**2. Samples and Methods**

***2.1 Participants***

We used data from N=36,931 participants from the UK BioBank (UKBB) for whom imaging and diagnostic information was available (Table 1) to train the ML classifier. We used participants’ diagnoses from inpatient records coded according to the International Classification of Disease version 10 (ICD-10). Major depressive disorder (MDD) cases were defined as participants with an ICD-10 code for major depressive episode (F32), recurrent depressive disorder (F33), or persistent mood/affective disorders (F34). MDD cases could not have other neuropsychiatric conditions.[ list either a full list of codes you define as neuropsychiatric conditions or something like this: “any other F or G diagnosis”. Please verify what you actually did]. Participants who were free of MDD or any other neuropsychiatric condition were treated as SMI controls. See Table 1 for descriptions of MDD cases and SMI controls..

|  |  |  |  |
| --- | --- | --- | --- |
| Table 1. Full neurological participant data in the training/testing set (UKBB) | | | |
| Group | N | Sex (M/F) | Age±std |
| Major Depressive Disorder | 3,848 | 1368/2480 | 62.1±6.4 |
| SMI Control | 33,083 | 15,844/17,239 | 63.8±6.9 |
|  |  |  |  |

We used data from N=347 participants from the the Amish Connectome Project (ACP) as the replication dataset. In the ACP, N=77 participants were diagnosed with MDD and N=270 controls without any Axis I diagnoses were included in the current analysis (Table 2). Participants with any other neuropsychiatric disorders were excluded.

|  |  |  |  |
| --- | --- | --- | --- |
| Table 2. Full neurological participant data in the replication set (ACP) | | | |
| Group | N | Sex (M/F) | Age±std |
| Major Depressive Disorder | 77 | 55/12 | 62.1±6.4 |
| SMI Control | 270 | 118/142 | 63.8±6.9 |
| Total | 347 | 140/197 | 63.6±7.5 |
|  |  |  |  |

Individuals were recruited from seventeen nuclear families from Lancaster County, PA, who could be combined into a single extended family that connected them across eight generations based on genealogical records maintained by the old-order Amish community and incorporated into the NIH Anabaptist Genealogy Database which traces back to the founders 2. Exclusion criteria included major medical and neurological conditions that might affect gross brain structures such as developmental disability, head trauma, seizure, stroke, or transient ischemic attack. All subjects provided written informed consent on forms approved by the Institutional Review Board (IRB) of the University of Maryland Baltimore. Diffusion data were preprocessed using ENIGMA Diffusion pipeline 3, 4

We also used two verification sets of SMI-free participants to (explain a bit why and what you are “verifying” here for readers less familiar with ML). This included 1,114 SMI-free participants from the Human Connectome Project-Young Adult (HCP) (cite), as well as 11,394 SMI-free participants from the Adolescent Brain Cognitive Development (ABCD).

|  |  |  |  |
| --- | --- | --- | --- |
| Table 3. Full neurological participant data in the first verification set (HCP) | | | |
| Group | N | Sex (M/F) | Age±std |
| SMI Control | 1,114 | 507/606 | 28.8±3.7 |

|  |  |  |  |
| --- | --- | --- | --- |
| Table 4. Full neurological participant data in the second verification set (ABCD) | | | |
| Group | N | Sex (M/F) | Age±std |
| SMI Control | 11,394 | 5,967/5,426 | 9.9±0.6 |

The ABCD Study is a study of US children born between 2006 and 2008. A total cohort of n=11,394 (5,967/5,426 M/F; average age 9.9±0.6) was recruited from 22 sites (with one site no longer active) and is being followed for at least ten years. Eligible children were recruited from the household populations in defined catchment areas for each of the study sites between September 2016 and October 2018.

***2.2 Imaging Protocols and Processing***

We selected 75 phenotype traits including 33 cortical gray matter thicknesses (GMT), 8 gray matter subcortical volumes (GMV), 24 white matter fractional anisotropies (WMFA) from three neuroimaging domains and extracted from ACP, HCP, UKBB and ABCD imaging data.

***2.2.1 UK BioBank (UKBB)***

The UKBB imaging data were collected using a Siemens Skyra 3T scanner and a standard 32-channel RF head coil. The T1-weighted (T1w) imaging data were acquired with the following settings (resolution=1x1x1 mm, FOV=208x256x256 mm, duration=5 minutes, 3D MPRAGE, sagittal, in-plane acceleration iPAT=2, prescan-normalize). The diffusion data were collected with a resolution of 2x2x2 mm and two diffusion shells of *b*=1000 and 2000 s/mm2 with 50 diffusion directions per shell and 5 *b*=0 images (FOV=104x104x72, duration=7 minutes). The neuroimaging traits were extracted using the UKBB processing pipeline. More information on image acquisition, processing and quality control could be found in the UKBB Brain Imaging Documentation (https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/brain\_mri.pdf) (Alfaro-Almagro et al., 2018; Miller et al., 2016).

***2.2.2 Amish Connectome Project (ACP)***

The ACP imaging data were collected at Maryland Psychiatric Research Center using a Siemens Prisma 3 Tesla scanner using a 64-channel coil. [Typically we describe T1 first.] DWI data were collected using an expansion of the HCP protocol that consisted of 6 shells of b-values (b=600, 900, 1200, 1500, 1800, and 3000 s/mm2) with 98 isotropically distributed diffusion-weighted directions per shell collected twice with the reversal of the phase encoding and readout gradients (anterior-to-posterior AP and posterior-to-anterior PA) to correct for spatial distortions, including twenty b=0 images interleaved within the acquisition. The T2??? data was collected using a multiband, echo-planar, spin-echo, T2-weighted sequence (TE/TR/Multiband Factor=97/4000ms/4 with the FOV=200 mm) with an isotropic spatial resolution of 1.6mm. that was combined with DESIGNER diffusion preprocessing tools including advance denoising, Gibbs ringing correction, and correction of the EPI distortions 5. FA maps were obtained by fitting the diffusion tensor model using the FSL-FDT toolkit 6.

***2.2.3 Human Connectome Project (HCP)***

The HCP imaging data were collected at Washington University in St. Louis using a customized Siemens Magnetom Connectome 3 Tesla scanner and a 32-channel head coil. The T1w imaging data were acquired with 0.7 mm isotropic, FOV=224x224 mm, 3D MPRAGE, duration=8 minutes, TR/TE/TI=2400/2.14/100. The diffusion imaging data were collected using a single-shot, single refocusing spin-echo, echo-planar imaging sequence with 1.25 mm isotropic spatial resolution (TE/TR=89.5/5520 ms, FOV=210x180 mm), 90 directions with right-to-left and left-to-right phase encoding polarities for each of the four diffusion weightings (b=1000, 2000, and 3000 s/mm2). Then T1w and diffusion imaging data were processed using the MR structural pipeline and diffusion pipeline, respectively. Details on the image acquisition, processing pipelines and quality controls are available online at (https://www.humanconnectome.org/study/hcp-young-adult/document/1200-subjects-data-release).

***2.2.4 Adolescent Brain Cognitive Development (ABCD)***

The ABCD imaging were collected at 21 research sites across the United States data using ABCD imaging protocol and three 3T Scanner platforms including Siemens Prisma, General Electric (GE) 750 and Philips. The ABCD imaging protocol is harmonized for three scanner platforms and use of multi-channel coils using a standard adult-size coil. The T1w imaging data were acquired in 1×1×1 mm resolution with the following parameters (Siemens Prisma: FOV=176×256×256, duration=7 minutes, axonal, TR/TE/TI =2500/2.88/1060 ms, flip angle=8 degrees; Philips: FOV=225×256×240, duration=6 minutes, axial, TR/TE/TI =6.31/2.9/1060 ms, flip angle=8; GE: FOV=208×256×256, duration=6 minutes, axial, TR/TE/TI =2500/2/1060, flip angle=8) and diffusion imaging data were collected in 1.7×1.7×1.7 mm resolution (Siemens Prisma: FOV=81×240×240, duration=7 minutes, TR/TE=4100/88, flip angle=90, diffusion directions=96, b=500, 1000, 2000 ,and 3000 s/mm²; Philips: FOV=81×240×240, duration=9 minutes, TR/TE=5300/89 ms, flip angle=78 degrees, diffusion directions=96, b=500, 1000, 2000 ,and 3000 s/mm²; GE: FOV=81×240×240, duration=8 minutes, TR/TE=4100/81.9 ms, flip angle=77, diffusion directions=96, b=500, 1000, 2000 ,and 3000 s/mm²). Then T1w and diffusion imaging data were processed using the freesurfer pipeline.

**Regional Vulnerability Index**

RVI scores were calculated using the 75 brain-wide structures from GMT, GMV, and WMFA domain based on the protocol documented in Kochunov, Zavaliangos-Petropuli, et al. (**2020**) using the “RVIpkg” in [R] software. Briefly, the effects of age, sex, intracranial volume, and/or scanning site were first regressed out from the imaging phenotypes using linear regression. The residuals from the linear model were inverse-normalized based on normal quantile, and then standardized to z-scores based on the average and standard deviation of the controls. The Pearson's correlation coefficient was then calculated between an individual's z-scores and corresponding effect sizes reported by the ENIGMA consortium in MDD and SSD. The RVIs were then Fisher's *z* transformed to enhance normality.

**2.3 Machine learning algorithms**

We used three machine learning algorithms, including linear model, SVM, and Extreme Gradient Boosting, to train models that differentiates MDD cases from SMI-controls from the UKBB training dataset.

**2.3.1 Linear Model**

The linear learning algorithm is similar in its approach to the RVI method where a vector of weights is used to derive a vulnerability index. The linear model derives this vector using a training dataset. For a subject i, the phenotype vector **xi** coded the 64 regional values from ENIGMA workflow. The training set of N subjects is coded as an N x 64 matrix **X** . The labels are stored as a vector **y** of length N where values 0 were assigned to Control and 1 indicated MDD diagnosis. The aim of the training step is to calculate the weight vector **w** of 1x64 such that the vector **s** obtained as the dot product of X and w eq 1

(1)

will minimize the difference between vectors **y** and s Eq 2.

Min ( (2)

A logarithm of the sigmoid function Eq 3 was used as the cost function during training [2].

where is

(3)

A Gradient Descent (GD) method to calculate the **w** that minimized the cost function [2][1].

***2.3.2 Support Vector Machines (SVM)***

SVM’s is a supervised non-linear learning technique that finds planes to separate the data points of various classes [7]. SVM can be used for classification of that data that is linearly inseparable and can model non-linear decision boundaries by performing a transformation of the vector **xi** , noted as . The solution is:

(4)

Where w is the normal direction of the decision hyperplane ( that separates the classes and where b is a threshold value. The algorithm is trained by maximizing the distance from the nearest point Φ(x0) to the decision boundary. It is found using a differential equation below with respect to w. refers to the Euclidean distance of the vector w.

(6)

Once the optimal weight is found, the decision function (Eq. 4) tells us how close an unknown point is to the line we are (close to the boundary means a low-confidence decision). Positive decision values mean true while negative decision values denote false. Platt scaling is then invoked to convert these distances from the hyperplane to probability values for a datapoint being in a positive class. A and B refer to scalar parameters that are learned by the algorithm during training [5].

(7)

***2.3.3 Extreme Gradient Boosting (XGBoost)***

XGBoost is a modification of the random forest decision tree algorithm that uses gradient boosting training approach [8][9]. The XGBoost uses the additive output of decision trees versus defining the combination of the outputs of multiple decision trees used in the original algorithm. The output value O (Eq. 8) is the probability of the data point belonging to the positive class. The equation is like the original gradient boosting algorithm, however is included to handle regularization and should be explored to reduce overfitting and improving performance. [8].

(8)

This value is found through minimization of the logistic loss function in Eq. 9 where yi refers to the observation and pi refers to the sigmoid function of the predicted ŷ.

If x is the input vector for an individual subject of a training set and y is the corresponding label, then the optimal O satisfies the condition in Eq 3, where is replaced with the initial prediction and the output value.

(10)

Using the first derivative (gradient), the second derivative (hessian) of Eq. 2 with respect to , and Taylor series approximations, we can expand Eq. 3:

Minimizing this equation is done by performing numerical differentiation with respect to O, setting the derivative to zero, and solving for O:

The optimal output value in Eq. 5 represents the probability that a certain leaf in the decision tree belongs to the positive class and is equivalent to Eq. 1. An optimal output value will be calculated for each leaf in the tree, one for every patient. We used GridSearch to find the optimal hyperparameters for the XGBoost algorithm [REF]. Hyperparameters are variables that provide a balance between regularization and generalization which refers to the model specificity of the training data. Using GridSearch, we construct a grid of potential hyperparameter values. Each iteration tries a set of hyperparameters in a certain sequence. It tracks the model performance when fitting the model with every conceivable set of hyperparameters. The configuration with the highest AUC is chosen as the final model to predict the remaining ACP, HCP, and ABCD datasets.

2.4 Model replication, verification, and comparison

The models were compared using a small subset of the UKBB data known as the testing set. Area under the curve (AUC) was the main evaluation tool used to compare the different models. Once the model was trained on the UKBB training set, the best performing model on the UKBB testing set was then deployed on the ACP, HCP and ABCD datasets (verification sets). The model will return a probability for each participant in the verification set and each sets respective ML probabilities were plotted against the RVI and AVI values. We can then see how these probabilities in each set correlate with the RVI and AVI values and whether there is a consensus between the verification sets.