NeuroImage 283 (2023) 120412

Contents lists available at ScienceDirect

NeuroImage

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

Neuroimaging-based classification of PTSD using data-driven computational approaches: A multisite big data study from the ENIGMA-PGC PTSD consortium

Xi Zhu [a,b], Yoojean Kim [b], Orren Ravid [b], Xiaofu He [a], Benjamin Suarez-Jimenez [c], Sigal Zilcha-Mano [d], Amit Lazarov [e], Seonjoo Lee [a,b], Chadi G. Abdallah [f,g], Michael Angstadt [h], Christopher L. Averill [f,g], C. Lexi Baird [i], Lee A. Baugh [j], Jennifer U. Blackford [k], Jessica Bomyea [l], Steven E. Bruce [m], Richard A. Bryant [n], Zhihong Cao [o], Kyle Choi [l], Josh Cisler [p], Andrew S. Cotton [q], Judith K. Daniels [r], Nicholas D. Davenport [s], Richard J. Davidson [t], Michael D. DeBellis [i], Emily L. Dennis [u], Maria Densmore [v], Terri deRoon-Cassini [w], Seth G. Disner [s], Wissam El Hage [x], Amit Etkin [y], Negar Fani [z], Kelene A. Fercho [aa], Jacklynn Fitzgerald [ab], Gina L. Forster [ac], Jessie L. Frijling [ad], Elbert Geuze [ae], Atilla Gonenc [af], Evan M. Gordon [ag], Staci Gruber [af], Daniel W Grupe [t], Jeffrey P. Guenette [ah], Courtney C. Haswell [i], Ryan J. Herringa [ai], Julia Herzog [aj], David Bernd Hofmann [ak], Bobak Hosseini [al], Anna R. Hudson [am], Ashley A. Huggins [i], Jonathan C. Ipser [an], Neda Jahanshad [ao], Meilin Jia-Richards [ap], Tanja Jovanovic [aq], Milissa L. Kaufman [ar], Mitzy Kennis [ae], Anthony King [h], Philipp Kinzel [as,at], Saskia B.J. Koch [au], Inga K. Koerte [as,at], Sheri M. Koopowitz [an], Mayuresh S. Korgaonkar [av], John H. Krystal [g], Ruth Lanius [aw], Christine L. Larson [ax], Lauren A.M. Lebois [ay,az], Gen Li [ba], Israel Liberzon [bb], Guang Ming Lu [bc], Yifeng Luo [o], Vincent A. Magnotta [bd], Antje Manthey [be], Adi Maron-Katz [y], Geoffery May [bf], Katie McLaughlin [bg], Sven C. Mueller [am], Laura Nawijn [bh], Steven M. Nelson [bi], Richard W.J. Neufeld [v], Jack B Nitschke [t], Erin M. O'Leary [q], Bunmi O. Olatunji [bj], Miranda Olff [ad], Matthew Peverill [bk], K. Luan Phan [bl], Rongfeng Qi [bc], Yann Quidé [n,bm], Ivan Rektor [bn], Kerry Ressler [ay,az], Pavel Riha [bn], Marisa Ross [bo], Isabelle M. Rosso [ay,az], Lauren E. Salminen [ao], Kelly Sambrook [bk], Christian Schmahl [aj], Martha E. Shenton [at], Margaret Sheridan [bp], Chiahao Shih [q], Maurizio Sicorello [aj], Anika Sierk [be], Alan N. Simmons [bq], Raluca M. Simons [br], Jeffrey S. Simons [br], Scott R. Sponheim [s,bs], Murray B. Stein [l], Dan J. Stein [an], Jennifer S. Stevens [z], Thomas Straube [ak], Delin Sun [i], Jean Théberge [v], Paul M. Thompson [ao], Sophia I. Thomopoulos [ao], Nic J.A. van der Wee [bt], Steven J.A. van der Werff [bt], Theo G.M. van Erp [bu], Sanne J.H. van Rooij [z], Mirjam van Zuiden [ad], Tim Varkevisser [ae], Dick J. Veltman [bh], Robert R.J.M. Vermeiren [bt], Henrik Walter [be], Li Wang [ba,bv], Xin Wang [q], Carissa Weis [w], Sherry Winternitz [ar], Hong Xie [q], Ye Zhu [ba], Melanie Wall [a,b], Yuval Neria [a,*], Rajendra A. Morey [i,*]

[a] *Department of Psychiatry, Columbia University Medical Center, New York, NY, USA*
[b] *New York State Psychiatric Institute, New York, NY, USA*
[c] *University of Rochester, Rochester, NY, USA*
[d] *University of Haifa, Haifa, Israel*
[e] *Tel-Aviv University, Tel Aviv, Israel*
[f] *Baylor College of Medicine, Houston, TX, USA*

* Corresponding author at: New York State Psychiatric Institute, Unit 69, 1051 Riverside Drive, New York, NY 10032, USA.

https://doi.org/10.1016/j.neuroimage.2023.120412
Received 15 March 2023; Received in revised form 10 September 2023; Accepted 16 October 2023
Available online 18 October 2023

[g] Yale University School of Medicine, New Haven, CT, USA
[h] University of Michigan, Ann Arbor, MI, USA
[i] Duke University, Durham, NC, USA
[j] Sanford School of Medicine, University of South Dakota, Vermillion, SD, USA
[k] Munroe-Meyer Institute, University of Nebraska Medical Center, Omaha, NE, USA
[l] University of California San Diego, La Jolla, CA, USA
[m] Center for Trauma Recovery, Department of Psychological Sciences, University of Missouri-St. Louis, St. Louis, MO, USA
[n] School of Psychology, University of New South Wales, Sydney, NSW, Australia
[o] Department of Radiology, The Affiliated Yixing Hospital of Jiangsu University, Yixing, Jiangsu, China
[p] Department of Psychiatry, University of Texas at Austin, Austin, TX, USA
[q] University of Toledo, Toledo, OH, USA
[r] University of Groningen, Groningen, The Netherlands
[s] Minneapolis VA Health Care System, Minneapolis, MN, USA
[t] University of Wisconsin-Madison, Madison, WI, USA
[u] University of Utah School of Medicine, Salt Lake City, UT, USA
[v] Departments of Psychology and Psychiatry, Neuroscience Program, Western University, London, ON, Canada; Department of Psychology, University of British Columbia, Okanagan, Kelowna, British Columbia, Canada
[w] Medical College of Wisconsin, Milwaukee, WI, USA
[x] UMR 1253, CIC 1415, University of Tours, CHRU de Tours, INSERM, France
[y] Stanford University, Stanford, CA, USA
[z] Emory University Department of Psychiatry and Behavioral Sciences, Atlanta, GA, USA
[aa] Civil Aerospace Medical Institute, US Federal Aviation Administration, Oklahoma City, OK, USA
[ab] Marquette University, Milwaukee, WI, USA
[ac] Brain Health Research Centre, Department of Anatomy, University of Otago, Dunedin, New Zealand
[ad] Department of Psychiatry, Amsterdam University Medical Centers, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands
[ae] Brain Research and Innovation Centre, Ministry of Defence, Utrecht, The Netherlands
[af] Cognitive and Clinical Neuroimaging Core, McLean Hospital, Belmont, MA, USA
[ag] Department of Radiology, Washington University School of Medicine, St. Louis, MO, USA
[ah] Division of Neuroradiology, Brigham and Women's Hospital, Boston, MA, USA
[ai] School of Medicine and Public Health, University of Wisconsin-Madison, Madison, WI, USA
[aj] Heidelberg University, Heidelberg, Germany
[ak] University of Münster, Münster, Germany
[al] University of Illinois at Chicago, Chicago, IL, USA
[am] Ghent University, Ghent, Belgium
[an] University of Cape Town, Cape Town, South Africa
[ao] Imaging Genetics Center, Mark and Mary Stevens Neuroimaging and Informatics Institute, Keck School of Medicine of the University of Southern California, Marina del Rey, CA, USA
[ap] Department of Psychology and Neuroscience, Baylor University, Waco, TX, USA
[aq] Wayne State University School of Medicine, Detroit, MI, USA
[ar] Division of Women's Mental Health, McLean Hospital, Belmont, MA, USA
[as] Department of Child and Adolescent Psychiatry, Psychosomatic and Psychotherapy, Ludwig Maximilian University of Munich, Munich, Germany
[at] Psychiatry Neuroimaging Laboratory, Brigham and Women's Hospital, Boston, MA, USA
[au] Donders Institute for Brain, Cognition and Behavior, Centre for Cognitive Neuroimaging, Radboud University Nijmegen, Nijmegen, The Netherlands
[av] Westmead Institute for Medical Research, Westmead, NSW, Australia
[aw] Department of Neuroscience, Western University, London, ON, Canada
[ax] University of Wisconsin-Milwaukee, Milwaukee, WI, USA
[ay] McLean Hospital, Belmont, MA, USA
[az] Harvard Medical School, Boston, MA, USA
[ba] Institute of Psychology, Chinese Academy of Sciences, Beijing, China
[bb] Psychiatry and Behavioral Science, Texas A&M University Health Science Center, College Station, TX, USA
[bc] Department of Medical Imaging, Jinling Hospital, Medical School of Nanjing University, Nanjing, Jiangsu, China
[bd] University of Iowa, Iowa City, IA, USA
[be] Charité Universitätsmedizin Berlin Campus Charite Mitte: Charite Universitatsmedizin Berlin, Berlin, Germany
[bf] VISN 17 Center of Excellence for Research on Returning War Veterans, Waco, TX, USA
[bg] Harvard University, Boston, MA, USA
[bh] Department of Psychiatry, Amsterdam University Medical Centers, VU University Medical Center, VU University, Amsterdam, The Netherlands
[bi] Department of Pediatrics, University of Minnesota, Minneapolis, MN, USA
[bj] Department of Psychology, Vanderbilt University, Nashville, TN, USA
[bk] University of Washington, Seattle, WA, USA
[bl] Department of Psychiatry and Behavioral Health, Ohio State University, Columbus, OH, USA
[bm] Neuroscience Research Australia, Randwick, NSW, Australia
[bn] Masaryk University, Brno, Czechia
[bo] Northwestern Neighborhood and Networks Initiative, Northwestern University Institute for Policy Research, Evanston, IL, USA
[bp] University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[bq] Center of Excellence for Stress and Mental Health, VA San Diego Healthcare System, San Diego, CA, USA
[br] University of South Dakota, Vermillion, SD, USA
[bs] University of Minnesota, Minneapolis, MN, USA
[bt] Leiden University Medical Center, Leiden, The Netherlands
[bu] University of California, Irvine, Irvine, CA, USA
[bv] Department of Psychology, University of Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Keywords:
Posttraumatic stress disorder
Multimodal MRI
Machine learning

ABSTRACT

Background: Recent advances in data-driven computational approaches have been helpful in devising tools to objectively diagnose psychiatric disorders. However, current machine learning studies limited to small homogeneous samples, different methodologies, and different imaging collection protocols, limit the ability to directly

Deep learning
Classification

compare and generalize their results. Here we aimed to classify individuals with PTSD versus controls and assess the generalizability using a large heterogeneous brain datasets from the ENIGMA-PGC PTSD Working group.
Methods: We analyzed brain MRI data from 3,477 structural-MRI; 2,495 resting state-fMRI; and 1,952 diffusion-MRI. First, we identified the brain features that best distinguish individuals with PTSD from controls using traditional machine learning methods. Second, we assessed the utility of the denoising variational autoencoder (DVAE) and evaluated its classification performance. Third, we assessed the generalizability and reproducibility of both models using leave-one-site-out cross-validation procedure for each modality.
Results: We found lower performance in classifying PTSD vs. controls with data from over 20 sites (60 % test AUC for s-MRI, 59 % for rs-fMRI and 56 % for d-MRI), as compared to other studies run on single-site data. The performance increased when classifying PTSD from HC without trauma history in each modality (75 % AUC). The classification performance remained intact when applying the DVAE framework, which reduced the number of features. Finally, we found that the DVAE framework achieved better generalization to unseen datasets compared with the traditional machine learning frameworks, albeit performance was slightly above chance.
Conclusion: These results have the potential to provide a baseline classification performance for PTSD when using large scale neuroimaging datasets. Our findings show that the control group used can heavily affect classification performance. The DVAE framework provided better generalizability for the multi-site data. This may be more significant in clinical practice since the neuroimaging-based diagnostic DVAE classification models are much less site-specific, rendering them more generalizable.

1. Introduction

Posttraumatic stress disorder (PTSD) is a prevalent and debilitating disorder, with a world-wide prevalence rate of 3.9 % (Kessler et al., 2017; Koenen et al., 2017). Current clinical assessments of PTSD rely solely on reported subjective experiences, overlooking objective biomarkers, which may lead to many cases of PTSD being undetected or misdiagnosed (Sumpter and McMillan, 2005). Recent advances in computational power and data-driven computational approaches, especially supervised machine learning, have been helpful in devising tools to objectively diagnose psychiatric disorders (Liu et al., 2015; van Loo et al., 2012; Bzdok and Meyer-Lindenberg, 2018). These approaches improve diagnosis by mining neuroimaging datasets, generating clinically relevant inferences at the individual level (Lama et al., 2017; Steardo et al., 2020; Gao et al., 2018). In recent years, the number of supervised machine learning studies in translational neuroimaging has grown dramatically (Woo et al., 2017), but many challenges still remain. First, most extant studies are single-site studies of small homogeneous samples. Although efforts have been made to deal with overfitting (Srivastava et al., 2014; Ying, 2019), single-site studies still tend to yield better performance than studies of larger samples, due to overfitting in the latter (Y Li et al., 2020; Lanka et al., 2020; Varoquaux, 2018). Second, methodological differences across these studies (e.g., machine learning approaches, scanners, acquisition parameters, and data processing pipelines) limit the ability to directly compare their results. Third, most studies estimated classification performance via cross-validation (i.e., all samples are used in building the prediction model), without testing classification performance using independent yet-to-be-seen test data. For example, a recent review in depression has shown that only 4 of 66 studies evaluated classification performance using a holdout dataset, with all four containing less than 200 samples (Gao et al., 2018). However, for machine learning models to be useful in real-world clinical settings, predictive models need large samples that enable the evaluation of model performance on an unseen holdout dataset or independent cohorts. In PTSD, only a handful of studies exist, with none exploring the reproducibility of findings using multimodal brain imaging across multiple sites.

In addition to the above-described challenges, the selection of reliable and sensitive biomarkers to classify patients relative to controls is also crucial. In PTSD, most studies conduct group-level univariate analysis to identify PTSD-related biomarkers using one, and rarely two imaging modalities (Ben-Zion et al., 2020). No published studies thus far have explored three common imaging modalities of structural Magnetic Resonance Imaging (s-MRI), resting state functional MRI (rs-fMRI), and diffusion MRI (d-MRI), each tapping specific facets of structure or function to provide comprehensive information about the brain. S-MRI provides information on regional tissue volume of gray or white matter. In PTSD, structural abnormalities have been reported in the hippocampus, amygdala (Morey et al., 2020), prefrontal cortex, anterior cingulate cortex (O'Doherty et al., 2017) and insula (Siehl et al., 2020). Rs-fMRI measures the functional connectivity (FC) between brain regions. FC abnormalities in PTSD have been reported mainly in the default mode network (DMN), ventral attention network (VAN), executive control network (ECN) and salience network (SN) (Koch et al., 2016; Daniels et al., 2010). Finally, d-MRI provides information on white matter microstructure and the brain's structural connectivity. White matter abnormalities in PTSD have been reported within the hippocampus, corpus callosum (Dennis et al., 2021), cingulate gyrus (CG), uncinate fasciculus (O'Doherty et al., 2018), and inferior fronto-occipital fasciculus (McCunn et al., 2021; Ju et al., 2020). However, as results from all three modalities are based on group-level analysis between PTSD and healthy controls (HC), or trauma exposed healthy controls (TEHC), it remains unclear whether PTSD can be discriminated at the single-subject level. Finally, most studies used only a single imaging modality among small samples (Liu et al., 2015; Im et al., 2017; Gong et al., 2014; Zilcha-Mano et al., 2020), limiting their broad-scale implications (Liu et al., 2015; Im et al., 2017; Li et al., 2014).

Recently, deep learning methods have received increasing attention in psychiatry because they are capable of learning subtle, latent patterns from high dimensional neuroimaging data. Deep learning methods have the potential to automatically diagnose different clinical disorders (Kim et al., 2016; Zhao et al., 2017), including PTSD (Sheynin et al., 2021), advancing the understanding of the neural basis of neuropsychiatric disorders (Arbabshirani et al., 2017). Of specific interest is autoencoder, which is a type of artificial neural network that seeks to learn the most efficient representations of the data at the individual level (Pinaya et al., 2019). Several neuroimaging studies show promising results for autoencoders in the classification of Alzheimer's disease (Suk et al., 2015; Ju et al., 2019), attention deficit hyperactivity disorder (ADHD) (Liu et al., 2021), autism spectrum disorder (ASD) (Eslami et al., 2019), and schizophrenia (Pinaya et al., 2019; G Li et al., 2020). Yet, the potential of autoencoders for multi-site classification of PTSD remains unknown.

To address the gaps in knowledge, here we used machine learning approaches in large-scale multimodal datasets from a heterogenous sample that obtained through the Enhancing Neuro-Imaging Genetics through Meta-Analysis (ENIGMA) PTSD and Psychiatric Genetics Consortium-(PGC) consortium PTSD working groups. First, we assessed classification performance between PTSD and controls using traditional machine learning methods; 2) assess the utility of the denoising variational autoencoder (DVAE) and evaluated its classification performance; and 3) assess the generalizability and reproducibility of both models for

each modality.

More specifically, first, we assessed the utility of neuroimaging biomarkers from s-MRI, rs-fMRI, and D-MRI in classifying PTSD from healthy controls, both with and without trauma exposure, as previous research has suggested unique neural signatures associated with trauma-exposure that are not present in trauma-unexposed individuals (Weng et al., 2019; Ke et al., 2018). To achieve this goal, we first identified the brain features that best distinguish PTSD from all non-PTSD controls. Next, we assessed the common and distinct neural features of PTSD versus controls with (TEHC) and without (HC) trauma exposure. Such information may provide valuable insight into underlying neural mechanisms in the pathophysiology of PTSD, and provide a baseline for machine learning classification of PTSD using large-scale data.

Second, we assessed the utility of deep learning models as a feature reduction method to improve classification performance. Neuroimaging studies usually make the predictive modeling task challenging because of the high dimensional feature set and relatively small sample size (Mwangi et al., 2014). Feature reduction methods can reduce feature dimensions to avoid overfitting, without losing important information needed for classification. Autoencoder approaches have an advantage over traditional feature reduction in suppressing noise from the input signal, leaving only a high-value representation of the input. Such an approach can automatically identify ways to transform raw imaging features into latent space variables, which are more suitable for machine learning algorithms, as well as capture the nonlinear representations of the input data. In this study, we built a DVAE for high dimensionality data reduction (Han et al., 2019). The latent variables were used as new features and input into traditional machine learning approaches for classification. Instead of developing a system capable only of classifying individuals into patients and controls, we sought to capture the key feature information in the latent space using the DVAE model. We first trained the model using controls, and subsequently applied the model to data from PTSD patients. Our intent was that the model would first learn the features representing healthy brain function and then retrieve the latent variables in PTSD patient data for capturing deviation of brain features from controls (Pinaya et al., 2019).

Third, we assessed the generalizability and reproducibility of the classification model across heterogeneous datasets from multiple sites. The generalizability of machine learning to classify neuroimaging data is of great concern. Tremendous variability across studies inhibits the creation of a clear body of reliable knowledge from distinct studies (Cai et al., 2020). The ENIGMA-PGC consortium combines multimodal imaging and clinical data from multiple sites, enabling the development of models based on large samples. This offers an unprecedented opportunity for testing the generalizability and reproducibility of classification models to unseen datasets with vastly different characteristics compared to the sample used for model building. We evaluated generalizability across sites by assessing the classification performance for each site, and then by using Leave-One-Site-Out Cross-Validation (LOSOCV) to test how well the model generalized to independent cohorts.

2. Methods

2.1. Participants

Table 1 summarizes the descriptive information for each imaging modality. We analyzed brain MRI data from 7925 individuals (3477 structural-MRI; 2495 resting state-fMRI; and 1953 diffusion-MRI). Of these 7925 individuals, 498 individuals had all 3 modalities, 736 had 2 and 6691 only had 1 modality. Demographic information for each imaging modality are summarized in *Supplemental Table 1~3*. Inclusion and exclusion criteria for each cohort are summarized in *Supplemental Table 4*.

Depending on the cohort, current PTSD was diagnosed according to the Diagnostic and Statistical Manual of Mental Disorders (DSM) IV or V criteria, using the following standard instruments: Clinician-Administered PTSD Scale-IV (CAPS-IV), CAPS-5 (DSM-V), Structured Clinical Interview (SCID-IV) (DSM-IV), Mini International Neuropsychiatric Interview (MINI) 6.0.0 (3 cohorts, DSM-IV), PTSD Checklist (PCL)−4 (DSM-IV), PCL-5 (DSM-V), Davidson Trauma Scale (DTS) IV (1 cohort, DSM-IV), PTSD Symptom Scale (PSS) (DSM-IV), and Anxiety Disorders Interview Schedule (ADIS) (DSM-IV). All participating sites obtained approval from their local institutional review boards and ethics committees, and all study participants provided written informed consent.

2.2. Image preprocessing

The brain features included in machine learning analysis are presented in Fig. 1. All imaging data were acquired at the contributing sites and processed with standardized protocols established by the ENIGMA Consortium (Nunes et al., 2020; Renteria et al., 2017). The specific set of imaging features used in this study are summarized in *supplemental Table 5~7*.

S-MRI: T1-weighted images were processed using the FreeSurfer processing stream to create individual subject thickness maps (http://surfer.nmr.mgh.harvard.edu/). The cortex of each hemisphere was parcellated into 34 cortical regions of interest (ROIs) using the Desikan–Killiany atlas (Klein and Tourville, 2012). To match what we excluded for rs-fMRI data, 10 ROIs that are part of the motor or occipital lobes were removed from further analysis. The volume of an ROI was calculated by multiplying cortical thickness at each vertex in the ROI by the surface area across all vertices (Wang et al., 2021). ROI volumes and intracranial volume (ICV) were derived from subjects' native spaces.

Table 1
Demographics of PTSD and control groups across s-MRI (T1), rs-fMRI (RS), and D-MRI (DTI).

		PTSD	Control	TEHC	HC	Difference between PTSD and Control
s-MRI	N (%)	1344 (38.7 %)	2133 (61.3 %)	1486 (42.7 %)	647 (18.6 %)	
	Female N (%)	544 (40.5 %)	907 (42.6 %)	564 (38.0 %)	343 (53.0 %)	$\chi2$=1.33, p = 0.24
	Age (years)	36.0 ± 14.1	34.0 ± 15.3	36.2 ± 15.1	28.9 ± 14.4	T = 14.16, p = 0.0001
	Age range (years)	6.0~82.0	6.0~85.0	6.0~85.0	6.0~69.0	
	N of sites	32	30	24	19	
rs-MRI	N (%)	1016 (40.7 %)	1479 (59.3 %)	1182 (47.4 %)	297 (11.9 %)	
	Female N (%)	546 (54.9 %)	761 (51.6 %)	571 (48.5 %)	190 (64.0 %)	$\chi2$=2.47, p = 0.12
	Age (years)	38.4 ± 14.0	37.0 ± 15.5	39.6 ± 15.2	26.4 ± 11.3	T = 5.48, p = 0.02
	Age range (years)	9.0~95.0	1.0~86.0	9.0~86.0	1.0~61.0	
	N of sites	26	26	23	14	
d-MRI	N (%)	830 (42.5 %)	1122 (57.5 %)	1027 (52.6 %)	95 (4.9 %)	
	Female N (%)	310 (37.3 %)	440 (39.2 %)	387 (37.7 %)	53 (55.8 %)	$\chi2$=0.63, p = 0.43
	Age (years)	38.6 ± 14.3	36.0 ± 15.4	36.6 ± 15.5	30.0 ± 12.3	T = 13.86, p = 0.0002
	Age range (years)	8.34~81.75	9.0~83.0	8.53~83.16	13.0~65.0	
	N of sites	20	19	16	9	

Abbreviations: PTSD: Posttraumatic stress disorder; TEHC: Trauma exposed healthy control; HC: healthy control.

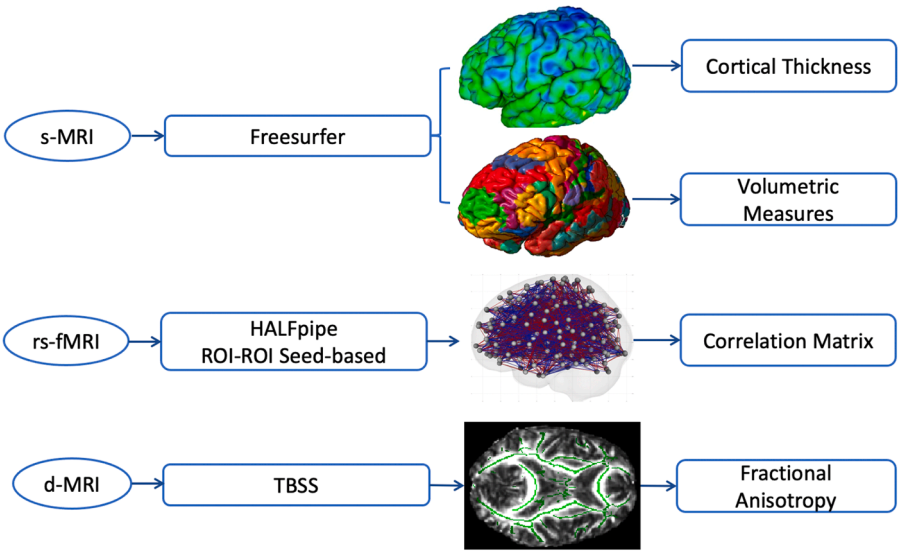

Fig. 1. Brain features from structural MRI (s-MRI), resting state fMRI (rs-fMRI), and DTI (d-MRI) used in this study. T1-weighted images were processed using the FreeSurfer pipeline, the final s-MRI features included 96 ROI cortical thicknesses (CT) and volumes for both left and right hemispheres. Rs-fMRI images were preprocessed using ENIGMA HALFpipe workflow, the final rs-fMRI features included 10,878 ROI-to-ROI functional connectivity measures. DTI data were preprocessed following ENIGMA-DTI protocols, 42 Tract-Based Spatial Statistics (TBSS) derived features of mean FA were included in the analysis.

Segmentations of gray and white matter and parcellations of ROIs were visually inspected using ENIGMA imaging quality control protocols (http://enigma.ini.usc.edu/protocols/). ROIs with segmentation or parcellation errors were excluded from the analysis. The final structural features included ROI cortical thicknesses (CT) and volumes for both left and right hemispheres, a total of 96 features (*Supplemental Table 5)*.

Rs-fMRI: Resting-state images were acquired at each site and preprocessed at a single location (Duke University). Preprocessing was implemented in ENIGMA HALFpipe workflow (https://github.com/HALFpipe/HALFpipe) based on fMRIPrep. Briefly, processing steps for T1w image include skull stripping, tissue segmentation, and spatial normalization to MNI space. Processing steps for functional images include motion correction using FSL MCFLIRT, slice time correction using AFNI 3dTshift for slice-timing correction, susceptibility distortion correction, and co-registration to the reference T1-weighted image using FSL FLIRT, and spatial normalization and warping to the template space using the MNI_2009 template. Each voxel was smoothed using signal from neighboring voxels with AFNI 3dBlurInMask followed by weighting by an isotropic Gaussian kernel. This method was repeated for each timepoint in the time series.

To ensure good quality of RS data, visual inspection was carried out on image registration, segmentation and brain extraction. To control confounding effects of motion artifact, several strategies were implemented: First, the top five aCompCor components were removed (Behzadi et al., 2007); second, frame-wise displacement (FD) was computed for each run, and subjects with more than 30 % frames have high levels of gross motion were excluded (FD$>$ 0.5 mm). Next, subjects with tSNR below 1.5 * IQR were excluded, and finally, subjects for whom more than 85 % of independent component analysis (ICA) components classified as noise were further removed. The ROI-to-ROI functional connectivity was calculated by extracting the average time series extracted from each of the 264 ROIs regions defined by the Power atlas (Jahanshad et al., 2013). A connectivity matrix between atlas regions was calculated using Pearson product moment correlation with PANDAS. We further reduced the number of features by only selecting 148 ROI regions that are part of known networks including default mode (DMN), ventral attention (VAN), frontal-parietal (FPN), salience (SN), subcortical (SCN), dorsal attention (DAN), cingulo-opercular networks (CON) (Gao et al., 2018). The final functional connectivity feature set contained 10,878 measures (*Supplemental Table 6*).

D-MRI: DTI data were preprocessed following ENIGMA-DTI protocols and quality control procedures at (Power et al., 2011). Processing steps include Eddy current correction, echo-planar imaging-induced distortion correction, motion correction, and tensor fitting. Fractional Anisotropy (FA) images generated from the estimated tensors were mapped to the ENIGMA DTI FA template and projected onto the skeleton FA template (FMRIB58_FA standard-space). FA values within ROIs were averaged within ROIs using JHU atlas for further analysis. 42 Tract-Based Spatial Statistics (TBSS) derived features of mean FA were extracted from D-MRI. Details and ROI abbreviations can be seen in *Supplemental Table 7*.

2.3. Data analysis

Overview: The overall analysis procedure is presented in Fig. 2. The analysis followed the structure of the three main aims and 2 supplementary aims of the paper. Goals 1 and 2 used data that was pooled across sites/scanners whereas goal 3 used site information to facilitate generalization performance assessment. Goal 1 investigated both Support vector machine (SVM) and random forest (RF) for classification of s-MRI data, which was repeated for rs-fMRI, and then for D-MRI. Goal 2 investigated DVAE in conjunction with SVM and in conjunction with RF (DVAE+SVM/RF), first using s-MRI features and then using rs-fMRI features. Goal 3 investigated performance of SVM or RF on single site data that was tested separately on s-MRI, rs-fMRI, and D-MRI brain features. Goal 3 also investigated performance of LOSOCV tested separately on s-MRI (SVM, DVAE+SVM), and rs-fMRI (SVM, DVAE+SVM) brain features.

For classification using SVM or RF (Goal 1) and DVAE+SVM or DVAE+RF (Goal 2) (Kingma, 2013), we used aggregated pooling methods for each modality. Pooling methods refer to techniques used to aggregate or combine data from multiple sites, so that the model can leverage information from diverse resources or sites to improve the overall model performance, and generalize the learned patterns across different sites. In this analysis, 70 % of all sites' data was used for cross-validation, and 30 % of all sites' data was used for independent testing. For the generalization test (Goal 3), we first tested the classification performance for each site across all three modalities, and then used LOSOCV procedure for each modality. In addition, we tested the impact of site, age and sex on classification performance.

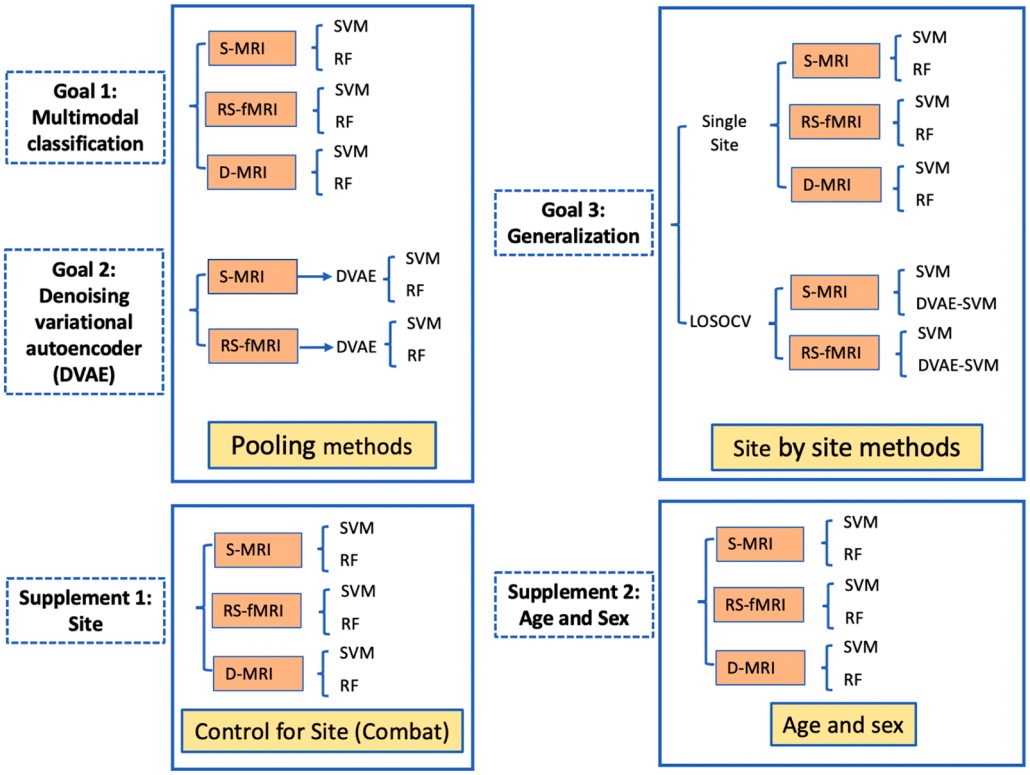

Fig. 2. Overall analysis procedure. The analysis followed the structure of the three main aims and 2 supplementary aims of the paper. Goals 1 and 2 used data that was pooled across sites/scanners whereas goal 3 used site information to facilitate performance assessment. Goal 1, investigated both Support vector machine (SVM) and random forest (RF) for classification of s-MRI data, which was repeated for rs-fMRI, and then for D-MRI. Goal 2 investigated DVAE in conjunction with SVM and in conjunction with RF (DVAE+SVM/RF), first using s-MRI features and then using rs-fMRI features. Goal 3 investigated performance of SVM or RF on single site data that was tested separately on s-MRI, rs-fMRI, and D-MRI brain features. Goal 3 also investigated performance of LOSOCV tested separately on s-MRI (SVM, DVAE+SVM), rs-fMRI (SVM, DVAE+SVM), and D-MRI (SVM) brain features. S-MRI: structural MRI; RS-fMRI: resting state fMRI; D-MRI: diffusion MRI; SVM: support vector machine; RF: random forest; DVAE: Denoising variational autoencoder.

3. Classification

We built three models for classifying PTSD relative to 1) all controls (HC and TEHC), 2) healthy controls with no trauma history (HC), and 3) those previously exposed to trauma who did not develop PTSD (TEHC) for each modality.

SVM and RF algorithms were used for classification (*Supplemental Material Methods*). Machine learning algorithms and Gridsearch were implemented in Python, and are available as part of the *scikit-learn* library (https://scikit-learn.org/stable/about.html#citing-scikit-learn). Our first task was to train classifiers that can differentiate patients with PTSD from control subjects using pooling methods. We randomly split all imaging data into two subsets: 70 % of the data was used for training and validation (cross-validation), and the remaining 30 % was used as a hold-out test dataset. The labeled training+validation data is used to train a machine learning model through cross-validation. The validation data, which is separate from the training data, is used for hyper-parameter tuning and assessing the model's performance during the cross-validation training process. The independent-test data is entirely separate from the training+validation data and is never involved in model training phase. Brain features with 30 % of missing data were dropped from further analysis (Madley-Dowd et al., 2019). RobustScaler from the *scikit-learn* library was used to scale the data for each modality, and missing values were imputed with the mean of the training dataset. The same scaling method was applied to the test set (Pedregosa et al., 2011). Gridsearch with stratified 10-fold cross-validation was used to select hyperparameters for both classifiers and to validate performance. Based on previous research (Gao et al., 2018), we used 10-fold cross validation, which generally provides better and more stable performance across different datasets, compared to Leave-One-Out Cross-Validation (LOOCV). To achieve an equal number of samples for each group, random under-sampling was applied to the imbalanced groups, with the under-sampling transform applied to the training dataset on each split of a repeated 10-fold cross-validation. The model's performance during training phase was evaluated by averaging across its the performance in the 10 fold. After model training, and the selection of the best hyperparameters, its performance is also assessed on the independent test set. Classification performance was measured using standard metrics including accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (ROC-AUC), which summarizes sensitivity and specificity at different thresholds.

Denoising variational autoencoder (DVAE): In our study, the feature dimension was very high for rs-fMRI data (148 ROIs, 10,878 ROI-to-ROI connectivity pathways), and relatively high for s-MRI data (96 regions). Researchers often use various feature reduction techniques for better performance and efficiency. Here, we implemented DVAE models using the PyTorch library (https://arxiv.org/abs/1912.01703). Gaussian noise with a mean of 0 and a standard deviation of 0.1 was applied to the input data. Our goal was to induce the model to learn to find more robust features of the data that are tolerant to noise and thus be able to reconstruct the noiseless data from noisy input data. This was the *denoising* aspect of the DVAE (Pinaya et al., 2019; Du et al., 2017).

Model Architecture: The autoencoder consists of an encoder and a decoder (*Supplemental Fig. 1*). The encoder has one input layer, x, one hidden layer, $h1$, and an encoding layer, z. The decoder consists of one hidden layer, $h2$, and one output layer $\hat{x}$. The size of h1, h2, and z was varied depending on the modality used. For s-MRI, a size of h1 = h2 = 250 and size of $z = 5$ were chosen. For rs-fMRI, a size of h1 = h2 = 400

and a size of $z = 10$ was chosen. The sizes of the respective layers were chosen by performing a sparse grid search for each of the layers' sizes independently and evaluating the performance of the model both with respect to the loss function and classification accuracy (Sheela, 2013). The grid search parameters for s-MRI and rs-fMRI included activation function (tanh, selu), latent size (Koenen et al., 2017; van Loo et al., 2012; Woo et al., 2017; Varoquaux, 2018; Koch et al., 2016; Kim et al., 2016), and hidden layer size ((Jahanshad et al., 2013), 100, 150, 200, 250, 400, 500).

The encoding layer z is referred to as the *latent space* of the model. This layer stores the model's reduced feature representation of the input data. In a general VAE framework, the features of the latent space z, referred to as *latent variables*, are drawn from Gaussian distributions determined by learned parameters (μ, $log(\sigma^2)$). These Gaussian distributions comprise an estimated distribution $q(z|x)$ to approximate the true underlying prior distribution $p(z)$. Once the encoded representation z is sampled, the values are reparameterized and fed into the decoder network. The decoder network then tries to reproduce the input using the reparameterized encoded data. The activation function for the layers was chosen as scaled exponential linear units (SELU) (Pinaya et al., 2019).

Loss Function of Model: For an autoencoder, the loss is usually determined solely by $L = MSE(x, \hat{x})$ where MSE is the mean squared error loss. This makes the autoencoder's sole objective to maximize its reconstruction accuracy. For a VAE the Kullback-Leibler Divergence (D_{KL}) is added to the loss function. The D_{KL} term is used to determine how much $q(z|x)$ and $p(z)$ differ. This constrains the way in which the parameters for the Gaussian distributions are updated and regularizes the latent space. Thus for a VAE, the loss function is generally

$$L = MSE(x, \hat{x}) + D_{KL}(q(x) \parallel p(z)).$$

For $p(z)$, usually the unit Gaussian or $N(0, 1)$ is chosen. This was the choice for our model as well.

Training of the DVAE model: The model was trained with ADAM optimizer (Kingma, 2014) using rs-fMRI or s-MRI data from controls only. Our intent was that the model would first learn the features representing salient aspects of healthy brain function and use the same features to represent PTSD. Prior to feeding the data to the model, the data was standardized by median and interquartile range. The total control sample was split into training (70 %) +validation (30 %) data. The labeled training data is used to train the VAE model. The validation data, which is separate from the training data, is used for hyperparameter tuning and assessing the model's performance in each epoch. The independent-test data is completely separate from the training+validation data and is never involved in model training. It is used to evaluate the generalization of the trained model to unseen PTSD data. Each epoch, the samples were fed as mini batches of size 128 to the network. L2 regularization (regularization parameter = 0.1) was applied to penalize high values of the network's weights and to avoid overfitting. The model was then trained for 500 epochs with the training+validation data. Once the training was completed, the model's performance was evaluated on the independent-test data, which provides an unbiased estimate of how the model generalizes to unseen data. The resulting VAE model learned to encode healthy patterns from the input brain features into its latent representation. Later, the brain features from patients with PTSD (PTSD test set) were input into the same VAE model, and the latent variables were extracted as new features for classification analysis (Fig. 3).

Convergence was measured by evaluating the *per-feature loss*, which we defined as L/n where n is the number of features of the input. This was done so as to be able to roughly compare the loss for models from different modalities, as each had a different number of features.

Encoding and Classification: After training the model, we used it to encode and reconstruct the data from both control subjects and the individuals with PTSD. For each subject, the values for their latent distributions, μ, $log(\sigma^2)$, were computed and extracted. Second, we compared the performance of the encoded features (DVAE+SVM/RF) to the original features using the SVM and RF classifiers (SVM/RF).

Calculating feature importance: To find features that are most

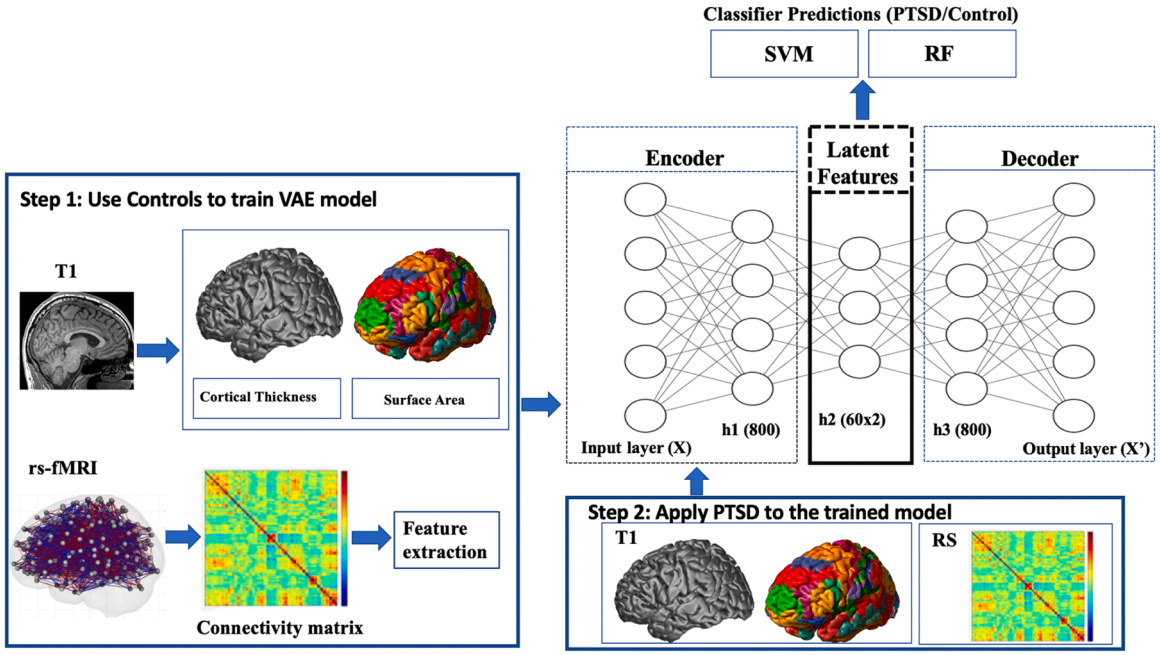

Fig. 3. Denoising Variational Autoencoder analysis pipeline: The model was trained using rs-fMRI or s-MRI data from controls only. The samples were then split into a training+validation (70 %) and independent-test (30 %) data. Then 20 % of the training data was set aside for validation and hyperparameter tuning. Once the training+validation was completed, the model's performance was evaluated on the independent-test data, which provides an unbiased estimate of how the model generalizes to unseen data. The resulting VAE model learned to encode healthy patterns from the input brain features into its latent representation. Later, the brain features from patients with PTSD (PTSD test set) were input into the same VAE model, and the latent variables were extracted as new features for classification analysis.

predictive of PTSD, we used a permutation-based feature-importance method on a RF classifier (Altmann et al., 2010; Breiman, 2001). After choosing the best RF model, we permuted the values of each feature and recomputed the accuracy. Predictor importance was then described by the difference between the baseline accuracy of the classifier and the difference in accuracy after permuting the feature. This method, while slower to compute, is more robust than the *Gini importance* (GI) method, which is a more commonly used method to calculate feature importance.

4. Generalization

Single Site: We evaluated the same SVM or RF parameterization used in previous analyses on each site's data, to shed light on its replicability. However, this method requires each site's sample size to be large enough to appropriately fit a machine-learning model. Thus, we only included sites that had more than 20 subjects in each group (PTSD and all control). For sites that have imbalanced samples, a down sampling approach was used to have a distributed sample across the two groups. To maximize generalizability and avoid overfitting, we applied the SVM or RF for each site using the default parameters, without grid search for optimal parameters, or feature reduction and selection. This method is stratified insofar as the proportion of cases and controls (in respective folds) is similar in both training and validation sets. The SVM or RF model was trained and evaluated using a 10-fold cross validation, and predictive performance was evaluated based on the cross validation.

Leave-one-site-out cross-validation (LOSOCV): Sites with sample size greater than 20 in each group were included in this analysis. In each fold of cross-validation, we used the DVAE trained latent features as described above. The DVAE model was trained based on control subjects' data only (exclude the controls from the independent test site), then the brain features from patients with PTSD across all sites were input into the same VAE model, and the latent variables were extracted as new features for machine learning analysis. Thus, the hold-out site was completely left out from the training procedure. For machine learning analysis (SVM),), in each iteration, one site was left out as test set, data from the rest of the sites was used in training procedure through 10 folds cross-validation. The training set was further randomly partitioned into 10 folds for cross-validation. Model performance was evaluated on the data from the hold-out site. The goal of this procedure was to assess the generalizability of the classifier to a totally independent data set that was sampled from a different sample and scanner. The LOSOCV performance was compared with an aggregated pooling method, in which data from all sites were included in training process.

ComBat: For a large multi-site study, it is important to consider whether a classifier can generalize well to new data coming from a different scanner or site. We used the ComBat method (Radua et al., 2020) to remove the site-specific information from the data and to test the generalizability of our classifier. The ComBat method models each imaging measure as a combination of three parts: variation of Y captured by the covariates such as age and sex, mean differences across sites, and the error term that contributes a different normal from each sites. Then the ComBat harmonized data can remove these additive and multiplicative effects due to site differences.

Biologically relevant covariates: To evaluate the contribution of confounding factors (age and sex) on the classification performance, we included age and sex as features and tested the impact on the overall performance.

5. Results

5.1. Classification performance between PTSD and controls for each imaging modality using traditional SVM and RF

The CV AUC and test AUC using RF and SVM are presented in Fig. 4 for brain features from s-MRI, rs-fMRI, and D-MRI modalities respectively. The performance for RF was similar to SVM. Accuracy, Sensitivity and Specificity are reported in the *Supplemental Table 8* and the receiver operating characteristic curve (ROC curve) are reported in the *Supplemental Figures 2–4*. First, our findings show that RF and SVM achieved similar performance when classifying PTSD from controls. Second, our models showed balanced CV AUC and test AUC, indicating that our models can generalize to an independent test set, which was not involved in model training, with no overfitting in these models. Third, all three modalities achieved comparable performance (using SVM: s-MRI: test AUC=0.60, Cohen's d = 0.354; rs-fMRI: test AUC=0.59, Cohen's d = 0.325; D-MRI: test AUC=0.56, Cohen's d = 0.212). Among the three contrasts (PTSD vs. HC; PTSD vs. TEHC; PTSD vs Controls), the performance of classifying PTSD from HC was the best across all three modalities (SVM: s-MRI: test AUC=0.72, Cohen's d = 0.82; rs-fMRI: test AUC=0.75, Cohen's d = 0.948; D-MRI: test AUC=0.78, Cohen's d = 1.09) (see *Supplemental Table 8*).

Some common and distinct features (*Supplemental Fig. 5*) that differentiate PTSD from both HC and TEHC are presented in *Supplemental Results*.

5.2. Classification performance between PTSD and controls using deep learning framework

Applying DVAE to rs-fMRI data reduced the number of features from 10,878 (latent variables) to 10. The performance of DVAE+SVM was CV

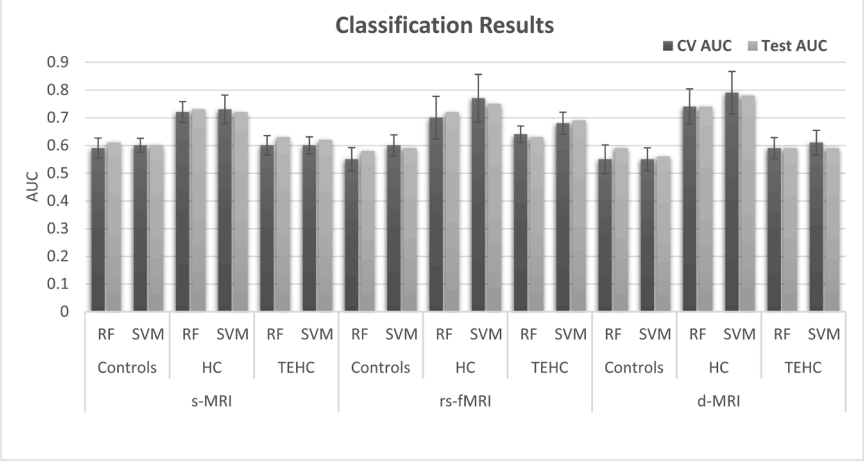

Fig. 4. The overall classification performance (measured by cross validation AUC [CV AUC], and test AUC) between PTSD and all controls, between PTSD and HC, and between PTSD and TEHC, for s-MRI, rs-fMRI, and D-MRI. Error bar represents standard deviation of the 10 fold cross validation results.

AUC mean=0.60, std=0.045; test AUC=0.62, Cohen's d = 0.424. Compared with the performance (CV AUC) using SVM of all features, the performance of DVAE+SVM or DVAE+RF (CV AUC) significantly improved (SVM: t (O'Doherty et al., 2017)=2.56, p = 0.016; RF: t (O'Doherty et al., 2017)=4.158, p = 0.0006). The classification performance between PTSD and controls using SVM was presented in Fig. 5, which achieved similar results. We also applied DVAE to s-MRI data (feature size: 96), which reduced the features to 5 latent variables. The performance of DVAE+SVM was CV AUC mean=0.60, std=0.045; test AUC=0.62, Cohen's d = 0.424. Compared with the performance (CV AUC) using SVM or RF of all features, the performance of DVAE+SVM or DVAE+RF (CV AUC) significantly improved (SVM: t (O'Doherty et al., 2017)=1.55, p = 0.019; RF: t (O'Doherty et al., 2017)=2.56, p = 0.0196).

The reconstruction loss function was used to assess whether the DVAE is a good predictor for classification of controls vs PTSD. The loss function during the training process for each modality is reported in Fig. 6.

5.3. Generalization and reproducibility

5.3.1. Assessing the classification performance for each site

S-MRI: The CV AUC in individual sites ranged from 0.36 to 0.83 using SVM. The average of individual site results yielded a CV AUC of 0.55 (std: 0.11) using SVM.**Rs-fMRI**: The CV AUC in individual sites ranged from 0.39 to 0.69 using SVM. The average of individual site results yielded a CV AUC of 0.54 (std: 0.08). **D-MRI**: The CV AUC of individual sites ranged from 0.24 to 0.68 using SVM. The average of individual site results yielded a CV AUC of 0.53 (std: 0.11) (Fig. 7, and *Supplemental Table 9)*. We also assessed the CV AUC using RF. There is no statistical differences between the CV AUC results using RF or SVM (s-MRI p = 0.97; rs-fMRI p = 0.32; D-MRI p = 0.65).

We further assessed the correlation between the sample size at each individual site and the CV AUC. No significant correlation was found, for all three modalities.

5.3.2. Leave one site out cross validation (LOSOCV)

The LOSOCV performance was compared with an aggregated pooling method (as in Results Section 1). For all three MRI modalities (s-MRI, rs-fMRI, and D-MRI), LOSOCV provided chance level classification (s-MRI: test AUC=0.56, Cohen's d = 0.212; rs-fMRI: test AUC=0.47, Cohen's d = 0; D-MRI: test AUC=0.49, Cohen's d = 0) *(Supplemental Table 10)*, and performed worse than the aggregate pooling method (Fig. 8). In s-MRI and rs-fMRI, we also compared LOSOCV and pooling methods using DVAE features, and assessed their generalizability. The DVAE achieved the same performance between LOSOCV and the pooling method, indicating a good generalization to unseen dataset using DVAE. Specifically, the LOSOCV method using SVM yielded an averaged test AUC of 0.61 (std:0.064) for s-MRI; and an averaged test AUC of 0.62 (std: 0.052) for rs-fMRI (*Supplemental Fig. 6*).

5.3.3. Effects of site, age, and sex

We evaluated the impact of site by first harmonizing each site using ComBat (Pomponio et al., 2020), and then assessed the classification performance between PTSD and all controls using RF and SVM. The site harmonization did not impact the classification performance using RF, but the performance dropped using SVM (s-MRI: before: test AUC=0.60, Cohen's d = 0.354; after: test AUC=0.52, Cohen's d = 0.071; rs-fMRI: before: test AUC=0.59, Cohen's d = 0.325; after: test AUC=0.46, Cohen's d = 0; D-MRI: before: test AUC=0.56, Cohen's d = 0.212; after: test AUC=0.52, Cohen's d = 0.071) (*Supplemental Fig. 7*).

We also evaluated the impact of age and sex on classification performance by including age and sex as features in the classification models. Age and sex did not impact the classification performance using either RF or SVM (*Supplemental Fig. 8*).

6. Discussion

The primary focus in the present study was to use machine learning techniques to create classifiers that leverage the complex multivariate patterns of structural and functional brain deficits. Specifically, we rigorously tested the classification performance on both cross-validation AUC and test AUC, in which a fully independent portion of the data was left out when selecting the model (both architectures and parameters). We found relatively poor classification performance in classifying PTSD vs. controls (60 % test AUC for s-MRI, 59 % for rs-fMRI and 56 % for D-MRI using SVM). This is lower than top-performing studies conducted at a single site, with sample size ranging from N = 30 to 89. These studies achieved accuracy ranging between 55.56 % (Y Li et al., 2020) and 97.1 % (Lanka et al., 2020) for rs-fMRI, and between 73 % (Im et al., 2017) and 80 % (Li et al., 2014) for studies focusing on multimodal biomarkers. Our single-site performance is comparable to other single-site studies (Fig. 7). Yet, single-site studies show poor generalization to independent datasets (Pereira et al., 2009), suggesting that performance might be adversely affected by small sample sizes, high-dimensional features, and use of complex models with a large number of parameters. Good performance on training data, with poor performance on test data, suggests overfitting, as most machine-learning studies are evaluated only on the basis of cross validation. Therefore, while our accuracy is relatively low (Y Li et al., 2020), the strength of our methods and sample size support the importance of our findings. Conversely, the present results are comparable with machine-learning studies using large scale imaging datasets in other psychiatric disorders based on s-MRI data – 65 % accuracy when classifying MDD from HC (Gao et al.,

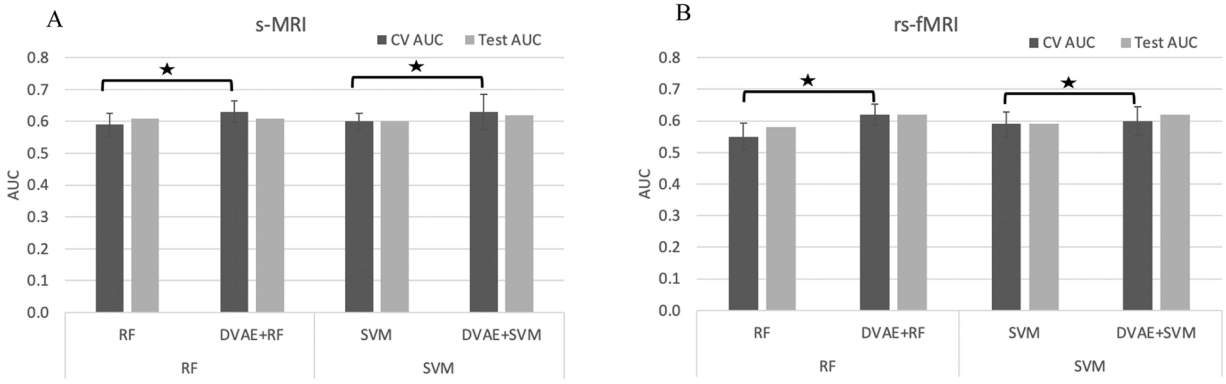

Fig. 5. Compare classification performance between PTSD and Controls using all features (labeled as RF or SVM) and DVAE-based latent variables (labeled as DVAE+RF or DVAE+SVM) in s-MRI (A) and rs-fMRI (B). Compared with the performance (CV AUC) using SVM of all features, the performance of DVAE+SVM (CV AUC) significantly improved for both s-MRI and rs-fMRI.

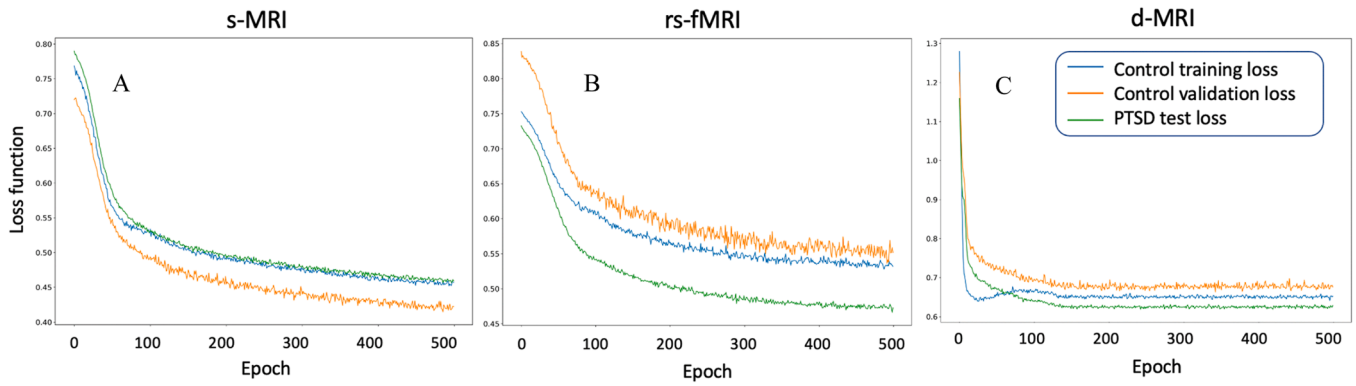

Fig. 6. The reconstruction loss function of the Denoising Variational Autoencoder model for s-MRI (A), rs-fMRI (B), and D-MRI (C), blue line: loss for the training set (from control data), orange line: loss for the validation set (from control data), green line: loss for the validation data (from PTSD data).

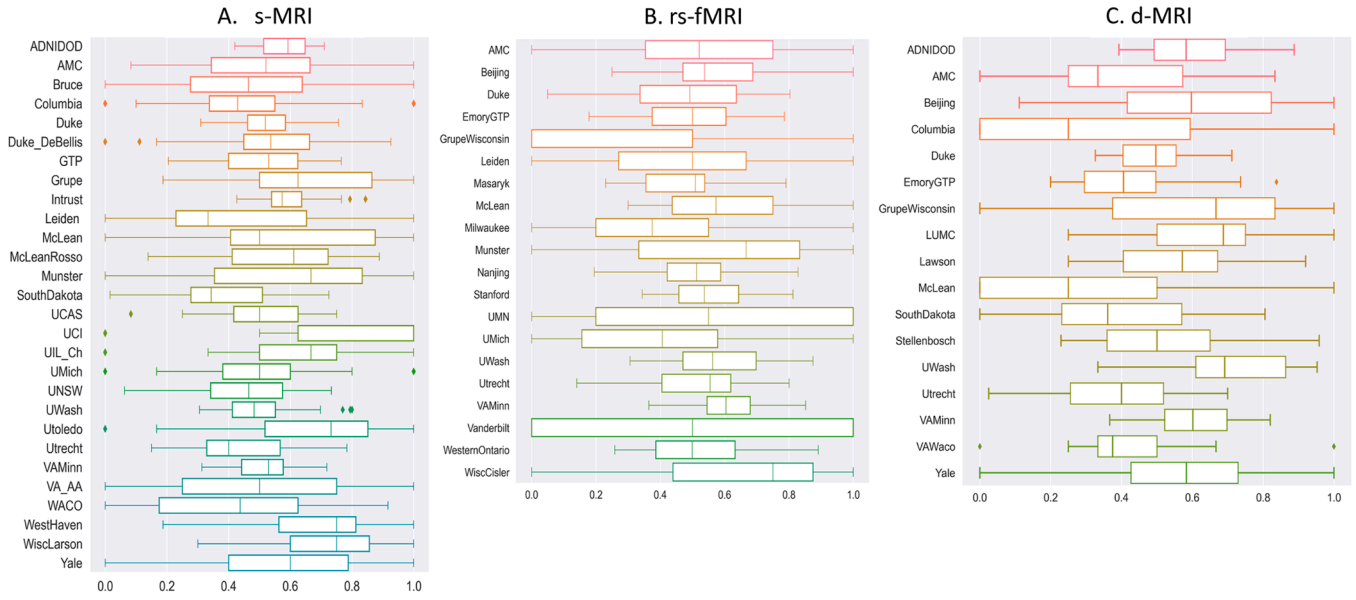

Fig. 7. s-MRI, rs-fMRI, and D-MRI single site performance for classification of PTSD from controls using SVM. The classification performance was measured by cross validation (CV) AUC, the dot indicates the average of the AUC of each fold in cross validation for each site, the line indicates the standard deviation of each fold in cross validation for each site. The boxplots were made by utilizing the boxplot() function from the seaborn library in Python. The box encompasses the interquartile interval, or the middle 50 % of the dataset. The upper and lower whiskers represent data points located in the top and bottom 25 % of the dataset. Data that fall outside this range are considered outliers and are plotted individually.

2018); 65.2 % accuracy in differentiating bipolar disorder from controls (Nunes et al., 2020); and a CV AUC of 0.57–0.61 when classifying OCD from controls (Bruin et al., 2020). Exploring the utility of a DVAE, improved classification results emerged as compared to traditional ML approaches. The DVAE successfully reduced feature dimensions, e.g. reduced the rs-fMRI features from 10,878 features to 10 latent variables, without losing information important for classification (SVM test AUC=59 %, Cohen's $d = 0.325$ using 10,878 features; test AUC=62 %, Cohen's $d = 0.424$ using 10 latent variables). Thus, the present results have the potential to provide a baseline classification performance for PTSD when using large scale imaging datasets.

When considering HCs and TEHCs as separate control groups, our results yielded a markedly improved discrimination standard (test AUC in the range of 72 % to 78 %) across the three modalities, with the discrimination between PTSD and HC outperforming that of PTSD and TEHC. These findings are in line with previous studies showing greater similarity in underlying neural circuits between PTSD and TEHC participants (Belleau et al., 2020; Sheynin et al., 2020), than when comparing PTSD to HC with no trauma exposure.

Evaluating the generalizability by assessing the model performance for each site and each modality, showed that the classification AUC at the individual sites across all three imaging modalities ranged from 40 % to 82 % using SVM. However, such a wide range in the performance across individual sites is expected in large-scale multi-site studies, also shown in other disorders (Nunes et al., 2020; Bruin et al., 2020), as samples are highly heterogeneous due to between-site differences (e.g., inclusion/exclusion criteria, demographic characteristics, clinical profiles, scanner used and scanning parameters, etc.). Furthermore, to avoid overfitting, we limited the scope of default parameters to SVM only, without hyperparameter tuning, which may have impacted the range of site performances compared to fine-tuning models using cross validation (Nunes et al., 2020).

Our results also indicated that LOSOCV performed using traditional machine learning and the aggerated pooling method performed worse than the performance using DVAE framework, as typically LOSOCV using traditional machine learning may result in large between-sample heterogeneity between training and test sets, resulting in roughly chance-level classification. Thus, imaging features do not provide strong biomarkers that enable generalization to new sites using traditional machine-learning methods. Previous studies have made an attempt at

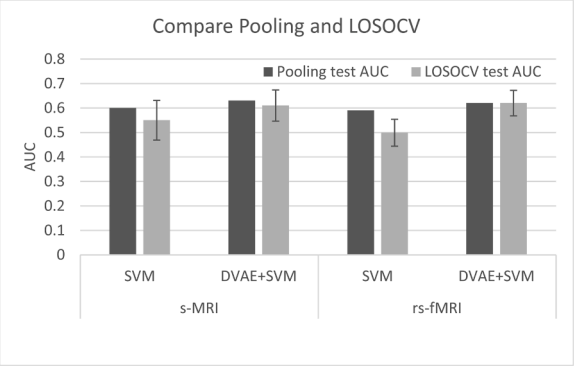

Fig. 8. The comparison of Leave One Site Out Cross Validation (LOSOCV) performance with aggregated pooling method on the independent-test data using SVM for classification between PTSD and Controls across s-MRI (T1), rs-fMRI (RS), LOSOCV: In each iteration, one independent test site was completely left out from the training partition. Then the training set was further randomly partitioned into 10 subfolds for cross validation. Predictive performance was evaluated on the data from the hold-out site (presented light gray in the figure). Aggregated pooling method: data from all sites was included in the training process. We randomly split all imaging data into two subsets: 70 % of the data was used for training+validation (cross validation), and the remaining 30 % was used as independent test data. Random under-sampling was applied to the imbalanced groups, with the under-sampling transform applied to the training dataset on each split of a repeated 10-fold cross validation. Predictive performance was evaluated on the data from the independent test set (presented dark gray in the figure).

LOSOCV, yielding average accuracies of around 75.0 % schizophrenia using s-MRI (Rozycki et al., 2018; Skatun et al., 2017), and an accuracy of 58.67 % when assessed LOSOCV in bipolar disorder (Nunes et al., 2020). These studies, however, used relatively small number of sites (3 to 5) for LOSOCV test, while we tested generalizability in 28 sites (s-MRI). Conversely, when extracting s-MRI and rs-fMRI features using DVAE models based on controls' data only, the LOSOCV method achieved the same performance as the pooling method, demonstrating better generalizability using the DVAE framework. Importantly, the LOSOCV may be more significant in clinical practice because when multi-site data is used for model training, the final neuroimaging-based diagnostic classification models are much less site-specific, rendering them more generalizable. Indeed, the VAE framework has been used for site harmonization and produced promising results. Site specific information can then be added to the latent representations to reconstruct the MRI data (Moyer et al., 2020; Dinsdale et al., 2021).

Assessing the effect of *site* on classification performance showed that discrimination remained the same when using a random forest classifier, and dropped when using the SVM classifier, and after site harmonization with ComBat. Previous literature suggests that statistical harmonization methods developed to reduce data heterogeneity have the potential to improve accuracy, but at the cost of generalizability. Such approaches may compromise the train/test separation and introduce additional assumptions. Our findings demonstrate that DVAE may be able to capture differences across sites, and can be better generalized to new sites data. More importantly, the DVAE model does not require *a priori* knowledge of site information. Taken together, our findings support reproducibility of the DVAE across heterogeneous datasets from multiple sites. Testing generalizability, we also assessed the effect of *age* and *sex* on performance by adding them as features to the model (Bruin et al., 2020), which did not affect classification performance. Neither did they emerge as informative features in classifying patients with PTSD from controls. Future studies should further assess the specific effects of *age* and *sex* on PTSD classification.

Several challenges still remain to be explored. First, combining biomarkers from different modalities with data fusion approaches is still in its infancy, and should be considered in future analyses to better detect potentially weak or latent effects hidden within high-dimensional datasets. Most deep-learning models are still being applied as black boxes, but serious efforts are underway to visualize latent variables and therefore improve the interpretability of results. Second, neuropsychiatric comorbidity was not consistently recorded across participating sites, so we could not evaluate it in the present study. Future studies should rectify this by also assessing comorbid conditions, exploring the underlying brain features that discriminate PTSD patients with and without comorbidity. Third, due to limited data of neurocognitive performance, we were not able to link emergent brain biomarkers to neurocognitive performance associated with the same brain circuits and/or regions. Fourth, while deep learning models typically give better performance than traditional machine learning, they are still perceived as black-box models, as they do not readily provide corresponding interpretations. However, deep learning need not be uninterpretable - as witnessed by the rapid expansion of methods for explainable deep learning (Bai et al., 2021; Singh et al., 2020), which uses new forms of visualization and representations of model outcomes. For example, VAE offers several advantages over the autoencoder and can be used for better interpretation of the latent representations. Specifically, VAE models the latent space as a probability distribution, thus it can generate new data by sampling from different parts of the latent distribution. This allows for meaningful interpolation and exploration of the latent space. Future studies should explore the latent representation discovered in this study. Fifth, the deep learning models were trained using data from all controls, not HC; future studies could generate separate models using HC and TEHC, and further explore the difference in latent variables generated by different control groups (HC and TEHC). Sixth, this study only used the DVAE model for ML classification. Future studies can assess different variations of autoencoder models such as VAE, sparse autoencoder, and adversarial autoencoder. Lastly, although our study benefited from a large sample size and advanced analytics, its value in predicting disease progression and treatment response needs to be investigated by future studies.

Taken together, our findings highlight the promise offered by machine learning and deep learning methods in diagnosing patients with PTSD using multimodal brain imaging data. Our findings show that the control group used can heavily affect classification performance. We also demonstrate the possibility of improving generalizability using DVAE models, which may provide valuable insight into the neural mechanisms underlying the pathophysiology of PTSD.

Author statement

Conceptualization: Xi Zhu, Rajendra A. Morey

Methodology, Formal analysis and software: Xi Zhu, Yoojean Kim, Orren Ravid, Xiaofu He

Writing - Original Draft: Xi Zhu, Yoojean Kim, Orren Ravid

Writing - Review & Editing: Benjamin Suarez-Jimenez, Sigal Zilcha-Mano, Amit Lazarov, Seonjoo Lee, Melanie Wall

Resources, Data Curation, Project administration, Funding acquisition, and Writing - Review & Editing:

Chadi G. Abdallah, Michael Angstadt, Christopher L. Averill, C. Lexi Baird, Lee A. Baugh, Jennifer U. Blackford, Jessica Bomyea, Steven E. Bruce, Richard A. Bryant, Zhihong Cao, Kyle Choi, Josh Cisler, Andrew S. Cotton, Judith K. Daniels, Nicholas D. Davenport, Richard J. Davidson, Michael D. DeBellis, Emily L. Dennis, Maria Densmore, Terri deRoon-Cassini, Seth G. Disner, Wissam El Hage, Amit Etkin, Negar Fani, Kelene A. Fercho, Jacklynn Fitzgerald, Gina L. Forster, Jessie L. Frijling, Elbert Geuze, Atilla Gonenc, Evan M. Gordon, Staci Gruber, Daniel W Grupe, Jeffrey P. Guenette, Courtney C. Haswell, Ryan J. Herringa, Julia Herzog, David Bernd Hofmann, Bobak Hosseini, Anna R. Hudson, Ashley A. Huggins, Jonathan C. Ipser, Neda Jahanshad, Meilin Jia-Richards, Tanja Jovanovic, Milissa L. Kaufman, Mitzy Kennis, Anthony King, Philipp Kinzel, Saskia B. J. Koch, Inga K. Koerte, Sheri M. Koopowitz, Mayuresh S. Korgaonkar, John H. Krystal, Ruth Lanius,

Christine L. Larson, Lauren A. M. Lebois, Gen Li, Israel Liberzon, Guang Ming Lu, Yifeng Luo, Vincent A. Magnotta, Antje Manthey, Adi Maron-Katz, Geoffery May, Katie McLaughlin, Sven C. Mueller, Laura Nawijn, Steven M. Nelson, Richard W.J. Neufeld, Jack B Nitschke, Erin M. O'Leary, Bunmi O. Olatunji, Miranda Olff, Matthew Peverill, K. Luan Phan, Rongfeng Qi, Yann Quidé, Ivan Rektor, Kerry Ressler, Pavel Riha, Marisa Ross, Isabelle M. Rosso, Lauren E. Salminen, Kelly Sambrook, Christian Schmahl, Martha E. Shenton, Margaret Sheridan, Chiahao Shih, Maurizio Sicorello, Anika Sierk, Alan N. Simmons, Raluca M. Simons, Jeffrey S. Simons, Scott R. Sponheim, Murray B. Stein, Dan J. Stein, Jennifer S. Stevens, Thomas Straube, Delin Sun, Jean Théberge, Paul M. Thompson, Sophia I. Thomopoulos, Nic J.A. van der Wee, Steven J.A. van der Werff, Theo G. M. van Erp, Sanne J. H. van Rooij, Mirjam van Zuiden, Tim Varkevisser, Dick J. Veltman, Robert R.J.M. Vermeiren, Henrik Walter, Li Wang, Xin Wang, Carissa Weis, Sherry Winternitz, Hong Xie, Ye Zhu, Yuval Neria, Rajendra A. Morey

Declaration of Competing Interest

Dr. Thompson received partial grant support from Biogen, Inc., and Amazon, Inc., for work unrelated to the current study; Dr. Lebois reports unpaid membership on the Scientific Committee for International Society for the Study of Trauma and Dissociation (ISSTD), grant support from the National Institute of Mental Health, K01 MH118467 and the Julia Kasparian Fund for Neuroscience Research, McLean Hospital. Dr. Lebois also reports spousal IP payments from Vanderbilt University for technology licensed to Acadia Pharmaceuticals unrelated to the present work. ISSTD and NIMH were not involved in the analysis or preparation of the manuscript; Dr. Etkin reports salary and equity from Alto Neuroscience, equity from Mindstrong Health and Akili Interactive. Other authors have no conflicts of interest to declare.

Data availability

All data examined in the manuscript are available upon request in deidentified format. Code is available on GitHub: https://github.com/ColumbiaNeriaLab/MultimodalPTSDClassification.

Acknowledgements

Dr. Zhu is supported by NIH K01MH122774 and by a NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation 27040; Dr. Dennis is supported by NIH R61NS120249; Dr. Jahanshad is supported by NIH R01MH117601; Dr. Thompson is supported by NIH U54 EB020403; Dr. Fani is supported by NIH AT011267 and MH111671; Dr. Bomyea is supppprted by NIH R61MH127005 and CX001600; Dr. Lebois is supported by NIH K01MH118467; Dr. Daniels is supported by German Research Foundation DA 1222/4–1; Dr. Disner is supported by VA RR&D Award IK2RX002922; Dr. Bruce is supported by NIH K23 MH090366; Dr. Bryant is supported by National Health and Medical Research Council #1073041; Dr. Ross is supported by the NIH T32MH018931, F31MH122047 and T32GM007507; Dr. Cisler is supported by NIMH MH119132 and MH097784; Dr. Morey is supported by NIMH MH111671 and MH129832.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2023.120412.

References

Altmann, A., Tolosi, L., Sander, O., et al., 2010. Permutation importance: a corrected feature importance measure. Bioinformatics 26, 1340–1347.
Arbabshirani, M.R., Plis, S., Sui, J., et al., 2017. Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. Neuroimage 145, 137–165.
Bai, X., Wang, X., Liu, X., et al., 2021. Explainable deep learning for efficient and robust pattern recognition: a survey of recent developments. Pattern Recognit. 120, 108102.
Behzadi, Y., Restom, K., Liau, J., et al., 2007. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. Neuroimage 37, 90–101.
Belleau, E.L., Ehret, L.E., Hanson, J.L., et al., 2020. Amygdala functional connectivity in the acute aftermath of trauma prospectively predicts severity of posttraumatic stress symptoms. Neurobiol. Stress 12, 100217.
Ben-Zion, Z., Zeevi, Y., Keynan, N.J., et al., 2020. Multi-domain potential biomarkers for post-traumatic stress disorder (PTSD) severity in recent trauma survivors. Transl. Psychiatry 10, 208.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Bruin, W.B., Taylor, L., Thomas, R.M., et al., 2020. Structural neuroimaging biomarkers for obsessive-compulsive disorder in the ENIGMA-OCD consortium: medication matters. Transl. Psychiatry 10, 342.
Bzdok, D., Meyer-Lindenberg, A., 2018. Machine learning for precision psychiatry: opportunities and challenges. Biol. Psychiatry Cogn. Neurosci. Neuroimaging 3, 223–230.
Cai, X.L., Xie, D.J., Madsen, K.H., et al., 2020. Generalizability of machine learning for classification of schizophrenia based on resting-state functional MRI data. Hum. Brain Mapp. 41, 172–184.
Daniels, J.K., McFarlane, A.C., Bluhm, R.L., et al., 2010. Switching between executive and default mode networks in posttraumatic stress disorder: alterations in functional connectivity. J. Psychiatry Neurosci. 35, 258–266.
Dennis, E.L., Disner, S.G., Fani, N., et al., 2021. Altered white matter microstructural organization in posttraumatic stress disorder across 3047 adults: results from the PGC-ENIGMA PTSD consortium. Mol. Psychiatry 26, 4315–4330.
Dinsdale, N.K., Jenkinson, M., Namburete, A.I.L., 2021. Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal. Neuroimage 228, 117689.
Du, B., Xiong, W., Wu, J., et al., 2017. Stacked convolutional denoising auto-encoders for feature representation. IEEE Trans. Cybern. 47, 1017–1027.
Eslami, T., Mirjalili, V., Fong, A., et al., 2019. ASD-DiagNet: a hybrid learning approach for detection of autism spectrum disorder using fMRI data. Front Neuroinform 13, 70.
Gao, S., Calhoun, V.D., Sui, J., 2018. Machine learning in major depression: from classification to treatment outcome prediction. CNS Neurosci. Ther. 24, 1037–1052.
Gong, Q., Li, L., Tognin, S., et al., 2014. Using structural neuroanatomy to identify trauma survivors with and without post-traumatic stress disorder at the individual level. Psychol. Med. 44, 195–203.
Han, K., Wen, H., Shi, J., et al., 2019. Variational autoencoder: an unsupervised model for encoding and decoding fMRI activity in visual cortex. Neuroimage 198, 125–136.
Im, J.J., Kim, B., Hwang, J., et al., 2017. Diagnostic potential of multimodal neuroimaging in posttraumatic stress disorder. PLoS One 12, e0177847.
Jahanshad, N., Kochunov, P.V., Sprooten, E., et al., 2013. Multi-site genetic analysis of diffusion images and voxelwise heritability analysis: a pilot project of the ENIGMA-DTI working group. Neuroimage 81, 455–469.
Ju, R., Hu, C., Zhou, P., et al., 2019. Early diagnosis of Alzheimer's disease based on resting-state brain networks and deep learning. IEEE/ACM Trans. Comput. Biol. Bioinform. 16, 244–257.
Ju, Y., Ou, W., Su, J., et al., 2020. White matter microstructural alterations in posttraumatic stress disorder: an ROI and whole-brain based meta-analysis. J. Affect. Disord. 266, 655–670.
Ke, J., Zhang, L., Qi, R., et al., 2018. Typhoon-related post-traumatic stress disorder and trauma might lead to functional integration abnormalities in intra- and inter-resting state networks: a resting-state fmri independent component analysis. Cell. Physiol. Biochem. 48, 99–110.
Kessler, R.C., Aguilar-Gaxiola, S., Alonso, J., et al., 2017. Trauma and PTSD in the WHO World Mental Health Surveys. Eur. J. Psychotraumatol. 8, 1353383.
Kim, J., Calhoun, V.D., Shim, E., et al., 2016. Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. Neuroimage 124, 127–146.
Kingma, D.P.W., 2013. M.: auto-Encoding Variational Bayes. arXiv 1312, 6114.
Kingma D.P.B., J.: adam: a Method for Stochastic Optimization. arXiv 2014; 1412.6980.
Klein, A., Tourville, J., 2012. 101 labeled brain images and a consistent human cortical labeling protocol. Front. Neurosci. 6, 171.
Koch, S.B., van Zuiden, M., Nawijn, L., et al., 2016. Aberrant resting-state brain activity in posttraumatic stress disorder: a meta-analysis and systematic review. Depress. Anxiety 33, 592–605.
Koenen, K.C., Ratanatharathorn, A., Ng, L., et al., 2017. Posttraumatic stress disorder in the World Mental Health Surveys. Psychol. Med. 47, 2260–2274.
Lama, R.K., Gwak, J., Park, J.S., et al., 2017. Diagnosis of Alzheimer's disease based on structural mri images using a regularized extreme learning machine and PCA features. J. Healthc. Eng. 2017, 5485080.
Lanka, P., Rangaprakash, D., Dretsch, M.N., et al., 2020. Supervised machine learning for diagnostic classification from large-scale neuroimaging datasets. Brain Imaging Behav. 14, 2378–2416.
Li, X., Zhu, D., Jiang, X., et al., 2014. Dynamic functional connectomics signatures for characterization and differentiation of PTSD patients. Hum. Brain Mapp. 35, 1761–1778.
Li, Y., Zhu, H., Ren, Z., et al., 2020a. Exploring memory function in earthquake trauma survivors with resting-state fMRI and machine learning. BMC Psychiatry 20, 43.
Li, G., Han, D., Wang, C., et al., 2020b. Application of deep canonically correlated sparse autoencoder for the classification of schizophrenia. Comput. Methods Programs Biomed. 183, 105073.

Liu, F., Xie, B., Wang, Y., et al., 2015. Characterization of post-traumatic stress disorder using resting-state fMRI with a multi-level parametric classification approach. Brain Topogr. 28, 221–237.
Liu, S., Zhao, L., Wang, X., et al., 2021. Deep spatio-temporal representation and ensemble classification for attention deficit/hyperactivity disorder. IEEE Trans. Neural Syst. Rehabil. Eng. 29, 1–10.
Madley-Dowd, P., Hughes, R., Tilling, K., et al., 2019. The proportion of missing data should not be used to guide decisions on multiple imputation. J. Clin. Epidemiol. 110, 63–73.
McCunn, P., Richardson, J.D., Jetly, R., et al., 2021. Diffusion tensor imaging reveals white matter differences in military personnel exposed to trauma with and without post-traumatic stress disorder. Psychiatry Res. 298, 113797.
Morey, R.A., Clarke, E.K., Haswell, C.C., et al., 2020. Amygdala nuclei volume and shape in military veterans with posttraumatic stress disorder. Biol. Psychiatry Cogn. Neurosci. Neuroimaging 5, 281–290.
Moyer, D., Ver Steeg, G., Tax, C.M.W., et al., 2020. Scanner invariant representations for diffusion MRI harmonization. Magn. Reson. Med. 84, 2174–2189.
Mwangi, B., Tian, T.S., Soares, J.C., 2014. A review of feature reduction techniques in neuroimaging. Neuroinformatics 12, 229–244.
Nunes, A., Schnack, H.G., Ching, C.R.K., et al., 2020. Using structural MRI to identify bipolar disorders - 13 site machine learning study in 3020 individuals from the ENIGMA Bipolar Disorders Working Group. Mol. Psychiatry 25, 2130–2143.
O'Doherty, D.C.M., Tickell, A., Ryder, W., et al., 2017. Frontal and subcortical grey matter reductions in PTSD. Psychiatry Res. Neuroimaging 266, 1–9.
O'Doherty, D.C.M., Ryder, W., Paquola, C., et al., 2018. White matter integrity alterations in post-traumatic stress disorder. Hum. Brain Mapp. 39, 1327–1338.
Pedregosa, F., Varoquaux, GG, A., Michel, V., Thirion, B., Grisel, O., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.
Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45, S199–S209.
Pinaya, W.H.L., Mechelli, A., Sato, J.R., 2019. Using deep autoencoders to identify abnormal brain structural patterns in neuropsychiatric disorders: a large-scale multi-sample study. Hum. Brain Mapp. 40, 944–954.
Pomponio, R., Erus, G., Habes, M., et al., 2020. Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. Neuroimage 208, 116450.
Power, J.D., Cohen, A.L., Nelson, S.M., et al., 2011. Functional network organization of the human brain. Neuron 72, 665–678.
Radua, J., Vieta, E., Shinohara, R., et al., 2020. Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA. Neuroimage 218, 116956.
Renteria, M.E., Schmaal, L., Hibar, D.P., et al., 2017. Subcortical brain structure and suicidal behaviour in major depressive disorder: a meta-analysis from the ENIGMA-MDD working group. Transl. Psychiatry 7, e1116.
Rozycki, M., Satterthwaite, T.D., Koutsouleris, N., et al., 2018. Multisite machine learning analysis provides a robust structural imaging signature of schizophrenia detectable across diverse patient populations and within individuals. Schizophr. Bull. 44, 1035–1044.
Sheela, KGDSN, 2013. Review on methods to fix number of hidden neurons in neural networks. Mathematical Problems in Engineering.
Sheynin, J., Duval, E.R., Lokshina, Y., et al., 2020. Altered resting-state functional connectivity in adolescents is associated with PTSD symptoms and trauma exposure. Neuroimage Clin. 26, 102215.
Sheynin, S., Wolf, L., Ben-Zion, Z., et al., 2021. Deep learning model of fMRI connectivity predicts PTSD symptom trajectories in recent trauma survivors. Neuroimage 238, 118242.
Siehl, S., Wicking, M., Pohlack, S., et al., 2020. Structural white and gray matter differences in a large sample of patients with Posttraumatic Stress Disorder and a healthy and trauma-exposed control group: diffusion tensor imaging and region-based morphometry. Neuroimage Clin 28, 102424.
Singh, A., Sengupta, S., Lakshminarayanan, V., 2020. Explainable deep learning models in medical image analysis. J. Imaging 6, 52.
Skatun, K.C., Kaufmann, T., Doan, N.T., et al., 2017. Consistent functional connectivity alterations in schizophrenia spectrum disorder: a multisite study. Schizophr. Bull. 43, 914–924.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Machine Learn. Res. 15, 1929–1958.
Steardo Jr., L., Carbone, E.A., de Filippis, R., et al., 2020. Application of support vector machine on fmri data as biomarkers in schizophrenia diagnosis: a systematic review. Front. Psychiatry 11, 588.
Suk, H.I., Lee, S.W., Shen, D., et al., 2015. Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. Brain Struct. Funct. 220, 841–859.
Sumpter, R.E., McMillan, T.M., 2005. Misdiagnosis of post-traumatic stress disorder following severe traumatic brain injury. Br. J. Psychiatry 186, 423–426.
van Loo, H.M., de Jonge, P., Romeijn, J.W., et al., 2012. Data-driven subtypes of major depressive disorder: a systematic review. BMC Med. 10, 156.
Varoquaux, G., 2018. Cross-validation failure: small sample sizes lead to large error bars. Neuroimage 180, 68–77.
Wang, X., Xie, H., Chen, T., et al., 2021. Cortical volume abnormalities in posttraumatic stress disorder: an ENIGMA-psychiatric genomics consortium PTSD workgroup mega-analysis. Mol. Psychiatry 26, 4331–4343.
Weng, Y., Qi, R., Zhang, L., et al., 2019. Disturbed effective connectivity patterns in an intrinsic triple network model are associated with posttraumatic stress disorder. Neurol Sci 40, 339–349.
Woo, C.W., Chang, L.J., Lindquist, M.A., et al., 2017. Building better biomarkers: brain models in translational neuroimaging. Nat. Neurosci. 20, 365–377.
Ying, X., 2019. An Overview of Overfitting and Its Solutions. IOP Publishing.
Zhao, Y., Dong, Q., Chen, H., et al., 2017. Constructing fine-granularity functional brain network atlases via deep convolutional autoencoder. Med. Image Anal. 42, 200–211.
Zilcha-Mano, S., Zhu, X., Suarez-Jimenez, B., et al., 2020. Diagnostic and predictive neuroimaging biomarkers for posttraumatic stress disorder. Biol. Psychiatry Cogn. Neurosci. Neuroimag. 5, 688–696.