Boston, MA, USA
2. Department of Psychiatry, Harvard Medical School, Boston, MA, USA
3. Center for Genomic Medicine, Mass General Brigham, Boston, MA, USA
4. Department of Biosystems and Engineering, ETH Zurich, Basel, Switzerland

Words: 9509 (**6001** - excluding Online Methods)

Tables: 5

Figures: 7

Target Journals: Nature Mental Health, Nature Human Behavior

Abstract

Major depressive disorder (MDD) affects millions worldwide, yet its neurobiological underpinnings remain elusive. Neuroimaging studies have yielded inconsistent results, hindered by small sample sizes and heterogeneous depression definitions. We sought to address these limitations by leveraging the UK Biobank's extensive neuroimaging data (n=30,122) to investigate how depression phenotyping depth influences neuroanatomic profiles of MDD. We examined 256 brain structural features, obtained from T1- and diffusion-weighted brain imaging, and nine depression phenotypes, ranging from self-reported symptoms (shallow definitions) to clinical diagnoses (deep). Multivariable logistic regression, machine learning classifiers, and feature transfer approaches were used to explore correlational patterns, predictive accuracy and the transferability of important features across depression definitions. For white matter microstructure, we observed widespread fractional anisotropy decreases and mean diffusivity increases. In contrast, cortical thickness and surface area were less consistently associated across depression definitions, and demonstrated weaker associations. Machine learning classifiers showed varying performance in distinguishing depression cases from controls, with shallow phenotypes achieving similar discriminative performance (AUC=0.807) and slightly higher positive predictive value (PPV=0.655) compared to deep phenotypes (AUC=0.831, PPV=0.456), when sensitivity was standardized at 80%. However, when shallow phenotypes were downsampled to match deep phenotype case/control ratios, performance degraded substantially (AUC=0.690). Together, these results suggest that while core white-matter alterations are shared across phenotyping strategies, shallow phenotypes require approximately twice the sample size of deep phenotypes to achieve comparable classification performance, underscoring the fundamental power-specificity tradeoff in psychiatric neuroimaging research.

Introduction

Major depressive disorder (MDD) is a leading cause of disability worldwide, affecting over 300 million people[1]. Despite decades of research[2], the neurobiological underpinnings of MDD remain poorly understood. While neuroimaging studies have attempted to elucidate the brain basis of depression (e.g., collaborative efforts from the ENIGMA consortium[3]), findings have varied considerably and often lack replication. This inability to identify robust correlates of MDD poses a significant challenge for advancing our understanding of its pathophysiology, ultimately impeding our ability to improve patient outcomes. Recent research suggests that inconsistent results may be partially explained by limited statistical power. For example, Winter and colleagues[4] demonstrated that brain structural and functional differences between individuals with MDD and healthy controls were remarkably small. This underscores the challenge of detecting subtle but important brain-behavior relationships in MDD, highlighting the need for larger sample sizes to increase precision[5-6]. While addressing the issue of sample size will likely require large-scale, collaborative efforts over time, another major contributor to inconsistent findings that may be more readily addressed is phenotypic heterogeneity[7-8].

Across subfields of psychiatry, differences in how depression is measured and defined not only contribute to inconsistent findings but also shape the scope of biological discoveries. In the field of psychiatric genetics, Cai and colleagues[9] demonstrated that eight different definitions of depression, ranging from "shallow" self-report measures to "deep" clinical diagnoses, yielded varying heritability and genetic correlates. Similarly, in psychiatric neuroimaging, researchers have shown that phenotypic heterogeneity complicates efforts to identify reliable neurobiological markers. For instance, Harris and colleagues[8,10] found that brain structural anatomy and connectivity correlates of depression varied depending on whether cases were defined with self-reports, questionnaire-derived 'probable' depression, or diagnostic phenotyping. Although the authors expected to find greater effect sizes for more in-depth phenotypes, they found stronger associations between depression and cortical thickness when using a minimal phenotyping approach.

While these findings represent important advances, they also highlight key gaps in our understanding of depression heterogeneity. First, it is unclear how associations with imaging-derived phenotypes may vary across a broader range of depression phenotypes, particularly those defined using more granular approaches[9]. Second, the degree to which brain-based predictive models of MDD are impacted by phenotyping considerations has yet to be explored. Understanding whether the discriminatory power of imaging-derived phenotypes (IDPs) depends on the measurement of MDD – and whether models built on "shallow" definitions can generalize to "deeper" definitions and vice versa – could provide critical insights for the clinical translation of psychiatric neuroimaging.

Here, we draw upon large-scale data from the UK Biobank (UKBB) to investigate associations between 256 IDPs and nine depression phenotypes of varying depth in 30,122 adults. These

phenotypes, as defined by Cai et al[9] and detailed in Table 1, range from minimal self-reported measures to more clinically validated diagnoses. The phenotypes are stratified into two categories: shallow phenotypes based primarily on self-reporting and help-seeking behavior, and deep phenotypes reflecting stricter or clinically confirmed definitions of depression. By characterizing associations with a diverse set of IDPs (white matter microstructure, gray matter morphology, subcortical volume, and global volume), we systematically evaluate which structural features are consistently linked to depression. To contextualize these findings, we compare our results with meta-analytic results from the ENIGMA consortium, providing insight into the robustness and specificity of neuroanatomical alterations observed for MDD relative to other forms of serious mental illness. In addition to charting neuroanatomical associations, we use advanced machine learning techniques to evaluate the predictive power of neuroimaging features and whether this depends on the depth of depression phenotyping. Finally, via a novel feature transfer approach, we assess the generalizability of predictive neuroanatomical features across depression phenotypes. Collectively, this integrative framework enables us to identify both shared and distinct neural correlates across depression phenotypes, bridging the gap between clinical heterogeneity and neurobiological findings in MDD.

Results

1. White matter microstructure shows consistent alterations across depression phenotypes

We investigated the neuroanatomical correlates of depression across various phenotyping strategies with a series of multivariable logistic regression models, each including one of 256 z-score transformed brain structural variables as the primary predictor, alongside covariates (age, sex, intracranial volume, assessment center, and head motion). We derived odds ratios (ORs) for depression in association with brain white matter microstructure (fractional anisotropy and mean diffusivity at major fiber bundles), gray matter morphometry (cortical thickness and surface area), subcortical volumes, and global volumetric measures. All p-values reported were adjusted for multiple comparisons using the False Discovery Rate (FDR) method across all 256 tests. For detailed information on each variable, refer to **Supplementary Table S1**. Depression operationalization across phenotypes is described in **Table 1** and in the Methods section.

All brain structural correlates across IDPs and depression phenotypes are summarized in **Supplementary Table S2**. To illustrate these associations, we focused on specific examples comparing shallow (e.g. GPPsy and SelfRepDep) and deep (e.g. LifetimeMDD and MDDRecur) phenotypes as defined by Cai and colleagues[9], highlighting their most robust associations and the number of significant regions identified. As shown in **Supplementary Table S2**, white matter microstructure in fractional anisotropy and mean diffusivity demonstrated the most pronounced associations with depression status across phenotypes. The spatial distribution of associations across regions for each IDP is shown in **Supplementary Figure S1**. Fractional anisotropy measures predominantly showed reduced values across depression phenotypes, with 85.88% of odds ratios falling below 1.0. Values ranged from 0.781 to 1.098 (median: 0.949, mean: 0.946, SD: 0.052). The most robust associations in fractional anisotropy were observed in major white

matter bundles, including the corpus callosum, corona radiata, and principal association fibers. For instance, in the GPpsy phenotype, which represents individuals who reported seeking help for depression from a general practitioner[9], fractional anisotropy was significantly lower in 31 white matter tracts (FDR-adjusted $p \leq 0.05$) out of the 48 examined. The strongest association was found in the superior fronto-occipital fasciculus (right), with an OR of 0.915 (95% CI: 0.891-0.939, FDR-adjusted p = 6.65e-09): For every standard deviation decrease in fractional anisotropy in this region, the odds of belonging to the GPpsy group (vs. healthy controls) increased by 8.5%, after adjusting for covariates. By comparison, for the MDDRecur phenotype, representing individuals with recurrent lifetime MDD, fractional anisotropy was significantly lower in 11 white matter tracts (FDR-adjusted $p \leq 0.05$), with the strongest association also being in the superior fronto-occipital fasciculus (right) (OR = 0.871, 95% CI; 0.824-0.921, FDR-adjusted p = 1.07e-04).

Complementing the fractional anisotropy findings, mean diffusivity showed widespread increases associated with depression across phenotypes, with 75.23% of the odds ratios falling above 1, indicating higher diffusivity in depression cases, compared to controls (Figure 1B, top middle). Mean diffusivity measures also demonstrated a broader range of associations compared to fractional anisotropy, with odds ratios spanning 0.870 to 1.282 (median: 1.032 mean: 1.042, SD: 0.066). The SelfRepDep phenotype, which includes individuals who self-reported depression during a nurse-led interview, exhibited significantly higher mean diffusivity in 36 white matter tracts (FDR-adjusted $p \leq 0.05$) out of the 48 examined. The strongest association was observed in the anterior limb of the internal capsule (left), with an OR of 1.192 (95% CI: 1.134-1.251, FDR-adjusted p = 4.37e-10). By comparison, MDDRecur demonstrated significantly higher mean diffusivity in 15 white matter tracts, with the most pronounced effect in the superior corona radiata (left) (OR = 1.153, 95% CI: 1.088-1.220), FDR-adjusted p = 1.07e-04).

In contrast to the widespread associations for white matter microstructure, cortical thickness and surface area measures generally showed few significant associations with depression (**Supplementary Figure S1, Supplementary Table S2)**. Cortical thickness odds ratios (ORs) ranged from 0.862 to 1.169 (median = 0.990, mean = 0.994, SD = 0.008), whereas surface area measures displayed a similarly narrow range (ORs 0.824–1.133; median = 0.981, mean = 0.991, SD = 0.019). Across depression phenotypes, significant associations in surface area were limited to GPpsy (4 regions), Psypsy (2 regions), and ICD10Dep (1 region). In GPpsy, representing individuals who sought help from their general practitioner for depressive symptoms, surface area was significantly associated with depression in 4 out of 68 cortical regions (FDR-adjusted $p \leq 0.05$; ORs 0.961–1.044), and cortical thickness showed a parallel pattern in 4 out of 68 regions (FDR-adjusted $p \leq 0.05$; ORs 0.945–1.036). For LifetimeMDD, defined by DSM-5 criteria for lifetime MDD, cortical thickness emerged as significant in 4 out of 68 regions, highlighted by the most robust association in the left rostral anterior cingulate (OR = 0.925, 95% CI 0.885–0.966, FDR-adjusted p = 0.01), whereas surface area yielded no significant associations.

Overall, MDDRecur exhibited the largest number of significant thickness associations (6 regions), followed by GPpsy, SelfRepDep, and LifetimeMDD (4 regions each), ICD10Dep (2 regions), and Psypsy and ICD10DepExclude (1 region each), with DepAll showing no significant cortical findings. In sum, across phenotyping approaches, fractional anisotropy and mean diffusivity showed stronger and more consistent associations with depression, whereas cortical thickness and surface area yielded fewer and more variable results. These findings suggest that white matter microstructure is more consistently implicated in depression than cortical morphology, irrespective of how depression is defined. Further details regarding uncorrected and FDR-adjusted p-values alongside odds ratios and confidence intervals for each cortical region across depression phenotypes can be found in Supplementary Table 2.

1.1. Self-reported depression shows stronger white-matter microstructural alterations than clinical definitions

To assess whether the strength of associations between IDPs and depression varies across phenotypic definitions, we conducted pairwise comparisons of ORs using the delta method. This calculates the variance of the difference in log ORs, accounting for the uncertainty in estimates [11,12]. We focused on significant IDPs across pairs of depression phenotypes (FDR-adjusted $p < 0.05$). The most substantial differences in effect sizes were observed for mean diffusivity measures especially for the SelfRepDep phenotype (self-reported depression) with other phenotypes. For instance, the posterior limb of the internal capsule (left) showed a significant difference in effect size between SelfRepDep and GPpsy phenotypes (odds ratio difference = 0.121, 95% CI: 0.060-0.185, adjusted $p = 0.005$). The external capsule (left) demonstrated a notable difference between SelfRepDep and GPpsy (odds ratio difference = 0.128, 95% CI: 0.067-0.193, adjusted $p = 0.003$). The medial lemniscus (left) showed a significant difference between SelfRepDep and LifetimeMDD (odds ratio difference = 0.152, 95% CI: 0.082-0.226, adjusted $p = 0.002$). When comparing SelfRepDep to MDDRecur, significant differences were observed in the middle cerebellar peduncle (odds ratio difference = 0.160, 95% CI: 0.080-0.245, adjusted $p = 0.005$) and the corticospinal tract (left) (odds ratio difference = 0.168, 95% CI: 0.090-0.252, adjusted $p = 0.002$). Fractional anisotropy, cortical thickness, and surface area measures showed no significant differences between phenotypes after adjusting for multiple comparisons. Overall, while white matter microstructure (especially mean diffusivity) displayed notable effect-size differences between shallow and deep depression definitions, measures of fractional anisotropy, cortical thickness, and surface area remained largely similar across phenotypes.

2. Comparison of Findings Between the UKB and ENIGMA Consortium

Our results were compared with meta-analytic estimates from ENIGMA MDD ("deep" clinical phenotypes)[13-14], BD[15-16], and SCZ[17-18] working groups (**Figure 2**, **Supplementary Figures S4-S6)**. While the UKB was examined in a secondary analysis of the ENIGMA MDD DTI study [14], it was not included in the primary meta-analysis, enabling independent comparison. This

allowed us to evaluate the consistency of neuroimaging correlates across different operational definitions of depression and additionally contextualized our findings within the broader neuropsychiatric literature.

2.1. Deeper phenotypes show strongest concordance with ENIGMA findings in white matter microstructure

To evaluate the consistency between our UKBB findings and prior meta-analytic evidence, we conducted a comparative analysis with effect size estimates from the ENIGMA consortium's major depressive disorder working group. For standardization, we converted our UK Biobank odds ratios to Cohen's d effect sizes using the formula $d = \ln(OR) \times \sqrt{3} / \pi$. We then calculated Spearman's rank correlations between these standardized effect sizes and the corresponding ENIGMA consortium meta-analytic effect size estimates. This correlation analysis was performed independently for each neuroimaging measure (cortical thickness, surface area, fractional anisotropy, and mean diffusivity) across our nine depression phenotype definitions. The resulting correlation coefficients quantify the convergence between brain structural alterations identified in our UK Biobank analyses and those documented in the ENIGMA consortium's meta-analysis.

None of the correlations between UKB and ENIGMA effect sizes were statistically significant for cortical measures, with weak correlations for both thickness ($r = -0.01$ to 0.31, $p = 0.100$-0.910) and surface area ($r = -0.08$ to 0.22, $p = 0.477$-0.739). In contrast, we found stronger concordance for measures of white matter integrity. For fractional anisotropy, the correlations between our UK Biobank results and ENIGMA MDD findings were generally stronger for deeper phenotypes such as LifetimeMDD ($r = 0.69$), MDDRecur ($r = 0.67$), and the intermediary phenotype DepAll ($r = 0.65$), reflecting those reporting at least one cardinal symptom of depression (low mood or apathy) for two or more weeks on the touchscreen questionnaire and seeking help from a GP or psychiatrist. For MD, the strongest correlation was observed for the Psypsy phenotype, reflecting those who sought help from a psychiatrist ($r = 0.45$), though some deeper phenotypes showing non-significant correlations (e.g., MDDRecur, $r = 0.36$, FDR-adjusted $p = 0.118$). Overall, the correlates with ENIGMA MDD results were the strongest for fractional anisotropy measures.

2.2. Effect sizes in UK Biobank are consistently smaller than meta-analytic effect size estimates in ENIGMA

We then conducted a comparison of the effect sizes of the five top regions found in the ENIGMA consortium for MDD with their corresponding results from the UKBB, and vice versa (**Figure 3)**. The ENIGMA MDD meta-analysis reported on 20 bilateral white matter tracts, while our UKBB analysis included 48 tracts with hemispheric distinction. Additionally, we identified the top five regions in the UKBB by ranking each region across all phenotypes by odds ratio, summing these ranks, then selecting the five regions with the highest total ranks.

For fractional anisotropy measures, we observed notable discrepancies between ENIGMA MDD and UKBB results. The top five tracts within ENIGMA showed Cohen's d values ranging from -0.2 to -0.4, indicating small to moderate reductions in fractional anisotropy in the superior fronto-occipital fasciculus, anterior limb of internal capsule, body of corpus callosum, genu of corpus callosum, and the anterior corona radiata. While our UKBB results showed much smaller effect sizes across all depression definitions, we observed significant effects in EMR-based depression definitions (ICD10Dep and ICD10Dep_exclpsych). Specifically in ICD10Dep, we observed significant reductions (FDR-adjusted $p < 0.05$) in the genu of corpus callosum ($d = -0.14$), body of corpus callosum ($d = -0.12$), and bilateral anterior limb of internal capsule ($d = -0.09$), with similar magnitudes in ICD10Dep_exclpsych (detailed in Supplementary Table S2).

For mean diffusivity measures, we observed a similar pattern of discrepancy. The ENIGMA results showed Cohen's d values ranging from approximately 0.1 to 0.2, indicating small increases in mean diffusivity in regions such as the external capsule (EC), superior fronto-occpital fasculus (SFO), fornix (cres) / stria terminalis (FXST), superior corona radiata (SCR), and body of corpus callosum. Again, our UKBB results showed much smaller effect sizes, closer to zero, across most depression phenotypes. Furthermore, comparisons of UKBB results and meta-analytic effect size estimates in ENIGMA BD and ENIGMA SCZ can be found in **Supplementary Figures S5-S6.** Together, while these results demonstrate some directional agreement with ENIGMA, the overall magnitudes of effect sizes in UKBB were smaller, which may be partly driven by sample sizes and demographic differences between UKBB and ENIGMA, rather than phenotyping depth alone.

3. Machine learning classification of depression phenotypes

For the purposes of our machine learning classification, we categorized eight UK Biobank depression phenotypes into two broader groups ("shallow" versus "deep"), as summarized in Table 1. GPNoDep was excluded from machine learning analyses as these individuals sought help for mental health concerns but did not meet depression criteria.

3.1.1. Model development and cross-validation

The dataset was partitioned into training (70%) and testing (30%) sets using stratified random sampling to preserve the relative proportion of depression cases and controls across both sets. Model development utilized 5-fold repeated cross-validation during hyperparameter tuning via Bayesian optimization [19]. To address class imbalance in the training set, we implemented inverse frequency class weighting, where weights were assigned to each class inversely proportional to their prevalence in the training data. This approach ensured that the model gave appropriate attention to the minority class without altering the original data distribution. Specifically, weights were calculated as the proportion of the opposite class relative to the total sample size, resulting in higher weights for less frequent classes. Detailed information about the cross-validation

procedure, weighting strategy, and hyperparameter optimization can be found in the Methods section.

3.1.2. Neuroanatomical correlates yield divergent classification metrics across depression phenotyping depth

We developed multiple machine learning classifiers to discriminate between depression cases and controls using 256 IDPs. Our primary analysis compared deep versus shallow depression classifications, while additional analyses explored multi-class classification and stacked ensemble methods. When distinguishing cases from controls in the testing set, shown within Table 2, the deep depression classifier achieved an area under the receiver operating characteristic curve (AUC) of 0.831 (95% CI: 0.819-0.844), while the shallow depression (broad) model showed a slightly lower AUC of 0.810 (95% CI: 0.802-0.819). At a standardized sensitivity of 80%, the deep depression model demonstrated a specificity of 68.2% (95% CI: 66.9%-69.4%), compared to 61.1% (95% CI: 59.7%-62.5%) for the shallow depression (broad) model. The positive predictive value (PPV) for the deep depression model was 45.6% (95% CI: 43.8%-47.4%), while the shallow depression (broad) model achieved a substantially higher PPV of 67.2% (95% CI: 66.0%-68.4%). Both models demonstrated high negative predictive values (NPV), with 91.1% (95% CI: 90.1%-91.9%) for deep depression and 75.4% (95% CI: 74.0%-76.8%) for shallow depression (broad). The shallow (broad) model also showed superior performance in F1 score (0.730 vs 0.581) and AUPRC (0.849 vs 0.716), indicating better precision-recall balance.

To further investigate neuroanatomical distinctions between depression phenotypes, we implemented multi-class random forest and Super Learner classifiers using all 256 candidate features. As shown in Table 4, the multi-class random forest classifier achieved a macro average AUC of 0.803 (95% CI: 0.795-0.811), with deep depression showing the highest class-specific AUC of 0.841 (95% CI: 0.829-0.853), followed by shallow depression (AUC = 0.797, 95% CI: 0.787-0.806) and controls (AUC = 0.771, 95% CI: 0.762-0.780). Controls demonstrated the highest F1 score (0.718, 95% CI: 0.708-0.728) and sensitivity (0.736, 95% CI: 0.723-0.747), while deep depression exhibited the highest specificity (0.883, 95% CI: 0.876-0.890) among the classes. Shallow depression showed the lowest sensitivity (0.567, 95% CI: 0.550-0.583) and intermediate specificity (0.880, 95% CI: 0.872-0.888). Unlike the random forest models in Table 2, our super learner classifiers (Table 5) using all 256 candidate features demonstrated superior performance for deep vs control classification (AUC = 0.911, 95% CI: 0.899 - 0.923) compared to shallow vs control (AUC = 0.765, 95% CI: 0.752-0.775). Separate random forest classifiers were also developed for individual depression phenotypes, as detailed in Supplementary Tables S5-S6 (comprising the deepest CIDI-derived phenotypes of LifetimeMDD and MDDRecur (2 or more MDD episodes), alongside help-seeking behavior phenotypes when seeking care from a general practitioner or psychiatrist, relative to controls). The importance of neuroanatomical features in classifying depression phenotypes was examined across our models. Figure 5 presents

feature rankings for shallow and deep depression models using all 256 neuroimaging features, while Figure 7 examines feature importance patterns in multi-class versus binary case-case classifications (shallow versus deep). In sum, the larger and less stringent depression definitions were easier to predict with neuroimaging features, while deeper definitions showed lower PPV across models, implying trade-off between clinical specificity and statistical power.

3.1.3. Deep and shallow depression classifiers show partly distinct neuroanatomical signatures

Detailed feature rankings and their relative importance across models can be found in Figure 5. The deep depression model relied more heavily on volumetric measures, with normalized gray matter volume and peripheral cortical gray matter volume as top predictors. In contrast, the shallow depression model placed greater emphasis on surface area measurements. Despite these differences, some features maintained consistent importance across both models, such as mean fractional anisotropy in the fornix. These patterns suggest partially distinct neuroanatomical signatures between deeper and shallower depression phenotypes, though these differences should be interpreted cautiously given varying class prevalences and statistical power considerations.

3.1.4. Neuroanatomical features derived from broader depression phenotypes generalize well to more stringent depression definitions

To investigate the specificity and generalizability of neuroanatomical predictors across depression phenotypes, we implemented a feature transfer approach (Figure 4). The dataset was initially partitioned into training (70%) and testing (30%) sets using stratified random sampling of the outcome variable, to preserve the relative proportions of depression phenotypes ("deep", "shallow", and control) in both sets. We first trained random forest classifiers (5-fold repeated cross-validation) on each classification task - a) "deep" (n=5,712) vs. control and b) "shallow" (narrow definition; n=11,321) vs. control - and ranked all 256 candidate features by their permutation-based feature importance. We then selected the top 30% (76/256) from each "outcome-specific" classifier and used these feature sets to train new models for the opposite classification task. Specifically, the 76 most important features identified from the shallow vs control random forest classifier were used to train a new deep vs control random forest classifier, and conversely, the 76 most important features from the deep vs control classifier were used to train a shallow vs control classifier. Importantly, we transferred model features but not weights.

As summarized in Table 3, feature transfer generally yielded performance metrics comparable to the outcome-specific models. For the deep vs control task, transferring top features derived from the shallow phenotype resulted in an AUC of 0.843 (95% CI: 0.831-0.855) and an AUPRC of 0.773 (95% CI: 0.756-0.789), compared to an AUC of 0.840 (95% CI: 0.828-0.852) and an AUPRC of 0.745 (95% CI: 0.726-0.763) achieved by models trained exclusively on deep depression and controls. In the shallow vs. control task, applying deep-derived features yielded an AUC of 0.764 (95% CI: 0.745-0.775) with an F1 score nearly identical to that of the outcome-specific shallow classifier (0.628 vs 0.629; sensitivity 0.799, specificity 0.502, and PPV

0.518), while the outcome-specific shallow classifier attained an AUC of 0.779 (95% CI: 0.769-0.790).

To evaluate whether any nominal performance differences between the feature transfer and outcome-specific models might be influenced by the larger sample size of the shallow cohort, we randomly downsampled the shallow cohort to match the case/control ratio (approximately 1:3 with 5,712 cases) of the deep vs. control classifier. In this downsampled analysis, the shallow vs control classifier (top 30% of features) yielded an F1 score of 0.432 (95% CI: 0.497-0.525), sensitivity of 0.799 (95% CI: 0.780-0.818), specificity of 0.365 (95% CI: 0.352-0.379), PPV of 0.296 (95% CI: 0.283-0.310), NPV of 0.845 (95% CI: 0.830-0.860), AUC of 0.690 (95% CI: 0.674-0.706), and AUPRC of 0.554 (95% CI: 0.531-0.575). When comparing these results to the non-downsampled shallow classifier (n=11,321; *see Methods and Table 2*), the downsampled model showed performance degradation across most metrics, including reduced PPV (ΔPPV = -0.245; 0.296 vs 0.541), AUC (ΔAUC = -0.094; 0.690 vs 0.784), F1 score (ΔF1 score = -0.213), and AUPRC (ΔAUPRC = -0.244). These results indicate that sample size reductions negatively impact classifier performance in the shallow vs control task, although this finding should be interpreted cautiously, as they reflect both the loss of training data (reduced from 11,321 to 5712 cases) and the artificially lowered prevalence, which impacts prevalence-sensitive metrics like PPV.

Furthermore, to characterize the cross-phenotype generalizability at the feature level observed within the feature transfer tasks, we examined the specific neuroanatomical correlates that contribute most to the model's classification performance. Analysis of the top 76 features from each transfer model revealed approximately 47% feature overlap (36 shared features), indicating that while deep and shallow phenotypes share certain neuroanatomical signatures, each also captures distinct patterns (Figure 6, Supplementary Figure S7). Across all models, including outcome-specific, feature transfer, and multi-class classifiers, global volumetric measures were consistently among the top rankings, with normalized gray matter volume and peripheral cortical gray matter volume ranking first and second in both deep and shallow outcome-specific models (Supplementary Table S8). When the shallow definition cohort was downsampled to match the deep phenotype case/control ratio, not only did overall performance degrade as previously noted, but the raw importance scores of these top volumetric features in the shallow model decreased by 31% and 24% respectively, despite maintaining their leading ranks. This suggests that shallow phenotypes require approximately twice the sample size of deep phenotypes to achieve comparable classification performance.

Discussion

Efforts to identify the neurobiologic basis of MDD have been complicated by the heterogeneity of clinical presentations and variability in ascertainment and assessment in research cohorts. Many studies have employed shallow (or "minimal") phenotyping to facilitate the collection of

large-scale samples and enhance power. Depth of phenotypic assessment has been shown to impact genomic analyses of MDD[9] but its implications for neuroimaging research have not been systematically studied. Here, we leveraged neuroimaging data from more than 30,000 UK Biobank participants to examine how associations between depression and brain structure vary with phenotypic depth. Three main findings emerged from our analyses. First, regardless of phenotyping depth, the most robust associations with depression case status were found for indices of white matter microstructure (specifically fractional anisotropy and mean diffusivity). Second, neuroanatomical associations were more modest than those observed by the large ENIGMA MDD consortium, possibly reflecting differences in sample composition, depression severity, or measurement approaches. Third—contrary to our hypothesis, neuroanatomical classifier performance was similar regardless of deep vs. shallow case definition. In addition, our feature transfer analyses found that random forest models using the top 30% of features selected from shallow phenotypes performed similarly (by AUC and F1 metrics) in discriminating deep phenotype cases from controls compared to models trained directly on features from deep phenotypes, and vice versa. That said, the ranks of top features did differ for the deep and shallow classifiers, suggesting that biological correlates are not entirely the same (Figure 5). Of note, the GPNoDep phenotype - help-seeking without depression - showed markedly fewer neuroanatomical associations than any depression phenotype (Supplementary Table S2), consistent with genetic evidence that GPNoDep is related to non-specific help-seeking rather than depression-specific biology[9].

As noted, white matter microstructure showed the strongest and most consistent associations with depression status across phenotype definitions, in line with prior large biobank and multi-site studies identifying white matter alterations in depression[14,20–23]. For example, Nothdurfter et al. (2024)[24] analyzed data from the UKBB across 19,183 individuals (including 5823 with varying levels of depressive symptoms), demonstrating reduced white matter integrity in thalamic and intracortical fiber tracts in depression (corresponding to our "self-reported depression" shallow phenotype), regardless of current symptom level. These findings contrast, however, with an earlier 12-site mega-analysis by Koshiyama et al. (2020)[21], which found no significant differences in DTI indices between MDD patients (n=398) and healthy controls (n=1,506), potentially highlighting the value of large sample sizes in detecting subtle neuroanatomical differences in depression.

The causal relationship of white matter changes in depression remains uncertain. However, longitudinal findings from Flinkenflügel et al. (2024)[25], analyzing diffusion-weighted imaging data from 418 MDD patients and 463 healthy controls over two years, demonstrated that MDD patients show steeper declines in fractional anisotropy in the superior longitudinal fasciculus compared to healthy controls; these changes were associated with cognitive decline. Complementary work by Marawi et al. (2023)[20] in 226 older adults found white matter integrity reductions correlated with executive functioning, working memory, and processing speed deficits, but only in participants with mild cognitive impairment (MCI) or combined remitted

depression-MCI and not in remitted depression alone. Collectively, these studies suggest that white matter integrity may be relevant to depression and related cognitive impairment in at least a subset of individuals. Evidence from Mendelian randomization studies has supported a causal role of depression on white matter integrity [26,27]. In contrast to the robust white matter findings, cortical thickness and surface area measures showed weaker and more variable associations across depression phenotypes. Though prior meta-analyses have identified gray matter reductions in regions such as the hippocampus, amygdala, and cingulate cortex in MDD[2,28,29], we did not find strong evidence of this. In another study using UKBB data, Harris et al. (2022)[10] similarly reported widespread reductions in fractional anisotropy and increases in mean diffusivity across multiple depression definitions using data from the UKBB; however, they also observed associations between cortical thickness and depression status, particularly in self-reported phenotypes. This discrepancy may reflect differences in analytical approaches. For example, Harris et al.[10] aggregated effects across several cortical regions structures, reducing the burden of correction for multiple comparisons and perhaps enhancing power to detect some associations. Our more stringent multiple comparison approach, combined with greater regional specificity, likely contributed to our more variable cortical associations.

The reproducible pattern of white matter microstructural alterations across independent cohorts and varied analytical approaches points to their potential relevance as therapeutic targets. Consistent with this broader literature[14,22–25], our analyses revealed increases in fractional anisotropy (FA) and decreases in mean diffusivity (MD) across multiple depression phenotypes, with significant differences and directional consistency in associations across regions including the posterior or anterior limb of the internal capsule, various corpus callosum subregions, and association fibers in the corona radiata. These pathways are related to circuits historically targeted by neurosurgical ablation, such as subcaudate tractotomy, which targeted dysfunctional frontal–subcortical circuits[30]. Modern interventions such as electroconvulsive therapy (ECT), repetitive transcranial magnetic stimulation (rTMS), and deep brain stimulation (DBS), now refine this rationale by leveraging neuroimaging to guide precise, non-ablative modulation of these circuits.

Emerging evidence suggests white matter remodeling as a mechanism of action across neuromodulatory depression treatments, reinforcing the clinical relevance of the white matter alterations we observed. ECT studies have demonstrated both transient increases in right-hemisphere mean diffusivity and progressive increases in fractional anisotropy in dorsal fronto-limbic pathways that correlate with symptom improvement[31,32]. Similarly, rTMS appears to induce microstructural changes in prefrontal white matter tracts, with studies reporting increased fractional anisotropy in anterior-medial prefrontal bundles following treatment, despite minimal changes in directly stimulated pathways[33,34]. DBS represents a more invasive approach for treatment-resistant depression (TRD), directly targeting subcortical regions and critical white matter tracts. Studies of subcallosal cingulate DBS have shown that pre-existing structural abnormalities in specific pathways, particularly the forceps minor, uncinate fasciculus, and

cingulum bundle, correlate with longer recovery times in depression[35], while supero-lateral medial forebrain bundle stimulation has demonstrated high response rates with sustained effects in TRD[36]. Collectively, these findings suggest that while different neuromodulatory treatments target distinct neural circuits, their therapeutic effects involve white matter remodeling, with evidence that both treatment-induced changes and baseline structural integrity of white matter tracts, measured through fractional anisotropy and mean diffusivity, relate to clinical outcomes in depression.

While ECT, rTMS, and DBS studies suggest white matter pathways mediate treatment responses, our comparison revealed that depression-related alterations in fractional anisotropy and mean diffusivity in UKBB were markedly smaller than those reported by the ENIGMA MDD consortium[14], after converting our odds ratios to Cohen's d effect sizes, as shown in Figure 3. Notably, this comparison was restricted to 20 white matter tracts common to both datasets, with UKBB hemisphere-averaged values aligned to ENIGMA's bilateral reporting framework (see Methods). Specifically, the ENIGMA consortium reported Cohen's d values from -0.2 to -0.4 for fractional anisotropy in key tracts including the superior fronto-occipital fasciculus and anterior limb of internal capsule, and 0.1 to 0.2 for mean diffusivity in regions such as the external capsule and superior corona radiata. In contrast, our analyses showed smaller depression-associated changes in these same regions. The strongest alignment with effect size estimates reported by the ENIGMA MDD consortium were observed in electronic medical record-based depression definitions as shown in Figure 3, though the magnitude of depression-associated neuroanatomical alterations remained smaller than those reported in their meta-analysis. These differences in the strength of association between depression status and neuroanatomical measures between UKBB and ENIGMA likely reflect a combination of sample composition, assessment approaches, and demographic factors rather than phenotyping depth alone—a finding reinforced by the similar performance of shallow and deep phenotyping approaches in our analyses.

Having characterized the pattern and magnitude of neuroanatomical associations across phenotyping approaches, we next evaluated their utility for depression classification through machine learning techniques. Using 256 neuroanatomical features, random forest classifiers performed similarly in distinguishing deep or shallow depression phenotypes from controls. In contrast, super learner ensemble classifiers achieved somewhat better performance for distinguishing deep phenotype cases from controls as opposed to shallow vs. controls on AUC (0.911 vs 0.764), specificity (0.862 vs 0.701), and sensitivity (0.785 vs 0.663), as shown in Table 5. Nevertheless, the comparable performance of features derived from shallow phenotypes in our feature transfer analyses combined with larger sample sizes facilitated by shallow phenotyping approaches, provides some justification for using simpler operational definitions when developing neuroanatomical classifiers of depression. However, after downsampling the shallow phenotype case prevalence to match the deep phenotype's case/control ratio, we observed notable performance degradation across multiple classification metrics, suggesting that shallow

phenotypes required approximately twice the sample size to achieve performance comparable to deep phenotypes. This finding relates to the fundamental power-specificity tradeoff documented in recent psychiatric genetics and neuroimaging literature, where broader phenotypic definitions can capture meaningful biological signals but necessitate substantially larger samples to compensate for reduced measurement precision and increased noise [9,37–39].

Several important limitations of our study warrant consideration. First, while the UK Biobank provides an unprecedented scale for neuroimaging analyses, it is not representative of the general population[40]. Second, our depression phenotypes, while spanning a broad spectrum, rely heavily on retrospective self-report and may be subject to recall bias. The absence of structured clinical interviews for all participants may have resulted in information bias in our phenotypic classifications. Third, our classification of ICD-10 diagnoses of depression as "deep" phenotypes warrants clarification. While we categorized these as deep because they do not constitute self-reported depression, ICD-10 codes have inherent limitations - we cannot determine the specific clinical characteristics that led to diagnosis, and codes may have been assigned for reasons other than depression (e.g., insurance purposes). This potentially positions ICD-10 phenotypes as intermediate between our shallow self-report measures and the more rigorous CIDI-SF derived phenotypes. Fourth, the UK Biobank sample is relatively homogeneous in terms of race/ethnicity; generalizability of these findings, especially to more diverse populations, may be limited. Fifth, although our multiple "shallow" and "deep" definitions capture different ways to measure or classify depression, these do not necessarily correspond to distinct biological subtypes of depression. The disorder itself is known to be heterogeneous, with subtypes that may differ in symptom profiles, course, and neurobiological underpinnings. Our findings thus speak more directly to the implications of measurement variability in depression rather than to intrinsic subgroups within the disorder. Consequently, while we can compare the predictive utility of different operational definitions (e.g., self-report versus ICD-10 diagnosis), we cannot disentangle all forms of heterogeneity that naturally exist within depression.

In summary, our large-scale analysis reveals that alterations in white matter microstructure are observed across a spectrum of depression phenotype definitions. While the neuroanatomical associations observed in UKBB were of lower magnitude than those reported by the ENIGMA MDD consortium, the directional pattern of white matter changes across different phenotyping approaches (with decreased fractional anisotropy and increased mean diffusivity), aligns with prior studies[14,22–25], and suggests these alterations represent a replicable biomarker of depression. Although shallow and deep phenotypes achieved comparable classification performance in our machine learning analyses, this was contingent on shallow phenotypes having approximately twice the sample size, underscoring the fundamental power-specificity tradeoff in psychiatric neuroimaging research.

Methods

Study Design and Participants

This cross-sectional study utilized data from the UK Biobank, (UKB; http://www.ukbiobank.ac.uk), a large population-based cohort study that recruited approximately 500,000 volunteers across the UK between 2006 and 2010. A subset of participants underwent brain MRI scans starting in 2014. The UK Biobank received ethical approval from the North West Centre Research Ethics Committee (REC number 11/NW/0382), and the current analyses were conducted under the approved UKB application (32568). We investigated the neuroanatomical correlates of depression using data from 30,122 UK Biobank Participants. Nine depression phenotypes of varying depth were examined, as previously described[9]. These phenotypes ranged from "shallow" definitions based on minimal criteria to "deeper" definitions meeting Major Depressive Disorder (MDD) thresholds according to a structured self-report (CIDI-SF) or ICD-10 diagnoses from EMR. The essential characteristics of these phenotypes, taken from Cai et al. 2020, and respective sample sizes of cases and controls across depression phenotypes can be found in **Table 1**, whereas a visual depiction can be found in **Supplementary Figure S1**. For machine learning analyses, we dichotomized "shallower" vs. "deeper" definitions, excluding the GPNoDep category. The deeper phenotypes were based on meeting MDD thresholds according to either the CIDI-SF or ICD-10 diagnoses, while the shallower phenotypes encompassed GPpsy, Psypsy, SelfRepDep, and DepAll. Although GPNoDep was excluded from machine learning analyses as these individuals did not meet criteria for depression, we retained it in other analyses to provide insights into the spectrum of mood-related help-seeking behaviors and their potential neurobiological correlates.

Image Acquisition and Processing

Brain imaging data were acquired using standardized MRI protocols on identical Siemens Skyra 3T scanners with a Siemens 32-channel RF receive head coil across three sites.

Raw imaging data were processed through an automated pipeline by the UK Biobank imaging team to generate various imaging-derived phenotypes (IDPs). (https://www.fmrib.ox.ac.uk/ukbiobank/fbp). Detailed information about the MRI protocol and processing steps for UKBB data is publicly available, and further details can also be found in the **Supplementary Methods**. T1-weighted structural MRI data were processed using FreeSurfer (version 6.0)[41]. A semi-automated quality control procedure was used to identify potentially low quality data that was later manually reviewed. Cortical parcellations were generated based on the Desikan-Killiany atlas[42], which divides the cortex into 34 regions per hemisphere. The temporal pole was absent in the processed UK Biobank data. This omission reflects known challenges in reliably segmenting the temporal pole due to its anatomical complexity (e.g., thin cortical folding and high inter-subject variability). We therefore analyzed cortical thickness and surface area for the remaining 33 regions. In addition to these 66 regional measures (averaged bilaterally), we included average thickness and total surface area measures for each hemisphere,

resulting in a total of 68 cortical thickness and 68 surface area measurements. Furthermore, global cortical structural measures, including total brain volume, gray matter volume, and white matter volume, were derived from T1-weighted images, providing 10 metrics for overall brain structure assessment. Subcortical volumes were obtained for 14 structures using FreeSurfer's automated segmentation algorithm (aseg), which identifies and quantifies the volume of subcortical gray matter structures[42,43].

Diffusion-tensor imaging (DTI) was used to generate white matter microstructural measurements (50 diffusion-encoding directions). This technique measures preferential water molecule diffusion along axonal axes, providing insights into white matter integrity. We analyzed 48 mean fractional anisotropy and 48 mean diffusivity measurements derived from DTI data. In total, our analysis included 256 brain imaging features: 68 cortical thickness, 68 surface area, 10 global cortical volumes, 14 subcortical volumes, 48 fractional anisotropy, and 48 mean diffusivity measurements. Detailed information about the MRI acquisition parameters, preprocessing steps, and derived metrics is available in the UK Biobank brain imaging documentation, and an overview can also be found in the Supplementary Methods (https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/brain_mri.pdf).

Genetic Data

To address potential confounding arising from population stratification[44] in our analyses, we utilized the top 20 genetic ancestry principal components (PCs) provided by the UK Biobank (Data-Field 22009). These PCs were derived from genome-wide genotype data as part of the UK Biobank's quality control process, and capture major axes of genetic variation that reflect ancestral population differences. By including these 20 PCs as covariates in our logistic regression models and residualizing our neuroimaging features against them in our machine learning classifiers, we controlled for systematic differences in genetic ancestry that could otherwise lead to spurious associations. This approach ensured consistent handling of population structure effects across our analytic methods.

Statistical Analyses

Feature preprocessing and normalization

Prior to conducting univariate logistic regression models, all 256 candidate neuroimaging features were z-score normalized to ensure comparability across different measurement scales and to facilitate interpretation of the resulting odds ratios (ORs). Z-score normalization[45] was performed by subtracting the mean and dividing by the standard deviation of each feature, resulting in a standardized scale with a mean of 0 and a standard deviation of 1. ORs were calculated by exponentiating the coefficients of the logistic regression models, representing the change in the odds of depression associated with one standard deviation change in the brain measure, adjusting for confounders.

Univariate and multivariable logistic regression analyses

Univariate logistic regression analyses were conducted using the *'glm'* function in R within the Stats package[46], with the presence or absence of depression as the outcome variable and neuroimaging features as predictor variables. Multivariable logistic regression models were adjusted for several covariates, including age, sex, assessment center, estimated total intracranial volume, head motion, and the top 20 genetic ancestry PCs to control for potential confounding due to population stratification. To calculate 95% confidence intervals (CIs) for the ORs, we used the profile likelihood method. Unlike Wald intervals, which assume that the uncertainty in parameter estimates follow a symmetric, bell-shaped distribution around the best estimate[47], profile likelihood intervals allow for asymmetry in this uncertainty[48]. This flexibility makes profile likelihood intervals more reliable when the true shape of the uncertainty distribution deviates from this ideal symmetric form. To evaluate overlap in case definitions across depression phenotypes, we computed pairwise Jaccard indices[48,49] as shown in Supplementary Table S7 (range: 0–1; higher values indicate greater overlap). These analyses revealed minimal overlap between shallow and deep depression phenotypes (e.g., GPNoDep vs. ICD10Dep: J = 0.014; DepAll vs. LifetimeMDD: J = 0.126), supporting their stratification into distinct analytical groups within ML analyses (Supplementary Figure S9).

Visualization of neuroanatomical correlates

Two complementary approaches were used to visualize the neuroanatomical correlates across depression phenotypes. First, for each of the four categories of imaging derived phenotypes (IDPs), corresponding to surface area, cortical thickness, fractional anisotropy, and mean diffusivity, we computed the mean OR and its variance across all nine depression definitions. Mean ORs deviating from 1 reflected an overall directional tendency (positive or negative) across phenotypes, whereas high variance indicated heterogeneity in the magnitude or direction of associations among phenotypes. Moreover, we generated separate plots displaying the odds ratios for each of the nine depression definitions across the four IDPs. Both visualization methods utilized the 'ggseg3d' package in R to create interactive 3D brain maps[50]. Neuroimaging features were mapped to regions in the corresponding ICBM 3D atlas using the 'brain_join' function. For the visualizations across all nine depression definitions, color mapping was based on mean ORs and variance of ORs, providing a composite view of the central-tendency and variability in brain-depression associations. For individual depression definitions, color mapping reflected the actual effect sizes (ORs) for each category of IDPs, allowing for direct comparison of neuroanatomical correlates across different phenotyping strategies.

Pairwise comparisons of effect sizes across depression phenotypes

To investigate the dependency of effect sizes across depression phenotypes, we performed pairwise comparisons of ORs, where for each pair of phenotypes, we identified common significant neuroimaging features and calculated the differences in their log ORs. We used a custom function to evaluate the magnitude of OR differences and the overlap of confidence

intervals between phenotypes. For each feature, we calculated the log OR difference and its standard error, which accounts for the uncertainty in each estimate. We then computed 95% confidence intervals for the difference in log ORs and converted these back to the OR scale. P-values were calculated using a two-tailed z-test based on the standardized log OR difference. To account for multiple comparisons, p-values were adjusted using the Benjamini-Hochberg false discovery rate (FDR) method[50,51]. We also examined the overlap of confidence intervals between phenotypes for each feature. Features with non-overlapping confidence intervals and significant adjusted p-values (FDR-adjusted $p < 0.05$) were considered to have meaningfully different effect sizes between the compared phenotypes. This approach allowed us to quantify the similarities and differences in neuroanatomical correlates across depression phenotypes, thereby characterizing the consistency of brain-depression associations across the various phenotyping strategies employed in this study.

Concordance of Neuroanatomical effect sizes with Meta-analytic estimates

In order to contextualize our findings within the broader literature and assess the impact of phenotyping depth on neuroimaging correlates, we compared our results with meta-analytic effect size estimates from various ENIGMA working groups. We used summary statistics from the latest and largest meta-analyses for Major Depressive Disorder (MDD), Bipolar Disorder, Schizophrenia. For each of our nine depression definitions in the UK Biobank, we calculated Spearman's rank correlations between our effect sizes and the corresponding meta-analytic effect estimates from the ENIGMA consortia. To enable this comparison, we first converted the odds ratios from our analyses to Cohen's d values using the formula $d = \ln(OR) \times \sqrt{3} / \pi$[52]. This analysis was performed separately for cortical thickness, surface area, fractional anisotropy, and mean diffusivity. We subsequently identified the depression definitions that showed the strongest correlations with ENIGMA MDD effect size estimates, to provide a more targeted examination of the correlation between effect sizes.

Comparative analysis of top effect sizes

To further investigate the specificity and magnitude of neuroimaging correlates, we conducted two comparative analyses. While the ENIGMA MDD meta-analysis examined 20 bilateral white matter tracts, our UK Biobank analysis included 48 tracts with hemispheric distinction. For direct comparison, we restricted analyses to the 20 white matter tracts common to both datasets. We then averaged left/right hemisphere values for UKBB tracts (e.g., superior fronto-occipital fasciculus) into single bilateral measures to align with ENIGMA's bilateral reporting. Following this, we compared the top 5 effect sizes from the ENIGMA MDD meta-analysis with the corresponding effect sizes for each of our nine depression definitions. Additionally, we identified the top 5 regions in the UK Biobank by ranking each region across phenotypes by odds ratio, summing these ranks, and selecting the five regions with the highest total ranks. This approach allowed us to assess both the concordance of the strongest meta-analytic findings across our phenotypes and the replicability of our strongest findings in the larger meta-analysis. Overall, these analyses quantify how different depression phenotyping strategies align with meta-analytic

findings both across psychiatric disorders and within MDD itself, providing insights into consistency and specificity of neuroimaging correlates. By comparing effect sizes across various depression definitions to large-scale meta-analytic results, we can assess how phenotyping depth influences the detection of brain structural alterations.

Brain-based Prediction Models

To complement our examination of brain-based associations across depression phenotypes using case-control binary logistic regression analyses, we developed a series of machine learning classifiers to investigate the predictive power and generalizability of these associations. Our aim was to determine whether neuroimaging features could accurately discriminate between different depression phenotypes and controls, and to assess how both the most important predictive features and the model performance metrics varied across phenotypes. To achieve this, we implemented a range of binary and multi-class classification models to explore various comparisons: deep depression versus controls, shallow depression versus controls, combined depression versus controls, deep versus shallow depression, and a multi-class classifier distinguishing between deep depression, shallow depression, and controls. This approach allowed us to evaluate how predictive performance metrics (such as AUC-ROC, sensitivity, and specificity) and feature importance rankings varied across different depression phenotypes and classification tasks. By examining both the performance metrics and the most important features in each model, we sought to understand which neuroanatomical characteristics were most predictive of different depression phenotypes and how generalizable these patterns were across the phenotyping spectrum.

Model development and implementation

To implement our classification approach, we primarily utilized random forest classifiers, leveraging the *mlr3*[53] and *ranger*[54] packages in R. Random forests were chosen for their ability to handle high-dimensional data, capture nonlinear relationships, and provide interpretable feature importance measures[55]. As a point of comparison and to assess the potential benefits of more complex modeling strategies, we also implemented stacked ensemble models, as further discussed in the **Supplementary Methods**. Briefly, stacked ensembles, also known as super learners, combine predictions from multiple base models to create a more robust meta-model with potential improvements in performance metrics[55,56]. Our ensemble consisted of five diverse base learners, including random forest, elastic net[57], *k*-nearest neighbors (kNN)[58–60], light gradient-boosting machine (lightGBM), and naïve bayes[61]. Random forest, already utilized as our primary classifier, excels in capturing non-linear relationships and providing feature importance measures[55]. LightGBM was incorporated for its ability to handle large-scale, high-dimensional data and capture nonlinear interactions through gradient boosting. Elastic net, a form of logistic regression combining L1 and L2 regularization, was chosen for its capacity to perform feature selection and handle multicollinearity, respectively, within a linear framework[62]. The k-nearest neighbors (kNN) algorithm was incorporated for its ability to model complex decision boundaries without making distributional assumptions, particularly useful for capturing

local patterns in the data [60]. Lastly, naive bayes was included for its efficiency in high-dimensional spaces and effectiveness with smaller training sets, despite its assumption of feature independence [61]. A key feature of our stacked ensemble architecture was using cross-validated predictions from the base learners as input features for the meta-learner, rather than the original features [63]. This approach, implemented through a 5-fold cross-validation pipeline, allows the ensemble to potentially capture a wider range of probability distributions and decision boundaries than any individual base learner alone.

Machine Learning feature preprocessing and normalization

Prior to model development, we applied z-score normalization to all features, which transforms each feature to have a mean of 0 and a standard deviation of 1, calculated as $(x - \mu) / \sigma$, where x is the original feature value, μ is the mean, and σ is the standard deviation of that feature. This normalization ensures that all features contribute equally to the model, preventing features with larger scales from dominating the learning process. To evaluate the impact of potential confounding factors on model performance, residualization was performed, where for each neuroimaging feature, we regressed out the effects of age, sex, assessment center, estimated total intracranial volume, head motion, and the top 20 genetic ancestry principal components using a linear model with the *lm()* function in R[46]. This process was performed separately for shallow and deep depression phenotypes.

Hyperparameter tuning, model evaluation, and feature importance

We implemented random forest classifiers using a 70-30 stratified split [64] (*rsample* package [65]) to preserve outcome class distributions between training and testing sets. Hyperparameter tuning was performed using Bayesian optimization[66] via the *mlr3mbo* package with 5-fold cross-validation with 5 repeats, balancing exploration of the hyperparameter space against computational efficiency[67]. To mitigate class imbalance[68], we implemented inverse frequency class weighting rather than data resampling techniques. This approach assigns weights to each class inversely proportional to their prevalence in the training data, without altering the ordinal data distribution. Specifically, for each classification task, we calculated weights as w_cases = f_controls / (f_controls + f_cases) and w_controls = f_cases / (f_controls + f_cases), where f_controls and f_cases represent the frequencies of controls and cases in the training set, respectively. This formula results in higher weights for the minority class, proportionally compensating for its underrepresentation. These weights were directly applied to the random forest classifier through the case.weights parameter in the ranger package, which adjusts the contribution of each observation during tree construction, ensuring that the model is more sensitive to the minority class while preserving the full information contained in the training set. We chose this weighting strategy over resampling methods, such as SMOTE[69], to avoid potential issues including overfitting or information loss. Model performance was assessed on the 30% holdout using AUC-ROC (primary metric), F1 score (harmonic mean of precision/recall), sensitivity, specificity, PPV, NPV, and PR-AUC (95% CIs via stratified bootstrapping). Sensitivity thresholds were standardized to 80% for cross-model comparison, with depression

cases designated as the positive class in case-control models and deep depression cases in deep vs shallow comparisons, respectively. Detailed information on the stacked ensemble methodology, including the specific base learners and meta-learners used, can be found in the supplementary material. For our random forest models, we calculated permutation-based feature importance using the *Ranger* package in R[54]. This approach measures each feature's importance by first computing the model's prediction accuracy on out-of-bag samples, then randomly permutes the feature's values to break any association with the outcome. The resulting decrease in prediction accuracy after permutation, averaged across all trees in the forest, represents the feature's importance - where larger decreases indicate greater importance. This method offers robust and unbiased estimates of feature importance, accounting for complex feature interactions without favoring variables with more categories or higher variances[54].

Feature transfer approach

To investigate the specificity and generalizability of brain-based associations across depression phenotypes, we implemented a novel feature transfer approach. We first trained separate random forest classifiers for deep vs controls and shallow vs controls classification tasks. Using recursive feature elimination[70] to rank features by importance, we identified the top 30% of features (76 features) from each model based on their predictive contribution to the respective classification task. We then trained two sets of models for comparison. Outcome-specific models used the top 30% of features derived from the same classification task, serving as our baseline comparison. Feature transfer models used the top 30% of features derived from the opposite classification task. Specifically, deep vs controls classifiers were trained using the 76 most important features identified by the shallow vs controls classifier, and shallow vs controls classifiers were trained using the 76 most important features identified by the deep vs control classifier, respectively. Model performance metrics, including AUC, sensitivity, specificity, PPV, NPV, and AUPRC were compared between outcome-specific features and feature transfer approaches This systematic comparison allowed for evaluating whether neuroanatomical features predictive of one depression phenotype were specific to deep or shallow definitions or exhibited generalizability across phenotyping depth.

Model comparison and statistical testing

Furthermore, given the possibility of overlapping confidence intervals in performance metrics across outcome-specific and feature transfer classifiers, stratified bootstrap tests with 10,000 iterations for F1 score and AUC metrics were used to ascertain whether there were statistically significant differences between models. The stratified bootstrap test involves repeatedly sampling the data with replacement, stratifying by class to maintain the original class distribution[71]. The stratification was implemented by first calculating the proportion of each class in the original data, then separately resampling from each class with replacement to preserve these proportions in every bootstrap iteration. This approach ensures consistent class distributions across all bootstrap samples, reducing variability that could otherwise arise from

random sampling in imbalance datasets. We implemented a two-tailed test by calculating p-values based on the proportion of bootstrap samples where the absolute difference in performance metrics (F1 score or AUC) between the two classifiers was greater than or equal to the observed difference. This approach accounts for performance differences favoring either classifier. Importantly, we performed simultaneous bootstrapping for both classifiers, calculating the difference in each iteration. This method preserves the paired nature of the observations and models the sampling distribution of the performance differences directly. We selected F1 score and AUC as informative metrics because the F1 score provides a balanced measure of precision and recall, while AUC evaluates the model's ability to discriminate between classes across various threshold settings. Additionally, we employed McNemar's test to compare the classification performance of the models. This test assesses the differences between two models' classification errors on the same dataset, considering the paired nature of the data[72].

Multi-class classification

Additionally, to capture the nuanced relationships between different depression phenotypes and healthy controls simultaneously, we implemented multi-class random forest classifiers. This approach allowed us to discriminate between shallow depression, deep depression, and healthy controls within a single model, potentially revealing patterns that may be obscured in binary classifications. We utilized the *mlr3* and *ranger* packages to maintain consistency with our binary classification methodology. Hyperparameter tuning was performed using Bayesian optimization with the *mlr3tuning* package, optimizing for multi-class AUC (MAUC) using the "classif.mauc_aunu" measure, which extends binary AUC to multi-class problems and provides a balanced assessment of the model's discriminative ability across all classes[73]. We applied the same data splitting, stratified sampling, feature residualization, and feature importance strategies used in our binary classifiers to ensure comparability. To account for class imbalance in the multi-class setting, we weighted performance metrics by class prevalence.

Code Availability

A static version of all custom code used for data analysis and visualization in this study is available upon request. This includes R scripts for logistic regression analyses, visualization of neuroanatomical correlates, pairwise comparisons of effect sizes across depression phenotypes, concordance analyses with meta-analytic estimates, and machine learning analyses. For the visualization of neuroanatomical correlates, we utilized the *ggseg3d* package, which allows for the creation of interactive 3D brain maps of effect sizes. This package is openly available and can be accessed at https://github.com/ggseg/ggseg3d. Our implementation of *ggseg3d* for this study is included in the visualization scripts within our code repository.

References

1. Smith, K. Mental health: a world of depression. *Nature* **515**, 181 (2014).
2. Gray, J. P., Müller, V. I., Eickhoff, S. B. & Fox, P. T. Multimodal Abnormalities of Brain Structure and Function in Major Depressive Disorder: A Meta-Analysis of Neuroimaging Studies. *Am J Psychiatry* **177**, 422–434 (2020).
3. Schmaal, L. *et al.* ENIGMA MDD: seven years of global neuroimaging studies of major depression through worldwide data sharing. *Transl Psychiatry* **10**, 172 (2020).
4. Winter, N. R. *et al.* Quantifying Deviations of Brain Structure and Function in Major Depressive Disorder Across Neuroimaging Modalities. *JAMA Psychiatry* **79**, 879–888 (2022).
5. Liu, S., Abdellaoui, A., Verweij, K. J. H. & van Wingen, G. A. Replicable brain–phenotype associations require large-scale neuroimaging data. *Nature Human Behaviour* **7**, 1344–1356 (2023).
6. Feng, C., Thompson, W. K. & Paulus, M. P. Effect sizes of associations between neuroimaging measures and affective symptoms: A meta-analysis. *Depress Anxiety* **39**, 19–25 (2022).
7. Goldberg, D. The heterogeneity of 'major depression'. *World Psychiatry* **10**, 226–228 (2011).
8. Lynch, C. J., Gunning, F. M. & Liston, C. Causes and Consequences of Diagnostic Heterogeneity in Depression: Paths to Discovering Novel Biological Depression Subtypes. *Biol Psychiatry* **88**, 83–94 (2020).
9. Cai, N. *et al.* Minimal phenotyping yields genome-wide association signals of low specificity for major depression. *Nat Genet* **52**, 437–447 (2020).

10. Harris, M. A. *et al.* Structural neuroimaging measures and lifetime depression across levels of phenotyping in UK biobank. *Transl Psychiatry* **12**, 157 (2022).
11. Anastasiou, A. & Gaunt, R. E. Multivariate normal approximation of the maximum likelihood estimator via the delta method. *Braz. J. Probab. Stat.* **34**, 136–149 (2020).
12. Beutner, E. Delta method, asymptotic distribution. *Wiley Interdiscip. Rev. Comput. Stat.* **16**, (2024).
13. Schmaal, L. *et al.* Cortical abnormalities in adults and adolescents with major depression based on brain scans from 20 cohorts worldwide in the ENIGMA Major Depressive Disorder Working Group. *Mol Psychiatry* **22**, 900–909 (2017).
14. van Velzen, L. S. *et al.* White matter disturbances in major depressive disorder: a coordinated analysis across 20 international cohorts in the ENIGMA MDD working group. *Mol Psychiatry* **25**, 1511–1525 (2020).
15. Hibar, D. P. *et al.* Cortical abnormalities in bipolar disorder: an MRI analysis of 6503 individuals from the ENIGMA Bipolar Disorder Working Group. *Mol Psychiatry* **23**, 932–942 (2018).
16. Favre, P. *et al.* Widespread white matter microstructural abnormalities in bipolar disorder: evidence from mega- and meta-analyses across 3033 individuals. *Neuropsychopharmacology* **44**, 2285–2293 (2019).
17. van Erp, T. G. M. *et al.* Cortical Brain Abnormalities in 4474 Individuals With Schizophrenia and 5098 Control Subjects via the Enhancing Neuro Imaging Genetics Through Meta Analysis (ENIGMA) Consortium. *Biol Psychiatry* **84**, 644–654 (2018).
18. Kelly, S. *et al.* Widespread white matter microstructural differences in schizophrenia across 4322 individuals: results from the ENIGMA Schizophrenia DTI Working Group. *Mol*

Psychiatry **23**, 1261–1269 (2018).

19. Nguyen, V. Bayesian optimization for accelerating hyper-parameter tuning. in *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)* (IEEE, 2019). doi:10.1109/aike.2019.00060.
20. Marawi, T. *et al.* Brain-Cognition Associations in Older Patients With Remitted Major Depressive Disorder or Mild Cognitive Impairment: A Multivariate Analysis of Gray and White Matter Integrity. *Biol Psychiatry* **94**, 913–923 (2023).
21. Koshiyama, D. *et al.* White matter microstructural alterations across four major psychiatric disorders: mega-analysis study in 2937 individuals. *Mol Psychiatry* **25**, 883–895 (2020).
22. Repple, J. *et al.* Severity of current depression and remission status are associated with structural connectome alterations in major depressive disorder. *Mol Psychiatry* **25**, 1550–1558 (2020).
23. Chen, G. *et al.* Disorganization of white matter architecture in major depressive disorder: a meta-analysis of diffusion tensor imaging with tract-based spatial statistics. *Sci Rep* **6**, 21825 (2016).
24. Nothdurfter, D., Jawinski, P. & Markett, S. White Matter Tract Integrity Is Reduced in Depression and in Individuals With Genetic Liability to Depression. *Biol Psychiatry* **95**, 1063–1071 (2024).
25. Flinkenflügel, K. *et al.* Associations between white matter microstructure and cognitive decline in major depressive disorder versus controls in Germany: a prospective case-control cohort study. *Lancet Psychiatry* **11**, 899–909 (2024).
26. Shen, X. *et al.* A phenome-wide association and Mendelian Randomisation study of polygenic risk for depression in UK Biobank. *Nat Commun* **11**, 2301 (2020).

27. Lu, T. *et al.* Major depressive disorder and the development of cerebral small vessel disease: A Mendelian randomization study. *J Affect Disord* **377**, 68–76 (2025).

28. Schmaal, L. *et al.* Subcortical brain alterations in major depressive disorder: findings from the ENIGMA Major Depressive Disorder working group. *Mol Psychiatry* **21**, 806–812 (2016).

29. Arnone, D. *et al.* Computational meta-analysis of statistical parametric maps in major depression. *Hum Brain Mapp* **37**, 1393–1404 (2016).

30. Sanchez, S. M. *et al.* Repetitive Negative Thinking-Specific and -Nonspecific White Matter Tracts Engaged by Historical Psychosurgical Targets for Depression. *Biol Psychiatry* **94**, 661–671 (2023).

31. Lyden, H. *et al.* Electroconvulsive therapy mediates neuroplasticity of white matter microstructure in major depression. *Transl Psychiatry* **4**, e380 (2014).

32. Repple, J. *et al.* Influence of electroconvulsive therapy on white matter structure in a diffusion tensor imaging study. *Psychol Med* **50**, 849–856 (2020).

33. Ning, L., Rathi, Y., Barbour, T., Makris, N. & Camprodon, J. A. White matter markers and predictors for subject-specific rTMS response in major depressive disorder. *J Affect Disord* **299**, 207–214 (2022).

34. Peng, H. *et al.* High-frequency rTMS treatment increases white matter FA in the left middle frontal gyrus in young patients with treatment-resistant depression. *J Affect Disord* **136**, 249–257 (2012).

35. Alagapan, S. *et al.* Cingulate dynamics track depression recovery with deep brain stimulation. *Nature* **622**, 130–138 (2023).

36. Coenen, V. A. *et al.* Frontal white matter architecture predicts efficacy of deep brain

stimulation in major depression. *Transl Psychiatry* **9**, 197 (2019).

37. Tiego, J. *et al.* Precision behavioral phenotyping as a strategy for uncovering the biological correlates of psychopathology. *Nat Ment Health* **1**, 304–315 (2023).
38. Sanchez-Roige, S. & Palmer, A. A. Emerging phenotyping strategies will advance our understanding of psychiatric genetics. *Nat Neurosci* **23**, 475–480 (2020).
39. Dahl, A. *et al.* Phenotype integration improves power and preserves specificity in biobank-based genetic studies of major depressive disorder. *Nat Genet* **55**, 2082–2093 (2023).
40. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol* **186**, 1026–1034 (2017).
41. Fischl, B. FreeSurfer. *Neuroimage* **62**, 774–781 (2012).
42. Desikan, R. S. *et al.* An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* **31**, 968–980 (2006).
43. Sederevičius, D. *et al.* Reliability and sensitivity of two whole-brain segmentation approaches included in FreeSurfer - ASEG and SAMSEG. *Neuroimage* **237**, 118113 (2021).
44. Hellwege, J. N. *et al.* Population Stratification in Genetic Association Studies. *Curr Protoc Hum Genet* **95**, 1.22.1–1.22.23 (2017).
45. Patro, S. G. K. & Sahu, K. K. Normalization: A Preprocessing Stage. *arXiv [cs.OH]* (2015) doi:10.48550/ARXIV.1503.06462.
46. stats package - RDocumentation. https://rdocumentation.org/packages/stats/versions/3.6.2.
47. Newcombe, R. G. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Stat Med* **17**, 857–872 (1998).

48. Royston, P. Profile likelihood for estimation and confidence intervals. *Stata J.* **7**, 376–387 (2007).

49. Real, R. & Vargas, J. M. The probabilistic basis of jaccard's index of similarity. *Syst. Biol.* **45**, 380–385 (1996).

50. Mowinckel, A. M. & Vidal-Piñeiro, D. Visualization of brain statistics with R packages *ggseg* and *ggseg3d*. *Adv. Methods Pract. Psychol. Sci.* **3**, 466–483 (2020).

51. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).

52. Sánchez-Meca, J., Marín-Martínez, F. & Chacón-Moscoso, S. Effect-size indices for dichotomized outcomes in meta-analysis. *Psychol Methods* **8**, 448–467 (2003).

53. Lang, M. *et al.* mlr3: A modern object-oriented machine learning framework in R. *J. Open Source Softw.* **4**, 1903 (2019).

54. Wright, M. N. & Ziegler, A. Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Softw.* **77**, (2017).

55. Louppe, G. Understanding random forests: From theory to practice. *arXiv [stat.ML]* (2014) doi:10.48550/ARXIV.1407.7502.

56. Naimi, A. I. & Balzer, L. B. Stacked generalization: an introduction to super learning. *Eur J Epidemiol* **33**, 459–464 (2018).

57. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.* **67**, 301–320 (2005).

58. Noble, W. S. What is a support vector machine? *Nat Biotechnol* **24**, 1565–1567 (2006).

59. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.* **67**, 301–320 (2005).

60. Zhang, Z. Introduction to machine learning: k-nearest neighbors. *Ann Transl Med* **4**, 218 (2016).

61. Zhang, Z. Naïve Bayes classification in R. *Ann Transl Med* **4**, 241 (2016).

62. Noble, W. S. What is a support vector machine? *Nat Biotechnol* **24**, 1565–1567 (2006).

63. van der Laan, M. J., Polley, E. C. & Hubbard, A. E. Super learner. *Stat Appl Genet Mol Biol* **6**, Article25 (2007).

64. Singh, R. & Mangat, N. S. Stratified Sampling. in *Kluwer Texts in the Mathematical Sciences* 102–144 (Springer Netherlands, Dordrecht, 1996).

65. Wickham, H. *et al.* Welcome to the tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).

66. Snoek, J., Larochelle, H. & Adams, R. P. Practical Bayesian optimization of machine learning algorithms. *arXiv [stat.ML]* (2012) doi:10.48550/ARXIV.1206.2944.

67. Frazier, P. I. A tutorial on Bayesian optimization. *arXiv [stat.ML]* (2018) doi:10.48550/ARXIV.1807.02811.

68. Guo, X., Yin, Y., Dong, C., Yang, G. & Zhou, G. On the Class Imbalance Problem. in *2008 Fourth International Conference on Natural Computation* vol. 4 192–201 (2008).

69. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).

70. Darst, B. F., Malecki, K. C. & Engelman, C. D. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genet* **19**, 65 (2018).

71. Bickel, P. J. & Freedman, D. A. Asymptotic normality and the bootstrap in stratified sampling. *Ann. Stat.* **12**, 470–482 (1984).

72. Dietterich, T. G. Approximate Statistical Tests for Comparing Supervised Classification

Learning Algorithms. *Neural Comput* **10**, 1895–1923 (1998).

73. Ghanem, A. S., Venkatesh, S. & West, G. Multi-class Pattern Classification in Imbalanced Data. in *2010 20th International Conference on Pattern Recognition* 2881–2884 (IEEE, 2010).

Phenotype	Description	Sample Size Cases	Sample Size Controls
GPNoDep (shallow depression)	Individuals who sought help for nerves, anxiety, tension or depression from a GP (fields 2090/2100), but did not meet criteria for depression based on the touchscreen questionnaire (fields 4631/4598).	683	4789
GPPsy (shallow depression)	Individuals who sought help for nerves, anxiety, tension, or depression from a GP (fields 2090/2100).	9529	20477
Psypsy (shallow depression)	Individuals who sought help for nerves, anxiety, tension, or depression from a psychiatrist (fields 2090/2100).	2706	27353
SelfRepDep (shallow depression)	Individuals who self-reported having depression (field 20002) during a verbal interview with a trained nurse.	1944	20421
DepAll (shallow depression)	Individuals who reported at least one of the two cardinal symptoms of depression (low mood or apathy, fields 4631/4598) for at least two weeks on the Touchscreen questionnaire and sought help from either a GP or psychiatrist. This provides a more comprehensive, though still minimal, phenotyping approach.	1994	5491
ICD10Dep (deep depression)	This phenotype includes individuals with primary (data field 41202) or secondary (data field 41204) ICD-10 diagnoses of mood or affective disorders (F32-F39) from linked hospital admission records in the UK Biobank. The specific diagnoses include: depressive episodes (F32), recurrent depressive disorder (F33), persistent mood disorders (F34), other mood disorders (F38), and unspecified mood disorders (F39). This definition aligns with the "ICD10-coded depression" classification used in Howard et al. 2018.	432	15812
ICD10Dep_exclpsych (deep depression)	This phenotype includes individuals with primary or secondary ICD-10 diagnoses of mood or affective disorders (F32-F39), specifically: depressive episodes (F32), recurrent depressive disorder (F33), persistent mood disorders (F34), other mood disorders (F38),	392	15432

	and unspecified mood disorders (F39). Both cases and controls exclude individuals with diagnoses of schizophrenia, bipolar disorder, and other psychiatric conditions.		
LifetimeMDD (deep depression)	Individuals who met the DSM-V criteria for lifetime Major Depressive Disorder (MDD) based on responses to the Composite International Diagnostic Interview Short Form (CIDI-SF) in the UK Biobank's follow-up online questionnaire. This uses full DSM-5 criteria for diagnosing lifetime MDD, providing a stricter phenotyping definition.	3019	9458
MDDRecur (deep depression)	Individuals from the LifetimeMDD group who reported having two or more depressive episodes (field 20442), indicating recurrent MDD.	1869	9105

Table 1 | Spectrum of depression phenotypes in UK Biobank: from minimal to comprehensive definitions. This table presents a categorization of depression phenotypes in the UK Biobank, focusing on individuals with structural neuroimaging data. The phenotypes are arranged hierarchically, from the most minimal definitions (e.g., GPNoDep) to structured diagnostic criteria based on DSM-5 with multiple episodes (e.g., MDDRecur), building upon the work of Nai et al. (2020). Data collection in the UK Biobank involved multiple stages. Initially, all participants completed the Touchscreen Questionnaire, a broad screening tool for various health conditions and behaviors. Subsequently, a subset of participants completed the Online Mental Health Questionnaire, a more targeted assessment employing validated diagnostic instruments such as the Composite International Diagnostic Interview Short Form (CIDI-SF) to provide more precise psychiatric phenotyping. For the purposes of machine learning classification, we dichotomized these phenotypes into 'shallow' and 'deep' depression categories. Shallow depression encompasses individuals who sought help from their general practitioner (GPpsy), or psychiatrist (Psypsy), those who self-reported depression during nurse-led interviews (SelfRepDep), and those who additionally reported at least one of two cardinal symptoms of depression (low mood or anhedonia) for a minimum of two weeks (DepAll). Deep depression was defined using more stringent criteria: either electronic medical records with ICD-10 codes for depressive disorders, or meeting full diagnostic criteria on the CIDI-SF administered in the online mental health questionnaire.

	Deep vs Controls	Shallow vs Controls (Broad)	Shallow vs Controls (Narrow)
F1 Score	0.581 (0.564, 0.597)	0.720 (0.711, 0.729)	0.635 (0.623, 0.646)
Sensitivity	0.799 (0.780, 0.818)	0.800 (0.788, 0.811)	0.800 (0.786, 0.813)
Specificity	0.682 (0.669, 0.694)	0.655 (0.643, 0.667)	0.519 (0.505, 0.533)
PPV	0.456 (0.438, 0.474)	0.655 (0.643, 0.667)	0.526 (0.512, 0.540)
NPV	0.911 (0.901, 0.919)	0.761 (0.748, 0.774)	0.795 (0.781 - 0.809)
AUC	0.831 (0.819, 0.844)	0.807 (0.799, 0.816)	0.782 (0.772, 0.793)
AUPRC	0.716 (0.696, 0.735)	0.841 (0.832, 0.849)	0.792 (0.781, 0.802)

Table 2 | Performance metrics of random forest classifiers for deep and shallow depression phenotypes.
Random forest classifiers were trained on 256 neuroimaging features to discriminate deep depression (defined by DSM-V criteria for MDD or ICD-10 diagnoses) vs controls and shallow depression (based on self-reported symptoms) vs. controls. Models were developed using stratified sampling (70% training, 30% testing) with Bayesian optimization for hyperparameter tuning. In our depression phenotyping approach, individuals could meet multiple operational definitions across the spectrum. For the shallow vs control classifiers, we distinguished between broad and narrow definitions: the broad definition includes individuals who meet criteria for shallow depression regardless of whether they also meet criteria for deeper phenotypes, capturing the full spectrum as typically defined in population cohorts; the narrow definition includes only individuals who exclusively meet criteria for shallow depression without meeting deeper phenotype criteria. This broad approach reflects how depression is typically defined in large health systems and population cohorts, capturing the full spectrum of self-reported depression. For the deep vs controls model, the training set consisted of 11,978 controls and 3998 deep depression cases, with a test set of 5134 controls and 1714 deep depression cases. For the shallow (broad) versus controls classifier, the training set comprised 11,978 controls and 11,321 shallow depression cases, with a test set of 5134 controls and 4852 shallow depression cases. The shallow (narrow) versus controls classifier used a training set of 11,978

controls and 7996 shallow depression cases, with a test set of 5134 controls and 1714 shallow depression cases. Features were residualized for potential confounders including age, sex, assessment center, estimated total intracranial volume, head motion, and the top 20 genetic ancestry principal components. This residualization was performed using linear regression models for each feature, with the confounders as predictors. The residuals from these models were then used as input features for the classifiers. Sensitivity was standardized to 80% for all models to facilitate direct comparisons. Performance metrics are reported with 95% confidence intervals. As summarized in Table 2, the deep depression model showed a slightly higher AUC (0.831 vs 0.807), while the shallow depression (broad) classifier demonstrated stronger performance across several other key classification metrics, including a higher F1 score (0.720 vs 0.581) and PPV (0.655 vs 0.456) compared to the deep phenotype classifier. These differences reflect a more favorable balance between precision and recall for the shallow (broad) model at the standardized sensitivity of approximately 80%. The higher PPV indicates that when the shallow depression (broad) model predicts a positive case, it is correct 65.5% of the time, compared to 45.6% for the deep depression model, representing a lower false positive rate. Additionally, the shallow (broad) model achieved a higher AUPRC (0.849 vs 0.716), indicating better precision across different recall levels despite its slightly lower AUC. The shallow (narrow) model showed intermediate performance on most metrics (F1 score: 0.635, PPV: 0.526, AUC: 0.782, AUPRC: 0.792), suggesting that including individuals who also meet deep phenotype criteria enhances the discriminative ability of shallow depression classification.

Deep vs Control Task		
Metrics	Outcome-Specific (Deep vs Control Features)	Feature Transfer (Shallow vs Control Features)
F1 Score	0.585 (0.568, 0.602)	0.579 (0.563, 0.596)
Sensitivity	0.799 (0.780, 0.818)	0.799 (0.780, 0.818)
Specificity	0.689 (0.676, 0.702)	0.680 (0.667, 0.693)
PPV	0.462 (0.444, 0.480)	0.455 (0.437, 0.473)
NPV	0.912 (0.902, 0.920)	0.911 (0.901, 0.919)
AUC	0.840 (0.828, 0.852)	0.843 (0.831, 0.855)
AUPRC	0.745 (0.726, 0.763)	0.773 (0.756, 0.789)

Stratified Bootstrap Test (10,000)	Observed Difference	Permuted Difference (95% CI)	P-value (FDR corrected)
F1 Score	0.211	0.188, 0.235	0.499

Shallow vs Control Task		
Metrics	Outcome-Specific (Shallow vs Control Features)	Feature Transfer (Deep vs Control Features)
F1 Score	0.629 (0.617, 0.641)	0.632 (0.620, 0.644)
Sensitivity	0.799 (0.786, 0.813)	0.799 (0.786, 0.813)
Specificity	0.504 (0.490, 0.518)	0.512 (0.498, 0.526)
PPV	0.519 (0.505, 0.532)	0.522 (0.509, 0.536)
NPV	0.791 (0.776, 0.804)	0.793 (0.779, 0.806)
AUC	0.779 (0.769, 0.790)	0.766 (0.755, 0.776)
AUPRC	0.790 (0.779, 0.801)	0.756 (0.744, 0.769)

Stratified Bootstrap Test (10,000)	Observed Difference	Permuted Difference (95% CI)	P-value (FDR corrected)
F1 Score	-0.130	-0.142,	0.497

							-0.118	
AUC	0.003	-0.003, 0.010	0.511		**AUC**	-0.013	-0.018, -0.009	0.513
McNemar's chi-squared test					**McNemar's chi-squared test**			
$\chi^2 = 237.84$		df = 1	< 2.2e-16		$\chi^2 = 386.68$		df = 1	< 2.2e-16

Table 3 | Feature Transfer Performance Metrics. This table presents a comparison of performance metrics between outcome-specific and feature transfer models for both deep vs control and shallow vs control classification tasks. The outcome-specific models utilize the top 30% of features (76 features) trained to discriminate the respective task (i.e., shallow vs control features for the shallow task, deep vs control features for the deep task), while the feature transfer models apply the top 30% of features trained on the alternative task. For the shallow vs control analysis, only cases meeting the narrow definition of shallow depression (i.e., exclusively meeting criteria for shallow depression without deeper phenotypes) were included. Performance metrics include F1 Score, Sensitivity, Specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), Area Under the Receiver Operating Characteristic Curve (AUC), and Area Under the Precision-Recall Curve (AUPRC). Values in parentheses reflect the 95% confidence intervals around the point estimates for each metric, providing a measure of statistical uncertainty. Confidence intervals were estimated using a bootstrap resampling method with 10,000 iterations. For the deep vs control task, the outcome-specific and feature transfer models yielded similar F1 scores (0.585 [95% CI: 0.568-0.602] vs 0.579 [95% CI: 0.563-0.596], respectively) and identical sensitivity (0.799). However, the feature transfer model exhibited marginal improvements in AUC point estimates (0.842 vs 0.840) and AUPRC (0.773 vs 0.745), alongside slight reductions in specificity (0.580 vs 0.689) and PPV (0.455 vs 0.462). In the shallow vs control task, both models achieved similar F1 scores (0.629 [95% CI: 0.617-0.641] vs 0.632 [95% CI: 0.620-0.644], and similar specificity (0.504 vs 0.512) and PPV (0.519 vs 0.522). The outcome-specific model showed a higher AUC (0.779 [95% CI: 0.769-0.790] vs 0.766 [95% CI: 0.755-0.776]) and AUPRC (0.790 vs 0.756). Stratified bootstrap tests (10,000 iterations) for F1 score and AUC indicated no statistically significant differences between feature transfer and outcome-specific models, for either deep vs control or shallow vs control tasks. McNemar's chi-squared tests were highly significant for both tasks (deep vs control: $\chi^2 = 237.84$, $p < 2.2e-16$; shallow vs control: $\chi^2 = 386.68$, $p < 2.2e-16$). These significant McNemar's test results reveal that despite comparable

aggregate performance metrics, the outcome-specific and feature transfer models use distinct classification strategies, making different decisions in individual cases.

Class-Based Metrics	**Controls**	**Shallow Depression**	**Deep Depression**
F1 Score	0.718 (0.708, 0.728)	0.628 (0.613, 0.641)	0.589 (0.570, 0.607)
Sensitivity	0.736 (0.723, 0.747)	0.567 (0.550, 0.583)	0.660 (0.638, 0683)
Specificity	0.688 (0.675, 0.700)	0.880 (0.872, 0.888)	0.883 (0.876, 0.890)
PPV	0.702 (0.689, 0.714)	0.703 (0.510, 0.552)	0.531 (0.510, 0.552)
NPV	0.723 (0.710, 0.735)	0.802 (0.923, 0.934)	0.929 (0.923, 0.934)
AUC	0.771 (0.762, 0.780)	0.797 (0.787, 0.806)	0.841 (0.829, 0.853)
AUPRC	0.712 (0.698, 0.727)	0.765 (0.754, 0.777)	0.735 (0.717, 0.753)

Macro Average Metrics	**Multi-Class Classification** (Controls vs Shallow Depression vs Deep Depression)
F1 Score	0.645 (0.635, 0.655)
Sensitivity	0.654 (0.644, 0.664)
Specificity	0.817 (0.812, 0.822)
PPV	0.645 (0.636, 0.655)
NPV	0.818 (0.813, 0.823)
AUC	0.803 (0.795, 0.811)
AUPRC	0.738 (0.728, 0.747)

Table 4 | Multi-class random forest classification of controls, shallow depression, and deep depression. This table presents the performance metrics of a multi-class random forest classifier simultaneously discriminating between controls, shallow depression, and deep depression classes using all 256 candidate neuroimaging features. The left side displays class-based metrics, evaluating each class in a one-versus rest framework with 95% confidence intervals. Controls demonstrated the highest F1 score at 0.718 (0.708-0.728) and sensitivity (0.723-0.747), however, controls showed the lowest specificity at 0.688 (0.675-0.700) and the lowest AUC at 0.771 (0.762-0.780) among the three classes, indicating some difficulty in correctly excluding non-control cases and somewhat lower overall discriminative ability. Shallow depression showed moderate performance with an F1 score of 0.628 (0.613-0.641) and the lowest sensitivity at 0.567 (0.550-0.583), suggesting that the model correctly identifies only 56.7% of shallow depression cases when compared against both controls and deep depression. Furthermore, shallow vs rest demonstrated an intermediate AUC at 0.797 (0.7897-0.806) and high specificity at 0.880 (0.0.872-0.888), indicating a reliable identification of non-shallow depression cases. Furthermore, deep depression achieved an F1 score of 0.589 (0.570-0.607) with a relatively balanced

sensitivity at 0.660 (0.638-0.683) and high specificity at 0.883 (0.876-0.890). The deep depression class demonstrated the highest AUC at 0.841 (0.829-0.853), indicating comparatively higher discriminative ability across classification thresholds, despite having the lowest positive predictive value at 0.531 (0.510-0.552).

Deep vs Control Task (Super Learner)		**Shallow vs Control Task (Super Learner)**		**Shallow vs Deep Task (Super Learner)**		**Combined Depression vs Control Task (Super Learner)**	
F1 Score	0.565 (0.539, 0.592)	F1 Score	0.628 (0.615, 0.641)	F1 Score	0.723 (0.713, 0.750)	F1 Score	0.722 (0.712, 0.731)
Sensitivity	0.785 (0.755, 0.814)	Sensitivity	0.663 (0.647, 0.679)	Sensitivity	0.708 (0.687, 0.730)	Sensitivity	0.783 (0.772, 0.795)
Specificity	0.862 (0.852, 0.872)	Specificity	0.701 (0.688, 0.713)	Specificity	0.874 (0.863, 0.885	Specificity	0.612 (0.599, 0.625)
PPV	0.442 0.415, 0.470)	PPV	0.597 (0.581, 0.612)	PPV	0.737 (0.716, 0.759)	PPV	0.669 (0.657, 0.681)
NPV	0.966 (0.961, 0.971)	NPV	0.757 (0.745, 0.769)	NPV	0.857 (0.845, 0.868)	NPV	0.738 (0.725, 0.751)
AUC	0.911 (0.899 - 0.923)	AUC	0.764 (0.752, 0.775)	AUC	0.896 (0.887, 0.905)	AUC	0.814 (0.806, 0.822)
AUPRC	0.720 (0.687, 0.750)	AUPRC	0.779 (0.767, 0.790)	AUPRC	0.856 (0.843, 0.869)	AUPRC	0.859 (0.852, 0.866)

Table 5 | Performance metrics of super learner classifiers for deep depression and shallow depression phenotypes.
This table presents the performance metrics of a stacked ensemble classifier, also known as a super learner, for distinguishing deep depression from controls using all 256 neuroimaging features shown in Supplementary Table 1. The model employs a diverse set of five base learners (elastic net, *k*-nearest neighbors, random forest, naïve bayes, and a light gradient-boosting machine) with an XGBoost meta-learner. This approach leverages the strengths and varying assumptions of multiple algorithms to capture different aspects of the data and potentially outperform any single model. The classifier was trained using nested cross-validation to provide an unbiased estimate of performance. The outer loop used 5-fold cross-validation to assess overall performance, while the inner loop used

another 5-fold cross-validation to generate out-of-fold predictions from base learners as input for the meta-learner. As discussed further in Table 2, the shallow vs control classifier only included cases who exclusively meet criteria for shallow depression (based on self-reported symptoms), excluding those who met criteria for deep depression (EMR and CIDI-SF definitions). The deep vs control model demonstrates a sensitivity of 0.785 (0.755, 0.814), correctly identifying 78.5% of deep depression cases, with a specificity of 0.862 (0.852, 0.872), correctly identifying 86.2% of control subjects. When the model classifies a subject as having deep depression, there is a 44.2% probability this classification is correct (PPV: 0.442 (0.415, 0.470)). Notably, its negative predictive value (NPV) of 0.966 (0.961, 0.971) indicates a 96.6% probability that subjects classified as controls are truly non-depressed. The model's F1 score of 0.565 (0.539, 0.592) and AUC of 0.911 (0.899, 0.923) indicate good discriminative performance despite the modest PPV. The shallow vs control model shows moderate discriminative ability with an F1 score of 0.628 (0.615, 0.641) and an AUC of 0.764 (0.752, 0.775). This model correctly identifies 66.3% of shallow depression cases (sensitivity: 0.663 (0.647, 0.679)), with a specificity of 0.701 (0.688, 0.713) for correctly identifying control subjects. Furthermore, the shallow vs deep depression model shows a sensitivity of 0.708 (0.687, 0.730), correctly identifying 70.8% of deep depression cases, with a specificity of 0.874 (0.863, 0.885), correctly identifying 87.4% of shallow depression cases. This model achieves a PPV of 0.737 (0.716, 0.759), an F1 score of 0.723 (0.713, 0.750), and an AUC of 0.896 (0.887, 0.905), demonstrating effective discrimination between depression phenotypes. The combined depression vs control model exhibits a sensitivity of 0.783 (0.772, 0.795), identifying 78.3% of all depression cases, with a specificity of 0.612 (0.599, 0.625). This model's F1 score of 0.722 (0.712, 0.731) and AUC of 0.814 (0.806, 0.822) reflect its ability to distinguish depression cases from controls, though with lower specificity than the deep depression classifier. Across all models, the deep vs control classifier demonstrates superior performance on nearly all metrics, particularly in specificity, NPV, and AUC, suggesting neuroimaging features may be most effective at discriminating deep depression from healthy controls. The shallow vs deep model also shows strong discriminative ability, while the shallow vs control model exhibits the most modest performance, potentially indicating greater neurobiological overlap between shallow depression and healthy controls.

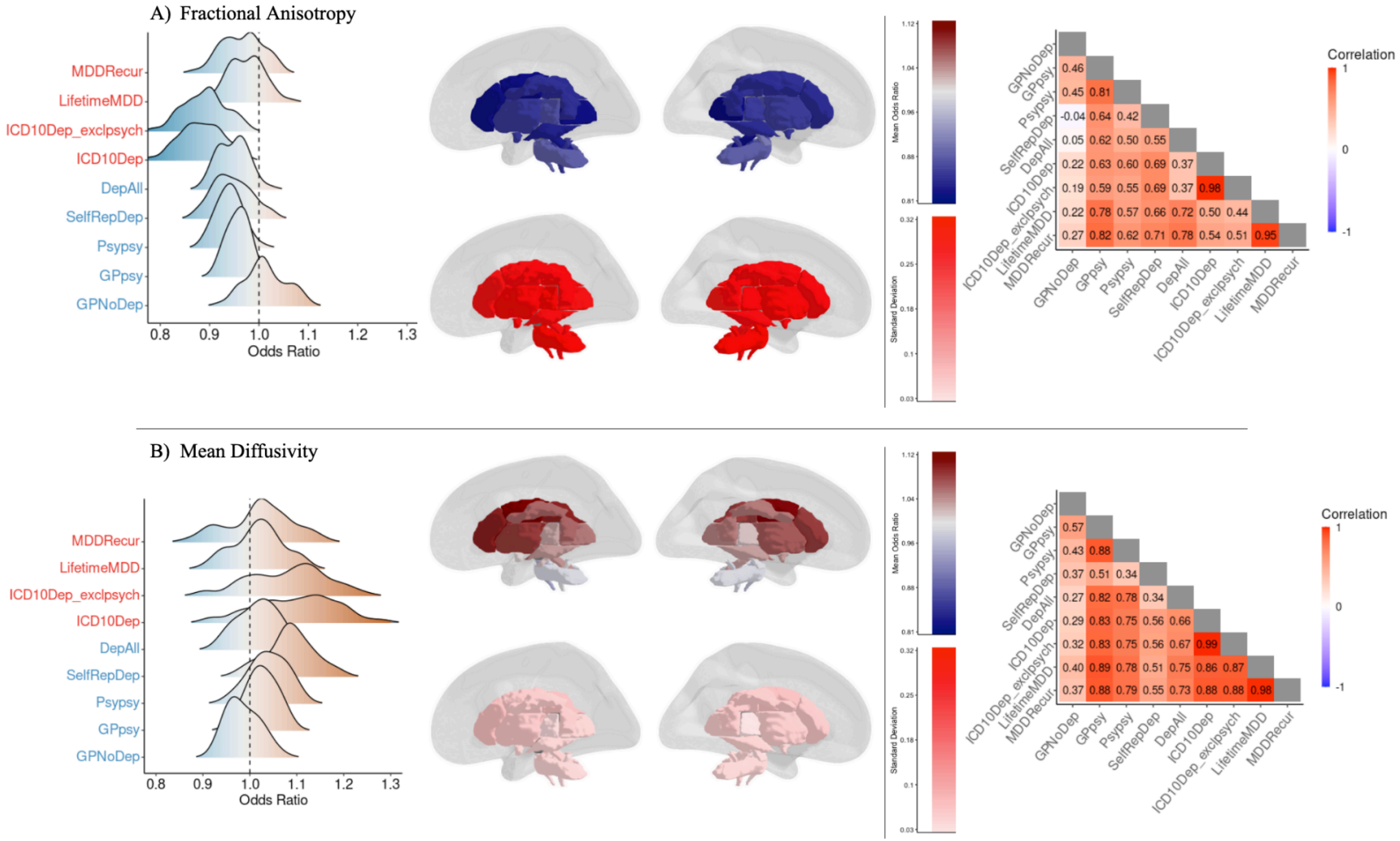
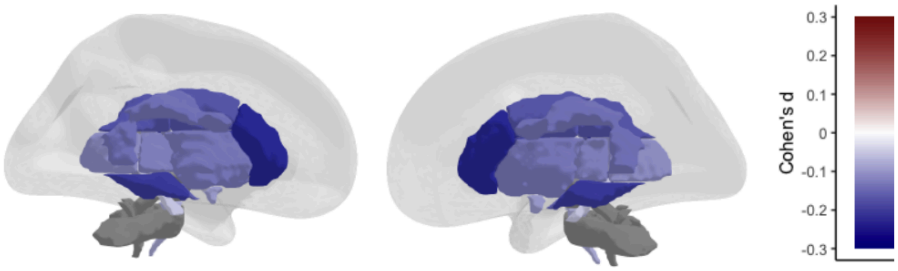


Figure 1 | Associations between white matter imaging-derived phenotypes and depression definitions. This figure illustrates the relationship between white matter microstructure and depression across varying phenotyping strategies. These results are derived from

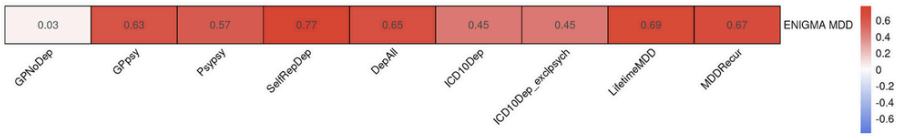
multivariable logistic regression models, where each included one of 256 z-score transformed neuroanatomical variables as the primary predictor, along with covariates (age, sex, intracranial volume, assessment center, and head motion), with depression status (case/control) as the outcome. The neuroanatomical variables compass white matter microstructure (FA and MD), gray matter morphometry (cortical thickness and surface area), subcortical volumes, and global brain volumetric measures.

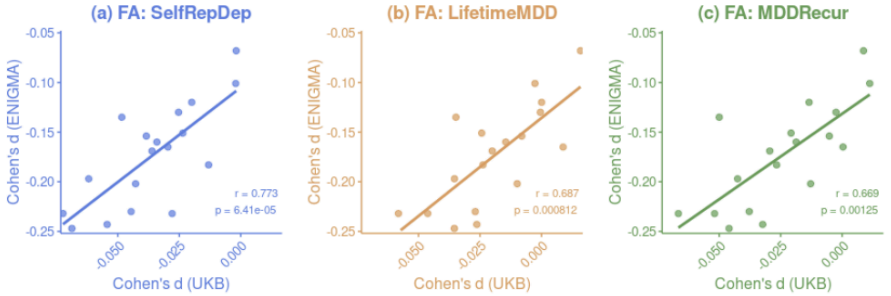
(A) Fractional anisotropy (FA) and **(B)** mean diffusivity (MD) associations are presented through density plots, 3D brain maps, and correlation heatmaps. Density plots (left) display odds ratios (ORs) for FA and MD across depression phenotypes, ranging from strictly defined (top) to minimally defined (bottom). Deeper phenotypes are depicted in red, shallower in blue. ORs < 1 are shown in light blue, >1 in light brown. The 3D brain maps (center) visualize mean ORs (top) and standard deviations (bottom) across all nine depression definitions. For FA, cooler colors (blues) indicate regions where increased FA is associated with decreased odds of depression. For MD, warmer colors (red) show areas where increased MD correlates with higher depression odds. Standard deviation maps highlight regions with the most variance in effects across definitions. Heatmaps (right) display Spearman's correlations of ORs across depression definitions, ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation). These heatmaps provide insight into the consistency of neuroanatomical associations across different depression phenotypes, with darker colors indicating stronger correlations.

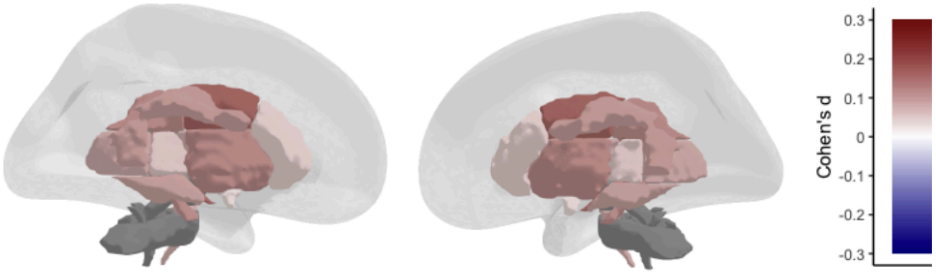
FA shows widespread reductions (ORs < 1) across white matter tracts, indicating that lower FA values are generally associated with higher odds of depression. The strongest associations are observed in the corpus callosum, corona radiata, and major association fibers. Concurrently, MD demonstrates widespread increases (ORs > 1), with the most pronounced effect in similar regions, particularly the corpus callosum, corona radiata, and internal capsule. Importantly, the variability in effects across phenotyping definitions differs between FA and MD. FA shows higher variability (standard deviations of ORs ranging from ~0.29 to ~0.32) compared to MD (standard deviations of ORs ranging from ~0.03 to ~0.09). This suggests that FA measures may be more sensitive to the depth and specificity of depression phenotyping strategies, while MD associations appear more consistent across different phenotyping approaches. These patterns suggest that altered white matter microstructure, characterized by both reduced FA and increased MD, is a pervasive feature across depression phenotypes, but with differing levels of consistency.



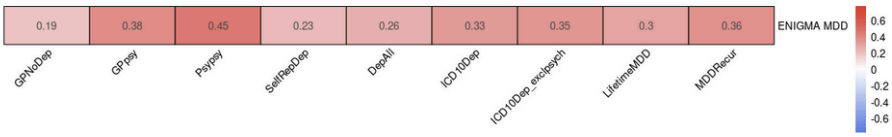
Fractional Anisotropy (FA)







Mean Diffusivity (MD)



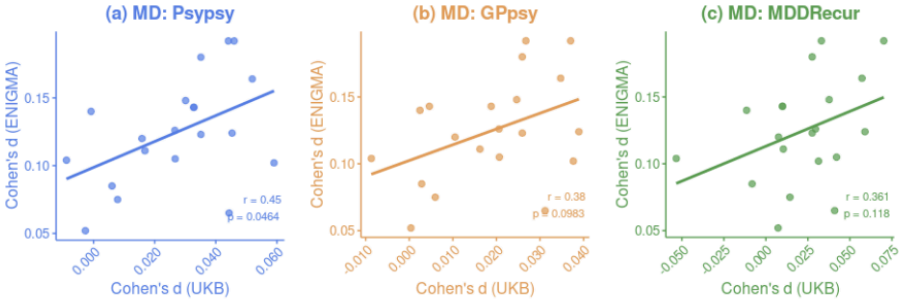


Figure 2 | Relationships between brain-psychopathology associations in UK Biobank and the ENIGMA MDD consortium in white-matter tracts. This figure illustrates the comparison of neuroimaging effect sizes for white matter diffusion metrics between the UK Biobank (UKBB) and effect size estimates from the latest and largest meta-analysis from the ENIGMA MDD Working Group. The figure is divided into two main sections: (A) Fractional Anisotropy (FA) and Mean Diffusivity (MD). Each section comprises three panels. The upper panels present 3D brain maps illustrating ENIGMA MDD Working Group effect size estimates (Cohen's d). For FA, cooler colors (blues) indicate regions where increased FA is associated with decreased odds of depression, while for MD, warmer colors (red) show areas where increased MD correlates with higher depression odds. The middle panels display heatmap vectors of Spearman's rank correlations between ENIGMA MDD effect size estimates and UKBB results across various depression phenotypes. For FA, stronger correlations are generally observed for more 'strictly' defined phenotypes such as LifetimeMDD and MDDRecur. In contrast, among MD measures, weaker correlations were observed between effect size estimates and UKBB results relative to FA, with the strongest positive correlations observed among shallower phenotypes, including Psypsy and GPpsy. The lower panels feature scatter plots for the three depression phenotypes showing the strongest correlations with ENIGMA MDD effect size estimates. These plots compare Cohen's d values from ENIGMA MDD (vertical axis) against those from UKBB (horizontal axis) across white matter tracts. For FA, the phenotypes displayed are SelfRepDep (blue, $r = 0.773$), LifetimeMDD (orange, $r = 0.687$), and MDDRecur (green, $r = 0.669$), demonstrating strong positive correlations across these phenotypes. For MD, the phenotypes are Psypsy (blue, $r = 0.45$), GPpsy (orange, $r = 0.38$), and MDDRecur (green, $r = 0.361$), showing moderate positive correlations. Each data point represents a specific white matter tract.

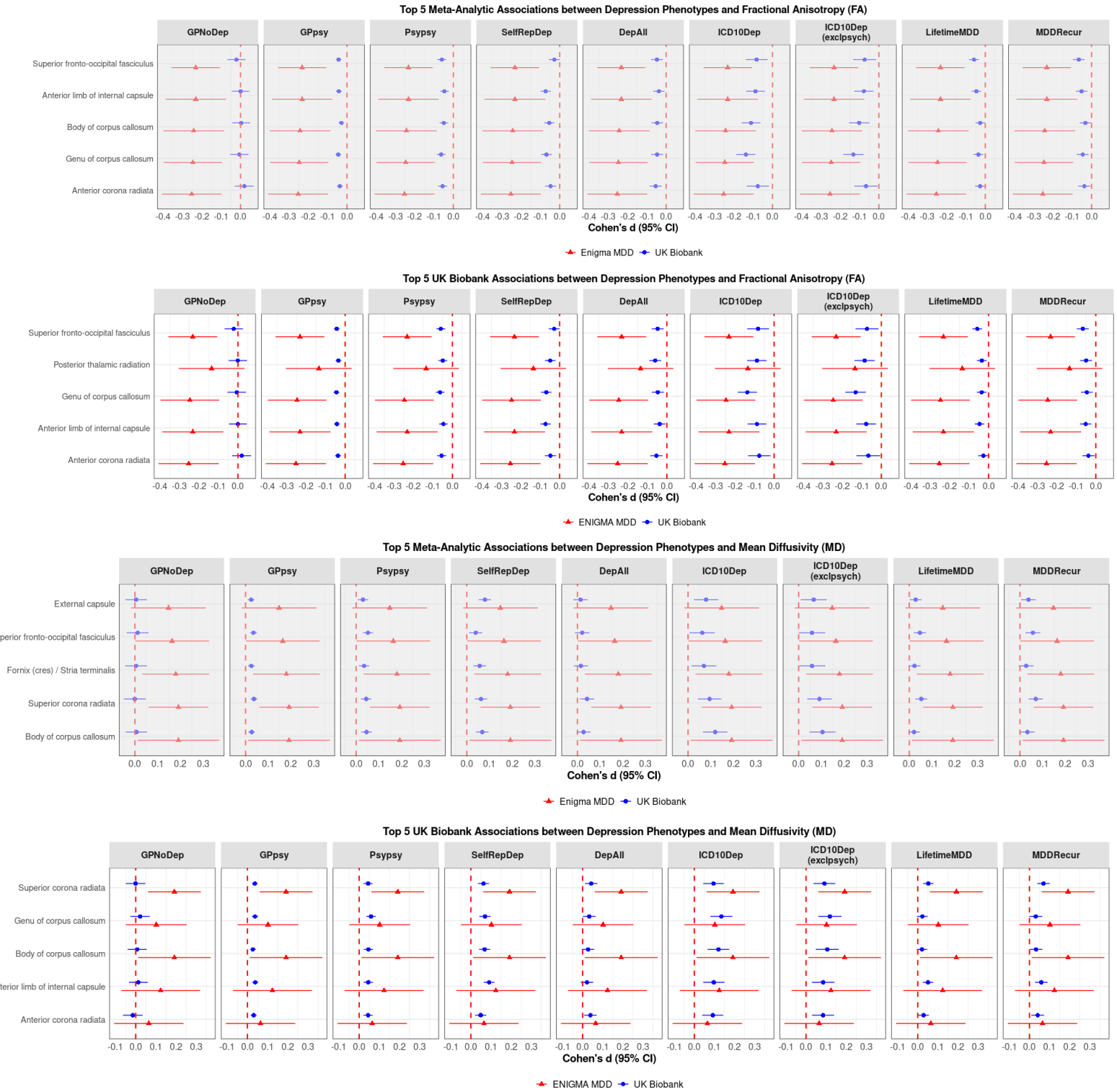


Figure 3 | Comparison of effect sizes between UK Biobank depression phenotypes and ENIGMA MDD meta-analysis for fractional anisotropy (FA) and mean diffusivity (MD) measures. This figure presents a comparison of effect sizes (Cohen's d) for white matter diffusion metrics between various depression phenotypes in the UK Biobank (UKBB) and the largest and most recent meta-analytic findings from the ENIGMA Major Depressive Disorder

(MDD) Working Group. The first panel presents the five strongest associations between depression and FA from ENIGMA MDD, juxtaposed with corresponding effect sizes across nine depression phenotypes in the UK Biobank. The meta-analytic effect size estimates demonstrate small to moderate reductions in fractional anisotropy in individuals with depression, with Cohen's d values ranging from -0.2 to -0.4. These reductions are observed in key white matter tracts, including the superior fronto-occipital fasciculus, anterior limb of internal capsule, body and genu of corpus callosum, and anterior corona radiata. In contrast, the UKBB data reveal substantially smaller effect sizes for most phenotypes among these tracts, clustering around zero. Notably, depression definitions based on electronic medical records (EMR) show slightly larger effect sizes, though still truncated relative to those reported in the meta-analysis. The second panel displays the five strongest associations between depression and fractional anisotropy in the UKBB, compared with their corresponding effect sizes in ENIGMA MDD. The UKBB effect sizes are generally small, with Cohen's d values close to zero across most phenotypes. The ENIGMA MDD effect size estimates for these regions, while also small, are negative across all five regions, indicating subtle reductions in fractional anisotropy associated with depression. The third panel, mirroring the first, presents the five strongest associations between depression and mean diffusivity from the ENIGMA meta-analysis, alongside the corresponding UKBB phenotypes. The ENIGMA effect sizes range from 0.1 to 0.2, indicating small increases in mean diffusivity in regions such as the external capsule, superior fronto-occipital fasciculus, fornix/stria terminalis, superior corona radiata, and body of corpus callosum. As with fractional anisotropy, the UK Biobank effect sizes are smaller, with depression definitions based on EMR showing the closest alignment with the ENIGMA results. The fourth panel illustrates the five strongest associations between depression and mean diffusivity in the UK Biobank, compared with their corresponding effect size estimates in ENIGMA MDD. The UK Biobank effect sizes vary across phenotypes but are generally small. The ENIGMA MDD effect sizes for these regions are consistently positive but small, suggesting subtle increases in mean diffusivity associated with depression. Across all panels, a consistent pattern emerges where effect sizes in the UK Biobank are generally smaller and more variable across phenotypes compared to the ENIGMA Major Depressive Disorder meta-analysis. Depression phenotypes defined using EMR in the UK Biobank show the closest alignment with the ENIGMA results, particularly for fractional anisotropy measures.

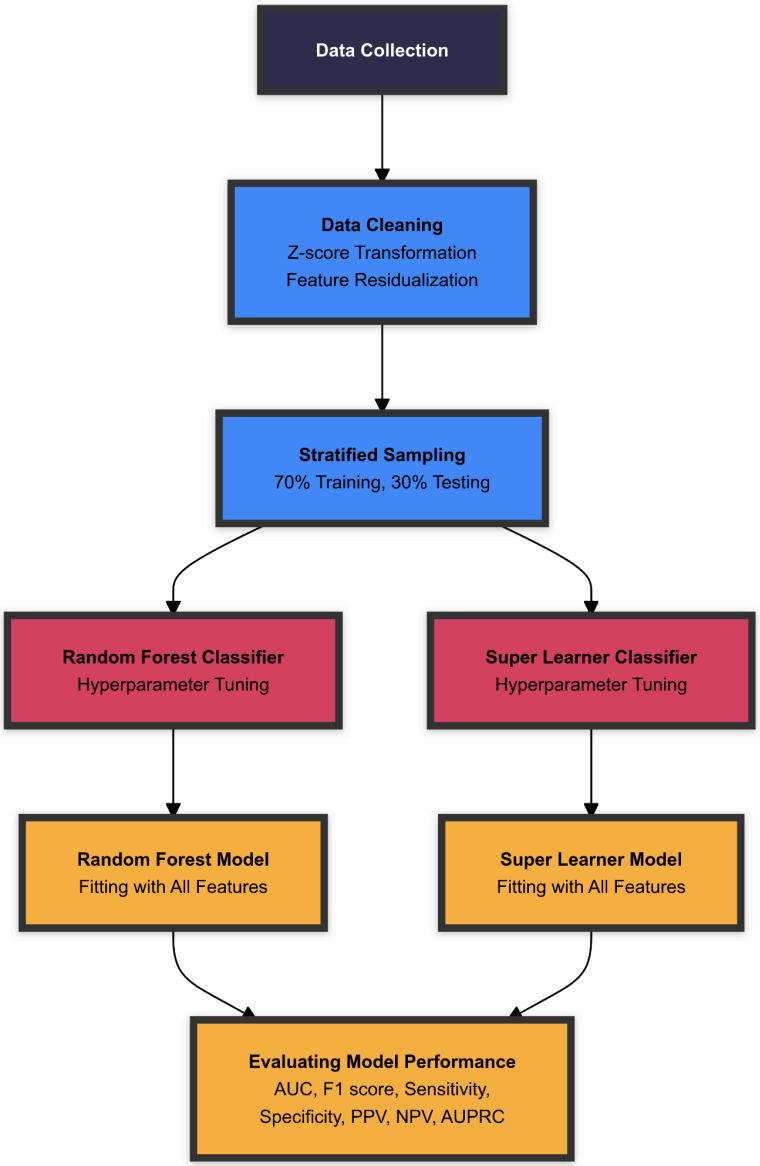


Figure 4 | Neuroimaging-based machine learning pipeline and feature transfer framework for depression classification. An illustration of the machine learning pipeline used to investigate the predictive power and generalizability of neuroanatomical features across various depression phenotypes. **a)** The primary machine learning pipeline begins with using preprocessed neuroanatomical data collected by the UK Biobank (detailed in Supplementary Table 1). After data cleaning through z-score transformation and feature residualization, stratified sampling splits the data into training (70%) and testing (30%) sets. Two classifiers are implemented - Random Forest and Super Learner - both using Bayesian Optimization for hyperparameter tuning. Model performance is evaluated using multiple metrics, including AUC, F1 Score (harmonic mean of precision and recall), sensitivity, specificity, PPV, NPV, and AUPRC. **b)** The feature transfer framework assesses neuroanatomical feature generalizability across depression phenotypes. From 256 neuroimaging features, recursive feature elimination

selects the top 30% (76) features for both deep and shallow depression versus controls models. These features are then applied reciprocally: deep depression features classify shallow depression and vice versa. Performance comparisons between original and transferred feature models undergo statistical testing via stratified bootstrap and McNemar's tests. This approach evaluates both predictive performance metrics and feature importance rankings across depression phenotypes, identifying which neuroanatomical correlates best predict different phenotypes and their cross-phenotype generalizability.

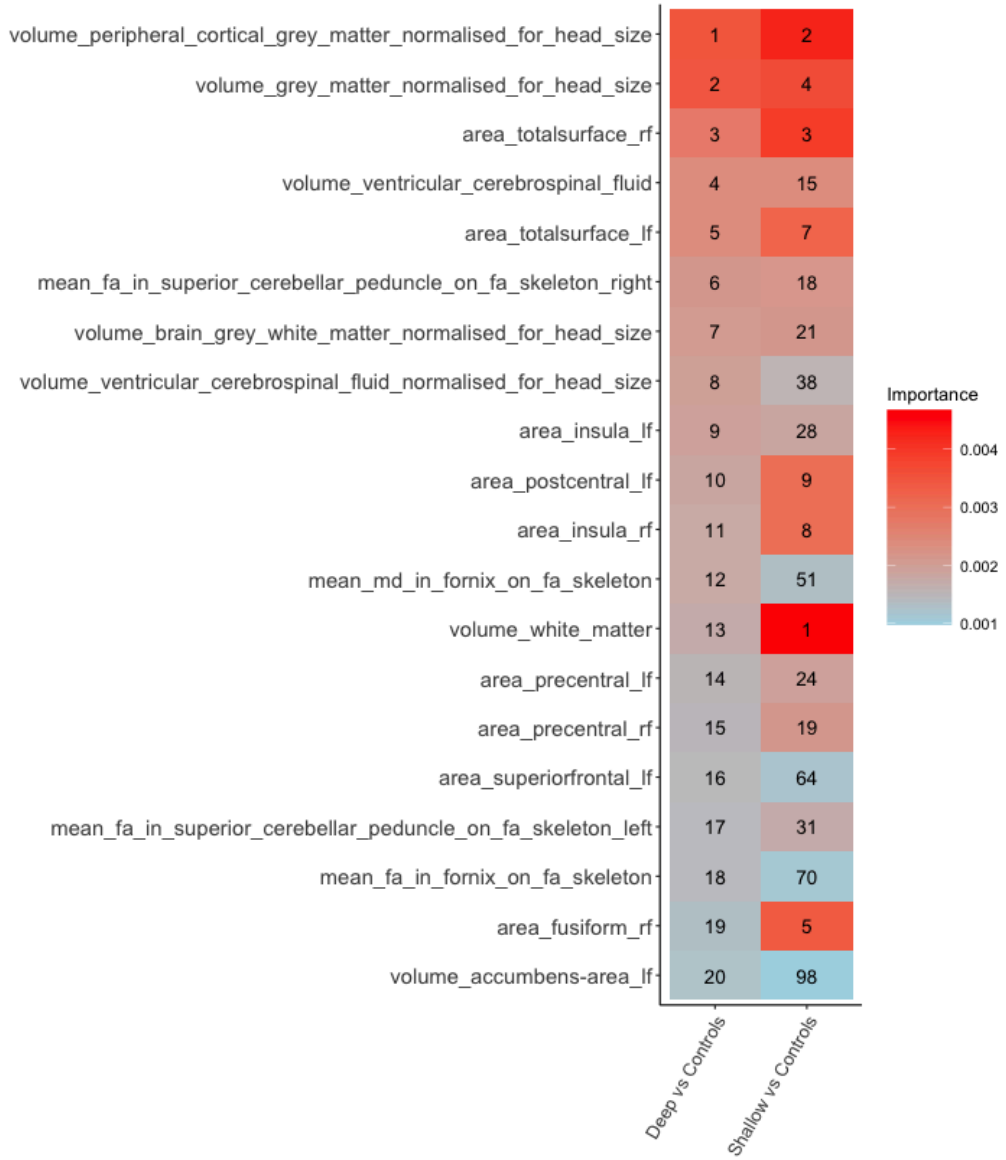


Figure 5 | Total feature importance ranking across shallow and deep depression models. This figure illustrates the relative importance of the top 20 neuroimaging features in discriminating deep depression from controls, and shallow depression from controls using random forest classifiers. Feature importance was calculated using permutation-based methods implemented in the Ranger package in R, following Altmann et al.'s (2010) approach. For each feature, the decrease in out-of-bag (OOB) prediction accuracy after random permutation of the feature's values was measured, with larger decreases indicating greater feature importance. Warmer colors (red) indicate higher importance, while cooler colors (blue) represent lower importance. Overall, the deep depression model prioritizes volume measurements, with normalized volume of peripheral cortical gray matter and normalized gray matter volume ranking 1st and 2nd respectively (2nd and 4th in shallow model). Five of the top eight features in the deep depression model are volume-related. In contrast, the shallow depression model shows

different prioritization, with volume of white matter as the top feature (ranked 13th in the deep model). The area measurements show varying importance between models - total surface area right ranks 3rd in both models, while total surface area left ranks 5th in deep but 7th in shallow. Some features show marked differences in importance - volume of ventricular cerebrospinal fluid ranks 4th in deep but 15th in shallow depression, and area fusiform right ranks 19th in deep but 5th in shallow. The volume of accumbens area left shows the most dramatic difference, ranking 20th in deep but 98th in shallow depression.

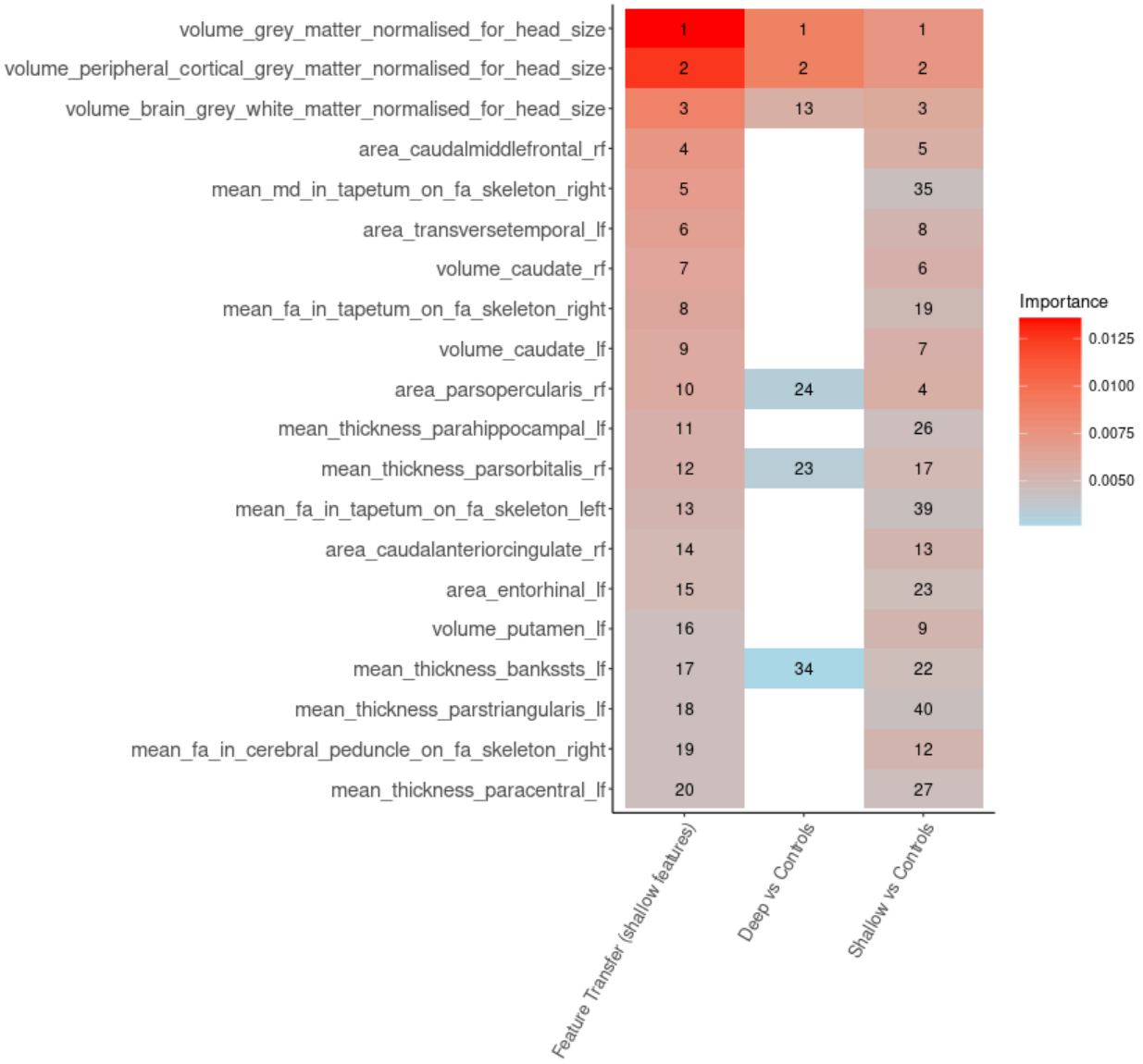


Figure 6 | Feature transfer framework using top shallow features and importance comparison across depression models. This figure illustrates the feature transfer approach comparing three models: shallow depression vs controls, deep depression vs controls, and a feature transfer model that applies shallow depression-derived features to deep depression classification. The analysis focuses on the top 30% of features (76 total) identified through recursive feature elimination (RFE) in the training set. Feature importance was calculated using permutation-based methods in the Ranger R package. This approach measures each feature's importance by first computing the model's prediction accuracy on out-of-bag samples, then randomly permutes the feature's values to break any association with the outcome. The resulting decrease in prediction accuracy after permutation, averaged across all trees in the forest, represents the feature's importance - where larger decreases indicate greater importance. The heatmap visualizes feature rankings and relative importance (red indicating higher importance, blue lower) across all three models. While the shallow depression and feature transfer models

share the same feature set, the deep depression model shows limited overlap, evidenced by numerous blank spaces indicating features absent from its top 30% (76 features in total). Notable ranking differences exist even between models sharing feature sets - for example, the 5th-ranked feature in the feature transfer model (mean md in tapetum on fa skeleton right) ranks 35th in the shallow depression model, and the 8th-ranked feature (mean fa in tapetum on fa skeleton right) ranks 19th. However, the top two features maintain consistent importance across all models.

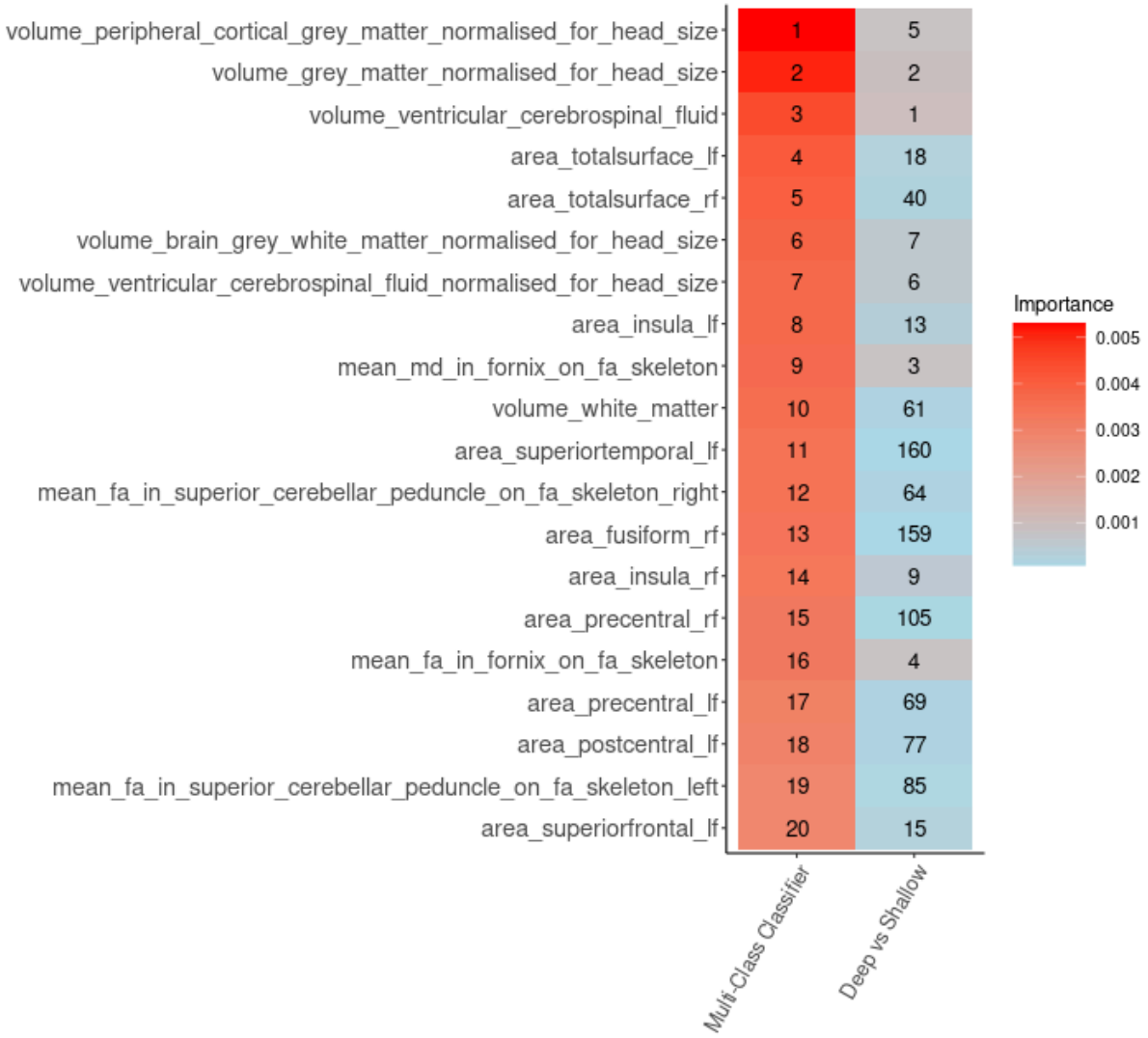


Figure 7 | Comparison of feature importance rankings between multi-class and deep vs shallow depression classifiers. This figure compares the relative importance of neuroimaging features between two random forest models: a multi-class classifier distinguishing between controls, shallow depression, and deep depression (left column), and a binary classifier differentiating deep from shallow depression (right column). Feature importance was calculated using permutation-based methods, where larger decreases in model accuracy after feature permutation indicates greater importance, represented by warmer colors in the heatmap. The visualization reveals both consistencies and divergences is how these models prioritize neuroanatomical features. Volume-based measurements demonstrate high relative importance in the multi-class classifier, with peripheral cortical gray matter, total gray matter, and ventricular cerebrospinal fluid volumes occupying the top three positions. Surface area measurements show notably different rankings between models - for example, total surface area (left) ranks 4th in the multi-class classifier, but 18th in the deep vs shallow comparison. White matter measurements, including both volumetric and microstructural (mean FA/MD) features, appear consistently

important but with varying rankings. This pattern suggests that while volume-based measurements are strongly associated with distinguishing depression from healthy controls, the differentiation between ‘shallower’ and ‘deeper’ depression phenotypes relies on a broader range of neuroanatomical characteristics.