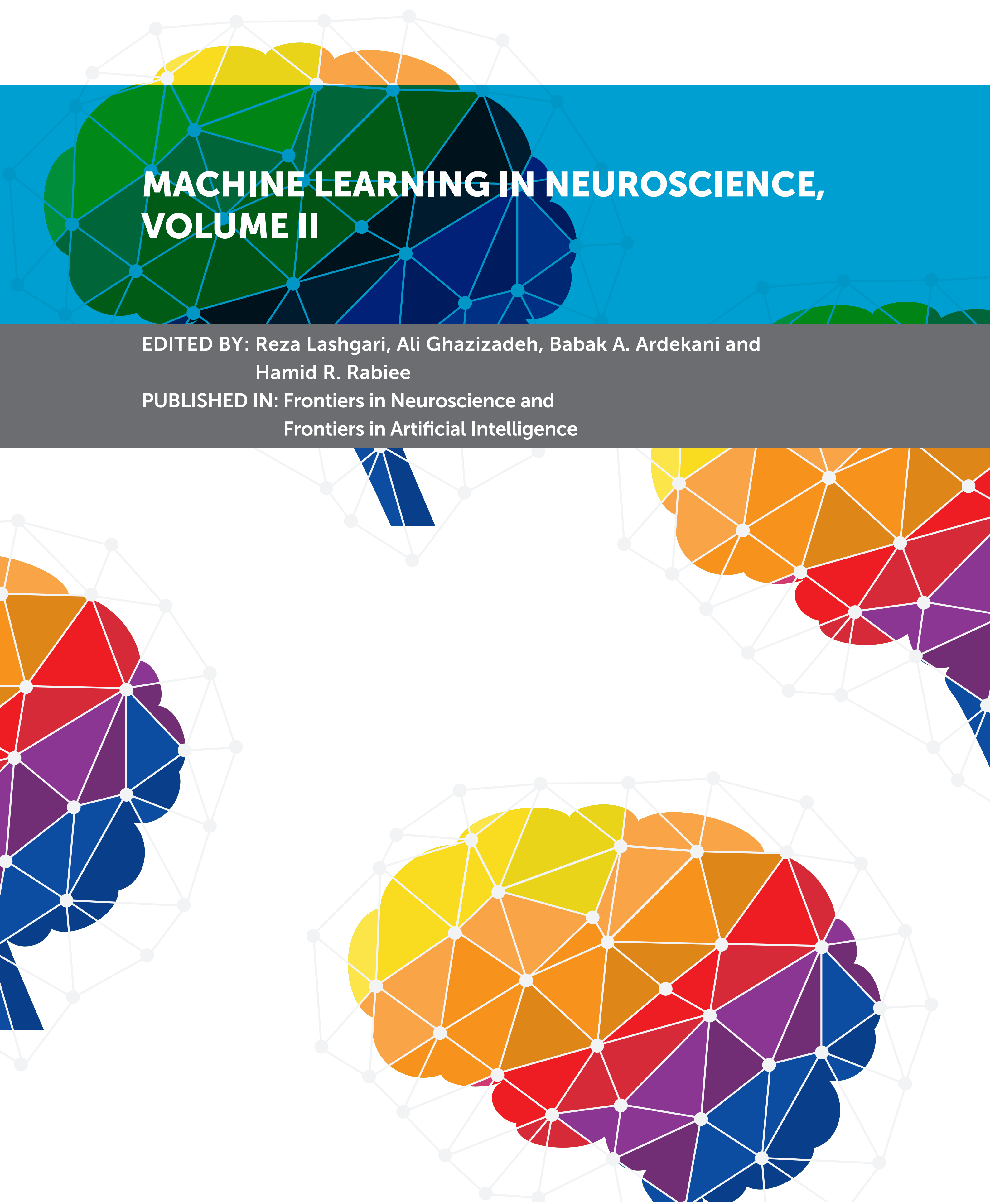


MACHINE LEARNING IN NEUROSCIENCE, VOLUME II

EDITED BY: Reza Lashgari, Ali Ghazizadeh, Babak A. Ardekani and
Hamid R. Rabiee

PUBLISHED IN: Frontiers in Neuroscience and
Frontiers in Artificial Intelligence

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-83250-549-6
DOI 10.3389/978-2-83250-549-6

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

MACHINE LEARNING IN NEUROSCIENCE, VOLUME II

Topic Editors:
Reza Lashgari, Shahid Beheshti University, Iran
Ali Ghazizadeh, Sharif University of Technology, Iran
Babak A. Ardekani, Nathan Kline Institute for Psychiatric Research, United States
Hamid R. Rabiee, Sharif University of Technology, Iran

Citation: Lashgari, R., Ghazizadeh, A., Ardekani, B. A., Rabiee, H. R., eds. (2022). Machine Learning in Neuroscience, Volume II. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83250-549-6

Table of Contents

frontiers in Neuroscience

OPINION
published: 15 June 2021
doi: 10.3389/fnins.2021.685005

Building the Precision Medicine for Mental Disorders via Radiomics/Machine Learning and Neuroimaging

Long-Biao Cui[1,2*†], *Xian Xu*[1†] and *Feng Cao*[3*]

[1] *Department of Radiology, The Second Medical Center, Chinese PLA General Hospital, Beijing, China,* [2] *Department of Clinical Psychology, School of Medical Psychology, Fourth Military Medical University, Xi'an, China,* [3] *The Second Medical Center, National Research Center for Geriatric Disease, Chinese PLA General Hospital, Beijing, China*

OPEN ACCESS

Edited by:
Reza Lashgari,
Institute for Research in Fundamental Sciences, Iran

Reviewed by:
Mingrui Xia,
Beijing Normal University, China
Amirmasoud Ahmadi,
Max Planck Institute of Ornithology, Germany

***Correspondence:**
Feng Cao
wind8828@gmail.com
Long-Biao Cui
lbcui@fmmu.edu.cn
orcid.org/0000-0002-0784-181X

† *These authors have contributed equally to this work*

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *24 March 2021*
Accepted: *10 May 2021*
Published: *15 June 2021*

Citation:
Cui L-B, Xu X and Cao F (2021) Building the Precision Medicine for Mental Disorders via Radiomics/Machine Learning and Neuroimaging. Front. Neurosci. 15:685005. doi: 10.3389/fnins.2021.685005

Keywords: radiomics, machine learning, mental disorders, magnetic resonance imaging, schizophrenia

INTRODUCTION

A pressing need of diagnostic, predictive, and prognostic markers exists in clinical settings. Unfortunately, in spite of evidence for underpinnings in mental disorders over the past decade, biologically based markers in reflection to the physiology for guiding clinical practice have obviously lagged behind. On the one hand, the updated Diagnostic and Statistical Manual of Mental Disorders remains to emphasize how mental disorders are expressed, although it provides a model that helps clinicians to perform better diagnosis and follow-up care (Kupfer et al., 2013), like cholesterol and blood pressure measurement. On the other hand, as the 23rd edition of the "Clinical Handbook of Psychotropic Drugs" has said, "we can provide current evidence-based and clinically relevant information to optimize patient care," but the information is derived from randomized controlled trials and leading clinical experts, etc. Therefore, clinical practice requests guidance of some objective, quantitative, and specific biomarker, reflecting its neurobiological substrates for diagnosis and treatment selection.

To this end, machine learning methods, as demonstrated by a sizable number of recent neuroimaging studies, hold great promise for improving the diagnosis, treatment, and prediction of prognosis in psychiatric domains, which will have an effect on personalized medicine. The term "machine learning" was coined in 1959 by Arthur Samuel (Samuel, 1959), and it is a science of the artificial intelligence, showing an evident capacity to reveal relationships between different variables used for classification (Tandon and Tandon, 2018). Furthermore, radiomics is a newly developed method to obtain high-dimensional features that might be options used for machine learning analysis. This Research Topic "Machine Learning in Neuroscience, Volume II" in "Frontiers in Neuroscience" provides new study strategy and applies radiomics/machine learning and distinct neuroimaging in mental disorders. Transforming existing clinical pathways toward optimizing care for the specific needs of each psychiatric patient, the significance is to achieve better diagnosis, treatment, and prognosis of mental disorders using radiomics/machine learning.

The field of psychiatry research remains a focus of medicine; in particular, mental health has arrived on the global health agenda. Currently, PubMed comprises more than 10 citations for literature involving radiomics and mental disorders, and efforts to develop radiomics/machine learning-based objective means have intensified for autism spectrum disorder (Chaddad et al., 2017), attention-deficit/hyperactivity disorder (Sun et al., 2018), schizophrenia spectrum and other psychotic disorders (Cui et al., 2018, 2021; Gong et al., 2020; Park et al., 2020; Xi et al., 2020), bipolar and related disorders (Wang et al., 2020), mild cognitive impairment, and Alzheimer's disease (Kai et al., 2018; Li et al., 2018; Ranjbar et al., 2019; Huang et al., 2020). Radiomics/machine learning enables the neuroimaging data of mental disorders to be extracted for improving clinical decision support (Wang et al., 2019).

Due to the excellent performance of radiomics analysis for feature selection and classification, it is regarded as the bridge between medical imaging and personalized medicine (Lambin et al., 2017). However, a critical issue is related to radiomics analysis in non-cancer field. In spite of a lack of lesions for conventional feature extraction, neuroimaging-based measures are features that could be extracted in radiomics/machine learning study. Taking schizophrenia as an example, previous studies on this topic illustrate the potential value in the application of radiomics/machine learning methods to disease definition and diagnosis and prediction of response to antipsychotics (APs) or electroconvulsive therapy (ECT).

RADIOMICS AND MAGNETIC RESONANCE IMAGING IN SCHIZOPHRENIA

Diagnosis and treatment of schizophrenia are pivotal clinical issues that need to be solved urgently. In a recent review, Kraguljac et al. (2021) highlighted and discussed the neuroimaging biomarkers in schizophrenia. Magnetic resonance imaging (MRI), as a non-invasive neuroimaging method, has been widely used in the study of schizophrenia. As we commented on a meta-analysis of the association of clinical and demographic characteristics and magnetic resonance spectroscopy in schizophrenia ("Targeting the Whole Clinical Course of Schizophrenia With Magnetic Resonance Imaging," https://jamanetwork.com/journals/jamapsychiatry/fullarticle/2778479), MRI combined with radiomics/machine learning could be the most important approach in schizophrenia research, involving predicting transition from clinical high risk to psychosis, providing evidence of macroscale neural mechanisms, delving into the nature behind symptoms, facilitating diagnosis and subtyping, predicting treatment response, detecting psychopharmacological effects, and guiding neuronavigation of neuromodulation, thereby managing very-late-onset schizophrenia-like psychosis. MRI-based studies are promising for clinical translation (Jiang et al., 2020).

It is of no doubt that we took the precision medicine view more seriously with the advent of radiomics/machine learning in the field of schizophrenia research (Wang et al., 2019). The number of publications on radiomics/machine learning via MRI has increased to five until this year, including one regression analysis (Gong et al., 2020) and four classification analyses (Cui et al., 2018, 2021; Park et al., 2020; Xi et al., 2020) (**Table 1**). There are two well-done MRI studies from Xi et al. (2020) and Gong et al. (2020), respectively. They used radiomics features on structural MRI (sMRI) to predict response to APs plus ECT. In the first study, the group of Yi-Bin Xi and colleagues extracted radiomics features from the regions of interest (ROIs) with differences of gray matter volume between responders and non-responders (Xi et al., 2020). Specifically, voxel-wise gray matter volume was compared between responders and non-responders, and then, 11 ROIs identified in the previous step entered first-order statistics feature extraction and classification analysis. A leave-one-out cross-validation (LOOCV) framework and support vector machine (SVM) was used to perform pattern classification analysis. This study built a fusion logistic regression model (LRM) with the least absolute shrinkage and selection operator (LASSO) with an accuracy of 93.18%, and the fixed features were from the right anterior cingulum, left supramarginal gyrus, and right hippocampus. In the second study, the group of Gong et al. (2020) examined whether the combination of gray and white matters can predict the outcome using sMRI and diffusion tensor imaging. They selected first-order statistics radiomics features from regions (gray and white matters) with strong electric field distribution under ECT in this regression analysis study. The prediction process was performed with a support vector regression model based on a LOOCV framework. Features in the left inferior frontal gyrus, right superior temporal gyrus, left temporal pole, right insula, and fibers connecting the frontal and temporal lobes were used in the final support vector regression model. The majority of ECT studies thus far has focused on identification of treatment response biomarkers in major depressive disorder. These are

TABLE 1 | Radiomics/machine learning and MRI in schizophrenia.

Articles	Subject number	MRI	Features selected	Accuracy	Sensitivity	Specificity
Identifying patients						
Park et al. (2020)	Training: Pat = 60/Con = 46 Testing: Pat = 26/Con = 20	sMRI	30 radiomics features from the bilateral hippocampal subfields	82.1%	76.9%	70%
Cui et al. (2018)	Training: Pat = 52/Con = 66 Testing: Pat = 56/Con = 55	fMRI	32 connections of the whole brain	87.09%	86.79%	87.22%
Predicting treatment response						
Cui et al. (2021)	Training: R = 47/*N* = 38 Testing: R = 41/*N* = 22	sMRI/fMRI	Nine functional connections and three cortical features	85.03%	92.04%	80.23%
Xi et al. (2020)	Training: R = 22/*N* = 22 Testing: R = 6/*N* = 7	sMRI	Three gray matter radiomics features	93.18%	95.45%	90.91%

Search terms: "schizophrenia and radiomics"
Con, controls; N, non-responder; Pat, patient; R, responder.

interesting studies that seek to predict treatment response in patients with schizophrenia undergoing electroconvulsive therapy. Based on the similar radiomics features, Park et al. (2020) used bilateral hippocampal subfields to differentiate patients with schizophrenia from healthy controls. ROIs were automatically segmented, and various combinations of classifiers (LRM, extra-trees, AdaBoost, XGBoost, or SVM) were trained, yielding an accuracy of 82.1%. These findings on the basis of structural differences with biological significance thus offer the potential to add new information to the literature.

In addition to conventional features, another two studies considered abnormal functional connectivity as features (Cui et al., 2018, 2021). In a disease identification study, a total of 137 connections determined by functional MRI (fMRI) were detected between patients with schizophrenia and healthy controls (Cui et al., 2018). Then, they reduced to 32 using the LASSO binary LRM. The accuracy of detecting patients was 87.09%. One of the strengths of this study is the two types of cross-validation (CV), i.e., intra- and inter-data set CV. In an early treatment response prediction study, both functional connectivity and cortical measures with group difference were used to obtain baseline features (Cui et al., 2021). They used CV-LASSO to conduct feature selection and dimension reduction and constructed an SVM model to predict response to treatment. The combined features obtained an outstanding accuracy with 85.03%. Likewise, biologically meaningful group differences of MRI reflect the pathophysiology of schizophrenia in relation to diagnosis and treatment. Although these two studies included non-conventional radiomics features, MRI analyses produce tens of thousands of functional connections and cortical measurements. In line with radiomics, high-throughput mining of quantitative features from medical imaging, we can call it the radiomics strategy in schizophrenia research.

Nevertheless, dozens of MRI studies using machine learning are emerging in schizophrenia. Many of them are characterized by large international multicenter samples (Chen et al., 2020), multimodal MRI fusions (Lei et al., 2020), elegant machine learning models (Rozycki et al., 2018), considerable accuracy with high generalizability (Koutsouleris et al., 2018), or enhanced understanding of brain circuits that can serve as potential biomarkers (Zhao et al., 2020). For this reason, MRI-based machine learning approaches may offer better individual-level diagnostic and predictive value in mental disorders (Keshavan et al., 2020).

DISCUSSION

MRI-based radiomics/machine learning studies hold several strengths with regard to schizophrenia, e.g., having biological underpinnings (structure/function), extension of treatment prediction (APs/ECT), and validation methods (intra-/inter-data set CV). However, many critical challenges exist in this field from both clinical and research perspectives. First, most current classification studies treat a clinical diagnosis as the gold standard; however, with MRI or psychopathology, some recent unsupervised (Jauhar et al., 2018; Matsubara et al., 2019) or supervised (Jacobs et al., 2021) machine learning studies have tended to explore transdiagnostic characteristics of mental disorders and try to break the boundary of classic diagnosis and establish bioinformation-based disorder classification. Second, many novel machine learning models, such as generative adversarial networks (GANs), have been applied to large multicenter MRI data in this field (Zhong et al., 2020; Ren et al., 2021). GANs contribute a lot to improving reproducibility of radiomics features across manufacturers and increasing diagnostic accuracy (Marcadent et al., 2020). The use of "radiomics" combining novel machine learning models is considered an initiative and an important development over prior work in the precision medicine of mental disorders.

Driven by the need for better management of patients, as well as advances in neuroimaging-based machine learning approach, a quest for accurate detection of convention to illness, identification of patients, and prediction of treatment response and outcome is noted. MRI-based radiomics/machine learning researchers should promote the generalizability of findings across patients and pave the way to facilitate the guidance of clinical decision making by means of these findings.

AUTHOR CONTRIBUTIONS

L-BC conceptualized and wrote the first draft of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the grant support of Fourth Military Medical University (2019CYJH), Project funded by China Postdoctoral Science Foundation (2019TQ0130), and Scientific Research Project 2020 of The Second Medical Center (ZXBJ2005) and Innovative Project of Chinese PLA General Hospital (CX19028).

ACKNOWLEDGMENTS

We thank Prof. Xiaopan Xu (Fourth Military Medical University) and Miss Jie Gong (Xidian University) for their help in revision.

REFERENCES

Chaddad, A., Desrosiers, C., Hassan, L., and Tanougast, C. (2017). Hippocampus and amygdala radiomic biomarkers for the study of autism spectrum disorder. *BMC Neurosci.* 18:52. doi: 10.1186/s12868-017-0373-0

Chen, J., Patil, K. R., Weis, S., Sim, K., Nickl-Jockschat, T., Zhou, J., et al. (2020). Neurobiological divergence of the positive and negative schizophrenia subtypes identified on a new factor structure of psychopathology using non-negative factorization: an international machine learning study. *Biol. Psychiatry* 87, 282–293. doi: 10.1016/j.biopsych.2019.08.031

Cui, L. B., Fu, Y. F., Liu, L., Wu, X. S., Xi, Y. B., Wang, H. N., et al. (2021). Baseline structural and functional magnetic resonance imaging predicts early treatment response in schizophrenia with radiomics strategy. *Eur. J. Neurosci.* 53, 1961–1975. doi: 10.1111/ejn.15046

Cui, L. B., Liu, L., Wang, H. N., Wang, L. X., Guo, F., Xi, Y. B., et al. (2018). Disease definition for schizophrenia by functional connectivity using radiomics strategy. *Schizophr. Bull.* 44, 1053–1059. doi: 10.1093/schbul/sby007

Gong, J., Cui, L. B., Xi, Y. B., Zhao, Y. S., Yang, X. J., Xu, Z. L., et al. (2020). Predicting response to electroconvulsive therapy combined with antipsychotics in schizophrenia using multi-parametric magnetic resonance imaging. *Schizophr. Res.* 216, 262–271. doi: 10.1016/j.schres.2019.11.046

Huang, K., Lin, Y., Yang, L., Wang, Y., Cai, S., Pang, L., et al. (2020). A multipredictor model to predict the conversion of mild cognitive impairment to Alzheimer's disease by using a predictive nomogram. *Neuropsychopharmacology* 45, 358–366. doi: 10.1038/s41386-019-0551-0

Jacobs, G. R., Voineskos, A. N., Hawco, C., Stefanik, L., Forde, N. J., Dickie, E. W., et al. (2021). Integration of brain and behavior measures for identification of data-driven groups cutting across children with ASD, ADHD, or OCD. *Neuropsychopharmacology* 46, 643–653. doi: 10.1038/s41386-020-00902-6

Jauhar, S., Krishnadas, R., Nour, M. M., Cunningham-Owens, D., Johnstone, E. C., and Lawrie, S. M. (2018). Is there a symptomatic distinction between the affective psychoses and schizophrenia? A machine learning approach. *Schizophr. Res.* 202, 241–247. doi: 10.1016/j.schres.2018.06.070

Jiang, J.-B., Cao, Y., An, N.-Y., Yang, Q., and Cui, L.-B. (2020). Magnetic resonance imaging-based connectomics in first-episode schizophrenia: from preclinical study to clinical translation. *Front. Psychiatry* 11:948. doi: 10.3389/fpsyt.2020.565056

Kai, C., Uchiyama, Y., Shiraishi, J., Fujita, H., and Doi, K. (2018). Computer-aided diagnosis with radiogenomics: analysis of the relationship between genotype and morphological changes of the brain magnetic resonance images. *Radiol. Phys. Technol.* 11, 265–273. doi: 10.1007/s12194-018-0462-5

Keshavan, M. S., Collin, G., Guimond, S., Kelly, S., Prasad, K. M., and Lizano, P. (2020). Neuroimaging in schizophrenia. *Neuroimaging Clin. N. Am.* 30, 73–83. doi: 10.1016/j.nic.2019.09.007

Koutsouleris, N., Wobrock, T., Guse, B., Langguth, B., Landgrebe, M., Eichhammer, P., et al. (2018). Predicting response to repetitive transcranial magnetic stimulation in patients with schizophrenia using structural magnetic resonance imaging: a multisite machine learning analysis. *Schizophr. Bull.* 44, 1021–1034. doi: 10.1093/schbul/sbx114

Kraguljac, N. V., McDonald, W. M., Widge, A. S., Rodriguez, C. I., Tohen, M., and Nemeroff, C. B. (2021). Neuroimaging biomarkers in schizophrenia. *Am. J. Psychiatry*. doi: 10.1176/appi.ajp.2020.20030340. [Epub ahead of print].

Kupfer, D. J., Kuhl, E. A., and Regier, D. A. (2013). DSM-5–the future arrived. *JAMA* 309, 1691–1692. doi: 10.1001/jama.2013.2298

Lambin, P., Leijenaar, R. T. H., Deist, T. M., Peerlings, J., de Jong, E. E. C., van Timmeren, J., et al. (2017). Radiomics: the bridge between medical imaging and personalized medicine. *Nat. Rev. Clin. Oncol.* 14, 749–762. doi: 10.1038/nrclinonc.2017.141

Lei, D., Pinaya, W. H. L., Young, J., van Amelsvoort, T., Marcelis, M., Donohoe, G., et al. (2020). Integrating machining learning and multimodal neuroimaging to detect schizophrenia at the level of the individual. *Hum. Brain Mapp.* 41, 1119–1135. doi: 10.1002/hbm.24863

Li, Y., Jiang, J., Shen, T., Wu, P., and Zuo, C. (2018). Radiomics features as predictors to distinguish fast and slow progression of Mild Cognitive Impairment to Alzheimer's disease. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* 2018, 127–130. doi: 10.1109/EMBC.2018.8512273

Marcadent, S., Hofmeister, J., Preti, M. G., Martin, S. P., Van De Ville, D., and Montet, X. (2020). Generative adversarial networks improve the reproducibility and discriminative power of radiomic features. *Radiol. Artif. Intell.* 2:e190035. doi: 10.1148/ryai.2020190035

Matsubara, T., Tashiro, T., and Uehara, K. (2019). Deep neural generative model of functional MRI images for psychiatric disorder diagnosis. *IEEE Trans. Biomed. Eng.* 66, 2768–2779. doi: 10.1109/TBME.2019.2895663

Park, Y. W., Choi, D., Lee, J., Ahn, S. S., Lee, S. K., Lee, S. H., et al. (2020). Differentiating patients with schizophrenia from healthy controls by hippocampal subfields using radiomics. *Schizophr. Res.* 223, 337–344. doi: 10.1016/j.schres.2020.09.009

Ranjbar, S., Velgos, S. N., Dueck, A. C., Geda, Y. E., and Mitchell, J. R. (2019). Brain MR radiomics to differentiate cognitive disorders. *J. Neuropsychiatry Clin. Neurosci.* 31, 210–219. doi: 10.1176/appi.neuropsych.17120366

Ren, M., Dey, N., Fishbaugh, J., and Gerig, G. (2021). Segmentation-renormalized deep feature modulation for unpaired image harmonization. *IEEE Trans. Med. Imaging* doi: 10.1109/TMI.2021.3059726. [Epub ahead of print].

Rozycki, M., Satterthwaite, T. D., Koutsouleris, N., Erus, G., Doshi, J., Wolf, D. H., 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. doi: 10.1093/schbul/sbx137

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM J. Res. Dev.* 3, 535–554. doi: 10.1147/rd.33.0210

Sun, H., Chen, Y., Huang, Q., Lui, S., Huang, X., Shi, Y., et al. (2018). Psychoradiologic utility of MR imaging for diagnosis of attention deficit hyperactivity disorder: a radiomics analysis. *Radiology* 287, 620–630. doi: 10.1148/radiol.2017170226

Tandon, N., and Tandon, R. (2018). Will machine learning enable us to finally cut the gordian knot of schizophrenia. *Schizophr. Bull.* 44, 939–941. doi: 10.1093/schbul/sby101

Wang, X. H., Yu, A., Zhu, X., Yin, H., and Cui, L. B. (2019). Cardiopulmonary comorbidity, radiomics and machine learning, and therapeutic regimens for a cerebral fMRI predictor study in psychotic disorders. *Neurosci. Bull.* 35, 955–957. doi: 10.1007/s12264-019-00409-1

Wang, Y., Sun, K., Liu, Z., Chen, G., Jia, Y., Zhong, S., et al. (2020). Classification of unmedicated bipolar disorder using whole-brain functional activity and connectivity: a radiomics analysis. *Cereb. Cortex* 30, 1117–1128. doi: 10.1093/cercor/bhz152

Xi, Y. B., Cui, L. B., Gong, J., Fu, Y. F., Wu, X. S., Guo, F., et al. (2020). Neuroanatomical features that predict response to electroconvulsive therapy combined with antipsychotics in schizophrenia: a magnetic resonance imaging study using radiomics strategy. *Front. Psychiatry* 11:456. doi: 10.3389/fpsyt.2020.00456

Zhao, W., Guo, S., Linli, Z., Yang, A. C., Lin, C. P., and Tsai, S. J. (2020). Functional, anatomical, and morphological networks highlight the role of basal ganglia-thalamus-cortex circuits in schizophrenia. *Schizophr. Bull.* 46, 422–431. doi: 10.1093/schbul/sbz062

Zhong, J., Wang, Y., Li, J., Xue, X., Liu, S., Wang, M., et al. (2020). Inter-site harmonization based on dual generative adversarial networks for diffusion tensor imaging: application to neonatal white matter development. *Biomed. Eng.* 19:4. doi: 10.1186/s12938-020-0748-9

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 05 July 2021
doi: 10.3389/fnins.2021.682777

Thalamus Radiomics-Based Disease Identification and Prediction of Early Treatment Response for Schizophrenia

Long-Biao Cui[1,2]†, Ya-Juan Zhang[3]†, Hong-Liang Lu[3]†, Lin Liu[4,5]†, Hai-Jun Zhang[6], Yu-Fei Fu[7], Xu-Sha Wu[7], Yong-Qiang Xu[7], Xiao-Sa Li[8], Yu-Ting Qiao[8], Wei Qin[4], Hong Yin[7] and Feng Cao[1]**

[1] The Second Medical Center, Chinese PLA General Hospital, Beijing, China, [2] Department of Clinical Psychology, Fourth Military Medical University, Xi'an, China, [3] Military Medical Psychology School, Fourth Military Medical University, Xi'an, China, [4] School of Life Sciences and Technology, Xidian University, Xi'an, China, [5] Peking University Sixth Hospital/Institute of Mental Health and Key Laboratory of Mental Health, Peking University, Beijing, China, [6] Department of Clinical Aerospace Medicine, School of Aerospace Medicine, Fourth Military Medical University, Xi'an, China, [7] Department of Radiology, Xijing Hospital, Fourth Military Medical University, Xi'an, China, [8] Department of Psychiatry, Xijing Hospital, Fourth Military Medical University, Xi'an, China

OPEN ACCESS

Edited by:
Ali Ghazizadeh,
Sharif University of Technology, Iran

Reviewed by:
Isaac Shiri,
Geneva University Hospitals (HUG), Switzerland
Seyed Amir Hossein Batouli,
Tehran University of Medical Sciences, Iran

***Correspondence:**
Long-Biao Cui
lbcui@fmmu.edu.cn
orcid.org/0000-0002-0784-181X
Feng Cao
wind8828@gmail.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *19 March 2021*
Accepted: *31 May 2021*
Published: *05 July 2021*

Citation:
Cui L-B, Zhang Y-J, Lu H-L, Liu L, Zhang H-J, Fu Y-F, Wu X-S, Xu Y-Q, Li X-S, Qiao Y-T, Qin W, Yin H and Cao F (2021) Thalamus Radiomics-Based Disease Identification and Prediction of Early Treatment Response for Schizophrenia. Front. Neurosci. 15:682777. doi: 10.3389/fnins.2021.682777

Background: Emerging evidence suggests structural and functional disruptions of the thalamus in schizophrenia, but whether thalamus abnormalities are able to be used for disease identification and prediction of early treatment response in schizophrenia remains to be determined. This study aims at developing and validating a method of disease identification and prediction of treatment response by multi-dimensional thalamic features derived from magnetic resonance imaging in schizophrenia patients using radiomics approaches.

Methods: A total of 390 subjects, including patients with schizophrenia and healthy controls, participated in this study, among which 109 out of 191 patients had clinical characteristics of early outcome (61 responders and 48 non-responders). Thalamus-based radiomics features were extracted and selected. The diagnostic and predictive capacity of multi-dimensional thalamic features was evaluated using radiomics approach.

Results: Using radiomics features, the classifier accurately discriminated patients from healthy controls, with an accuracy of 68%. The features were further confirmed in prediction and random forest of treatment response, with an accuracy of 75%.

Conclusion: Our study demonstrates a radiomics approach by multiple thalamic features to identify schizophrenia and predict early treatment response. Thalamus-based classification could be promising to apply in schizophrenia definition and treatment selection.

Keywords: schizophrenia, thalamus, radiomics, machine learning, diagnosis, treatment

INTRODUCTION

Driven by the need for precision medicine, a quest for accurate diagnosis and treatment was recently noted in the management of schizophrenia. A variety of abnormalities in the thalamus is associated with this disorder, including reduced volume (Pergola et al., 2015; Brugger and Howes, 2017; Dietsche et al., 2017; Dorph-Petersen and Lewis, 2017) and disrupted structural and functional connections to the cortices (Pergola et al., 2015; Giraldo-Chica and Woodward, 2017; Murray and Anticevic, 2017), as well as increased perfusion (Scheef et al., 2010; Zhu et al., 2015) and weaker correlation between glucose metabolism and dopaminergic state (Mitelman et al., 2019). Copious neuroimaging studies suggest thalamic association with schizophrenia, ranging from region to network level.

Task-state studies have found increased blood oxygenation level-dependent response to retrieval in the thalamus among schizophrenia patients (Stolz et al., 2012). Meanwhile, resting-state functional connectivity studies have reported thalamic abnormal connectivity with the bilateral cerebellum, anterior cingulate cortex, and multiple sensory-motor regions (Ferri et al., 2018). Effective connectivity by means of dynamic causal modeling revealed a deficit sensitivity of auditory cortex to its thalamic afferents in schizophrenia (Li et al., 2017). In addition, disrupted coactivation within resting-state networks analysis has been observed in the thalamus (Cui et al., 2017a). Both functional and structural imaging findings support dysconnectivity of the thalamus and cerebellum (Liu et al., 2011). As the neuroanatomical and neurochemical theories implicated in the pathophysiology of schizophrenia, the notion of emphasizing psychopathological processes mediated by the thalamus (Parnaudeau et al., 2018) should also be paralleled by identifying patients and predicting treatment response *via* multi-dimensional thalamic features.

A number of studies indicate that magnetic resonance imaging (MRI) techniques have provided insights into the classification and prediction in schizophrenia. MRI combined with machine learning technique represents a promising approach to distinguish patients with schizophrenia from healthy population, and responders from non-responders (de Filippis et al., 2019; Wang et al., 2019). In general, previous studies have related to the classification of schizophrenia using resting-state functional MRI (Anderson et al., 2010; Shen et al., 2010; Anderson and Cohen, 2013; Skatun et al., 2017; Cui et al., 2018; Huang et al., 2019; Zeng et al., 2018), structural MRI (Liang et al., 2018; Mikolas et al., 2018; Cui et al., 2019b; Liu et al., 2020), or their combination (Cui et al., 2021a). More importantly, classification approaches are able to aid subtyping symptoms of schizophrenia (Dwyer et al., 2018) and *trans*-diagnostic discrimination between schizophrenia and bipolar disorder (Arribas et al., 2010; Schnack et al., 2014; Rashid et al., 2016). In particular, MRI may be able to predict the response of treatments in schizophrenia, including structural (Fung et al., 2014; Hutcheson et al., 2014; Molina et al., 2014; Morch-Johnsen et al., 2015; Premkumar et al., 2015; Altamura et al., 2017; Dusi et al., 2017; Francis et al., 2018) and functional (Hadley et al., 2014; Kraguljac et al., 2016a,b; Sarpal et al., 2016; Doucet et al., 2018; Shafritz et al., 2018; Cui et al., 2019a) MRI (see Cui et al., for review; Cui et al., 2019a). These studies involved MRI features such as gray matter or white matter volume, cortical thickness, morphology of gyrus, and brain activation and connectivity with time of outcome assessment arranging from 6 weeks to 3 years. The structural MRI findings have shown a linkage between clinical improvements and higher gray matter volume [e.g., bilateral caudate (Hutcheson et al., 2014), bilateral lentiform and striatum (Fung et al., 2014), orbitofrontal cortex (Premkumar et al., 2015), and total brain (Altamura et al., 2017)], thinner right prefrontal (Molina et al., 2014) and thicker left caudal middle frontal cortical thickness (Francis et al., 2018), and rightward orbitofrontal cortex (Premkumar et al., 2015). In contrast, poor response has been linked to thinner left orbitofrontal cortex and left anterior cingulate cortex (Morch-Johnsen et al., 2015), and decreased right dorsolateral prefrontal cortex white matter volume (Dusi et al., 2017). Several functional MRI studies have reported greater activation in the anterior cingulate cortex in a simple response conflict task (Shafritz et al., 2018), and increased regional activity in the left postcentral gyrus/inferior parietal lobule (Cui et al., 2019a) and distinctive striatal functional connectivity (Sarpal et al., 2016) for responders. Hippocampal connectivity (Kraguljac et al., 2016b), connectivity within the dorsal attention network (Kraguljac et al., 2016a), and connectivity between ventral tegmental area/midbrain and the dorsal anterior cingulate cortex (Hadley et al., 2014) have been found to positively correlate to changes in symptoms. However, these potential predictors are inordinately heterogeneous and, to our knowledge, much earlier prediction of treatment response has not been identified.

Emerging evidence suggests structural and functional disruptions of the thalamus in schizophrenia, but whether thalamus abnormalities are able to be used for classification and prediction in schizophrenia remains to be determined. Thalamic features with successful level prediction of electroconvulsive therapy (ECT) response have been identified by radiomics (Xi et al., 2020). An opinion article in this journal illustrates the application of MRI and radiomics/machine learning methods to the study of schizophrenia (see Cui et al. for review; Cui et al., 2021b). Therefore, we aimed to validate a method of classification for schizophrenia and prediction of treatment response by multi-dimensional thalamic features derived from structural MRI using radiomics approaches. Relying on the thalamic association with schizophrenia, we hypothesized that thalamus-based classification and prediction could play a role in individualized diagnosis and treatment of schizophrenia as an objective and useful tool in this study.

MATERIALS AND METHODS

This study was approved by the Institutional Ethics Committee, First Affiliated Hospital (Xijing Hospital) of the Fourth Military Medical University. All participants (or their parents for those under age of 18 years) gave written informed consent after a full description of the aims and design of the study. **Table 1** provides further details on the two patient and control populations.

Participants

The inclusion and exclusion criteria are shown in previous studies (Cui et al., 2018, 2019a,b). The first dataset included 100 patients with schizophrenia patients and 92 healthy controls. The structural clinical interview for Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision (DSM-IV-TR) was used, and consensus diagnoses were made using all the available information. The second dataset included 91 patients and 107 healthy controls, and DSM, Fifth Edition (DSM-5) was used. Each patient was assessed by using the Positive and Negative Syndrome Scale (PANSS) at the time of imaging (Cui et al., 2018, 2019a,b).

Data were collected from May 2011 to December 2013 (dataset 1) and from April 2015 to December 2017 (dataset 2) in the Department of Psychiatry, Xijing Hospital, respectively, including inpatients undergoing their first or single hospitalization and outpatients seeking help. Inclusion criteria for patients are as follows: (1) they were assessed by two senior clinical psychiatrists, and consensus diagnosis of schizophrenia was made; (2) PANSS score was not less than 60 at the time of imaging; (3) all subjects were right handed, and their biological parents were of the Han Chinese ethnic group. Two groups of healthy controls without any reported psychotic syndrome (as assessed by psychiatrists) were recruited by advertisement from the local community.

Exclusion criteria for patients included the following: (1) presence of another psychiatric disorder; (2) history of repetitive transcranial magnetic or current stimulation, or a history of behavioral treatment; (3) history of clinically significant neurological, neurosurgical, or medical illnesses; (4) substance abuse within the prior 30 days or substance dependence within the prior 6 months; and (5) pregnancy or any other MRI contraindications, e.g., cardiac pacemakers and other metallic implants. Exclusion criteria for healthy controls included the following: (1) presence of any psychotic syndrome; (2) history of receiving antipsychotics, repetitive transcranial magnetic stimulation, transcranial current stimulation, or behavioral treatment; (3) history of clinically significant neurological, neurosurgical, or medical illnesses; (4) substance abuse within the prior 30 days or substance dependence within the prior 6 months; and (5) pregnancy or MRI contraindications, e.g., cardiac pacemakers and other metallic implants.

A total of 109 patients (67 from the first dataset; 42 from the second dataset) had clinical data of early treatment response. The majority of patients received second-generation antipsychotics, and the minority of patients received first-generation antipsychotics. Treatment response at discharging was assessed using percentage change of PANSS score: PANSS percentage change = (total score_1 − total score_0) × 100 ÷ (total score_0 − 30). Responders were defined as 30% reduction in PANSS total scores previously used (Cui et al., 2019a).

Image Acquisition

High-resolution structural imaging was acquired on a Siemens 3.0 T Magnetom Trio Tim MR scanner (the first dataset) or General Electric (GE) Discovery MR750 3.0 T scanner (the second dataset) using protocols described elsewhere (Xi et al., 2016; Cui et al., 2017a,b). All the imaging data were collected in the Department of Radiology, Xijing Hospital. A custom-built head coil cushion and earplugs were used to minimize head motion and dampen scanner noise. During data acquisition, subjects were asked to remain alert with eyes closed and keep

TABLE 1 | Clinical and demographical data.

Characteristics	Patients (*n* = 191)	Healthy controls (*n* = 199)	*P*-values	Responders (*n* = 61)	Non-responders (*n* = 48)	*P*-values
Age (years)	25 ± 7	29 ± 9	<0.001	24 ± 6	27 ± 8	0.036
Gender (M/F)	107/84	109/90	0.804	40/21	26/22	0.226
Education level (years)	12 ± 3	14 ± 4	<0.001	12 ± 2	13 ± 3	0.579
Duration of illness (months)	19 ± 26	–	–	17 ± 21	21 ± 31	0.368
PANSS score at baseline						
Total score	90 ± 17	–	–	90 ± 20	89 ± 14	0.774
Positive score	23 ± 6	–	–	23 ± 7	23 ± 7	0.847
Negative score	21 ± 8	–	–	21 ± 8	22 ± 8	0.548
General score	46 ± 9	–	–	46 ± 10	45 ± 7	0.329
PANSS score at discharging						
Total score	–	–	–	60 ± 15	80 ± 12	<0.001
Positive score	–	–	–	14 ± 5	20 ± 5	<0.001
Negative score	–	–	–	14 ± 6	20 ± 7	<0.001
General score	–	–	–	32 ± 8	40 ± 6	<0.001
Changes in PANSS score (%)	–	–	–	51 ± 16	16 ± 11	<0.001
Stay in hospital (days)	–	–	–	17 ± 5	15 ± 5	0.115
Antipsychotic dose (mg/day)[a]	–	–	–	10 ± 4	10 ± 4	0.388

Data are means ± standard deviations.
[a] *Defined Daily Dose (DDD).*

their head still. Participants in dataset 1 underwent scanning using a 3.0-T Siemens Magnetom Trio Tim scanner and an eight-channel phased array head coil (Siemens, Germany). Participants in dataset 2 underwent scans on a GE Discovery MR750 3.0-T scanner and an eight-channel phased array head coil (Milwaukee, WI, United States). Detailed parameters of high-resolution T1-weighted anatomical data are listed in **Table 2**. As performed previously (Cui et al., 2018), steps for the following analysis are shown in **Figure 1**.

Imaging Data Preprocessing and Extracting Thalamus

T1-weighted image processing was performed using the FreeSurfer image analysis suite (version 6.0.0)[1]. Data preprocessing was to register the original high-resolution structural image of each subject to standard template, and project it back to each subject to extract thalamus tissue. The preprocessing process is the standard process of the FreeSurfer toolkit.

Briefly, preprocessing was performed with the following steps: (i) skull stripping, (ii) normalization to a standard anatomical template (Talairach and Tournoux, 1988), (iii) correction for bias-field inhomogeneity, (iv) segmentation of subcortical white matter and deep gray matter volumetric structures (Fischl et al., 2002, 2004), (v) gray–white matter boundary tessellation and a series of deformation procedures that consist of surface inflation (Dale et al., 1999), and (vi) registration to a spherical atlas (Fischl et al., 1999) and parcellation of the cerebral cortex into units based on the gyral and sulcal structures (Fischl et al., 2004). In line with previous studies using the radiomics features from the bilateral structures in mental disorders (Chaddad et al., 2017; Park et al., 2020), we considered the bilateral thalami as regions of interest. In addition, the workflow of extracting thalamus was as follows: (i) the T1 images after preprocessing were matched to Anatomical Automatic Labeling (AAL) cortical and subcortical 1 mm × 1 mm × 1 mm atlas, and got the transformation matrix; (ii) use the inverse matrix of the transformation matrix to register AAL to individual space. After preprocessing, each subject's thalamus was registered to the standard space with consistent resolution. In this study, we did not perform interpolation in image processing[2].

TABLE 2 | Scanning parameters of T1-weighted imaging.

	The first dataset	The second dataset
Scanner	Siemens	GE
TR (ms)	2530	8.2
TE (ms)	3.5	3.2
Flip angle (°)	7	12
FOV (mm^2)	256 × 256	256 × 256
Matrix	256 × 256	256 × 256
Slice thickness (mm)	1	1
Section gap (mm)	0	0
Number of slices	192	196

Radiomics Features

The following analysis is based on the guidelines in radiomics (Lambin et al., 2017; Vallieres et al., 2018). Each image feature calculation formula is provided in the **Supplementary Material**, and they were based on the image biomarker standardization initiative[3]. Four types of radiomics features were used to quantify thalamic characteristics (Aerts et al., 2014): (i) first-order features, (ii) second-order features, (iii) texture features, and (iv) wavelet features, which have been used in previous studies (Gong et al., 2020; Xi et al., 2020). The first group quantified thalamus intensity characteristics using first-order statistics, calculated from the histogram of all thalamus voxel intensity values (14 radiomic features: energy, entropy, kurtosis, maximum, mean, mean absolute deviation, median, minimum, range, root mean square, skewness, standard deviation, uniformity, and variance). Group 2 consists of features based on the shape of the thalamus (eight radiomics features: compactness 1 and 2, maximum 3D diameter, spherical disproportion, sphericity, surface area, surface-to-volume ratio, and volume). Group 3 consists of textual features that are able to quantify intra-thalamus heterogeneity differences in the texture that is observable within the thalamus volume. These features are calculated in all three-dimensional directions within the thalamus volume, taking the spatial location of each voxel compared with the surrounding voxels into account. In this research, texture features describing patterns or the spatial distribution of voxel intensities were calculated from, respectively, gray level co-occurrence (GLCM) and gray level run-length (GLRLM) texture matrices. Texture matrices were determined considering 26 connected voxels. Group 4 wavelet transform effectively decouples textural information by decomposing the original image in low and high frequencies. Here, the first- and second-order features and textural features of eight directions (the original images were decomposed into eight directions) were calculated. All feature algorithms were implemented in Matlab 2016a (MathWorks, Natick, MA, United States). In the process of feature extraction, we performed the discretization and used 2 mm × 2 mm × 2 mm as voxels to extract the imaging features of the thalamus and take the mean value.

Feature Selection, Classification Model, and Efficacy Prediction

Ten-fold cross-validation (CV) was used to assess the reliability of the classification model (**Figure 1**). Briefly, 390 subjects (190 patients) were randomly separated into 10 groups. Each time, one group in turn was used as a test group and the other nine groups were used as training group.

A total of 4019 radiomics features were selected as initial features. After that, we used a 10-fold CV-based Least Absolute Shrinkage and Selection Operator (CV-LASSO) method to further select features. Briefly, subjects in the training group

[1] http://surfer.nmr.mgh.harvard.edu/

[2] https://ibsi.readthedocs.io/en/latest/

[3] https://arxiv.org/abs/1612.07003

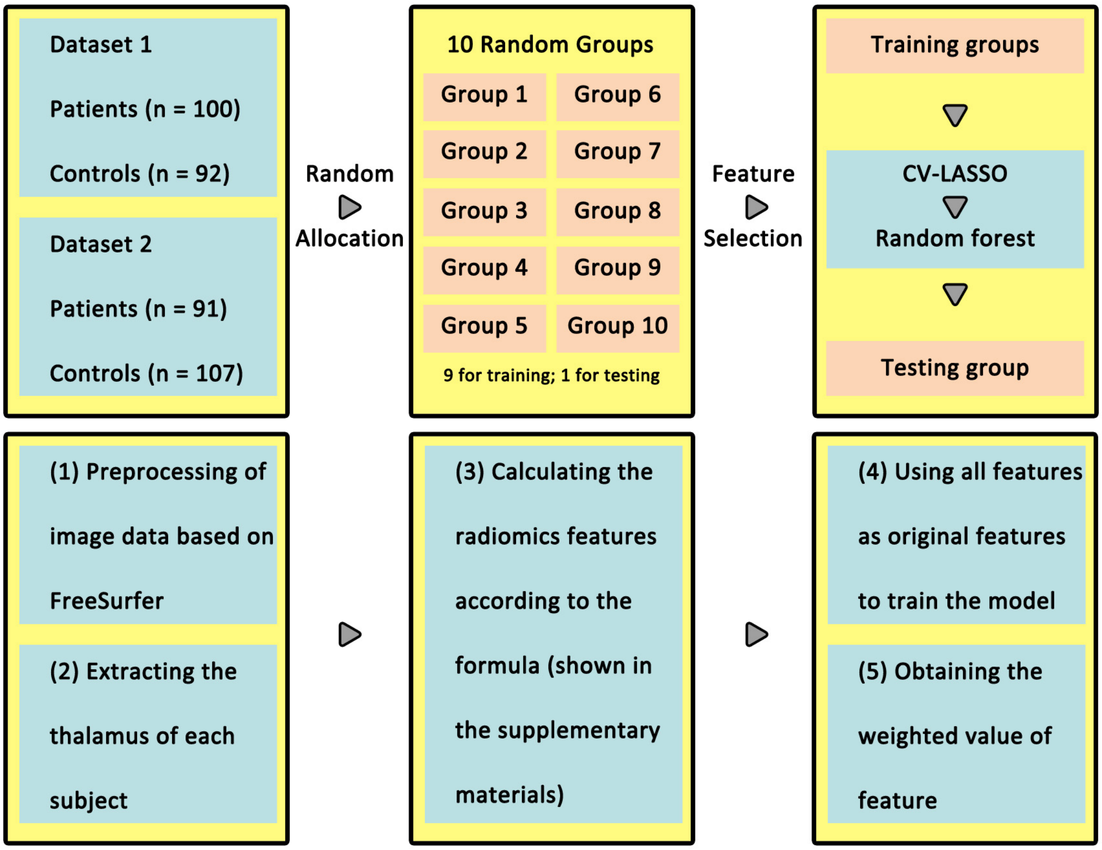

Dataset 1
Patients (n = 100)
Controls (n = 92)
Dataset 2
Patients (n = 91)
Controls (n = 107)
Random
Allocation
10 Random Groups
Group 1
Group 6
Group 2
Group 7
Group 3
Group 8
Group 4
Group 9
Group 5
Group 10
9 for training; 1 for testing
Feature
Selection
Training groups
CV-LASSO
Random forest
Testing group
(1) Preprocessing of image data based on FreeSurfer
(2) Extracting the thalamus of each subject
(3) Calculating the radiomics features according to the formula (shown in the supplementary materials)
(4) Using all features as original features to train the model
(5) Obtaining the weighted value of feature

FIGURE 1 | Workflow for analysis in classification of patients and healthy controls. In the **upper panel**, all of the participants were randomly divided into 10 groups, nine for training and one for testing. The **lower panel** summarizes radiomics steps. The radiomics features were extracted using CV-LASSO in the training group and validated in the testing group using random forest.

were again randomly separated into 10 groups. Each time, one group in turn was excluded from the dataset, and the LASSO (Sauerbrei et al., 2007) method with mean of square error (MSE) as the cost function was used on the remaining nine groups to narrow down the initial features into the most important features according to the MSE + 1SE criteria (Sauerbrei et al., 2007). This step was repeated 10 times, which resulted in 10 different groups of selected features. Finally, the edges that were included in the selected feature group at least N times (i.e., occurring N times) were selected as LASSO features for further analysis. Next, the random forest (RF) method was used to construct the classification model based on LASSO features in training group. The accuracy, sensitivity, specificity, and recall indices of the constructed model were calculated using testing group. Considering any confound factors due to data from two scanners, the differences of features selected between participants in the two datasets were compared.

All these steps above were repeated 10 times. As for the setting of P0, N, and the number of trees t in RF, we used grid-search method to find them. These parameters were set at a group of specific values when the accuracy index of the constructed classification model achieved the maximum. The P0 was set from 0.01 to 0.1 with a step of 0.01. The N was set from 1 to 10 with a step of 1. The t was set from 5 to 100 with a step of 5.

To avoid the random group effect, we repeated the 10-fold CV 100 times. For each time, a new random group was split. The mean ± standard deviation of each index across the 1000 testing groups (10 × 100) was used to assess the performance and stability of the constructed model. Finally, 1000 times permutation test (group label permutation) was performed to check if our results were significantly different from random labels.

Validation

Finally, we used another machine learning method, support vector machine, to estimate the status of each participant (schizophrenia or control; responder or non-responder) *via* intra- and inter-dataset CV (Cui et al., 2018).

RESULTS

Clinical Characteristics

Table 1 shows the full description of demographic and clinical characteristics of patients and healthy controls. No significant difference was found in gender between patients and healthy controls. For patients, there was statistical difference in being younger ($P < 0.001$) and having a lower education level ($P < 0.001$).

Feature Selection

The RF was performed for the high-resolution T1-weighted imaging. In this study, 4019 radiomics features were extracted (**Figure 2** and see the **Supplementary Material**), resulting in 12 features for identifying patients ("W1.Mid," "W1.SRE_8," "W2.LRHGLE_8," "W3.Min," "W4.Co_Corr_12," "W4.Co_Var_13," "W5.Co_Corr_11," "W5.RLN_9," "W6.Co_Corr_2," "W6.Co_Corr_7," "W7.IMC1_9," and "W9.Co_Corr_12") and four features for predicting treatment response ("W1.LRE_9," "W3.Min," "W6.Co_Corr_7," and "W6.Co_Var_7"). For the selected features, we performed comparison between subjects on two scanners, e.g., patients/healthy controls on Siemens scanner and GE scanner, and no significant difference was found between two scanners by *t*-tests.

Classification Performance

Figure 3 and **Table 3** show the classification performance. Using 12 features, the RF classifier accurately discriminated patients from healthy controls on the basis of the receiver operating characteristic (ROC) curve, with an accuracy of 68%. Four features were further confirmed in the prediction of treatment response, with an accuracy of 75%. The DeLong test suggested that the model of the area under curve (AUC) of the ROC analysis for response prediction was superior to that for diagnosis ($P = 0.015$).

Validation

Combining radiomics and support vector machine method, thalamic features had an accuracy arranging from 63 to 71% for classification with intra- and inter-dataset CVs (**Table 4**).

DISCUSSION

Using radiomics approach and RF, we explored whether multi-dimensional thalamic features define patients with schizophrenia/patients who responded to treatment in this study, resulting in an accuracy of 68% for distinguishing patients with schizophrenia from healthy population and an accuracy of 75% for prediction of early treatment response. Furthermore, support vector machine method revealed similar results through intra- and inter-dataset CV. Our findings might help to facilitate objective diagnosis and treatment selection based on quantitative

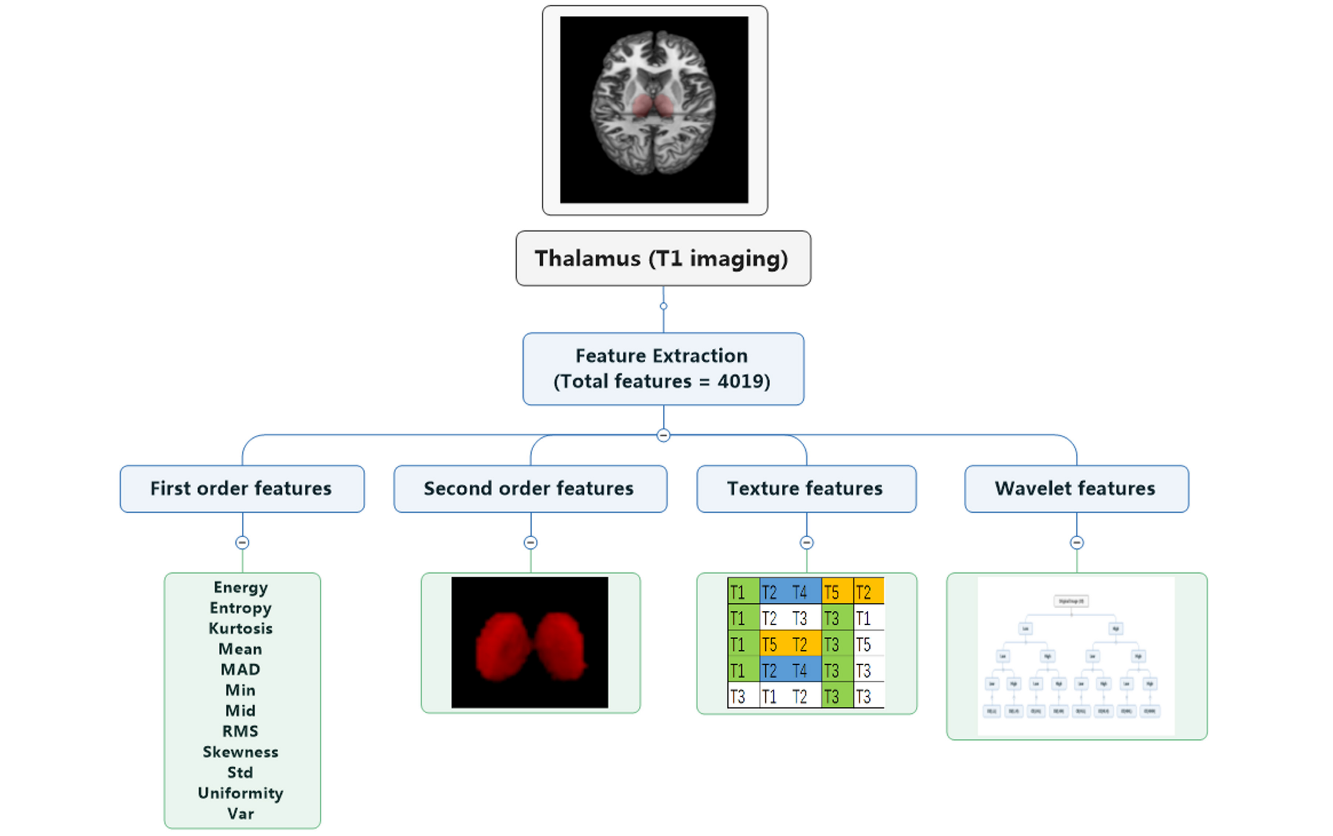


FIGURE 2 | Extraction of radiomics features. Four groups of radiomics features include first-order features, second-order features, texture features, and wavelet features. A total of 4019 features were extracted.

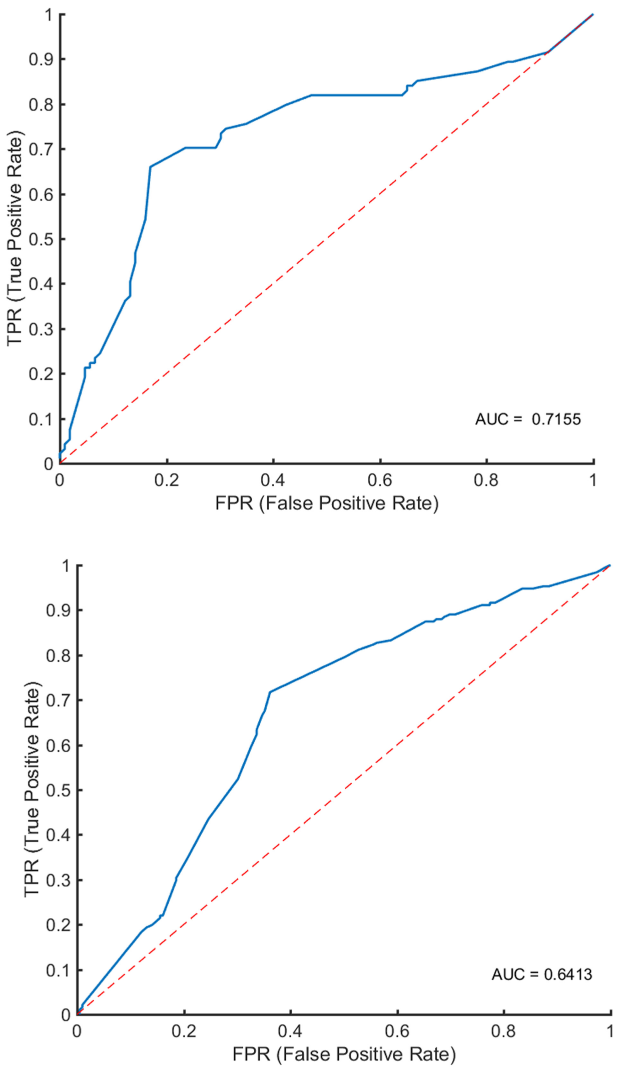


FIGURE 3 | Classification performance. In the **upper panel**, ROC analyses showed an AUC of 0.7155 for predicting early treatment response. In the **lower panel**, ROC analyses showed an AUC of 0.6413 for identifying patients with schizophrenia.

and specific thalamic signature, reflecting its pathophysiology underlying schizophrenia (Tomaszewski and Gillies, 2021).

With the exception of showing conventional features of the thalamus, we also provide newly developed high-throughput features on structural imaging. Findings from imaging and postmortem studies of whole thalamus volume and other structural measures are mixed in schizophrenia and may be influenced by methods, disease state, and the fact that the thalamus is an exceptionally heterogeneous structure. Convergent findings based on multimodality MRI provide support for these neural substrates mediated by the thalamus in schizophrenia (Huang et al., 2015; Xi et al., 2016), suggesting that thalamic abnormalities are implicated in the pathophysiology of this mental disorder. Detecting schizophrenia based on functional connectome is driven by a distributed bilateral network including the thalamus and temporal regions (Lei et al., 2019). As for predicting treatment response, higher baseline glutamate/creatine in the thalamus was seen in non-responders on aripiprazole monotherapy at week 6 and on naturalistic antipsychotic treatment at week 26 compared with healthy controls (Bojesen et al., 2019). Extending previous findings, this evidence is the fundamental basis for disease definition and treatment selection by means of thalamic features using radiomics approach.

Neuroimaging findings have not been used for psychotic disorders clinically, because they are “not sufficiently sensitive or specific for reliable diagnosis in individual patients” (Lieberman and First, 2018). Therefore, from the perspective of methodology, *via* the quickly developed radiomics strategy, the diagnostic performance of multi-dimensional thalamic features is proved liable for identifying individual patients with schizophrenia and predicting early treatment response. The accuracy varied from 82.1 to 87.09% in two previous similar studies by radiomic features from the bilateral hippocampal subfields (Park et al., 2020) and whole brain functional connectivity (Cui et al., 2018). We obtained an accuracy of 68% using the high-throughput thalamic features, in comparison to the diagnostic performance with an accuracy of 78.3% by resting-state networks features (Skatun et al., 2017) and 73.0–81.3% by resting-state connectivity (Mikolas et al., 2016; Huang et al., 2019). Radiomics is considered as the bridge between medical imaging and personalized medicine, and promising to play a central role in the context of psychiatry.

For the cutoff of less than 25% PANSS/Brief Psychiatric Rating Scale (BPRS) reduction, the overall non-response is 43%, and for the cutoff of less than 50% reduction, it is 66.5% (Samara et al., 2019). In line with the finding from randomized controlled trials that the response was assessed at 4–6 weeks, 48 out of 109 (44%) were non-responders for less than 30% PANSS reduction assessed at 2–3 weeks (15–17 days) in this study. Moreover, the olanzapine equivalent was 10 ± 4 mg/day for both responders and non-responders. Our results demonstrate a radiomics approach by multiple thalamic features to predict early treatment response with an accuracy of 75%, an increased level relative to diagnosis. In addition to MRI, genetic evidence indicates schizophrenia polygenic risk score as a predictor of response to antipsychotics in patients with first-episode psychosis (Zhang et al., 2019). An analysis combining neuroimaging and genetics is needed to facilitate the prediction of antipsychotic efficacy in the future. Moreover, radiomics risk modeling combined with time-to-event analysis will contribute to clarifying treatment response (Leger et al., 2017).

An issue of this study that needs to be pointed out is the absence of validation in an independent cohort. Validation could help to confirm the discriminating capacity from different scanners and sites with heterogeneity. A previous study combined independent data of KaSP (Karolinska Schizophrenia Project) and HUBIN (Human Brain Informatics) (Skatun et al., 2017), supporting generalizability across heterogeneous samples. Features across MRI scanners with no difference suggest the repeatability (Cui et al., 2018, 2021a). In the next step, the combination of data from different scanners could consolidate

TABLE 3 | Classification performance.

	Accuracy	Sensitivity	Specificity	AUC	Features
Diagnosis (191 patients and 199 controls)	0.68 ± 0.04	0.60 ± 0.31	0.61 ± 0.30	0.64 ± 0.23	"W1.Mid"; "W1.SRE_8"; "W2.LRHGLE_8"; "W3.Min"; "W4.Co_Corr_12"; "W4.Co_Var_13"; "W5.Co_Corr_11"; "W5.RLN_9"; "W6.Co_Corr_2"; "W6.Co_Corr_7"; "W7.IMC1_9"; "W9.Co_Corr_12"
Prediction (61 responders and 48 non-responders)	0.75 ± 0.08	0.65 ± 0.25	0.80 ± 0.23	0.72 ± 0.12	"W1.LRE_9"; "W3.Min"; "W6.Co_Corr_7"; "W6.Co_Var_7"

AUC, area under the curve; "W1"–"W9", Wavelet features; "_1"–"_13", direction of each feature; Mid, middle; SRE, short run emphasis; LRHGLE, long-run high gray-level emphasis; Min, minimum; Co_Corr, correlation; RLN, run length non-uniformity; IMC, informational measure of correlation; LRE, long-run emphasis; Var, variance.

TABLE 4 | Classification performance using intra- and inter-dataset cross-validation.

	Accuracy	Sensitivity	Specificity	Features
Intra-dataset cross-validation (80% dataset 1 and dataset 2 for training and the other 20% for testing)				
Diagnosis (17 features)	68.37%	71.15%	70.62%	W1.Mid; W1.Min; W1.Mid; W2.RMS; W2.Surface; W2.SVR; W2.Volume; W2.SRE_8; W2.Homo2_13; W3.Min; W4.Co_Corr_12; W4.Co_Var_13; W5.Co_Corr_11; W5.RLN_9; W6.Co_Corr_2; W6.Co_Corr_7; W7.IMC1_9
Prediction (7 features)	71.01%	72.53%	71.69%	W1.SRLGLE_1; W1.Compactness1; W2.Energy; W2.MAD; W3.Min; W6.Cluster_Shade_mean; W8.Cluster_Shade_8
Inter-dataset cross-validation (dataset 1 training, dataset 2 testing)				
Diagnosis (12 features)	65.19%	63.21%	68.55%	W1.Mid; W1.SRE_8; W1.Min; W2.RMS; W2.Surface; W2.Homo2_13; W3.Min; W4.Co_Corr_12; W5.Co_Corr_11; W5.RLN_9; W6.Co_Corr_2; W9.SRHGLE_5
Prediction (5 features)	68.36%	65.75%	69.73%	W1.LRE_9; W1.HGLRE_3; W2.Energy_GLCM_3; W7.Max_GLCM_1; W8.AutoCorr_2
Inter-dataset cross-validation (dataset 2 training, dataset 1 testing)				
Diagnosis (10 features)	63.88%	67.56%	66.46%	W1.LGLRE_11; W1.SRE_6; W2.Sum_var_mean; W2.SRLGLE_6; W4.Dissimilarity_1; W5.Dissimilarity_mean; W5.SRLGLE_4; W7.Diff_entropy_13; W7.Homo2_13; W9.IMC1_mean
Prediction (4 features)	65.21%	69.02%	64.35%	W1.Min; W2.Uniformity; W8.Energy_GLCM_1; W9.LRHGLE_mean

"W1"–"W9", Wavelet features; "_1"–"_13", direction of each feature; AutoCorr, autocorrelation; Co_Corr, correlation; GLCM, gray-level co-occurrence matrix; Homo, homogeneity; IMC, informational measure of correlation; LGLRE, low gray-level run emphasis; LRE, long-run emphasis; LRHGLE, long-run high gray-level emphasis; MAD, median absolute deviation; Max, maximum; Mid, middle; Min, minimum; RLN, run length non-uniformity; RMS, root mean square; SRE, short-run emphasis; SRHGLE, short-run high gray-level emphasis; SRLGLE, short-run low gray-level emphasis; Var, variance.

and promote the generalizability of MRI findings in clinical practice. As DSM-5 stated (American Psychiatric Association, 2013), "The peak age at onset for the first psychotic episode is in the early- to mid-20s for males and in the late-20s for females." Our sample included patients with a wide age range, so potential confounders of brain development could not be excluded. A mixed group of high school students and young adults in the study reflects the clinical heterogeneity of schizophrenia. Besides, because of a very small number of patients with relatively long medical history and no precise boundary between short and long duration of illness, we were unable to perform meaningful subgroup analyses, which may introduce an effect on the results owing to the design of a naturalistic study. Finally, connectomics defined by MRI and genomics in neuropathology gain ground on brain disorder (Jiang et al., 2020); hence, *trans*-omics is promising to shape a refined diagnosis and prediction in schizophrenia. No "one fits all" omics approach exists in this field (Shiri et al., 2020). It depends on the study design.

Our study demonstrates a radiomics approach by multiple thalamic features to diagnose schizophrenia and predict early treatment response with a comparable accuracy. Combining novel machine learning models, radiomics studies try to break the boundary and tend to explore transdiagnostic characteristics of mental disorders (Cui et al., 2021b), transforming the guidance of diagnosis and treatment selection for mental disorders in the future.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: There was no relevant provision concerning public access to data when participants were included, so the data in this study could not be publicly available. Requests to access these datasets should be directed to the corresponding author (L-BC).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institutional Ethics Committee, First Affiliated Hospital (Xijing Hospital) of the Fourth Military Medical University. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

L-BC, HY, WQ, and FC: guarantors of integrity of entire study. L-BC, Y-JZ, H-LL, and LL: literature research and experimental studies. L-BC and LL: statistical analysis. All authors: study concepts/study design or data acquisition or data analysis/interpretation, manuscript drafting or manuscript revision for important intellectual content, approval of final version of submitted manuscript, clinical studies, and manuscript editing.

FUNDING

This study was supported by the Project funded by the China Postdoctoral Science Foundation (2019TQ0130), the Fourth Military Medical University (2019CYJH), NSFC grant 81801675, Wisdom in Creation to L-BC, Innovative Project of Chinese PLA General Hospital (CX19028) to FC, the National Basic Research Program of China grants 2015CB856403 and 2014CB543203, Science and Technology Projects of Xi'an grant 201809170CX11JC12, NSFC grants 81771918, 81471811, and 81471738 to WQ. LL was supported in part by the Postdoctoral Fellowship of Peking-Tsinghua Center for Life Sciences.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2021.682777/full#supplementary-material

REFERENCES

Aerts, H. J., Velazquez, E. R., Leijenaar, R. T., Parmar, C., Grossmann, P., Carvalho, S., et al. (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* 5:4006. doi: 10.1038/ncomms5006

Altamura, A. C., Delvecchio, G., Paletta, S., Di Pace, C., Reggiori, A., Fiorentini, A., et al. (2017). Gray matter volumes may predict the clinical response to paliperidone palmitate long-acting in acute psychosis: a pilot longitudinal neuroimaging study. *Psychiatry Res.* 261, 80–84. doi: 10.1016/j.pscychresns.2017.01.008

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders*, Fifth Edition. American Psychiatric Publishing.

Anderson, A., and Cohen, M. S. (2013). Decreased small-world functional network connectivity and clustering across resting state networks in schizophrenia: an fMRI classification tutorial. *Front. Hum. Neurosci.* 7:520. doi: 10.3389/fnhum.2013.00520

Anderson, A., Dinov, I. D., Sherin, J. E., Quintana, J., Yuille, A. L., and Cohen, M. S. (2010). Classification of spatially unaligned fMRI scans. *Neuroimage* 49, 2509–2519. doi: 10.1016/j.neuroimage.2009.08.036

Arribas, J. I., Calhoun, V. D., and Adali, T. (2010). Automatic Bayesian classification of healthy controls, bipolar disorder, and schizophrenia using intrinsic connectivity maps from FMRI data. *IEEE Trans. Biomed. Eng.* 57, 2850–2860. doi: 10.1109/TBME.2010.2080679

Bojesen, K. B., Ebdrup, B. H., Jessen, K., Sigvard, A., Tangmose, K., Edden, R. A. E., et al. (2019). Treatment response after 6 and 26 weeks is related to baseline glutamate and GABA levels in antipsychotic-naive patients with psychosis. *Psychol. Med.* 50, 2182–2193. doi: 10.1017/S0033291719002277

Brugger, S. P., and Howes, O. D. (2017). Heterogeneity and homogeneity of regional brain structure in schizophrenia: a meta-analysis. *JAMA Psychiatry* 74, 1104–1111. doi: 10.1001/jamapsychiatry.2017.2663

Chaddad, A., Desrosiers, C., Hassan, L., and Tanougast, C. (2017). Hippocampus and amygdala radiomic biomarkers for the study of autism spectrum disorder. *BMC Neurosci.* 18:52. doi: 10.1186/s12868-017-0373-0

Cui, L. B., Cai, M., Wang, X. R., Zhu, Y. Q., Wang, L. X., Xi, Y. B., et al. (2019a). Prediction of early response to overall treatment for schizophrenia: a functional magnetic resonance imaging study. *Brain Behav.* 9:e01211. doi: 10.1002/brb3.1211

Cui, L. B., Wei, Y., Xi, Y. B., Griffa, A., De Lange, S. C., Kahn, R. S., et al. (2019b). Connectome-based patterns of first-episode medication-naive patients with schizophrenia. *Schizophr. Bull.* 45, 1291–1299. doi: 10.1093/schbul/sbz014

Cui, L. B., Fu, Y. F., Liu, L., Wu, X. S., Xi, Y. B., Wang, H. N., et al. (2021a). Baseline structural and functional magnetic resonance imaging predicts early treatment response in schizophrenia with radiomics strategy. *Eur. J. Neurosci.* 53, 1961–1975. doi: 10.1111/ejn.15046

Cui, L. B., Liu, L., Guo, F., Chen, Y. C., Chen, G., Xi, M., et al. (2017a). Disturbed brain activity in resting-state networks of patients with first-episode schizophrenia with auditory verbal hallucinations: a cross-sectional functional MR imaging study. *Radiology* 283, 810–819. doi: 10.1148/radiol.2016160938

Cui, L. B., Wang, L. X., Tian, P., Wang, H. N., Cai, M., Guo, F., et al. (2017b). Aberrant perfusion and its connectivity within default mode network of first-episode drug-naive schizophrenia patients and their unaffected first-degree relatives. *Sci. Rep.* 7:16201. doi: 10.1038/s41598-017-14343-7

Cui, L. B., Liu, L., Wang, H. N., Wang, L. X., Guo, F., Xi, Y. B., et al. (2018). Disease definition for schizophrenia by functional connectivity using radiomics strategy. *Schizophr. Bull.* 44, 1053–1059. doi: 10.1093/schbul/sby007

Cui, L. B., Xu, X., and Cao, F. (2021b). Building the precision medicine for mental disorders via radiomics/machine learning and neuroimaging. *Front Neurosci.* 15:685005. doi: 10.3389/fnins.2021.685005

Dale, A. M., Fischl, B., and Sereno, M. I. (1999). Cortical surface-based analysis ☆ : I. segmentation and surface reconstruction. *Neuroimage* 9:179.

de Filippis, R., Carbone, E. A., Gaetano, R., Bruni, A., Pugliese, V., Segura-Garcia, C., et al. (2019). Machine learning techniques in a structural and functional MRI diagnostic approach in schizophrenia: a systematic review. *Neuropsychiatr. Dis. Treat.* 15, 1605–1627. doi: 10.2147/NDT.S202418

Dietsche, B., Kircher, T., and Falkenberg, I. (2017). Structural brain changes in schizophrenia at different stages of the illness: a selective review of longitudinal magnetic resonance imaging studies. *Aust. N. Z. J. Psychiatry* 51, 500–508. doi: 10.1177/0004867417699473

Dorph-Petersen, K. A., and Lewis, D. A. (2017). Postmortem structural studies of the thalamus in schizophrenia. *Schizophr. Res.* 180, 28–35. doi: 10.1016/j.schres.2016.08.007

Doucet, G. E., Moser, D. A., Luber, M. J., Leibu, E., and Frangou, S. (2018). Baseline brain structural and functional predictors of clinical outcome in the early course of schizophrenia. *Mol. Psychiatry*. 25, 863–872. doi: 10.1038/s41380-018-0269-0

Dusi, N., Bellani, M., Perlini, C., Squarcina, L., Marinelli, V., Finos, L., et al. (2017). Progressive disability and prefrontal shrinkage in schizophrenia patients with poor outcome: a 3-year longitudinal study. *Schizophr. Res.* 179, 104–111. doi: 10.1016/j.schres.2016.09.013

Dwyer, D. B., Cabral, C., Kambeitz-Ilankovic, L., Sanfelici, R., Kambeitz, J., Calhoun, V., et al. (2018). Brain subtyping enhances the neuroanatomical discrimination of schizophrenia. *Schizophr. Bull.* 44, 1060–1069. doi: 10.1093/schbul/sby008

Ferri, J., Ford, J. M., Roach, B. J., Turner, J. A., van Erp, T. G., Voyvodic, J., et al. (2018). Resting-state thalamic dysconnectivity in schizophrenia and relationships with symptoms. *Psychol. Med.* 48, 2492–2499. doi: 10.1017/S003329171800003X

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., et al. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron* 33:341.

Fischl, B., Sereno, M. I., and Dale, A. M. (1999). Cortical surface-based analysis. II: inflation, flattening, and a surface-based coordinate system. *Neuroimage* 9, 195–207.

Fischl, B., Van, D. K. A., Destrieux, C., Halgren, E., SãcGonne, F., Salat, D. H., et al. (2004). Automatically parcellating the human cerebral cortex. *Cerebral Cortex* 14, 11–22.

Francis, M. M., Hummer, T. A., Vohs, J. L., Yung, M. G., Visco, A. C., Mehdiyoun, N. F., et al. (2018). Cognitive effects of bilateral high frequency repetitive transcranial magnetic stimulation in early phase psychosis: a pilot study. *Brain Imaging Behav*. 13, 852–861. doi: 10.1007/s11682-018-9902-4

Fung, G., Cheung, C., Chen, E., Lam, C., Chiu, C., Law, C. W., et al. (2014). MRI predicts remission at 1 year in first-episode schizophrenia in females with larger striato-thalamic volumes. *Neuropsychobiology* 69, 243–248. doi: 10.1159/000358837

Giraldo-Chica, M., and Woodward, N. D. (2017). Review of thalamocortical resting-state fMRI studies in schizophrenia. *Schizophr. Res.* 180, 58–63. doi: 10.1016/j.schres.2016.08.005

Gong, J., Cui, L. B., Xi, Y. B., Zhao, Y. S., Yang, X. J., Xu, Z. L., et al. (2020). Predicting response to electroconvulsive therapy combined with antipsychotics in schizophrenia using multi-parametric magnetic resonance imaging. *Schizophr. Res.* 216, 262–271. doi: 10.1016/j.schres.2019.11.046

Hadley, J. A., Nenert, R., Kraguljac, N. V., Bolding, M. S., White, D. M., Skidmore, F. M., et al. (2014). Ventral tegmental area/midbrain functional connectivity and response to antipsychotic medication in schizophrenia. *Neuropsychopharmacology* 39, 1020–1030. doi: 10.1038/npp.2013.305

Huang, J., Zhu, Q., Hao, X., Shi, X., Gao, S., Xu, X., et al. (2019). Identifying resting-state multi-frequency biomarkers via tree-guided group sparse learning for schizophrenia classification. *IEEE J. Biomed. Health Inform.* 23, 342–350. doi: 10.1109/JBHI.2018.2796588

Huang, P., Xi, Y., Lu, Z. L., Chen, Y., Li, X., Li, W., et al. (2015). Decreased bilateral thalamic gray matter volume in first-episode schizophrenia with prominent hallucinatory symptoms: a volumetric MRI study. *Sci. Rep.* 5, 14505. doi: 10.1038/srep14505

Hutcheson, N. L., Clark, D. G., Bolding, M. S., White, D. M., and Lahti, A. C. (2014). Basal ganglia volume in unmedicated patients with schizophrenia is associated with treatment response to antipsychotic medication. *Psychiatry Res.* 221, 6–12. doi: 10.1016/j.pscychresns.2013.10.002

Jiang, J.-B., Cao, Y., An, N.-Y., Yang, Q., and Cui, L.-B. (2020). Magnetic resonance imaging-based connectomics in first-episode schizophrenia: from preclinical study to clinical translation. *Front. Psychiatry* 11:948. doi: 10.3389/fpsyt.2020.565056

Kraguljac, N. V., White, D. M., Hadley, J. A., Visscher, K., Knight, D., ver Hoef, L., et al. (2016a). Abnormalities in large scale functional networks in unmedicated patients with schizophrenia and effects of risperidone. *Neuroimage Clin.* 10, 146–158. doi: 10.1016/j.nicl.2015.11.015

Kraguljac, N. V., White, D. M., Hadley, N., Hadley, J. A., Ver Hoef, L., Davis, E., et al. (2016b). Aberrant hippocampal connectivity in unmedicated patients with schizophrenia and effects of antipsychotic medication: a longitudinal resting state functional MRI study. *Schizophr. Bull.* 42, 1046–1055. doi: 10.1093/schbul/sbv228

Lambin, P., Leijenaar, R. T. H., Deist, T. M., Peerlings, J., de Jong, E. E. C., van Timmeren, J., et al. (2017). Radiomics: the bridge between medical imaging and personalized medicine. *Nat. Rev. Clin. Oncol.* 14, 749–762. doi: 10.1038/nrclinonc.2017.141

Leger, S., Zwanenburg, A., Pilz, K., Lohaus, F., Linge, A., Zophel, K., et al. (2017). A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling. *Sci. Rep.* 7:13206. doi: 10.1038/s41598-017-13448-3

Lei, D., Pinaya, W. H. L., van Amelsvoort, T., Marcelis, M., Donohoe, G., Mothersill, D. O., et al. (2019). Detecting schizophrenia at the level of the individual: relative diagnostic value of whole-brain images, connectome-wide functional connectivity and graph-based metrics. *Psychol. Med.* 50, 1852–1861. doi: 10.1017/S0033291719001934

Li, B., Cui, L. B., Xi, Y. B., Friston, K. J., Guo, F., Wang, H. N., et al. (2017). Abnormal effective connectivity in the brain is involved in auditory verbal hallucinations in schizophrenia. *Neurosci. Bull.* 33, 281–291. doi: 10.1007/s12264-017-0101-x

Liang, S., Li, Y., Zhang, Z., Kong, X., Wang, Q., Deng, W., et al. (2018). Classification of first-episode schizophrenia using multimodal brain features: a combined structural and diffusion imaging study. *Schizophr. Bull.* 45, 591–599. doi: 10.1093/schbul/sby091

Lieberman, J. A., and First, M. B. (2018). Psychotic disorders. *N. Engl. J. Med.* 379, 270–280. doi: 10.1056/NEJMra1801490

Liu, H., Fan, G., Xu, K., and Wang, F. (2011). Changes in cerebellar functional connectivity and anatomical connectivity in schizophrenia: a combined resting-state functional MRI and diffusion tensor imaging study. *J. Magn. Reson. Imaging* 34, 1430–1438. doi: 10.1002/jmri.22784

Liu, L., Cui, L.-B., Wu, X.-S., Fei, N.-B., Xu, Z.-L., Wu, D., et al. (2020). Cortical abnormalities and identification for first-episode schizophrenia via high-resolution magnetic resonance imaging. *Biomark Neuropsychiatry* 3:100022. doi: 10.1016/j.bionps.2020.100022

Mikolas, P., Hlinka, J., Skoch, A., Pitra, Z., Frodl, T., Spaniel, F., et al. (2018). Machine learning classification of first-episode schizophrenia spectrum disorders and controls using whole brain white matter fractional anisotropy. *BMC Psychiatry* 18:97. doi: 10.1186/s12888-018-1678-y

Mikolas, P., Melicher, T., Skoch, A., Matejka, M., Slovakova, A., Bakstein, E., et al. (2016). Connectivity of the anterior insula differentiates participants with first-episode schizophrenia spectrum disorders from controls: a machine-learning study. *Psychol. Med.* 46, 2695–2704. doi: 10.1017/S0033291716000878

Mitelman, S. A., Buchsbaum, M. S., Christian, B. T., Merrill, B. M., Buchsbaum, B. R., Mukherjee, J., et al. (2019). Positive association between cerebral grey matter metabolism and dopamine D2/D3 receptor availability in healthy and schizophrenia subjects: an (18)F-fluorodeoxyglucose and (18)F-fallypride positron emission tomography study. *World J. Biol. Psychiatry* 21, 368–382. doi: 10.1080/15622975.2019.1671609

Molina, V., Taboada, D., Aragues, M., Hernandez, J. A., and Sanz-Fuentenebro, J. (2014). Greater clinical and cognitive improvement with clozapine and risperidone associated with a thinner cortex at baseline in first-episode schizophrenia. *Schizophr. Res.* 158, 223–229. doi: 10.1016/j.schres.2014.06.042

Morch-Johnsen, L., Nesvag, R., Faerden, A., Haukvik, U. K., Jorgensen, K. N., Lange, E. H., et al. (2015). Brain structure abnormalities in first-episode psychosis patients with persistent apathy. *Schizophr. Res.* 164, 59–64. doi: 10.1016/j.schres.2015.03.001

Murray, J. D., and Anticevic, A. (2017). Toward understanding thalamocortical dysfunction in schizophrenia through computational models of neural circuit dynamics. *Schizophr. Res.* 180, 70–77. doi: 10.1016/j.schres.2016.10.021

Park, Y. W., Choi, D., Lee, J., Ahn, S. S., Lee, S. K., Lee, S. H., et al. (2020). Differentiating patients with schizophrenia from healthy controls by hippocampal subfields using radiomics. *Schizophr. Res.* 223, 337–344. doi: 10.1016/j.schres.2020.09.009

Parnaudeau, S., Bolkan, S. S., and Kellendonk, C. (2018). The mediodorsal thalamus: an essential partner of the prefrontal cortex for cognition. *Biol. Psychiatry* 83, 648–656. doi: 10.1016/j.biopsych.2017.11.008

Pergola, G., Selvaggi, P., Trizio, S., Bertolino, A., and Blasi, G. (2015). The role of the thalamus in schizophrenia from a neuroimaging perspective. *Neurosci. Biobehav. Rev.* 54, 57–75. doi: 10.1016/j.neubiorev.2015.01.013

Premkumar, P., Fannon, D., Sapara, A., Peters, E. R., Anilkumar, A. P., Simmons, A., et al. (2015). Orbitofrontal cortex, emotional decision-making and response to cognitive behavioural therapy for psychosis. *Psychiatry Res.* 231, 298–307. doi: 10.1016/j.pscychresns.2015.01.013

Rashid, B., Arbabshirani, M. R., Damaraju, E., Cetin, M. S., Miller, R., Pearlson, G. D., et al. (2016). Classification of schizophrenia and bipolar patients using static and dynamic resting-state fMRI brain connectivity. *Neuroimage* 134, 645–657. doi: 10.1016/j.neuroimage.2016.04.051

Samara, M. T., Nikolakopoulou, A., Salanti, G., and Leucht, S. (2019). How many patients with schizophrenia do not respond to antipsychotic drugs in the short term? an analysis based on individual patient data from randomized controlled trials. *Schizophr. Bull.* 45, 639–646. doi: 10.1093/schbul/sby095

Sarpal, D. K., Argyelan, M., Robinson, D. G., Szeszko, P. R., Karlsgodt, K. H., John, M., et al. (2016). Baseline striatal functional connectivity as a predictor of response to antipsychotic drug treatment. *Am. J. Psychiatry* 173, 69–77. doi: 10.1176/appi.ajp.2015.14121571

Sauerbrei, W., Royston, P., and Binder, H. (2007). Selection of important variables and determination of functional form for continuous predictors in multivariable model building. *Stat. Med.* 26, 5512–5528. doi: 10.1002/sim.3148

Scheef, L., Manka, C., Daamen, M., Kuhn, K. U., Maier, W., Schild, H. H., et al. (2010). Resting-state perfusion in nonmedicated schizophrenic patients:

a continuous arterial spin-labeling 3.0-T MR study. *Radiology* 256, 253–260. doi: 10.1148/radiol.10091224

Schnack, H. G., Nieuwenhuis, M., van Haren, N. E., Abramovic, L., Scheewe, T. W., Brouwer, R. M., et al. (2014). Can structural MRI aid in clinical classification? A machine learning study in two independent samples of patients with schizophrenia, bipolar disorder and healthy subjects. *Neuroimage* 84, 299–306. doi: 10.1016/j.neuroimage.2013.08.053

Shafritz, K. M., Ikuta, T., Greene, A., Robinson, D. G., Gallego, J., Lencz, T., et al. (2018). Frontal lobe functioning during a simple response conflict task in first-episode psychosis and its relationship to treatment response. *Brain Imaging Behav*. 13, 541–553. doi: 10.1007/s11682-018-9876-2

Shen, H., Wang, L., Liu, Y., and Hu, D. (2010). Discriminative analysis of resting-state functional connectivity patterns of schizophrenia using low dimensional embedding of fMRI. *Neuroimage* 49, 3110–3121. doi: 10.1016/j.neuroimage.2009.11.011

Shiri, I., Maleki, H., Hajianfar, G., Abdollahi, H., Ashrafinia, S., Hatt, M., et al. (2020). Next-Generation radiogenomics sequencing for prediction of EGFR and KRAS mutation status in NSCLC patients using multimodal imaging and machine learning algorithms. *Mol. Imaging Biol.* 22, 1132–1148. doi: 10.1007/s11307-020-01487-8

Skatun, K. C., Kaufmann, T., Doan, N. T., Alnaes, D., Cordova-Palomera, A., Jonsson, E. G., et al. (2017). Consistent functional connectivity alterations in schizophrenia spectrum disorder: a multisite study. *Schizophr. Bull.* 43, 914–924. doi: 10.1093/schbul/sbw145

Stolz, E., Pancholi, K. M., Goradia, D. D., Paul, S., Keshavan, M. S., Nimgaonkar, V. L., et al. (2012). Brain activation patterns during visual episodic memory processing among first-degree relatives of schizophrenia subjects. *Neuroimage* 63, 1154–1161. doi: 10.1016/j.neuroimage.2012.08.030

Tomaszewski, M. R., and Gillies, R. J. (2021). The biological meaning of radiomic features. *Radiology* 298, 505–516. doi: 10.1148/radiol.2021202553

Talairach, J., and Tournoux, P. (1988). *Co-Planar Stereotaxic Atlas of the Human Brain*. G. Thieme.

Vallieres, M., Zwanenburg, A., Badic, B., Cheze Le Rest, C., Visvikis, D., and Hatt, M. (2018). Responsible radiomics research for faster clinical translation. *J. Nucl. Med.* 59, 189–193. doi: 10.2967/jnumed.117.200501

Wang, X. H., Yu, A., Zhu, X., Yin, H., and Cui, L. B. (2019). Cardiopulmonary comorbidity, radiomics and machine learning, and therapeutic regimens for a cerebral fMRI predictor study in psychotic disorders. *Neurosci. Bull.* 35, 955–957. doi: 10.1007/s12264-019-00409-1

Xi, Y. B., Cui, L. B., Gong, J., Fu, Y. F., Wu, X. S., Guo, F., et al. (2020). Neuroanatomical features that predict response to electroconvulsive therapy combined with antipsychotics in schizophrenia: a magnetic resonance imaging study using radiomics strategy. *Front. Psychiatry* 11:456. doi: 10.3389/fpsyt.2020.00456

Xi, Y. B., Guo, F., Li, H., Chang, X., Sun, J. B., Zhu, Y. Q., et al. (2016). The structural connectivity pathology of first-episode schizophrenia based on the cardinal symptom of auditory verbal hallucinations. *Psychiatry Res.* 257, 25–30. doi: 10.1016/j.pscychresns.2016.09.011

Zeng, L. L., Wang, H., Hu, P., Yang, B., Pu, W., Shen, H., et al. (2018). Multi-Site diagnostic classification of schizophrenia using discriminant deep learning with functional connectivity MRI. *EBioMedicine* 30, 74–85. doi: 10.1016/j.ebiom.2018.03.017

Zhang, J. P., Robinson, D., Yu, J., Gallego, J., Fleischhacker, W. W., Kahn, R. S., et al. (2019). Schizophrenia polygenic risk score as a predictor of antipsychotic efficacy in first-episode psychosis. *Am. J. Psychiatry* 176, 21–28. doi: 10.1176/appi.ajp.2018.17121363

Zhu, J., Zhuo, C., Qin, W., Xu, Y., Xu, L., Liu, X., et al. (2015). Altered resting-state cerebral blood flow and its connectivity in schizophrenia. *J. Psychiatr. Res.* 63, 28–35. doi: 10.1016/j.jpsychires.2015.03.002

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 19 August 2021
doi: 10.3389/fnins.2021.716051

Enhancing Performance of SSVEP-Based Visual Acuity via Spatial Filtering

Xiaowei Zheng[1], *Guanghua Xu*[1,2]*, *Chengcheng Han*[1], *Peiyuan Tian*[1], *Kai Zhang*[1], *Renghao Liang*[1], *Yaguang Jia*[1], *Wenqiang Yan*[1], *Chenghang Du*[1] *and Sicong Zhang*[1]

[1] *School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an, China,* [2] *State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an, China*

The purpose of this study was to enhance the performance of steady-state visual evoked potential (SSVEP)-based visual acuity assessment with spatial filtering methods. Using the vertical sinusoidal gratings at six spatial frequency steps as the visual stimuli for 11 subjects, SSVEPs were recorded from six occipital electrodes (O1, Oz, O2, PO3, POz, and PO4). Ten commonly used training-free spatial filtering methods, i.e., native combination (single-electrode), bipolar combination, Laplacian combination, average combination, common average reference (CAR), minimum energy combination (MEC), maximum contrast combination (MCC), canonical correlation analysis (CCA), multivariate synchronization index (MSI), and partial least squares (PLS), were compared for multielectrode signals combination in SSVEP visual acuity assessment by statistical analyses, e.g., Bland–Altman analysis and repeated-measures ANOVA. The SSVEP signal characteristics corresponding to each spatial filtering method were compared, determining the chosen spatial filtering methods of CCA and MSI with a higher performance than the native combination for further signal processing. After the visual acuity threshold estimation criterion, the agreement between the subjective Freiburg Visual Acuity and Contrast Test (FrACT) and SSVEP visual acuity for the native combination (0.253 logMAR), CCA (0.202 logMAR), and MSI (0.208 logMAR) was all good, and the difference between FrACT and SSVEP visual acuity was also all acceptable for the native combination (−0.095 logMAR), CCA (0.039 logMAR), and MSI (−0.080 logMAR), where CCA-based SSVEP visual acuity had the best performance and the native combination had the worst. The study proved that the performance of SSVEP-based visual acuity can be enhanced by spatial filtering methods of CCA and MSI and also recommended CCA as the spatial filtering method for multielectrode signals combination in SSVEP visual acuity assessment.

Keywords: visual acuity, steady-state visual evoked potential, spatial filtering, multielectrode signals combination, canonical correlation analysis

OPEN ACCESS

Edited by:
Hamid R. Rabiee,
Sharif University of Technology, Iran

Reviewed by:
Yu Zhang,
Lehigh University, United States
Vassiliy Tsytsarev,
University of Maryland, College Park, United States

****Correspondence:***
Guanghua Xu
ghxu@xjtu.edu.cn

Specialty section:
This article was submitted to Neural Technology, a section of the journal Frontiers in Neuroscience

Received: *28 May 2021*
Accepted: *20 July 2021*
Published: *19 August 2021*

Citation:
Zheng X, Xu G, Han C, Tian P, Zhang K, Liang R, Jia Y, Yan W, Du C and Zhang S (2021) Enhancing Performance of SSVEP-Based Visual Acuity via Spatial Filtering. Front. Neurosci. 15:716051. doi: 10.3389/fnins.2021.716051

INTRODUCTION

Visual acuity, one of the most necessary parameters to test visual function, is a measure of the spatial resolution of the visual processing. In general, it is mainly tested by psychophysical methods, e.g., Sloan letters and tumbling E charts (Ricci et al., 1998). However, these methods require the subjects to have sufficient intelligence to comply with the test process and are hard for preverbal or infantile children, the mentally disabled, and malingerers (Incesu and Sobaci, 2011; Zheng et al., 2020c).

Noninvasive electroencephalography (EEG), e.g., steady-state visual evoked potentials (SSVEPs), has been proved to provide an alternative method to estimate visual acuity objectively (Regan, 1973; Norcia et al., 2015). By varying the spatial frequency of the visual stimuli, visual acuity can be measured by a threshold determination criterion by establishing the mathematical model between spatial frequency and SSVEP signals (Hamilton et al., 2021a). Besides, previous studies proved that a larger number of posterior electrodes was relevant to optimize visual function assessment (Hemptinne et al., 2018) and recommended multielectrode montage, e.g., six-electrode of O1, Oz, O2, PO3, POz, and PO4 (Zheng et al., 2020a, 2021), rather than single-electrode in SSVEP visual acuity assessment (Hamilton et al., 2021b). However, in SSVEP visual acuity assessment, SSVEPs are mainly collected at only one active electrode, e.g., Oz at the midline over the occiput (McBain et al., 2007; Odom et al., 2016; Ridder, 2019), except for some other electrode montages, e.g., the bipolar electrodes of Oz and O1 (Norcia and Tyler, 1985a,b; Skoczenski and Norcia, 1999), which was sometimes used to enhanced signal-to-noise-ratio (SNR), especially close to the threshold (Hamilton et al., 2021b).

The spatial filtering technique combining the multielectrode signals into single- or multichannel signals offers a better method for extracting SSVEP features and eliminating nuisance signals in SSVEP studies (Yan et al., 2018). Since scalp EEG is usually regarded to be a linear mixture of multiple time series from various cortical sources (Onton et al., 2006), the weight coefficients can be applied for multielectrode scalp EEG signals to estimate the cortical source activities (Nakanishi et al., 2018b). On the basis of this idea, several methods of extracting optimal spatial filters to reconstruct source activities from scalp EEG signals have been carried out to enhance the SNR of SSVEPs. For instance, the basic spatial filtering methods [e.g., Laplacian combination (Friman et al., 2007) and common average reference (CAR) (Zheng et al., 2020d)] and the model-based spatial filtering methods [e.g., minimum energy combination (MEC) (Friman et al., 2007), canonical correlation analysis (CCA) (Bin et al., 2009; Zheng et al., 2020b; Li et al., 2021), and multivariate synchronization index (MSI) (Zhang et al., 2014a)] have been applied to improve the performance of SSVEPs. However, to date, little is known about whether there is an enhancement of the spatial filtering technique from multielectrode signals on SSVEP visual acuity.

On the basis, in this study, 10 commonly used training-free spatial filtering methods, i.e., native combination (i.e., single-electrode) (Friman et al., 2007), bipolar combination (Hamilton et al., 2021b), Laplacian combination, average combination (Friman et al., 2007), CAR, MEC, maximum contrast combination (MCC) (Friman et al., 2007), CCA, MSI, and partial least squares (PLS) (Ge et al., 2017), were compared for multielectrode signals combination in SSVEP visual acuity assessment. First, SSVEPs were induced by the vertical sinusoidal gratings at six spatial frequency steps and recorded from six occipital electrodes (O1, Oz, O2, PO3, POz, and PO4) for 11 subjects. Next, the SSVEP signal characteristics corresponding to each spatial filtering method were compared to determine the chosen spatial filtering methods with good performance for further signal processing. Then, SSVEP visual acuity can be obtained by the threshold estimation criterion for each chosen spatial filtering method, and the statistical analyses, e.g., Bland–Altman analysis and repeated-measures ANOVA, were used to explore the performance of the spatial filtering technique from multielectrode signals on SSVEP visual acuity. The main purpose of this study was to enhance the performance of SSVEP visual acuity with spatial filtering methods.

MATERIALS AND METHODS

SSVEP Model

For the visual stimulus with a temporal frequency of f, the SSVEP signal, $y_i(t)$, measured as the voltage between a reference electrode and the ith electrode at time t, can be modeled as (Friman et al., 2007; Zerafa et al., 2018):

$$y_i(t) = \sum_{h=1}^{N_h} a_{i,h} \sin\left(2\pi hft + \phi_{i,h}\right) + e_i(t) \tag{1}$$

This linear model consists of two parts: the evoked SSVEP response signal and the noise signal. The evoked SSVEP response consists of many sinusoids with the frequency given by the stimulus frequency f and its harmonic frequencies. N_h is the number of harmonic frequencies. Each sinusoid is determined by its specific amplitude $a_{i,h}$ and phase $\phi_{i,h}$. The noise signal $e_i(t)$ is composed of other signals that are unrelated to SSVEP response, such as electromyography (EMG), electrooculogram (EOG), and other components.

Hence, the SSVEP signal for a time segment of N_t samples with a sampling frequency F_s can be defined in vector form:

$$\boldsymbol{y_i} = \boldsymbol{X_f g_i} + \boldsymbol{e_i} \tag{2}$$

where $\boldsymbol{y_i} = \left[y_i(1), \ldots, y_i(N_t)\right]^{\mathrm{T}} \in \mathbb{R}^{N_t \times 1}$ contains the SSVEP signal of the ith electrode in one segment of N_t samples, and $\boldsymbol{e_i} \in \mathbb{R}^{N_t \times 1}$ is the noise vector. The SSVEP reference signals model $\boldsymbol{X_f} \in \mathbb{R}^{N_t \times 2N_h}$ is defined by Nakanishi et al. (2018b):

$$\boldsymbol{X}_f = \begin{pmatrix} \sin\left(2\pi f \frac{m}{F_s}\right) \\ \cos\left(2\pi f \frac{m}{F_s}\right) \\ \vdots \\ \sin\left(2\pi N_h f \frac{m}{F_s}\right) \\ \cos\left(2\pi N_h f \frac{m}{F_s}\right) \end{pmatrix}^{\mathrm{T}}, \ m = 1, \ldots, N_t. \tag{3}$$

The vector $\boldsymbol{g_i} \in \mathbb{R}^{2N_h \times 1}$ contains the corresponding amplitude $a_{i,h}$ and phase $\phi_{i,h}$.

Finally, for SSVEP signals recorded from N_e electrodes, the model $\boldsymbol{Y}$ can be further defined as:

$$\boldsymbol{Y} = \boldsymbol{X}_f \boldsymbol{G} + \boldsymbol{E} \tag{4}$$

where $\boldsymbol{Y} = \left[\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{N_e}\right] \in \mathbb{R}^{N_t \times N_e}$ contains the sampled SSVEP signals from all electrodes, with each column corresponding to an electrode. $\boldsymbol{E} \in \mathbb{R}^{N_t \times N_e}$ is the noise matrix, and $\boldsymbol{G} \in \mathbb{R}^{2N_h \times N_e}$ contains the amplitudes and phases for all sinusoids.

Spatial Filtering Model

In SSVEPs, the method of linearly combining the multielectrode signals into single- or multichannel signals is called spatial filtering (Yan et al., 2018) to enhance the SNR of SSVEP response. Given N_e-electrode SSVEP signals $\boldsymbol{Y}$ as expressed in **Equation (4)**, single-channel $\boldsymbol{s} \in \mathbb{R}^{N_t \times 1}$ can be created by combining $\boldsymbol{Y}$ linearly using weights $\boldsymbol{w} \in \mathbb{R}^{N_e \times 1}$ (Friman et al., 2007):

$$\boldsymbol{s} = \boldsymbol{Y}\boldsymbol{w}. \tag{5}$$

More generally, multichannel signals $\boldsymbol{S}$ can be created by combining $\boldsymbol{Y}$ linearly using weights $\boldsymbol{W}$ (Friman et al., 2007):

$$\boldsymbol{S} = \boldsymbol{Y}\boldsymbol{W} \tag{6}$$

where $\boldsymbol{S} = [\boldsymbol{s_1}, , \boldsymbol{s_{N_c}}] \in \mathbb{R}^{N_t \times N_c}$ are the spatially filtered signals, and N_c is the number of the channels considered for further signal analysis. When N_c is 1, **Equation (5)** is the same as **Equation (6)**. $\boldsymbol{W} = [\boldsymbol{w_1}, \ldots, \boldsymbol{w_{N_c}}] \in \mathbb{R}^{N_e \times N_c}$ is the weight matrix for spatial filtering. Below, 10 commonly used spatial filtering methods for choices of $\boldsymbol{W}$ were introduced.

Spatial Filtering Methods

Here, we aimed to compare the effect on visual acuity assessment by SSVEPs with different spatial filtering methods to combine multielectrode signals into a single-channel signal. The visual acuity results depend on the SSVEP amplitude changes versus spatial frequencies (Zheng et al., 2020c), and the SSVEP amplitude is usually obtained from single-channel SSVEP by using Fourier analysis to transform an SSVEP signal from the time domain to the frequency domain and extracting the specific SSVEP amplitude at the fundamental frequency of the visual stimulus from the resulting spectrum (Hamilton et al., 2021a,b). Hence, here, we only focused on the single-channel spatial filtering methods, i.e., $N_c = 1$, and $\boldsymbol{W} = \boldsymbol{w} \in\ R^{N_e \times 1}$.

Native Combination

The native combination is also called the monopolar combination where only the SSVEP signals from one of the electrodes are analyzed (Friman et al., 2007; Zerafa et al., 2018). In the SSVEP analysis, the most used electrode is Oz (Yan et al., 2021; Zheng et al., 2020c). Assuming that the SSVEP signals from the Oz electrode are corresponding to the first column in N_e-electrode SSVEP signals $\boldsymbol{Y}$ (same below), the spatial filtering weights $\boldsymbol{w}$ can be expressed as:

$$\boldsymbol{w} = [1, 0, \ldots, 0]^{\mathrm{T}}. \tag{7}$$

Bipolar Combination

The bipolar combination is used to reduce the common noise signals by measuring the voltage of two closely placed electrodes (Friman et al., 2007). In SSVEP visual acuity assessment, the bipolar combination sometimes is also used (Hamilton et al., 2021b). According to the previous studies (Norcia and Tyler, 1985a,b), we chose the commonly used electrode pair (Oz–O1). Hence, assuming that the SSVEP signals from the O1 electrode are from the second column in $\boldsymbol{Y}$, $\boldsymbol{w}$ can be expressed as:

$$\boldsymbol{w} = [1, -1, 0, \ldots, 0]^{\mathrm{T}}. \tag{8}$$

Laplacian Combination

The Laplacian combination is the improvement of the bipolar combination by using the mean voltage of the surrounding electrodes from one center electrode as the reference voltage (Hamilton et al., 2021b). Laplacian combination is mainly divided into two types in SSVEP visual acuity studies: one- and two-dimensional Laplacian combination (Hamilton et al., 2021b). One-dimensional Laplacian combination in SSVEP acuity studies is carried out by using voltage from Oz − 1/2(O1 + O2) as the signal (Bach and Heinrich, 2019; Knotzele and Heinrich, 2019; Kurtenbach et al., 2013). A two-dimensional Laplacian combination, i.e., the fourth Laplacian combination of Oz − 1/4(O1 + O2 + POz + Iz) (Hamilton et al., 2013), is also used in the relevant study. Here, assuming that the SSVEP signals from the O1, O2, POz, and Iz electrode are the second, the third, the fourth, and the fifth column in $\boldsymbol{Y}$, respectively, $\boldsymbol{w}$ for one-dimensional Laplacian combination can be expressed as:

$$\boldsymbol{w} = \left[1, -\frac{1}{2}, -\frac{1}{2}, 0, \ldots, 0\right]^{\mathrm{T}}. \tag{9}$$

and $\boldsymbol{w}$ for two-dimensional Laplacian combination can be expressed as:

$$\boldsymbol{w} = \left[1, -\frac{1}{4}, -\frac{1}{4}, -\frac{1}{4}, -\frac{1}{4}, 0, \ldots, 0\right]^{\mathrm{T}}. \tag{10}$$

Average Combination

The average combination is used by taking the average signals from all electrodes to amplify the SSVEP component and cancel the electrode-specific noise (Friman et al., 2007), where the weights $\boldsymbol{w}$ can be expressed as:

$$\boldsymbol{w} = \left[\frac{1}{N_e}, \ldots, \frac{1}{N_e}\right]^{\mathrm{T}}. \tag{11}$$

Common Average Reference

Common average reference, a commonly used spatial filtering method, is achieved by subtracting the mean signals of all electrodes from the selected electrode signals (Zheng et al., 2020d). Here, also choosing the Oz electrode, the weights $\boldsymbol{w}$ can be expressed as:

$$\boldsymbol{w} = \left[\frac{N_e - 1}{N_e}, -\frac{1}{N_e}, \ldots, -\frac{1}{N_e}\right]^{\mathrm{T}}. \tag{12}$$

Minimum Energy Combination

The MEC-based spatial filtering is proposed by Friman et al. (2007) to minimize the energy from nuisance signals. First, by removing any potential SSVEP activity from N_e-electrodes signals $\boldsymbol{Y}$ by projecting them onto the orthogonal complement of the

SSVEP model matrix $\boldsymbol{X_f}$ in **Equation (3)**, the nuisance signals $\widetilde{\boldsymbol{Y}}_f \in \mathbb{R}^{N_t \times N_e}$ can be expressed as (Friman et al., 2007):

$$\widetilde{\boldsymbol{Y}}_f = \boldsymbol{Y} - \boldsymbol{X}_f\left(\boldsymbol{X}_f^{\mathrm{T}}\boldsymbol{X}_f\right)^{-1}\boldsymbol{X}_f^{\mathrm{T}}\boldsymbol{Y}. \tag{13}$$

where $\widetilde{\boldsymbol{Y}}_f$ contains only nuisance signals and noise. In other words, $\widetilde{\boldsymbol{Y}}_f \approx \boldsymbol{E}$.

Next is to find a weight vector $\widehat{\boldsymbol{w}}_{\boldsymbol{f}} \in \mathbb{R}^{N_e \times 1}$ to minimize the energy of the combination of electrode signals $\widetilde{\boldsymbol{Y}}_{\boldsymbol{f}}\widehat{\boldsymbol{w}}_{\boldsymbol{f}}$:

$$\widehat{\boldsymbol{w}}_{\boldsymbol{f}} = \underset{\widehat{\boldsymbol{w}}_{\boldsymbol{f}}}{\operatorname{argmin}} \parallel \widetilde{\boldsymbol{Y}}_{\boldsymbol{f}}\widehat{\boldsymbol{w}}_{\boldsymbol{f}} \parallel^2 = \underset{\widehat{\boldsymbol{w}}_{\boldsymbol{f}}}{\operatorname{argmin}}\, \widehat{\boldsymbol{w}}_{\boldsymbol{f}}^{\mathrm{T}}\widetilde{\boldsymbol{Y}}_{\boldsymbol{f}}^{\mathrm{T}}\widetilde{\boldsymbol{Y}}_{\boldsymbol{f}}\widehat{\boldsymbol{w}}_{\boldsymbol{f}}. \tag{14}$$

The above minimization problem can be solved by decomposing the eigenvalues of the matrix $\widetilde{\boldsymbol{Y}}_{\boldsymbol{f}}^{\mathrm{T}}\widetilde{\boldsymbol{Y}}_{\boldsymbol{f}}$, and the spatial filter weights $\widehat{\boldsymbol{w}}_{\boldsymbol{f}}$ are defined by the eigenvector $\boldsymbol{v}_1$ corresponding to the smallest eigenvalue λ_1 (Friman et al., 2007; Yan et al., 2019):

$$\widehat{\boldsymbol{w}}_{\boldsymbol{f}} = \frac{\boldsymbol{v}_1}{\sqrt{\lambda_1}} \tag{15}$$

Maximum Contrast Combination

Maximum contrast combination is realized by maximizing the SSVEP energy and minimizing the nuisance noise energy simultaneously. Hence, MCC can be achieved as follows (Friman et al., 2007):

$$\widehat{\boldsymbol{w}}_{\boldsymbol{f}} = \underset{\widehat{\boldsymbol{w}}_{\boldsymbol{f}}}{\operatorname{argmax}} \frac{\parallel \boldsymbol{Y}\widehat{\boldsymbol{w}}_{\boldsymbol{f}} \parallel^2}{\parallel \widetilde{\boldsymbol{Y}}_{\boldsymbol{f}}\widehat{\boldsymbol{w}}_{\boldsymbol{f}} \parallel^2} = \underset{\widehat{\boldsymbol{w}}_{\boldsymbol{f}}}{\operatorname{argmax}} \frac{\widehat{\boldsymbol{w}}_{\boldsymbol{f}}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}\widehat{\boldsymbol{w}}_{\boldsymbol{f}}}{\widehat{\boldsymbol{w}}_{\boldsymbol{f}}^{\mathrm{T}}\widetilde{\boldsymbol{Y}}_{\boldsymbol{f}}^{\mathrm{T}}\widetilde{\boldsymbol{Y}}_{\boldsymbol{f}}\widehat{\boldsymbol{w}}_{\boldsymbol{f}}}. \tag{16}$$

The above maxima can be found by a generalized eigen-decomposition of the matrices $\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}$ and $\widetilde{\boldsymbol{Y}}_{\boldsymbol{f}}^{\mathrm{T}}\widetilde{\boldsymbol{Y}}_{\boldsymbol{f}}$, and the spatial filter weights $\widehat{\boldsymbol{w}}_{\boldsymbol{f}}$ are defined as the eigenvector corresponding to the largest eigenvalue (Zerafa et al., 2018).

Canonical Correlation Analysis

Canonical correlation analysis, a statistical way to measure the underlying correlation between two sets of multidimensional variables, was first used in SSVEP analysis by Lin et al. (2007). Till now, CCA has become the most widely used method in SSVEPs as a result of its effectiveness, robustness, and simple implementation (Bin et al., 2009; Zheng et al., 2020b; Li et al., 2021). Here, CCA finds the weights $\boldsymbol{w_y} \in \mathbb{R}^{N_e \times 1}$ and $\boldsymbol{w_{xf}} \in \mathbb{R}^{2N_h \times 1}$ to maximize the linear combinations between $\boldsymbol{y} = \boldsymbol{Y}\boldsymbol{w_y} \in \mathbb{R}^{N_t \times 1}$ and $\boldsymbol{x} = \boldsymbol{X_f}\boldsymbol{w_{xf}} \in \mathbb{R}^{N_t \times 1}$ representing the multichannel SSVEP signals and the SSVEP reference signals. Hence, the weight vectors $\boldsymbol{w_y}$ and $\boldsymbol{w_{xf}}$ can be obtained as follows:

$$\begin{aligned}\boldsymbol{w}_y, \boldsymbol{w}_{xf} &= \underset{W_y, W_{xf}}{\operatorname{argmax}}\ \rho\left(\boldsymbol{y}, \boldsymbol{x}\right) = \frac{\mathrm{E}\left[\boldsymbol{y}^{\mathrm{T}}\boldsymbol{x}\right]}{\sqrt{\mathrm{E}\left[\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}\right]\mathrm{E}\left[\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}\right]}}\\ &= \frac{\mathrm{E}\left[\boldsymbol{w}_{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{X_f}\boldsymbol{w}_{xf}\right]}{\sqrt{\mathrm{E}\left[\boldsymbol{w}_{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{w_y}\right]\mathrm{E}\left[\boldsymbol{w}_{\boldsymbol{xf}}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{f}}^{\mathrm{T}}\boldsymbol{X_f}\boldsymbol{w_{xf}}\right]}}.\end{aligned} \tag{17}$$

The maximum of ρ is the maximum canonical correlation. The spatial filter weights $\boldsymbol{w_y}$ is defined as the eigenvector corresponding to the largest eigenvalue after transforming the above optimization problem into the eigenvalue decomposition problem (Yan et al., 2019).

Multivariate Synchronization Index

Multivariate synchronization index, introduced by Zhang et al. (2014a), is another multichannel detection method for SSVEPs. Assuming that the reference signal $\boldsymbol{X_f}$ is synchronized to the SSVEP signals $\boldsymbol{Y}$, MSI is used for estimating the synchronization between $\boldsymbol{Y}$ and $\boldsymbol{X_f}$. First, the matrices of $\boldsymbol{Y}$ and $\boldsymbol{X_f}$ are normalized to have a zero mean and unitary variance. Then, a correlation matrix $\boldsymbol{C}$ is estimated as (Zhang et al., 2014a):

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{C_{YY}} & \boldsymbol{C_{YX_f}} \\ \boldsymbol{C_{X_fY}} & \boldsymbol{C_{X_fX_f}} \end{bmatrix} \tag{18}$$

where

$$\boldsymbol{C_{YY}} = \frac{1}{N_t}\boldsymbol{YY}^{\mathrm{T}},\ \ \boldsymbol{C_{X_fX_f}} = \frac{1}{N_t}\boldsymbol{X_fX_f}^{\mathrm{T}},\ \boldsymbol{C_{YX_f}} = \boldsymbol{C_{X_fY}} = \frac{1}{N_t}\boldsymbol{YX_f}^{\mathrm{T}}. \tag{19}$$

To weaken the effect from the autocorrelation on the synchronization measure, the following linear transformation is adopted:

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{C_{YY}^{-1/2}} & 0 \\ 0 & \boldsymbol{C_{X_fX_f}^{-1/2}} \end{bmatrix} \tag{20}$$

The transformed correlation matrix $\boldsymbol{C}'$ is as follows after canceling out the autocorrelation:

$$\boldsymbol{C}^{'} = \boldsymbol{UCU}^{\mathrm{T}} \tag{21}$$

Here, rather than the previous studies using the synchronization index S-estimator in MSI-based frequency recognition in SSVEPs (Zhang et al., 2014a; Zerafa et al., 2018), the spatial filter weights $\boldsymbol{w}$ is directly obtained by the eigenvector corresponding to the largest eigenvalue of the matrix $\boldsymbol{C}^{'}$.

Partial Least Squares

Partial least squares is a commonly used multiple linear regression method to compute the linear regression between multidimensional predicted variables and multidimensional observable variables (Trejo et al., 2006; Wang et al., 2014a). Wang et al. (2014a) and Ge et al. (2017) proposed a double PLS-based recognition method in SSVEPs, where the first step is to use PLS as a spatial filter to enhance the SNR. Here, we mainly focused on the first step.

In PLS, the SSVEP signals $\boldsymbol{Y}$ and the reference signal $\boldsymbol{X_f}$ are first decomposed into bilinear terms by an iterative procedure to extract the latent variables with maximal correlation (Rosipal and Krämer, 2006):

$$\boldsymbol{Y} = \boldsymbol{TP}^{\mathrm{T}} + \boldsymbol{E} \tag{22}$$

$$\boldsymbol{X_f} = \boldsymbol{UQ}^{\mathrm{T}} + \boldsymbol{F} \tag{23}$$

where matrices $\boldsymbol{T} = \{t_i\}_{i=1}^{D}$ and $\boldsymbol{U} = \{u_i\}_{i=1}^{D}$ are the extracted D latent vectors (i.e., score vectors), $\boldsymbol{P}$ and $\boldsymbol{Q}$ are loading matrices, and $\boldsymbol{E}$ and $\boldsymbol{F}$ are residual matrices. Since $\boldsymbol{Y}$ can be regarded as a linear mixture of $\boldsymbol{X_f}$ and noise [see **Equation (4)**], $\boldsymbol{X_f}$ can be decomposed by $\boldsymbol{Y}$:

$$\boldsymbol{X}_f = \boldsymbol{Y}\boldsymbol{W}_f + \boldsymbol{F}_f \tag{24}$$

where $\boldsymbol{F}_f$ is the residual matrix. $\boldsymbol{W}_f$ is the matrix of linear regression coefficients, which can be defined as (Rosipal and Krämer, 2006):

$$\boldsymbol{W}_f = \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{U}\left(\boldsymbol{T}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{U}\right)^{-1}\boldsymbol{T}^{\mathrm{T}}\boldsymbol{X}_f \tag{25}$$

The spatially filtered SSVEP signals $\boldsymbol{S}$ can be obtained by removing the residual matrix $\boldsymbol{F}_f$:

$$\boldsymbol{S} = \boldsymbol{Y}\boldsymbol{W}_f \tag{26}$$

Here, the spatial filter weights $\boldsymbol{w}$ is obtained by the eigenvector corresponding to the largest eigenvalue of the matrix $\boldsymbol{W}_f$.

EXPERIMENT

Participants

Eleven healthy volunteers (four female, ages 22–27 years) were recruited from Xi'an Jiaotong University. The subjective visual acuity was evaluated by Freiburg Visual Acuity and Contrast Test (FrACT) monocularly (Bach, 1996). The experimental protocol was approved by the Human Ethics Committee of Xi'an Jiaotong University, conforming to the Declaration of Helsinki. All subjects also submitted the written consent after informed of the contents of the experiment.

Experimental Equipment

Electroencephalography was recorded by an EEG system (g.USBamp and g.GAMMAbox, g.tec, Schiedlberg, Austria) with a sampling frequency of 1,200 Hz. According to the previous studies (Hemptinne et al., 2018; Zheng et al., 2020a), six occipital electrodes (O1, Oz, O2, PO3, POz, and PO4) were used to acquire EEG signals, as shown in **Figure 1**. The ground electrode was placed on the forehead (Fpz), and the reference electrode was placed on the left earlobe (A1). Besides, a notch filter from 48 to 52 Hz was applied to eliminate the power line interference. A 24.5-in LCD monitor (PG258Q, ASUS, Taipei, China) with a resolution of 1,920 × 1,080 pixels, and a refresh rate of 240 Hz was used to present visual stimuli.

Visual Stimuli

In this study, the vertical sinusoidal gratings with a reversal frequency of 7.5 Hz were used as the visual stimuli with the Michelson contrast of 50% and the mean background luminance of 80 cd/m^2 (Kurtenbach et al., 2013; Zheng et al., 2020c). The visual angle of the stimulus pattern with a side length of 720 pixels was set as four degrees by adjusting the distance between the display and subjects. Six spatial frequencies in logarithmically equidistant steps of 3.0, 4.8, 7.5, 12.0, 19.0, and 30.0 cycles per degree (cpd) corresponding to the optotypes of 1.0, 0.8, 0.6, 0.4, 0.2, and 0.0 logMAR were presented to subjects in each run (Zheng et al., 2019). Each run contained six blocks corresponding to six spatial frequency steps. Each block contained five trials, and each trial lasted 5 s with a 2-s interval between two trials. The right eye was tested first and then the left eye. Besides, four subjects accomplished two eyes' experiments, while the others only accomplished the right eye's experiment. The visual stimuli were developed by MATLAB (MathWorks, Natick, MA, United States) using the Psychophysics Toolbox (Brainard, 1997).

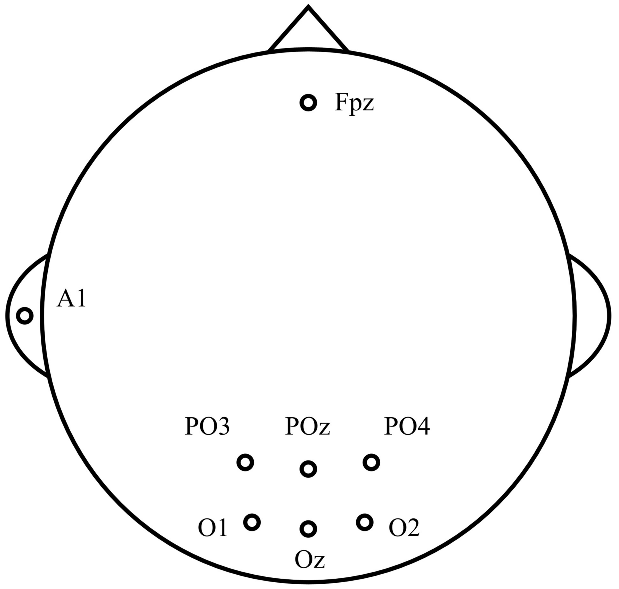


FIGURE 1 | Location of scalp electrodes.

Signal Processing

Data Preprocessing

Following the start and end times of each trial, the SSVEP data segments were extracted. Then, a band-pass filter from 3 to 40 Hz was imposed to exclude the high-frequency interferences and low-frequency drifts. The five data segments of the same spatial frequency corresponding to five trials in one block were averaged to a 5-s data epoch for further data processing.

Spatial Filtering and Feature Extraction

The above 10 spatial filtering methods were used to linearly combine the 5-s six-electrode data epoch into 5-s single-channel signals, respectively. Since there was only one stimulus frequency, i.e., 7.5 Hz, in stimulus presentation, the SSVEP reference signals model $\boldsymbol{X_f} \in \mathbb{R}^{N_t \times 2N_h}$ in this study was defined as:

$$\boldsymbol{X}_f = \begin{pmatrix} \sin\left(2\pi f \dfrac{m}{F_s}\right) \\ \cos\left(2\pi f \dfrac{m}{F_s}\right) \end{pmatrix}^{\mathrm{T}}, \; m = 1, \ldots, N_t \tag{27}$$

where f was set as 7.5 Hz, and the number of harmonic frequencies N_h was set as 1. The number of sampling points, N_t, was 6,000 in a 5-s data segment with a sampling frequency of 1,200 Hz.

Then, the SSVEP feature was extracted by the Fourier transform to obtain the frequency-domain spectrum, and the amplitude at the fundamental reversal frequency of 7.5 Hz was considered as the SSVEP amplitude.

Signal-to-Noise Ratio

The noise was defined by the mean value of the 20 adjacent amplitudes of either side of the fundamental frequency of 7.5 Hz on the frequency-domain spectrum (Bach and Meigen, 1999; Zheng et al., 2020b). Hence, the SNR can be determined by the ratio of SSVEP amplitude at 7.5 Hz to noise:

$$SNR = \frac{\text{SSVEP amplitude}}{\text{noise}} = \frac{a(f)}{\frac{1}{10} * \sum_{k=1}^{k=10} a\left(fk * \triangle f\right) a(f - k * \triangle f)} \tag{28}$$

where $a(f)$ denotes the amplitude on the frequency-domain spectrum at frequency f, and frequency resolution Δf is 0.1 Hz.

Visual Acuity Determination Criterion

Figure 2 shows an example of the tuning curve for the SSVEP visual acuity estimation criterion used in this study. SSVEP amplitude can be plotted versus spatial frequency, and then a regression line can be extrapolated from the last significant SSVEP peak to a noise level baseline (Zheng et al., 2020b). The range for the regression line was between the last significant SSVEP peak and the last data point with an SNR higher than the preset SNR level, and the noise level baseline for each visual stimulus was defined as the mean of the noise of the six spatial frequency steps (Hamilton et al., 2021b). Then, the SSVEP visual acuity was defined as the spatial frequency corresponding to the intersection point between the regression line and the noise level baseline (Zheng et al., 2020b; Hamilton et al., 2021a). Besides, the whole diagram of signal processing in this study is shown in **Figure 3**.

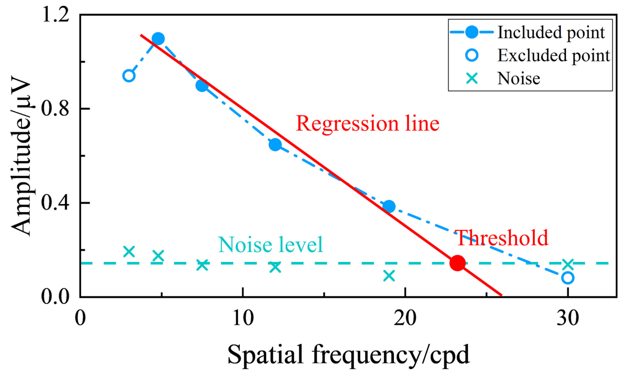


FIGURE 2 | Example of tuning curve for steady-state visual evoked potential (SSVEP) visual acuity estimation criterion. The green " × " represents the noise corresponding to each spatial frequency step, and the green dashed line represents the noise level baseline defined by the mean of the noise of the six spatial frequency steps. The data points included in the linear regression have an signal-to-noise ratio (SNR) higher than the preset SNR level, while the excluded points do not. The red solid line represents the regression line between the SSVEP amplitude and spatial frequency extrapolating from the last significant SSVEP peak to the last data point with an SNR higher than the preset SNR level. The red point is the intersection of the regression line and the noise level baseline, with its corresponding spatial frequency value defined as the visual acuity threshold.

Statistical Analysis

Bland–Altman was used to describe the agreement and difference between the psychophysical FrACT and objective SSVEP visual acuity for each spatial filtering method. Besides, one-way repeated-measures ANOVA was also employed to evaluate the difference among the FrACT and SSVEP visual acuity results for each spatial filtering method, and the *post-hoc* analysis with Bonferroni correction for multiple comparisons was subsequently employed.

RESULTS

Comparison of the SSVEP Signal Characteristics

Figure 4 shows an example of the time-domain, frequency-domain, and time–frequency-domain analyses of SSVEPs after each spatial filtering method. First, the 5-s single-channel SSVEP signals corresponding to each spatial filtering method were obtained according to the abovementioned signal processing flow in **Figure 3**. Then, the time-domain waveforms were obtained by averaging the 0.53-s nonoverlapping data segments subdivided by the 5-s single-channel SSVEP signals, with each segment containing four periods of the reversal process (Zheng et al., 2020a). The frequency-domain spectrums were obtained by the Fourier transform of the 5-s single-channel SSVEP signals. As for the time–frequency-domain analysis, the 2.0-s window length with 0.1-s sliding length over the 5-s single-channel signals was used to obtain the time–frequency-domain characteristics (Zheng et al., 2020a).

The time-domain waveforms in **Figure 4A** show that an obvious main periodicity was the fundamental reversal frequency of 7.5 Hz for all spatial filtering methods except for the two-dimensional Laplacian combination, while some other periodic components also existed in some waveforms, such as the native, bipolar, and one-dimensional Laplacian combination. Both the frequency-domain waveforms in **Figure 4B** and the time–frequency-domain analyses in **Figure 4C** show clear significant peaks at the fundamental reversal frequency of 7.5 Hz and the second harmonic frequency of 15 Hz for all spatial filtering methods except for the two-dimensional Laplacian combination, indicating that all these spatial filtering methods except for the two-dimensional Laplacian combination can obtain obvious signal characteristics by combining the multielectrode signals into single-channel signals.

Comparison of Spatial Filtering Effect

The main purpose of spatial filtering is to strengthen the SSVEP components and suppress the non-SSVEP components in EEG signals (Wong et al., 2020) and thus to enhance the

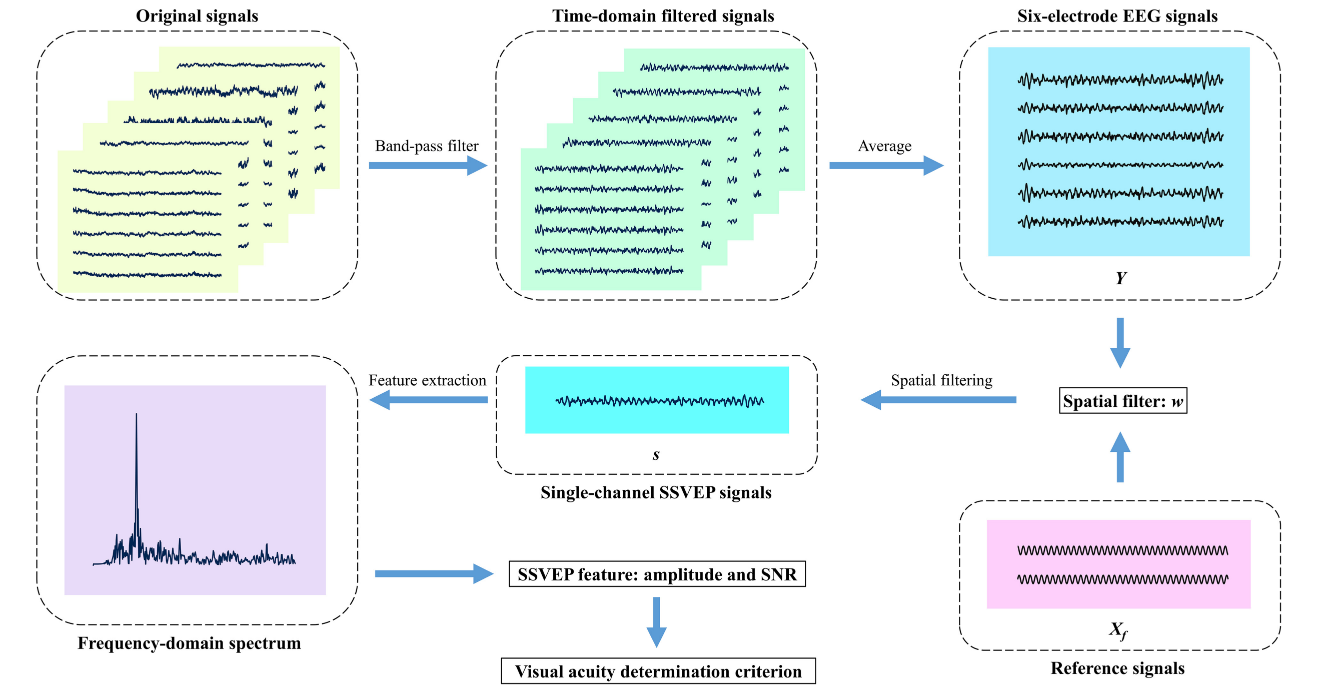

Original signals
Band-pass filter
Time-domain filtered signals
Average
Six-electrode EEG signals
Y
Feature extraction
Spatial filtering
Spatial filter: w
s
Single-channel SSVEP signals
Frequency-domain spectrum
SSVEP feature: amplitude and SNR
Visual acuity determination criterion
X_f
Reference signals

FIGURE 3 | Diagram of signal processing in this study. First, the original signals for five trials in one block are filtered by a band-pass filter from 3 to 40 Hz, and subsequently, the signal segments corresponding to five trials are averaged to a 5-s data epoch for each electrode. Then, each spatial filtering method linearly combines the six-electrode signals into one single-channel signals, respectively. Next, the Fourier transform extracts the steady-state visual evoked potential (SSVEP) amplitude and signal-to-noise ratio (SNR). Finally, the visual acuity determination criterion is carried out after six blocks in one run complete. Besides, the model-based spatial filter is obtained by mathematical transformation of six-electrode EEG signals Y and reference signals X_f.

SNR (Friman et al., 2007). Hence, the spatial filtering effect was evaluated by comparing the SNR values of the single-channel SSVEP signals corresponding to various spatial filtering methods. Since the visual stimuli at the spatial frequency of 3.0 cpd were the clearest to all subjects, the comparison of the SNR values corresponding to various spatial filtering methods at 3.0 cpd over all subjects was obtained, as shown in **Figure 5**. **Figure 5** shows that the SNR values of CCA (4.849 ± 1.101) and MSI (4.115 ± 1.372) were higher than that of the native combination (3.861 ± 1.188), with other spatial filtering methods had lower or close SNR values to that of the native combination. Since the native combination actually utilized only single-electrode signals from Oz and was widely used in SSVEP visual acuity assessment, here, the spatial filtering methods of CCA and MSI were compared to the native combination in the further visual acuity evaluation by SSVEPs.

SSVEP Visual Acuity Threshold Determination Criterion

SSVEP visual acuity was defined by the intersection point between the noise level baseline and the regression line extrapolating from the last significant SSVEP peak to the last data point with an SNR higher than the preset SNR level. For the native combination, previous studies have given the recommended value of SNR level, i.e., 1.0 (Yadav et al., 2009; Zheng et al., 2020b). However, as shown in **Figure 6**, CCA and MSI often obtained the higher SNR of SSVEPs than the native combination, especially in high spatial frequencies close to the visual acuity threshold. Hence, for the spatial filtering methods of CCA and MSI, the SNR level of 1.0 may not be applicable since both CCA and MSI enhanced the SNR of SSVEPs.

Here, first, the five SNR levels, i.e., 1.0, 1.5, 2.0, 2.5, and 3.0 (Zheng et al., 2019), were preselected for CCA and MSI. Then, as shown in **Figure 7**, corresponding to **Figure 6**, the tuning curves of the SSVEP visual acuity estimation criterion for the native combination, CCA, and MSI with various SNR levels of 1.0, 1.5, 2.0, 2.5, and 3.0, respectively, can be obtained. Next, the range for the linear regression of the native combination in **Figure 7A** was from the first data point with the amplitude peak of 1.140 μV to the last data point with an SNR of 1.508 higher than the SNR level of 1.0, and the SSVEP visual acuity for the native combination was determined as the spatial frequency of the intersection point of the regression line and the noise level baseline, i.e., 26.554 cpd. Similar to this, as shown in **Figures 7B,C**, the SSVEP visual acuities for CCA and MSI with various SNR levels were 32.470 cpd for CCA with the SNR levels of 1.0, 1.5, 2.0, and 2.5; 26.097 cpd for CCA with the SNR level of 3.0; 25.237 cpd for MSI with the SNR levels of 1.0, 1.5, 2.0, and 2.5; and 20.892 cpd for MSI with the SNR level of 3.0.

The unit of logMAR was used in the final visual acuity expression for its uniformity in spatial frequency (Bach, 2007). Finally, after SSVEP visual acuities for CCA and MSI at various SNR levels over all subjects were obtained, the Bland–Altman

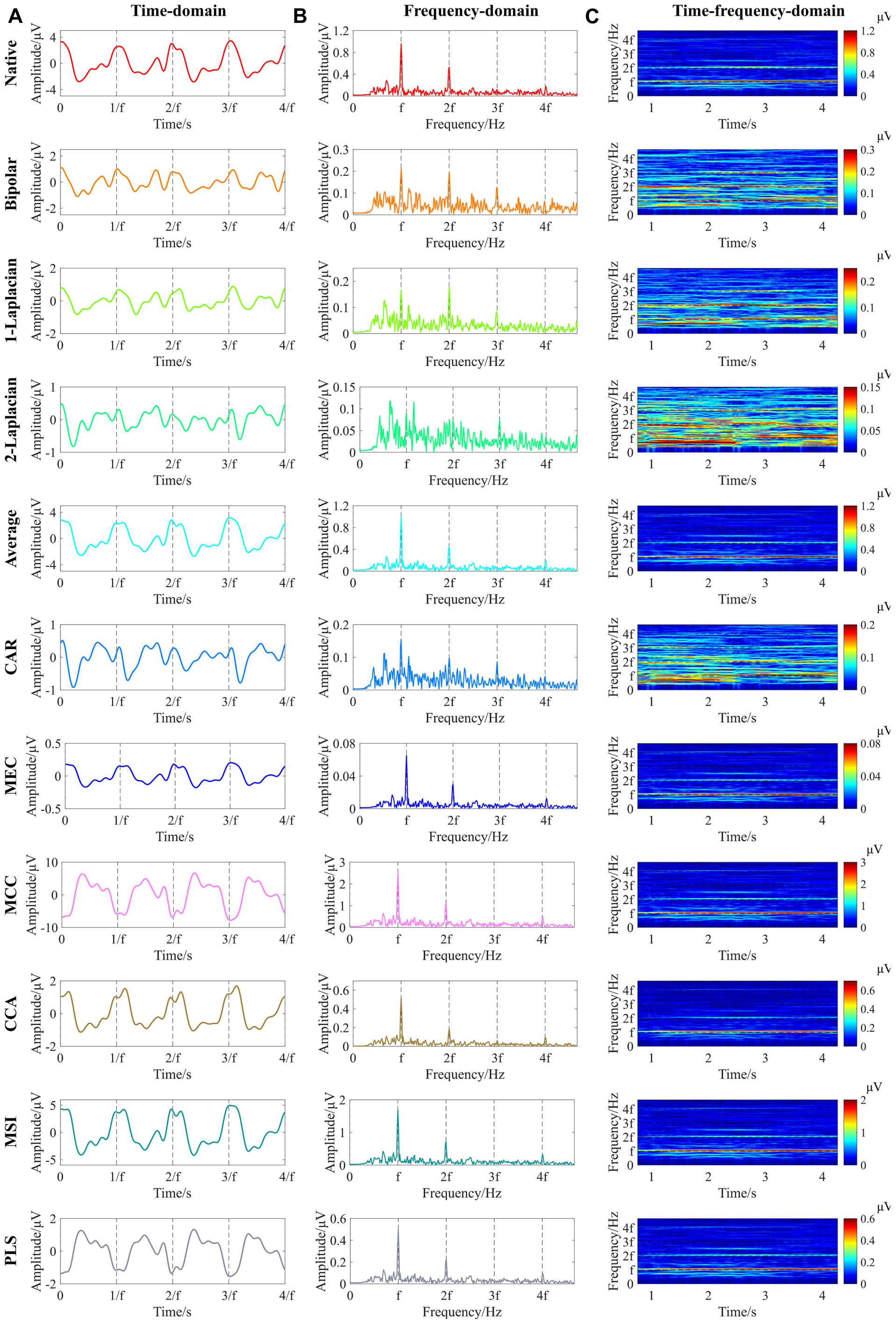


FIGURE 4 | Example of the time-domain, frequency-domain, and time–frequency-domain analyses of steady-state visual evoked potentials (SSVEPs) at 3.0 cpd after various spatial filtering methods (right eye of subject S9, Freiburg Visual Acuity and Contrast Test (FrACT) acuity = 0.00 logMAR). **(A)** Time-domain analysis. The vertical dashed lines correspond to four periods of the reversal process. **(B)** Frequency-domain analysis. **(C)** Time–frequency-domain analysis. The vertical dashed lines in Panel **(B)** and the horizontal dashed lines in Panel **(C)** correspond to the reversal frequency of 7.5 Hz and the second, third, and fourth harmonic frequencies of 15, 22.5, and 30 Hz, respectively. "f" in all subfigures represents the reversal frequency of 7.5 Hz.

analysis was used to analyze the difference and agreement between subjective FrACT visual acuity and objective SSVEP visual acuity for CCA and MSI at each SNR level, as shown in **Table 1**. Hence, the SNR level of 2.0 was chosen for CCA with a low 95% limit of agreement (i.e., 0.202 logMAR) and a low difference (i.e., 0.039 logMAR). Similarly, the SNR level of 1.5

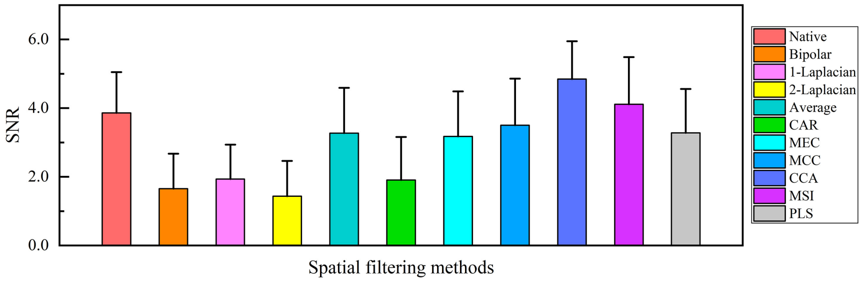


FIGURE 5 | Comparison of the mean values and SD of the signal-to-noise ratio (SNR) of the single-channel steady-state visual evoked potential (SSVEP) signals corresponding to various spatial filtering methods at 3.0 cpd over all subjects.

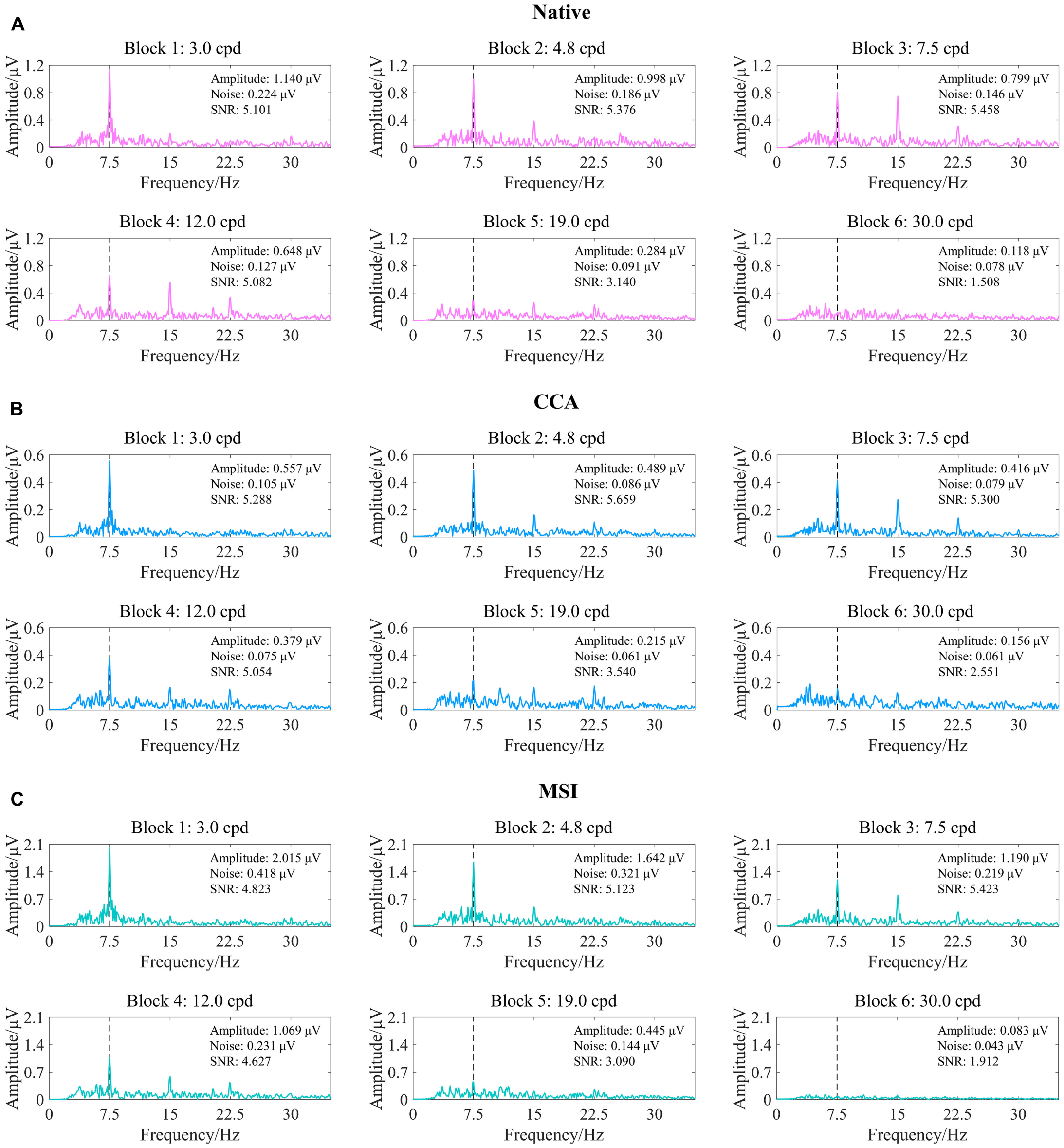


FIGURE 6 | Examples of the steady-state visual evoked potential (SSVEP) response to six spatial frequency steps after three spatial filtering methods of native combination, canonical correlation analysis (CCA), and multivariate synchronization index (MSI) (right eye of subject S2, FrACT acuity = −0.06 logMAR). **(A)** Native combination. **(B)** CCA. **(C)** MSI. The vertical dashed lines correspond to the reversal frequency of 7.5 Hz.

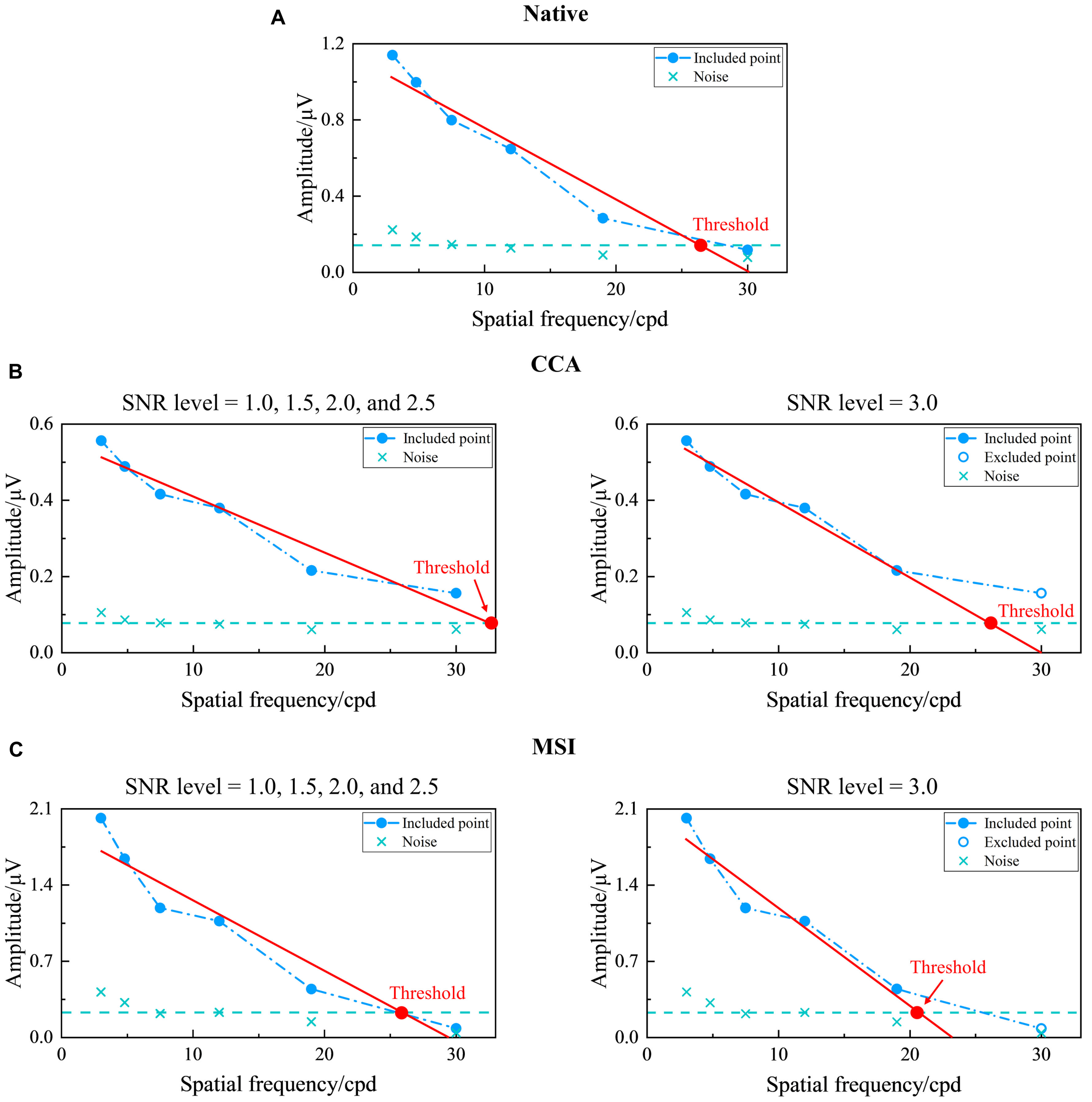


FIGURE 7 | Examples of the tuning curves for steady-state visual evoked potential (SSVEP) visual acuity estimation criterion corresponding to the native combination, canonical correlation analysis (CCA), and multivariate synchronization index (MSI) with various signal-to-noise ratio (SNR) levels of 1.0, 1.5, 2.0, 2.5, and 3.0, respectively (right eye of subject S2, FrACT acuity = −0.06 logMAR). **(A)** Native combination with the SNR level of 1.0. **(B)** CCA with the SNR levels of 1.0, 1.5, 2.0, and 2.5 for the left subpanel and 3.0 for the right subpanel. **(C)** MSI with the SNR levels of 1.0, 1.5, 2.0, and 2.5 for the left subpanel and 3.0 for the right subpanel. The representations of the symbols and lines are the same as in **Figure 2**.

was chosen for MSI with a low 95% limit of agreement (i.e., 0.208 logMAR) and a low difference (i.e., −0.080 logMAR).

Comparison of Visual Acuity Results

Figure 8 shows the Bland–Altman analysis between subjective FrACT visual acuity and final objective SSVEP visual acuity over all subjects for the native combination, CCA, and MSI, respectively. The 95% limits of agreement for the native combination, CCA, and MSI were 0.253 logMAR, 0.202 logMAR, and 0.208 logMAR, respectively, indicating that SSVEP visual acuity of the spatial filtering methods of CCA and MSI had better accuracy than the native combination.

Figure 9 shows the comparison in visual acuity estimated by four methods, i.e., FrACT and SSVEPs for three spatial filtering methods of the native combination, CCA, and MSI, over all subjects. One-way repeated-measures ANOVA found a significant difference in visual acuity among these four methods [$F_{(3,45)} = 10.277, p < 0.001$]. Then, Bonferroni *post-hoc* analysis showed no difference between psychophysical FrACT visual acuity and each SSVEP visual acuity for the native combination, CCA, and MSI ($p > 0.05$), as shown in **Table 2**, demonstrating that the SSVEP visual acuity obtained by these three spatial filtering methods all had a good agreement and a similar performance with subjective FrACT visual acuity. Besides, a

TABLE 1 | Results of Bland–Altman analysis between subjective Freiburg Visual Acuity and Contrast Test (FrACT) visual acuity and objective steady-state visual evoked potential (SSVEP) visual acuity for the native combination at signal-to-noise ratio (SNR) level of 1.0, and canonical correlation analysis (CCA) and multivariate synchronization index (MSI) at each SNR level of 1.0, 1.5, 2.0, 2.5, and 3.0, respectively.

	SNR level	Difference/logMAR	LoA/logMAR
Native	1.0	−0.095	0.253
CCA	1.0	0.057	0.204
	1.5	0.050	0.215
	2.0	0.039	0.202
	2.5	−0.011	0.230
	3.0	−0.065	0.348
MSI	1.0	−0.902	0.254
	1.5	−0.080	0.208
	2.0	−0.158	0.290
	2.5	−0.269	0.304
	3.0	−0.298	0.256

LoA, 95% limit of agreement.

significantly higher SSVEP visual acuity was found in CCA than the native combination ($p = 0.005$) and MSI ($p < 0.001$), indicating CCA had a better performance in combining the multielectrode signals in SSVEPs, especially when the spatial frequency near the psychophysical threshold, causing the higher SNR, i.e., higher SSVEP amplitude and lower noise, as shown in **Figure 5**.

In summary, compared to FrACT visual acuity, SSVEP visual acuity for the native combination, CCA, and MSI all had a good agreement with it, demonstrating that these three spatial filtering methods all had a good performance in SSVEP visual acuity assessment. Besides, CCA-based SSVEP visual acuity had a better performance than MSI and the native combination, with a difference and a limit of agreement of 0.039 logMAR and 0.202 logMAR, respectively, lower than −0.080 logMAR and 0.208 logMAR for MSI and −0.095 logMAR and 0.253 logMAR for the native combination, as shown in **Table 1**. Hence, this study recommended CCA as the spatial filtering method for multielectrode signals combination in the SSVEP visual acuity assessment.

DISCUSSION

In this study, to enhance the performance of visual acuity by SSVEPs, 10 commonly used spatial filtering methods, i.e., native combination, bipolar combination, Laplacian combination, average combination, CAR, MEC, MCC, CCA, MSI, and PLS, were compared to combine multielectrode SSVEP signals into single-channel SSVEP signals for the vertical sinusoidal gratings, finding that the Fourier analysis of SSVEP signals after these 10 spatial filtering methods all had a significant peak at the fundamental reversal frequency, where CCA- and MSI-based SSVEP signals had a higher SNR than the traditional single-electrode from Oz, i.e., the native combination. Then, CCA and MSI were used in the further SSVEP visual acuity evaluation. Compared to the SNR level of 1.0 for the native combination, according to the Bland–Altman analysis, the SNR levels of 2.0 and 1.5 were chosen for CCA and MSI, respectively, to determine the regression range for visual acuity determination criterion. After the calculation of SSVEP visual acuity over all subjects, SSVEP visual acuity for the native combination, CCA, and MSI all had a good agreement with subjective FrACT visual acuity, with CCA-based SSVEP visual acuity realizing the best performance, recommending CCA as the spatial filtering method for multielectrode signals combination in SSVEP visual acuity assessment.

The CCA-based SSVEP visual acuity achieved a difference of 0.039 logMAR and a limit of agreement of 0.202 logMAR from FrACT visual acuity, and that for MSI-based SSVEP visual acuity were −0.080 logMAR and 0.208 logMAR, which was all lower than them of SSVEP visual acuity for the native combination with a difference and a limit of agreement of −0.095 logMAR and 0.253 logMAR. Since the spatial filtering methods can enhance the SNR of SSVEPs and suppress the non-SSVEP noise (Nakanishi et al., 2018b), this result illustrated that the unrelated noise, e.g., EMG and EOG (Friman et al., 2007; Zhang et al., 2021), was one of the reasons for the difference between SSVEP and behavioral visual acuity (Hamilton et al., 2021b), and the other methods of enhancing the SNR, such as signal preprocessing (Kołodziej et al., 2016), e.g., time-domain filtering (Zheng et al., 2021) and blind source separation (BSS) (Ji et al., 2019), and SSVEP recognition algorithms (Zhang et al., 2021), e.g., wavelet transform (WT) (Rejer, 2017) and empirical mode decomposition (EMD) (Huang et al., 2013; Tello et al., 2014), may also have the property to improve the agreement between SSVEP and behavioral visual acuity.

The 10 commonly used spatial filtering methods in this study can be divided into two categories. One is the basic spatial filtering methods canceling the common noise of each electrode via averaging or subtracting (Friman et al., 2007), such as native combination, bipolar combination, Laplacian combination, average combination, and CAR, and the other is called model-based spatial filtering methods using the mathematical transformation between multielectrode SSVEP signals and the SSVEP reference signals l (Zerafa et al., 2018), such as MEC, MCC, CCA, MSI, and PLS. **Figure 5** shows that the model-based spatial filtering methods generally had a better performance than the basic spatial filtering methods in vertical sinusoidal gratings except for the average combination (Friman et al., 2007), and the reason for this may be that the model-based spatial filtering methods can adjust the weight coefficients to each electrode adaptively for various SSVEP signals.

All the spatial filtering methods used in this study were the training-free methods (Wong et al., 2020), which did not require any training data, and a new user can use this brain–computer interface (BCI) system immediately (Zerafa et al., 2018). Because of the fast and accurate requirement and infrequent testing for visual acuity assessment (Zheng et al., 2021), the training-free methods were adequate here. The filter bank strategy in training-free methods, such as filter bank CCA (FBCCA) (Chen et al., 2015) and filter bank MSI (FBMSI) (Qin et al., 2021), may be also used to enhance the performance of SSVEP-based

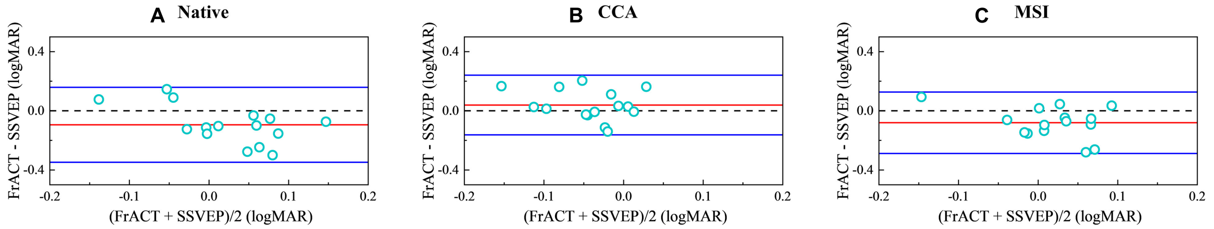


FIGURE 8 | Bland–Altman analysis between psychophysical Freiburg Visual Acuity and Contrast Test (FrACT) visual acuity and objective steady-state visual evoked potential (SSVEP) visual acuity over all subjects for the native combination, canonical correlation analysis (CCA), and multivariate synchronization index (MSI), respectively. **(A)** Native combination. **(B)** CCA. **(C)** MSI. In each panel, the red solid line represents the average value of the difference. The blue solid lines represent the 95% limit of agreement. The dashed line represents the difference of zero.

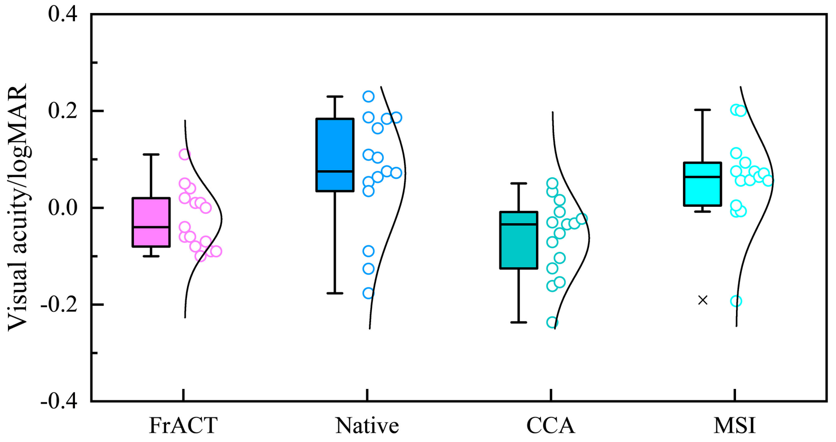


FIGURE 9 | Comparison of the visual acuity assessed by Freiburg Visual Acuity and Contrast Test (FrACT) and steady-state visual evoked potentials (SSVEPs) from three spatial filtering methods of the native combination, canonical correlation analysis (CCA), and multivariate synchronization index (MSI) over all subjects.

visual acuity assessment in future work. In contrast, the subject-specific training methods with the best performance (Zerafa et al., 2018), requiring training data from the specific user and needing the cost of long and tiring training sessions, such as individual template-based CCA (itCCA) (Bin et al., 2011), combined-CCA (Nakanishi et al., 2014; Wang et al., 2014b), multiway CCA (Zhang et al., 2011), multiset CCA (Zhang et al., 2014b), and task-related component analysis (TRCA) (Nakanishi et al., 2018a), may be more suitable for the situation where the subjects need long-term use of BCI system, such as the vision training with SSVEP biofeedback in amblyopia (Lapajne et al., 2020). Besides, the subject-independent training methods requiring training data from various subjects, providing a good trade-off between training effort and performance (Zerafa et al., 2018), such as transfer template CCA (ttCCA) (Yuan et al., 2015) and combined-tCCA (Waytowich et al., 2016), may be further applied in SSVEP visual acuity assessment.

As for the threshold determination criterion in this study, the extrapolation technique by extrapolating a regression line between significant SSVEP amplitudes and spatial frequencies to a noise level baseline was used. Compared to the threshold determination criterion of the finest spatial frequency evoking a significant SSVEP (Hamilton et al., 2021a), where the precision depends on the sampling density of spatial frequency when near the threshold (Hamilton et al., 2021b), this extrapolation technique is more practical (Zheng et al., 2020b). Compared to the other stimulus paradigms, such as concentric rings with oscillating expansion and contraction (Zheng et al., 2019), the visual stimulus paradigm of vertical sinusoidal gratings in this study can easily be realized, as recommended by the International Society for Clinical Electrophysiology of Vision (ISCEV) standard (Hamilton et al., 2021a).

Here, the basic spatial filtering methods used the fixed reference electrode, Oz, for all subjects, but this may not necessarily be the best choice for each subject (Yan et al., 2021), so an adaptive reference electrode selection method may be explored

TABLE 2 | Bonferroni *post hoc* analysis of visual acuity among Freiburg Visual Acuity and Contrast Test (FrACT) and steady-state visual evoked potentials (SSVEPs) from three spatial filtering methods of the native combination, canonical correlation analysis (CCA), and multivariate synchronization index (MSI).

Method	Native	CCA	MSI
FrACT	$p = 0.061$	$p = 0.522$	$p = 0.096$
Native	–	$p = 0.005^{**}$	$p = 1.000$
CCA	–	–	$p < 0.001^{***}$

$^{***}p < 0.001$; $^{**}p < 0.01$.

in future work to improve the performance. In the model-based spatial filtering methods, only the eigenvector corresponding to one extreme value was chosen as spatial filter weights, e.g., the spatial filter weights corresponding to the largest eigenvalue in CCA, and there may be also some more signal information at eigenvectors of the second largest eigenvalue or even the latter eigenvalues (Zhao et al., 2020). Hence, future work may propose more algorithm strategies to make full use of the information from the spatial filtering methods. Finally, some subjects with lower visual acuity rather than the normal visual acuity may be also required for further research.

CONCLUSION

This study introduced the spatial filtering methods in SSVEP-based visual acuity assessment, finding that CCA-based SSVEP visual acuity had a better performance with an agreement of 0.202 logMAR and a difference of 0.039 logMAR, compared to the single electrode and other spatial filtering methods. The study proved that the performance of SSVEP-based visual acuity can be enhanced by spatial filtering methods and also recommended CCA as the spatial filtering method for multielectrode signals combination in the SSVEP visual acuity assessment.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Human Ethics Committee of Xi'an Jiaotong University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

XZ contributed to the study design, data acquisition, analysis, interpretation, manuscript writing, and revision. GX contributed to the study design and the approval of the final version for publication. CH and PT contributed to the statistical analysis and manuscript drafting. KZ and RL contributed to the data analysis and interpretation. YJ and WY contributed to the manuscript writing and revision. CD provided the experimental equipment and approved the final version for publication. SZ conceptualized the study. All authors contributed to the article and approved the submitted version.

FUNDING

This research was supported by the Special Guidance Funds for the Construction of World-Class Universities (Disciplines) and Characteristic Development in Central Universities (PY3A071), the Science and Technology Plan Project of Xi'an (20KYPT0001-10), and the Xi'an Key Laboratory of Brain–Computer Interaction and Neuroregulation Digital Medicine.

REFERENCES

Bach, M. (1996). The Freiburg Visual Acuity test–automatic measurement of visual acuity. *Optom. Vis. Sci.* 73, 49–53. doi: 10.1097/00006324-199601000-00008

Bach, M. (2007). The Freiburg Visual Acuity Test-variability unchanged by post-hoc re-analysis. *Graefes Arch. Clin. Exp. Ophthalmol.* 245, 965–971. doi: 10.1007/s00417-006-0474-4

Bach, M., and Heinrich, S. P. (2019). Acuity VEP: improved with machine learning. *Doc. Ophthalmol.* 139, 113–122. doi: 10.1007/s10633-019-09701-x

Bach, M., and Meigen, T. (1999). Do's and don'ts in Fourier analysis of steady-state potentials. *Doc. Ophthalmol.* 99, 69–82. doi: 10.1023/a:1002648202420

Bin, G., Gao, X., Wang, Y., Li, Y., Hong, B., and Gao, S. (2011). A high-speed BCI based on code modulation VEP. *J. Neural. Eng.* 8:025015. doi: 10.1088/1741-2560/8/2/025015

Bin, G., Gao, X., Yan, Z., Hong, B., and Gao, S. (2009). An online multi-channel SSVEP-based brain-computer interface using a canonical correlation analysis method. *J. Neural. Eng.* 6:046002. doi: 10.1088/1741-2560/6/4/046002

Brainard, D. H. (1997). The psychophysics toolbox. *Spat. Vis.* 10, 433–436. doi: 10.1163/156856897x00357

Chen, X., Wang, Y., Gao, S., Jung, T. P., and Gao, X. (2015). Filter bank canonical correlation analysis for implementing a high-speed SSVEP-based brain-computer interface. *J. Neural. Eng.* 12:046008. doi: 10.1088/1741-2560/12/4/046008

Friman, O., Volosyak, I., and Graser, A. (2007). Multiple channel detection of steady-state visual evoked potentials for brain-computer interfaces. *IEEE Trans. Biomed. Eng.* 54, 742–750. doi: 10.1109/TBME.2006.889160

Ge, S., Wang, R., Leng, Y., Wang, H., Lin, P., and Iramina, K. (2017). A Double-Partial Least-Squares Model for the Detection of Steady-State Visual Evoked Potentials. *IEEE J. Biomed. Health Inform.* 21, 897–903. doi: 10.1109/JBHI.2016.2546311

Hamilton, R., Bach, M., Heinrich, S. P., Hoffmann, M. B., Odom, J. V., McCulloch, D. L., et al. (2021a). ISCEV extended protocol for VEP methods of estimation of visual acuity. *Doc. Ophthalmol.* 142, 17–24. doi: 10.1007/s10633-020-09780-1

Hamilton, R., Bach, M., Heinrich, S. P., Hoffmann, M. B., Odom, J. V., McCulloch, D. L., et al. (2021b). VEP estimation of visual acuity: a systematic review. *Doc. Ophthalmol.* 142, 25–74. doi: 10.1007/s10633-020-09770-3

Hamilton, R., Bradnam, M. S., Dutton, G. N., Lai Chooi Yan, A. L., Lavy, T. E., Livingstone, I., et al. (2013). Sensitivity and specificity of the step VEP in suspected functional visual acuity loss. *Doc. Ophthalmol.* 126, 99–104. doi: 10.1007/s10633-012-9362-x

Hemptinne, C., Liu-Shuang, J., Yuksel, D., and Rossion, B. (2018). Rapid objective assessment of contrast sensitivity and visual acuity with sweep visual evoked potentials and an extended electrode array. *Invest. Ophthalmol. Vis. Sci.* 59, 1144–1157. doi: 10.1167/iovs.17-23248

Huang, L., Huang, X., Wang, Y., Wang, Y., Jung, T., and Cheng, C. (2013). "Empirical mode decomposition improves detection of SSVEP," in *Proceedings of the 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Osaka), 3901–3904.

Incesu, A. I., and Sobaci, G. (2011). Malingering or simulation in ophthalmology-visual acuity. *Int. J. Ophthalmol.* 4, 558–566. doi: 10.3980/j.issn.2222-3959.2011.05.19

Ji, H., Chen, B., Petro, N. M., Yuan, Z., Zheng, N., and Keil, A. (2019). Functional source separation for EEG-fMRI Fusion: application to steady-state visual evoked potentials. *Front. Neurorobot.* 13:24. doi: 10.3389/fnbot.2019.00024

Knotzele, J., and Heinrich, S. P. (2019). Can VEP-based acuity estimates in one eye be improved by applying knowledge from the other eye? *Doc. Ophthalmol.* 139, 161–168. doi: 10.1007/s10633-019-09700-y

Kołodziej, M., Majkowski, A., Lukasz, O., and Rak, R. J. (2016). "Comparison of EEG signal preprocessing methods for SSVEP recognition," in *Proceedings* *of the 2016 39th International Conference on Telecommunications and Signal Processing (TSP)* (Vienna), 340–345.

Kurtenbach, A., Langrova, H., Messias, A., Zrenner, E., and Jagle, H. (2013). A comparison of the performance of three visual evoked potential-based methods to estimate visual acuity. *Doc. Ophthalmol.* 126, 45–56. doi: 10.1007/s10633-012-9359-5

Lapajne, L., Roskar, S., Tekavcic Pompe, M., Svetina, M., Jarc-Vidmar, M., and Hawlina, M. (2020). Vision training with VEP biofeedback in amblyopia after the critical period. *Doc. Ophthalmol.* 141, 269–278. doi: 10.1007/s10633-020-09774-z

Li, M., He, D., Li, C., and Qi, S. (2021). Brain-Computer interface speller based on steady-state visual evoked potential: a review focusing on the stimulus paradigm and performance. *Brain Sci.* 11:450. doi: 10.3390/brainsci11040450

Lin, Z., Zhang, C., Wu, W., and Gao, X. (2007). Frequency recognition based on canonical correlation analysis for SSVEP-based BCIs. *IEEE Trans. Biomed Eng.* 54, 1172–1176. doi: 10.1109/tbme.2006.889197

McBain, V. A., Robson, A. G., Hogg, C. R., and Holder, G. E. (2007). Assessment of patients with suspected non-organic visual loss using pattern appearance visual evoked potentials. *Graefes Arch. Clin. Exp. Ophthalmol.* 245, 502–510. doi: 10.1007/s00417-006-0431-2

Nakanishi, M., Wang, Y., Chen, X., Wang, Y. T., Gao, X., and Jung, T. P. (2018a). Enhancing Detection of SSVEPs for a high-speed brain speller using task-related component analysis. *IEEE Trans. Biomed Eng.* 65, 104–112. doi: 10.1109/TBME.2017.2694818

Nakanishi, M., Wang, Y., Jung, T.-P., Tanaka, T., and Arvaneh, M. (2018b). "Spatial filtering techniques for improving individual template-based SSVEP detection," in *Signal Processing and Machine Learning for Brain-Machine Interface*, eds T. Tanaka and M. Arvaneh (London: IET), 219–242.

Nakanishi, M., Wang, Y., Wang, Y. T., Mitsukura, Y., and Jung, T. P. (2014). A high-speed brain speller using steady-state visual evoked potentials. *Int. J. Neural. Syst.* 24:1450019. doi: 10.1142/S0129065714500191

Norcia, A. M., Appelbaum, L. G., Ales, J. M., Cottereau, B. R., and Rossion, B. (2015). The steady-state visual evoked potential in vision research: a review. *J. Vis.* 15:4. doi: 10.1167/15.6.4

Norcia, A. M., and Tyler, C. W. (1985a). Infant VEP acuity measurements: analysis of individual differences and measurement error. *Electroencephalogr. Clin. Neurophysiol.* 61, 359–369. doi: 10.1016/0013-4694(85)91026-0

Norcia, A. M., and Tyler, C. W. (1985b). Spatial frequency sweep VEP: visual acuity during the first year of life. *Vis. Res.* 25, 1399–1408. doi: 10.1016/0042-6989(85)90217-2

Odom, J. V., Bach, M., Brigell, M., Holder, G. E., McCulloch, D. L., Mizota, A., et al. (2016). ISCEV standard for clinical visual evoked potentials: (2016 update). *Doc. Ophthalmol* 133, 1–9. doi: 10.1007/s10633-016-9553-y

Onton, J., Westerfield, M., Townsend, J., and Makeig, S. (2006). Imaging human EEG dynamics using independent component analysis. *Neurosci. Biobehav. Rev.* 30, 808–822. doi: 10.1016/j.neubiorev.2006.06.007

Qin, K., Wang, R., and Zhang, Y. (2021). Filter bank-driven multivariate synchronization index for training-free SSVEP BCI. *IEEE Trans. Neural. Syst. Rehabil. Eng.* 29, 934–943. doi: 10.1109/TNSRE.2021.3073165

Regan, D. (1973). Rapid objective refraction using evoked brain potentials. *Invest. Ophthalmol.* 12, 669–679.

Rejer, I. (2017). "Wavelet transform in detection of the subject specific frequencies for SSVEP-Based BCI," in *Hard and Soft Computing for Artificial Intelligence, Multimedia and Security*, eds A. Piegat, I. El Fray, J. Kacprzyk, J. Pejaś, and S.-y Kobayashi (Cham: Springer International Publishing), 146–155.

Ricci, F., Cedrone, C., and Cerulli, L. (1998). Standardized measurement of visual acuity. *Ophthalmic Epidemiol.* 5, 41–53. doi: 10.1076/opep.5.1.41.1499

Ridder, W. H. III (2019). A comparison of contrast sensitivity and sweep visual evoked potential (sVEP) acuity estimates in normal humans. *Doc. Ophthalmol.* 139, 207–219. doi: 10.1007/s10633-019-09712-8

Rosipal, R., and Krämer, N. (2006). "Overview and recent advances in partial least squares," in *Subspace, Latent Structure and Feature Selection. SLSFS 2005. Lecture Notes in Computer Science*, Vol. 3940, eds C. Saunders, M. Grobelnik, S. Gunn, and J. Shawe-Taylor (Berlin: Springer Berlin Heidelberg), 34–51.

Skoczenski, A. M., and Norcia, A. M. (1999). Development of VEP Vernier acuity and grating acuity in human infants. *Invest. Ophthalmol. Vis. Sci.* 40, 2411–2417.

Tello, R. M. G., Müller, S. M. T., Bastos-Filho, T., and Ferreira, A. (2014). "Comparison of new techniques based on EMD for control of a SSVEP-BCI," in *Proceedings of the 2014 IEEE 23rd International Symposium on Industrial Electronics (ISIE)* (Istanbul), 992–997.

Trejo, L. J., Rosipal, R., and Matthews, B. (2006). Brain-computer interfaces for 1-D and 2-D cursor control: designs using volitional control of the EEG spectrum or steady-state visual evoked potentials. *IEEE Trans. Neural. Syst. Rehabil. Eng.* 14, 225–229. doi: 10.1109/TNSRE.2006.875578

Wang, R., Leng, Y., Yang, Y., Wu, W., Iramina, K., and Ge, S. (2014a). "A partial least squares-based stimulus frequency recognition model for steady-state visual evoked potentials detection," in *Proceedings of the 2014 7th International Conference on Biomedical Engineering and Informatics* (Dalian), 699–703.

Wang, Y., Nakanishi, M., Wang, Y., and Jung, T. (2014b). "Enhancing detection of steady-state visual evoked potentials using individual training data," in *Proceedings of the 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, (Chicago, IL), 3037–3040.

Waytowich, N. R., Faller, J., Garcia, J. O., Vettel, J. M., and Sajda, P. (2016). "Unsupervised adaptive transfer learning for Steady-State Visual Evoked Potential brain-computer interfaces," in *Proceedings of the 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (Budapest), 004135–004140.

Wong, C. M., Wang, B., Wang, Z., Lao, K. F., Rosa, A., and Wan, F. (2020). Spatial Filtering in SSVEP-Based BCIs: unified framework and new improvements. *IEEE Trans. Biomed Eng.* 67, 3057–3072. doi: 10.1109/TBME.2020.2975552

Yadav, N. K., Almoqbel, F., Head, L., Irving, E. L., and Leat, S. J. (2009). Threshold determination in sweep VEP and the effects of criterion. *Doc. Ophthalmol.* 119, 109–121. doi: 10.1007/s10633-009-9177-6

Yan, W., Du, C., Luo, D., Wu, Y., Duan, N., Zheng, X., et al. (2021). Enhancing detection of steady-state visual evoked potentials using channel ensemble method. *J. Neural. Eng.* 18:046008. doi: 10.1088/1741-2552/abe7cf

Yan, W., Xu, G., Chen, L., and Zheng, X. (2019). Steady-State motion visual evoked potential (SSMVEP) enhancement method based on time-frequency image fusion. *Comput. Intell. Neurosci.* 2019:9439407. doi: 10.1155/2019/9439407

Yan, W., Xu, G., Xie, J., Li, M., and Dan, Z. (2018). Four novel motion paradigms based on steady-state motion visual evoked potential. *IEEE Trans. Biomed Eng.* 65, 1696–1704. doi: 10.1109/TBME.2017.2762690

Yuan, P., Chen, X., Wang, Y., Gao, X., and Gao, S. (2015). Enhancing performances of SSVEP-based brain-computer interfaces via exploiting inter-subject information. *J. Neural. Eng.* 12:046006. doi: 10.1088/1741-2560/12/4/046006

Zerafa, R., Camilleri, T., Falzon, O., and Camilleri, K. P. (2018). To train or not to train? A survey on training of feature extraction methods for SSVEP-based BCIs. *J. Neural. Eng.* 15:051001. doi: 10.1088/1741-2552/aaca6e

Zhang, Y., Xie, S. Q., Wang, H., and Zhang, Z. (2021). Data analytics in steady-state visual evoked potential-based brain–computer interface: a review. *IEEE Sensors J.* 21, 1124–1138. doi: 10.1109/jsen.2020.3017491

Zhang, Y., Xu, P., Cheng, K., and Yao, D. (2014a). Multivariate synchronization index for frequency recognition of SSVEP-based brain-computer interface. *J. Neurosci. Methods* 221, 32–40. doi: 10.1016/j.jneumeth.2013.07.018

Zhang, Y., Zhou, G., Jin, J., Wang, X., and Cichocki, A. (2014b). Frequency recognition in SSVEP-based BCI using multiset canonical correlation analysis. *Int. J. Neural. Syst.* 24:1450013. doi: 10.1142/S0129065714500130

Zhang, Y., Zhou, G., Zhao, Q., Onishi, A., Jin, J., Wang, X., et al. (2011). *Multiway Canonical Correlation Analysis for Frequency Components Recognition in SSVEP-Based BCIs. Neural Information Processing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 287–295.

Zhao, J., Zhang, W., Wang, J. H., Li, W., Lei, C., Chen, G., et al. (2020). Decision-making selector (DMS) for integrating CCA-Based methods to improve performance of SSVEP-based BCIs. *IEEE Trans. Neural. Syst. Rehabil. Eng.* 28, 1128–1137. doi: 10.1109/TNSRE.2020.2983275

Zheng, X., Xu, G., Du, C., Yan, W., Tian, P., Zhang, K., et al. (2021). Real-time, precise, rapid and objective visual acuity assessment by self-adaptive step SSVEPs. *J. Neural. Eng.* 18, 046047. doi: 10.1088/1741-2552/abfaab

Zheng, X., Xu, G., Wang, Y., Han, C., Du, C., Yan, W., et al. (2019). Objective and quantitative assessment of visual acuity and contrast sensitivity based on steady-state motion visual evoked potentials using concentric-ring paradigm. *Doc. Ophthalmol.* 139, 123–136. doi: 10.1007/s10633-019-09702-w

Zheng, X., Xu, G., Wu, Y., Wang, Y., Du, C., Wu, Y., et al. (2020a). Comparison of the performance of six stimulus paradigms in visual acuity assessment based on steady-state visual evoked potentials. *Doc. Ophthalmol.* 141, 237–251. doi: 10.1007/s10633-020-09768-x

Zheng, X., Xu, G., Yan, W., Liang, R., Zhang, K., Tian, P., et al. (2020b). Threshold determination criterion in steady-state visual evoked potential-based acuity assessment: a comparison of four common methods. *IEEE Access.* 8, 188844–188852. doi: 10.1109/Access.2020.3032129

Zheng, X., Xu, G., Zhang, K., Liang, R., Yan, W., Tian, P., et al. (2020c). Assessment of human visual acuity using visual evoked potential: a review. *Sensors (Basel)* 20:5542. doi: 10.3390/s20195542

Zheng, X., Xu, G., Zhang, Y., Liang, R., Zhang, K., Du, Y., et al. (2020d). Anti-fatigue Performance in SSVEP-based visual acuity assessment: a comparison of six stimulus paradigms. *Front. Hum. Neurosci.* 14:301. doi: 10.3389/fnhum.2020.00301

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 27 August 2021
doi: 10.3389/fnins.2021.696853

Deep Feature Extraction for Resting-State Functional MRI by Self-Supervised Learning and Application to Schizophrenia Diagnosis

*Yuki Hashimoto[1], Yousuke Ogata[2], Manabu Honda[1] and Yuichi Yamashita[1]**

[1] Department of Information Medicine, National Center of Neurology and Psychiatry, National Institute of Neuroscience, Kodaira, Japan, [2] Institute of Innovative Research, Tokyo Institute of Technology, Yokohama, Japan

In this study, we propose a deep-learning technique for functional MRI analysis. We introduced a novel self-supervised learning scheme that is particularly useful for functional MRI wherein the subject identity is used as the teacher signal of a neural network. The neural network is trained solely based on functional MRI-scans, and the training does not require any explicit labels. The proposed method demonstrated that each temporal volume of resting state functional MRI contains enough information to identify the subject. The network learned a feature space in which the features were clustered per subject for the test data as well as for the training data; this is unlike the features extracted by conventional methods including region of interests (ROIs) pooling signals and principal component analysis. In addition, applying a simple linear classifier to the per-subject mean of the features (namely "identity feature"), we demonstrated that the extracted features could contribute to schizophrenia diagnosis. The classification accuracy of our identity features was comparable to that of the conventional functional connectivity. Our results suggested that our proposed training scheme of the neural network captured brain functioning related to the diagnosis of psychiatric disorders as well as the identity of the subject. Our results together highlight the validity of our proposed technique as a design for self-supervised learning.

Keywords: deep-learning, functional MRI, neural network, feature extraction, psychiatric diagnosis, self-supervised learning

OPEN ACCESS

Edited by:
Hamid R. Rabiee,
Sharif University of Technology, Iran

Reviewed by:
Sadia Shakil,
Institute of Space Technology, Pakistan
Katharina Glomb,
Lausanne University Hospital (CHUV), Switzerland

****Correspondence:***
Yuichi Yamashita
yamay@ncnp.go.jp

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *18 April 2021*
Accepted: *20 July 2021*
Published: *27 August 2021*

Citation:
Hashimoto Y, Ogata Y, Honda M and Yamashita Y (2021) Deep Feature Extraction for Resting-State Functional MRI by Self-Supervised Learning and Application to Schizophrenia Diagnosis. Front. Neurosci. 15:696853. doi: 10.3389/fnins.2021.696853

INTRODUCTION

In this study, we propose a novel deep-learning technique which extracts a feature from brain functional magnetic resonance images (fMRIs). Our proposed method solely depends on MRI-scans and does not require any additional data regarding the subjects (e.g., diseases or cognitive impairments), whereas the extracted features effectively capture the psychopathological characteristics of the subjects. Recent advances in machine learning have demonstrated its capability for medical sciences. Skin cancers have been successfully diagnosed from skin images (Esteva et al., 2017) and retinal diseases from three-dimensional optical coherence tomography

(OCT) images (De Fauw et al., 2018). In addition, Titano et al. (2018) reported that machine learning with three-dimensional brain computed tomography (CT) images performed well in terms of detection of acute neurologic events including stroke, hemorrhage, and hydrocephalus. These studies suggested the further potential of the deep neural networks, especially for the analysis of spatially structured data, including MRIs and functional MRIs. These studies trained a neural network to directly infer diseases from the input. This framework is called fully supervised learning and is known to be effective when a large training dataset with accurate labels is available. Titano et al. (2018), who aimed to classify acute neurological events, collected 37,236 brain images with clinical annotations for training and used 96,303 extra clinical reports to make the clinical annotations more suitable for training. In the supervised learning framework, the network is specialized for the target diseases, which further enhances the performance. However, the requirement of a vast amount of training data is not always practical; the number of patients is sometimes too small to train a neural network (Durstewitz et al., 2019; Khosla et al., 2019), and the accurate diagnoses require expert skills (Durstewitz et al., 2019). These drawbacks are remarkable especially for psychiatric disorders because the sample size tends to be small, accurate diagnoses are especially difficult, and the underlying mechanisms are still under discussion. In contrast, self-supervised learning does not require any explicit labels for training. Instead, the teacher signals (i.e., labels) are generated from the original input data in self-supervised learning. For example, Noroozi and Favaro (2016) proposed a self-supervised learning scheme for natural image processing, in which the input image was divided into nine pieces, and the network was trained to infer the original position of each piece. The intermediate outputs of the network were subsequently fed into another linear classifier, which resulted in comparable performance to fully supervised deep neural networks. The advantages of self-supervised training potentially overcome the shortages of clean labels for psychiatric disorders, although the teacher signal must be carefully designed. In many previous deep-learning studies for MRIs without additional labels (Suk et al., 2016; Aghdam et al., 2017; Heinsfeld et al., 2018; Oh et al., 2019; Yamaguchi et al., 2021), the teacher signal was the same as the input, namely auto-encoder. Such an auto-encoder tends to suffer from the bias-variance trade-off, wherein the network either underfits or overfits the teacher signals due to a lack of constraints to the feature manifold. In contrast, this study proposes a novel self-generated teacher signal for resting-state functional MRI; we used the temporal volumes as input, and the subject ID as the teacher signal. The explicit labels enable the network to generate a compact feature that represents a conceptual distance from the owner of the input to the subjects used in the training. In this study, we experimentally showed that: (i) each temporal volume of functional MRI contains enough information to identify the subject, (ii) the network learned a feature space in which the features cluster subject-by-subject for test data as well as for training data, and (iii) the extracted feature contributes to a schizophrenia diagnosis. These experiments together exhibit the validity of our proposed method as a design for self-supervised learning.

MATERIALS AND METHODS

Dataset

We used a dataset from the Center for Biomedical Research Excellence (COBRE) (Aine et al., 2012). The dataset is composed of anatomical and resting-state functional MRI scans; 72 scans were from schizophrenia patients and 75 from healthy controls. The anatomical and functional scans were acquired by MPRAGE and EPI by 3.0-Tesla Siemens Trio scanner (Siemens Healthineers, Erlangen, Germany). Each functional scan was composed of 150 timepoints, and the repetition time was 2 s. Each timepoint was originally composed of 64 × 64 × 32 voxels ($3 \times 3 \times 4$ mm^3), which was transformed to 91 × 109 × 91 voxels in MNI coordinates by the preprocessing (**Supplementary Section 1**). We excluded subjects without meta-data and controls with other psychiatric diseases, resulting in 69 patients (56 males and 13 females, 37.8 ± 14.0 years old) and 72 controls (51 males and 21 females, 35.9 ± 11.7 years old). We divided the patients and controls into training 1, training 2, and test dataset with random sampling stratified over present illness, age, and gender. The training 1 dataset was used for training the neural network, and training 2 was used for training the linear regressor for inferring the subject attributes. The number of patients p and controls c was $(p,\ c) = (51,\ 54)$ in training 1, $(9,\ 9)$ in training 2, and $(9,\ 9)$ in test datasets. The mean and standard deviation of the ages were 37.0 ± 13.4 in training 1, 36.4 ± 11.6 in training 2, and 36.3 ± 11.9 in test datasets. The number of males m and females f was $(m,\ f) = (78,\ 28)$ in training 1, $(15,\ 3)$ in training 2, $(14,\ 4)$ in test datasets. We unequally allocated samples to the three datasets because the neural network in training 1 possessed a huge number of optimization parameters (about 2 million), while the linear regressor/classifier used in training 2 has a relatively small number of optimization parameters (3–10,000).

Training 1

The input of the network was a batch of temporal MRI volumes, whose size was set to (80, 96, 80) by trimming outside of the brain. The network included four convolutional blocks, followed by two convolutional layers and one dense layer. Each block consisted of two three-dimensional convolutions and one average pooling layer (**Figure 1**). The number of convolutional blocks was preliminarily explored. The less number of convolution blocks resulted in underfitting, wherein the training accuracy was almost the same as the chance, while the training did not converge for the network with more convolution blocks.

The kernel size k and stride s were $(k,\ s) = (3,\ 1)$ for each convolutional layer, and $(2,\ 2)$ for each pooling layer. The number of output channels was set as 8 at the first convolutional layer and doubled before the pooling layers, resulting in 128 before the dense layer. We used softmax cross-entropy as the loss function, which was computed against a 105-dimensional one-hot vector of subject ID. The network was optimized using Adam (Kingma and Ba, 2015) with $\alpha = 0.0001$ for the first 17,000 iterations and $\alpha = 0.00001$ for the following 110,000 iterations, with a batch size of 32.

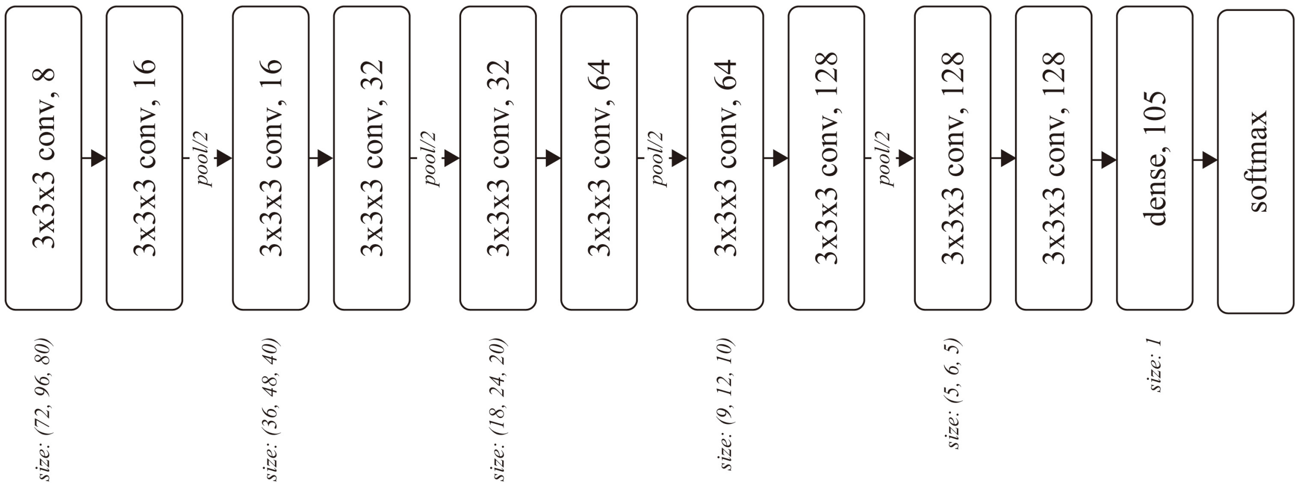


FIGURE 1 | Network architecture. Each rounded square represents a layer with the weight parameters. The number after the comma denotes the number of channels for the layer.

Training 2

The output of the dense layer was extracted for each timepoint as a feature vector. Subsequently, the feature vectors were averaged for each subject, yielding an identity feature for each subject. The identity features for the training 2 dataset were then fed into a linear classifier (regressor) to learn schizophrenia diagnosis and age regression.

We also trained a linear classifier with a slightly modified version of the feature vector, in which the average of all elements in the feature vector was subtracted from each element. This operation was naturally introduced by the formulation of the softmax function, in which the subtraction of the average does not affect the output of the function or the training process. In the following sections, we call the original feature vector (the output of the dense layer) as *classification*, and the modified one as *classification+*.

Experiment 1: Training Convergence

The training accuracy of the subject classification was computed to evaluate training. Reporting training accuracy is slightly unconventional in studies on neural networks because the convergence of training is now trivial in conventional two-dimensional natural scene image processing. However, to the best of our knowledge, this is the first report which trained networks to classify the subject from a single timepoint of functional MRI by stacked three-dimensional convolutions, and we concluded that the training convergence is worth reporting.

Experiment 2: Qualitative Analysis of Extracted Features

The characteristics of the acquired feature space were first qualitatively analyzed. We plotted the feature vectors in the training 2 and test datasets by t-distributed stochastic neighbor embedding (t-SNE) (Maaten and Hinton, 2008). The clusters were then quantitatively evaluated by precision@150 for each identity feature. Because the number of timepoints was 150 for each subject, precision@150 would be 1 if all the feature vectors for a subject clustered around his identity feature. The formula of precision@150 is given in **Supplementary Section 2**, wherein the feature vectors are ranked by Euclid distance to each identity feature in the original feature space. We applied these qualitative and quantitative analyses to the features of the *classification* and *classification+* feature vectors as well as the signals averaged over the ROI defined by the automated anatomical labeling (AAL) atlas (Suk et al., 2016; see Tzourio-Mazoyer et al., 2002 as the reference to AAL), namely ROI-pooled signals, and the top three and 10,000 principal components (Damaraju et al., 2014). For these features, the "identity" feature was also defined as the centroid for each subject.

We explored all the principal components up to 10,000 where the rational upper limit of computational resources for experiments 2 and 3. The results demonstrated that the number of principal components did not affect precision@150 in experiment 2 and the statistical significance of schizophrenia diagnosis in experiment 3, while the number affected the age regression performance in experiment 3 wherein the top three principal components performed the best and showed statistical significance in correlation between predicted and actual age. Therefore, we reported the results of the best (top three) and the maximum (10,000) principal components for experiment 2 and 3.

Experiment 3: Relation to Subject's Attributes

The schizophrenia classifier and age regressor developed in section "Training 2" were applied to the test dataset. The classification accuracy was computed and tested using a sign test, which evaluates a probability parameter of the binomial distribution underlying that the classification is significantly larger than chance (50%). This procedure was also applied to the identity features of the ROI-pooled signals, the top three and 10,000 principal components, similar to that in Experiment 2. In addition, the procedure was applied to the functional connectivity matrix (Liang et al., 2006; Kim et al., 2016), defined as the correlation coefficients among the time series of the ROI-pooled signals.

Ethics Statement

All experiments in this study were performed in accordance with the Ethical Guidelines for Medical and Health Research Involving Human Subjects in Japan.

RESULTS

Experiment 1: Training Convergence

The network was trained to classify the subject ID from each time point of fMRI. The training accuracy at the 127,000 iteration was 97.85%, which was considerably improved over the chance rate, suggesting that the training successfully converged.

Experiment 2: Qualitative Analysis of Extracted Features

The distributions of the feature vectors extracted by our proposed neural network, ROI-pooling, and PCA were visualized by t-SNE, and are depicted in **Figure 2A**. It should be noted that t-SNE preserves local adjacence well but it does not necessarily retain the global structure.

The features extracted by the network clustered for each subject, unlike the features extracted by ROI-pooling and PCA. The clustering performance was quantitatively evaluated using precision@150 around the identity feature for each subject. The precision@150 was 81.5 and 61.4% for our proposed *classification* and *classification*+ feature vectors, whereas it was 5.6% for the ROI-pooled feature and the top three and 10,000 principal components (**Figure 2B**). The precision@150 for each subject is shown in **Supplementary Section 3**.

Experiment 3: Schizophrenia Diagnosis

The average of the features was computed as the identity feature for each subject, and the identity features were fed into a linear classifier for schizophrenia diagnosis with a logistic loss function. The accuracies were 61.1 and 77.8% for the identity feature of our proposed *classification* and *classification*+ feature vectors, respectively. The performance of *classification*+ was significantly better than the chance ($p = 0.015$). The accuracy was 72.2% for the connectivity matrix, which was marginally higher above the chance ($p = 0.048$). The identity features of the top three and 10,000 principal components and the ROI-pooled signals did not significantly discriminate between the schizophrenia and control group (acc. = 27.8, 50, and 61.1%, respectively), as shown in **Figure 3A**.

Similarly, subject age was regressed from the identity feature. The correlations between the predicted and actual age were not significant ($r = 0.128$ and 0.115 for *classification* and *classification*+), while the top three principal components showed significant correlation ($r = 0.57$, $p = 0.013$). The other conditions (i.e., the top 10,000 principal components, ROI-pooled signals, and functional connectivity matrix) did not show significant correlation ($r = -0.21$, -0.29, and 0.34, respectively), as shown in **Figure 3B**.

DISCUSSION

We have shown that: (i) the self-supervised learning scheme led our neural network to acquire the projection from the high ($\sim 10^6$) dimensional signal space to the lower dimensional ($\sim 10^2$) feature space in which each dimension represented subject identity in the training dataset, (ii) the capability of the subject identification was generalized to the unknown subjects in the test dataset, and (iii) the temporal average of the extracted feature vector reflected the psychiatric status of the subjects. Surprisingly, our proposed method performed comparable to or even better than the functional connectivity matrix for schizophrenia diagnosis, which has been regarded as a promising biomarker of cognitive functions (Liang et al., 2006; Kim et al., 2016) and reported to reflect the cognitive trait in subjects (Finn et al., 2015).

The transferred capability from the subject identification to schizophrenia diagnosis can be regarded as a kind of "deep feature extraction." In the natural scene image processing, the intermediate output in a neural network pre-trained with a large-dataset classification often works well in another task, known as a "deep feature extraction" (Oquab et al., 2014). The underlying mechanism of the transferability is still under debate; however, one of the dominant hypotheses is that the stacked two-dimensional convolution itself works as the statistical prior of the natural scene images, regardless of the training task (Ulyanov et al., 2018). Our results showed that the transference also occurred with the combination of the human-brain T2* images and the stacked three-dimensional convolutions.

Our feature did not correlate with the subject's age, unlike the psychiatric status. This result suggests that subjects with similar psychiatric status are adjacent on the feature space, whereas similar age subjects are not. Given this discussion, the linear-decomposition-based features (i.e., the principal/independent components) and the functional connectivity matrix might have potentially ignored the discontinuity on the signal-space, yielding the results in the subject's age regression different from our identity feature.

Our identity feature and the functional connectivity exhibited a significant performance on schizophrenia diagnosis. The functional connectivity has been reported to be a good subject identifier (Finn et al., 2015), and thus, the features that classified patients from controls were those which behaved as the identifier of the subjects. The linkage between subject identification and the subject's mental condition should be investigated in future works. Although the difference in diagnosis accuracy was not statistically tested due to the shortage of samples, the diagnosis accuracy of our identity feature was slightly greater than that of the functional connectivity. A potential reason behind this superiority is the local interactions of the signal. In the functional

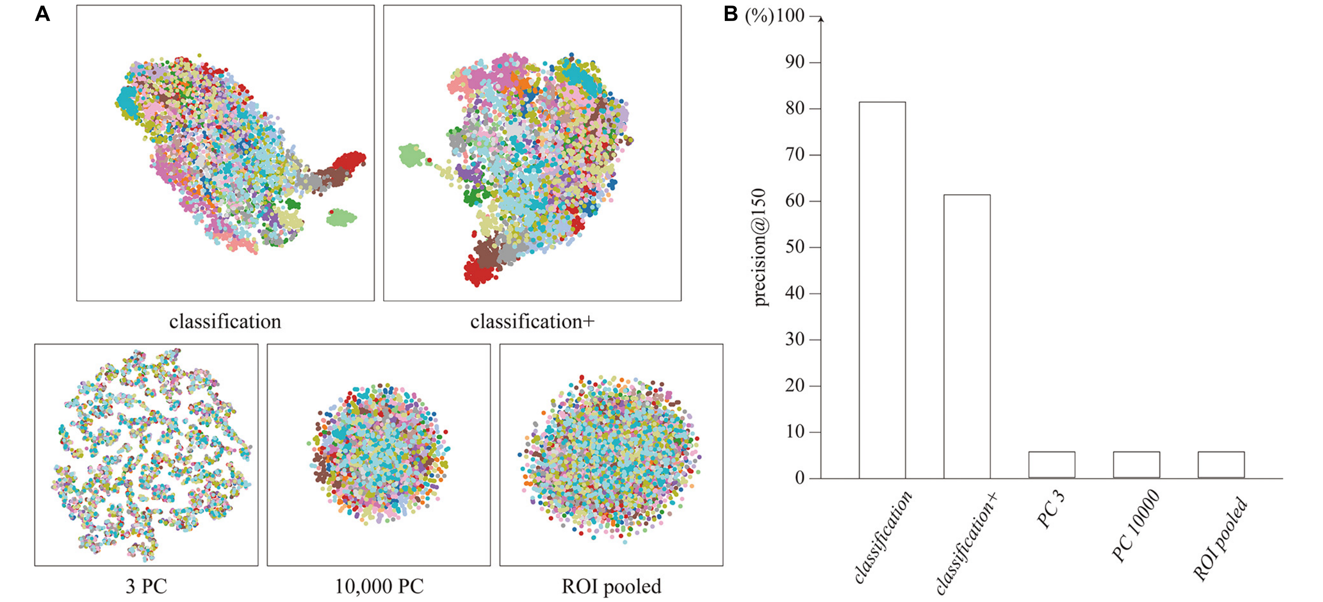


FIGURE 2 | (A) Distribution of feature vectors, visualized by t-SNE. Each dot represents a feature vector for a single timepoint, colored for each subject. **(B)** Precision@150 of the cluster for each feature vector.

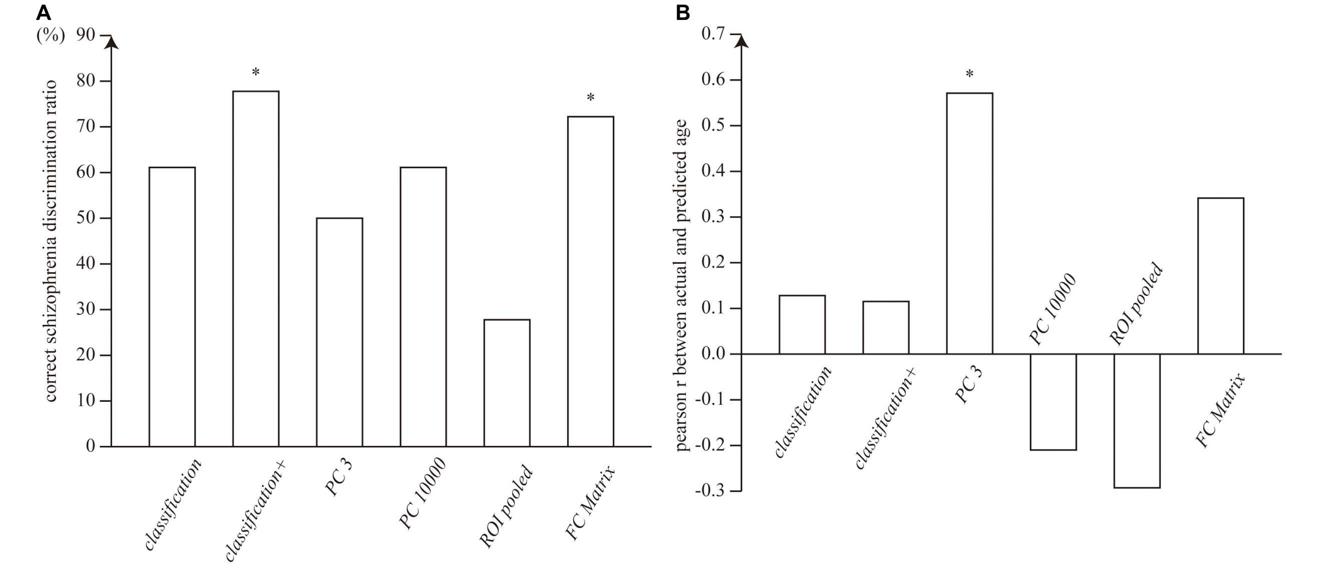


FIGURE 3 | (A) Accuracy of the schizophrenia discrimination. **(B)** Pearson's correlation coefficient of the age regression. The single asterisks show the statistical significance at $\alpha = 0.05$.

connectivity analysis, the signals are averaged for each ROI, discarding the local signal interactions. In contrast, previous studies have reported that both global and local activities in the brain lead to our cognitive functions (see Panzeri et al., 2015 for review). Both of the local and global interactions are modeled in the neural network, and it might have led to a positive effect in schizophrenia discrimination.

We introduced two versions of identity feature in this study, namely "classification" and "classification+." Both the "classification" and "classification+" feature vectors are the intermediate output of our neural network but the characteristics of these feature vectors were slightly different: "classification" feature vectors clustered more cohesive around the subject's identity feature than "classification+" feature vectors, while "classification+" identity feature performed better for schizophrenia diagnosis. The better performance of "classification+" in the schizophrenia diagnosis might be attributed to the small training dataset. The "classification+"

feature can be regarded as the projected space from the "classification" feature space to a hyperplane tangential to **1** (vector of all ones), which reduces the degree of freedom and potentially regularizes the feature space. The regularization of the feature space might have positively affected the training with small samples for the schizophrenia diagnosis. The relation between these two types of feature vectors should be investigated in future work with a larger dataset.

In this study, we introduced a novel self-supervised learning scheme and highlighted some of the characteristics of the extracted feature, especially in terms of the relation to schizophrenia. A few parameters, such as the optimal number of subjects in the training, the optimal neural network architecture, more detailed relations between the feature and the subject's attributes, and the mathematical analyses about the feature space will be addressed in the future work. Furthermore, for the clinical application, it is essential to evaluate the diagnosis accuracy and robustness more precisely with larger dataset as well as to explore better regressors rather than a simple linear regressor. We hope these will be uncovered in future works along the further accumulation of available datasets and with the advancement in the field of machine learning.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html, the Center for Biomedical Research Excellence.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Committee on Medical Ethics of the National Center of Neurology and Psychiatry. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

YH and YY conceived and designed the research and drafted the manuscript. YH conducted the experiments and analyzed the data. YO supported preprocessing of MRI data. YO and MH provided critical revisions. All authors contributed to and have approved the final manuscript.

FUNDING

This work was partially supported by the JSPS KAKENHI (Grant Nos. JP17H06039, JP18KT0021, JP18K07597, JP19H04998, and JP20H00625) and the JST CREST (Grant No. JPMJCR16E2).

ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at bioRxiv (Hashimoto et al., 2020).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2021.696853/full#supplementary-material

REFERENCES

Aghdam, M. A., Sharifi, A., and Pedram, M. M. (2017). Diagnosing autism spectrum disorder from brain resting-state functional connectivity patterns using a deep neural network with a novel feature selection method. *J. Digit. Imaging* 32, 899–918. doi: 10.1007/s10278-019-00196-1

Aine, C., Calhoun, V., Canive, J., Hanlon, F., Jung, R., Kiehl, K., et al. (2012). *Data From: The Mind Research Network & the University of New Mexico. The Center for Biomedical Research Excellence*. Available online at: http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html (accessed July 29, 2021).

Damaraju, E., Allen, E. A., Belger, A., Ford, J. M., McEwen, S., Mathalon, D. H., et al. (2014). Dynamic functional connectivity analysis reveals transient states of dysconnectivity in schizophrenia. *NeuroImage* 5, 298–308. doi: 10.1016/j.nicl.2014.07.003

De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., et al. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* 24, 1342–1350. doi: 10.1038/s41591-018-0107-6

Durstewitz, D., Koppe, G., and Meyer-Lindenberg, A. (2019). Deep neural networks in psychiatry. *Mol. Psychiatry* 24, 1583–1598. doi: 10.1038/s41380-019-0365-9

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 115–118. doi: 10.1038/nature21056

Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., et al. (2015). Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nat. Neurosci.* 18, 1664–1671. doi: 10.1038/nn.4135

Hashimoto, Y., Ogata, Y., Honda, M., and Yamashita, Y. (2020). Deep feature extraction for resting-state functional MRI by self-supervised learning and application to schizophrenia diagnosis. *bioRxiv* [Preprint] doi: 10.1101/2020.08.22.260406

Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., and Meneguzzi, F. (2018). Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *NeuroImage Clin.* 17, 16–23. doi: 10.1016/j.nicl.2017.08.017

Khosla, M., Jamison, K., Ngo, G. H., Kuceyeski, A., and Sabuncu, M. R. (2019). Machine learning in resting-state fMRI analysis. *Magn. Reson. Imaging* 64, 101–121. doi: 10.1016/j.mri.2019.05.031

Kim, J., Calhoun, V. D., Shim, E., and Lee, J. H. (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. doi: 10.1016/j.neuroimage.2015.05.018

Kingma, D. P., and Ba, L. J. (2015). "Adam: a method for stochastic optimization [Conference presentation abstract]," in *Proceedings of the International Conference on Learning Representations* (San Diego, CA).

Liang, M., Zhou, Y., Jiang, T., Liu, Z., Tian, L., Liu, H., et al. (2006). Widespread functional disconnectivity in schizophrenia with resting-state functional magnetic resonance imaging. *NeuroReport* 17, 209–213. doi: 10.1097/01.wnr.0000198434.06518.b8

Maaten, L. V. D., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

Noroozi, M., and Favaro, P. (2016). "Unsupervised learning on visual representations by solving jigsaw puzzles," in *Lecture Notes in Computer Science*, Vol. 9910, eds B. Leibe, J. Matas, N. Sebe, and M. Welling (Cham: Springer), 69–84. doi: 10.1007/978-3-319-46466-4_5

Oh, K., Kim, W., Shen, G., Piao, Y., Kang, N. I., Oh, I. S., et al. (2019). Classification of schizophrenia and normal controls using 3D convolutional neural network and outcome visualization. *Schizophr. Res.* 212, 186–195. doi: 10.1016/j.schres.2019.07.034

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. (2014). "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Colombus, OH), 1717–1724. doi: 10.1109/CVPR.2014.222

Panzeri, S., Macke, J. H., Gross, J., and Kayser, C. (2015). Neural population coding: combining insights from microscopic and mass signals. *Trends Cogn. Sci.* 19, 162–172. doi: 10.1016/j.tics.2015.01.002

Suk, H. I., Wee, C. Y., Lee, S. W., and Shen, D. (2016). State-space model with deep learning for functional dynamics estimation in resting-state fMRI. *Neuroimage* 129, 292–307. doi: 10.1016/j.neuroimage.2016.01.005

Titano, J. J., Badgeley, M., Schefflein, J., Pain, M., Su, A., Cai, M., et al. (2018). Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nat. Med.* 24, 1337–1341. doi: 10.1038/s41591-018-0147-y

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., et al. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* 15, 273–289. doi: 10.1006/nimg.2001.0978

Ulyanov, D., Vedaldi, A., and Ulyanov, D. (2018). "Deep image prior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 9446–9454.

Yamaguchi, H., Hashimoto, Y., Sugihara, G., Miyata, J., Murai, T., Takahashi, H., et al. (2021). Three-dimensional convolutional autoencoder extracts features of structural brain images with a "diagnostic labelfree" approach: application to schizophrenia datasets. *Front. Neurosci.* 15:804.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 15 September 2021
doi: 10.3389/fnins.2021.727448

Sub-Optimality of the Early Visual System Explained Through Biologically Plausible Plasticity

Tushar Chauhan[1,2]*, *Timothée Masquelier*[1,2] *and Benoit R. Cottereau*[1,2]

[1] Centre de Recherche Cerveau et Cognition, Université de Toulouse, Toulouse, France, [2] Centre National de la Recherche Scientifique, Toulouse, France

The early visual cortex is the site of crucial pre-processing for more complex, biologically relevant computations that drive perception and, ultimately, behaviour. This pre-processing is often studied under the assumption that neural populations are optimised for the most efficient (in terms of energy, information, spikes, etc.) representation of natural statistics. Normative models such as Independent Component Analysis (ICA) and Sparse Coding (SC) consider the phenomenon as a generative, minimisation problem which they assume the early cortical populations have evolved to solve. However, measurements in monkey and cat suggest that receptive fields (RFs) in the primary visual cortex are often noisy, blobby, and symmetrical, making them sub-optimal for operations such as edge-detection. We propose that this suboptimality occurs because the RFs do not emerge through a global minimisation of generative error, but through locally operating biological mechanisms such as spike-timing dependent plasticity (STDP). Using a network endowed with an abstract, rank-based STDP rule, we show that the shape and orientation tuning of the converged units are remarkably close to single-cell measurements in the macaque primary visual cortex. We quantify this similarity using physiological parameters (frequency-normalised spread vectors), information theoretic measures [Kullback–Leibler (KL) divergence and Gini index], as well as simulations of a typical electrophysiology experiment designed to estimate orientation tuning curves. Taken together, our results suggest that compared to purely generative schemes, process-based biophysical models may offer a better description of the suboptimality observed in the early visual cortex.

Keywords: vision, cortex, plasticity, suboptimality, Independent Component Analysis, Sparse Coding, STDP, natural statistics

OPEN ACCESS

Edited by:
Ali Ghazizadeh,
Sharif University of Technology, Iran

Reviewed by:
Anthony N. Burkitt,
The University of Melbourne, Australia
Michael Robert DeWeese,
University of California, Berkeley,
United States

****Correspondence:***
Tushar Chauhan
research@tusharchauhan.com

Specialty section:
This article was submitted to
Neural Technology,
a section of the journal
Frontiers in Neuroscience

Received: *18 June 2021*
Accepted: *25 August 2021*
Published: *15 September 2021*

Citation:
Chauhan T, Masquelier T and Cottereau BR (2021) Sub-Optimality of the Early Visual System Explained Through Biologically Plausible Plasticity.
Front. Neurosci. 15:727448.
doi: 10.3389/fnins.2021.727448

INTRODUCTION

The human visual system processes an enormous throughput of sensory data in successive operations to generate percepts and behaviours necessary for biological functioning (Anderson et al., 2005; Raichle, 2010). Computations in the early visual cortex are often explained through unsupervised normative models which, given an input dataset with statistics similar to our surroundings, carry out an optimisation of criteria such as energy consumption and information-theoretic efficiency (Olshausen and Field, 1996; Bell and Sejnowski, 1997;

van Hateren and van der Schaaf, 1998; Hoyer and Hyvärinen, 2000; Zhaoping, 2006; Bruce et al., 2016). While such approaches could explain why many properties of the early visual system are closely related to characteristics of natural scenes (Olshausen and Field, 1996; Bell and Sejnowski, 1997; Lee and Seung, 1999; Geisler, 2008; Hunter and Hibbard, 2015; Beyeler et al., 2019), they are not equipped to answer questions such as how cortical structures which support complex computational operations implied by such optimisation may emerge, how these structures adapt, even in adulthood (Wandell and Smirnakis, 2010; Hübener and Bonhoeffer, 2014), and why some neurones possess receptive fields (RFs) which are sub-optimal in terms of information processing (Jones and Palmer, 1987; Ringach, 2002).

It is now well established that locally driven synaptic mechanisms such as spike-timing dependent plasticity (STDP) are natural processes which play a pivotal role in shaping the computational architecture of the brain (Markram et al., 1997; Delorme et al., 2001; Caporale and Dan, 2008; Larsen et al., 2010; Masquelier, 2012; Brito and Gerstner, 2016). Indeed, locally operating implementations of generative schemes have been shown to be closer to biological measurements (see, e.g., Rozell et al., 2008). Therefore, it is only natural to hypothesise that locally operating, biologically plausible models of plasticity must offer a better description of RFs in early visual cortex. However, such line of reasoning leads to the obvious question: what exactly constitutes a "better description" of a biological system, and more specifically, the early visual cortex. Here, we use a series of criteria spanning across electrophysiology, information theory, and machine learning, to investigate how descriptions of early visual RFs provided by a local, abstract STDP model compare to biological data from the macaque. We also compare these results to two classical, and important normative schemes – Independent Component Analysis (ICA), and Sparse Coding (SC). Our results demonstrate that a local process-based model of experience-driven plasticity may be better suited to capturing the RFs of simple-cells, thus suggesting that biological preference does not always concur with forms of global, generative optimality.

More specifically, we show that STDP units are able to better capture the characteristic sub-optimality in RF shape reported in literature (Jones and Palmer, 1987; Ringach, 2002), and their orientation tuning closely matches measurements in the macaque primary visual cortex (V1) (Ringach et al., 2002). Taken together, our findings suggest that while the information carrying capacity of an STDP ensemble is not optimal when compared to generatively optimal schemes, it is precisely this sub-optimality which may make process-based, local models more suited for describing the initial stages of sensory processing.

MATERIALS AND METHODS

Dataset

The Hunter–Hibbard dataset of natural images was used (Hunter and Hibbard, 2015) for training. It is available under the MIT license at https://github.com/DavidWilliamHunter/Bivis, and consists of 139 stereoscopic images of natural scenes captured using a realistic acquisition geometry and a 20° field of view. Only images from the left channel were used, and each image was resized to a resolution of 5 px/° along both horizontal and vertical directions. Inputs to all encoding schemes were 3 × 3° patches (i.e., 15 × 15 px) sampled randomly from the dataset (**Figure 1A**).

Encoding Models

Samples from the dataset were used to train and test three models corresponding to the ICA, SC, and STDP encoding schemes. Each model consisted of three successive stages (**Figure 1B**). The first stage represented retinal activations. This was followed by a pre-processing stage implementing operations which are typically associated with processing in the lateral geniculate nucleus (LGN), such as whitening and decorrelation. In the third stage, LGN output was used to drive a representative V1 layer.

During learning, 10^5 patches (3 × 3°) were randomly sampled from the dataset to simulate input from naturalistic scenes. In this phase, the connections between the LGN and V1 layers were plastic, and modified in accordance with one of the three encoding schemes. Care was taken to ensure that the sequence of inputs during learning was the same for all three models. After training, the weights between the LGN and V1 layers were no longer allowed to change. The implementation details of the three models are described below.

Sparse Coding

Sparse Coding algorithms are based on energy-minimisation, which is typically achieved by a "sparsification" of activity in the encoding population. We used a now-classical SC scheme proposed by Olshausen and Field (1996, 1997). The pre-processing in this scheme consists of an initial whitening of the input using low pass filtering, followed by a trimming of higher frequencies. The latter was employed to counter artefacts introduced by high frequency noise, and the effects of sampling across a uniform square grid. In the frequency domain the pre-processing filter was given by a zero-phase kernel:

$$H\left(f\right) = f \cdot e^{-\left(\frac{f}{f_0}\right)^4} \tag{1}$$

Here, $f_0 = 10$ cycles/° is the cut-off frequency. The outputs of these LGN filters were then used as inputs to the V1 layer composed of 225 units (3° × 3° RF at 5 px/°). The total number of weights in the model was 50,625. Retinal projections of the converged RFs were recovered by an approximate reverse-correlation algorithm (Ringach, 2002; Ringach and Shapley, 2004) derived from a linear-stability analysis of the SC objective function about its operating point. The RFs (denoted as columns of a matrix, say $\boldsymbol{\xi}$) were given by:

$$\boldsymbol{\xi} = \mathbf{A}\left[\mathbf{A}^T\mathbf{A} + \lambda S''(0)\,\mathbf{I}\right]^{-1} \tag{2}$$

Here, $\mathbf{A}$ is the matrix containing converged sparse components as column vectors, λ is the regularisation parameter (for the reconstruction, it is set to 0.14σ, where σ^2 is the variance in the input dataset), and $S(x)$ is the shape-function for the prior distribution of the sparse coefficients [this implementation uses $\log(1 + x^2)$].

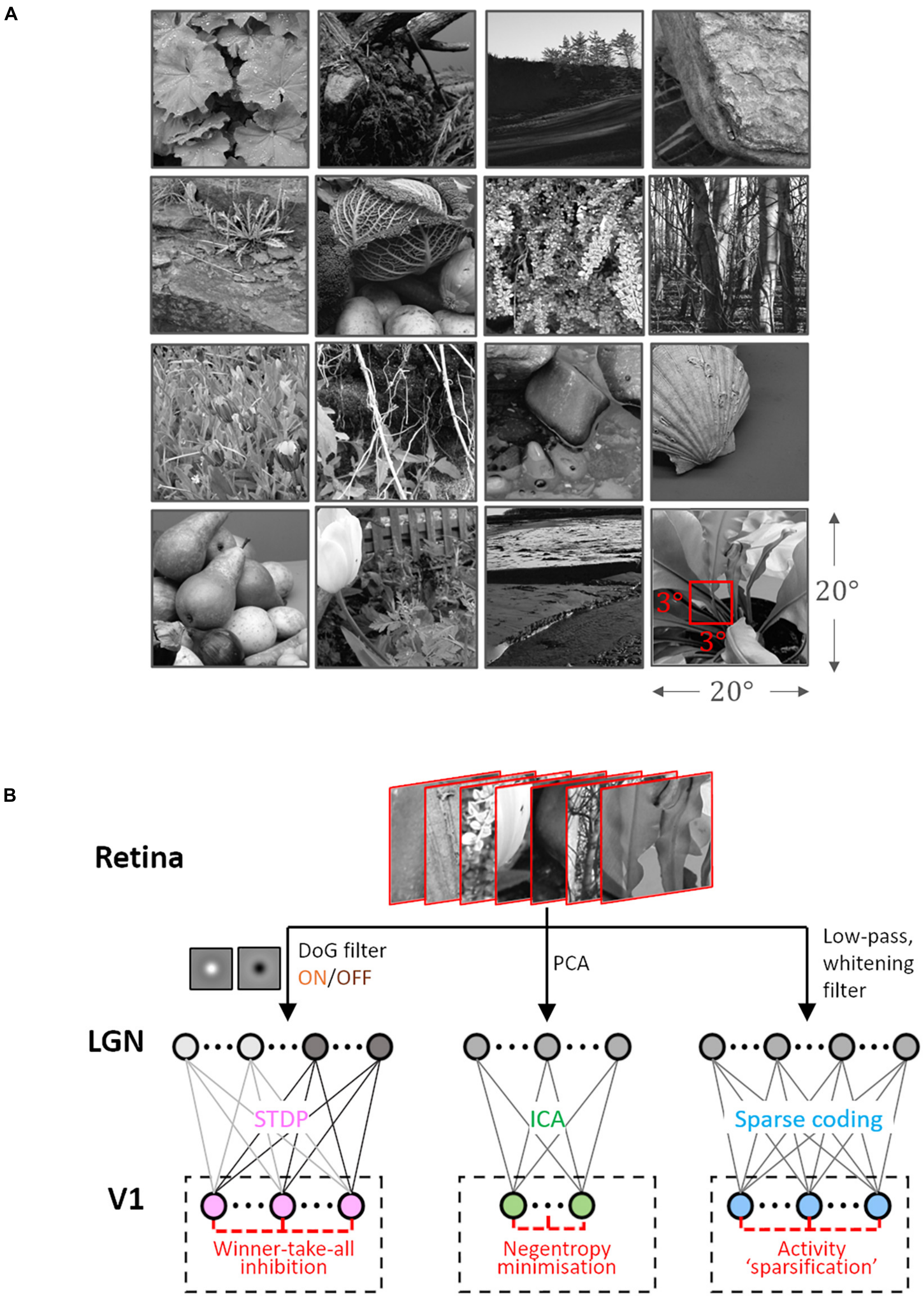


FIGURE 1 | Dataset and the computational pipeline. **(A)** Training data. The Hunter–Hibbard dataset of natural images was used. The images in the database have a 20° × 20° field of view. Patches of size 3° × 3° were sampled from random locations in the images (overlap allowed). The same set of 100,000 randomly sampled patches was used to train three models: Spike-timing dependent plasticity (STDP), Independent Component Analysis (ICA), and Sparse Coding (SC). **(B)** Modelling the early visual pathway. Three representative stages of early visual computation were captured by the models: retinal input, processing in the lateral geniculate nucleus (LGN), and the activity of early cortical populations in the primary visual cortex (V1). Each input patch represented a retinal input. This was followed by filtering operations generally associated with the LGN, such as decorrelation and whitening. Finally, the output from the LGN units/filters was connected to the V1 population through all-to-all (dense) plastic synapses which changed their weights during learning. Each model had a specific optimisation strategy for learning: the STDP model relied on a local rank-based Hebbian rule, ICA minimised mutual information (approximated by the negentropy), and SC enforced sparsity constraints on V1 activity. DoG, difference of Gaussian; PCA, Principal Component Analysis.

Independent Component Analysis

Independent Component Analysis algorithms are based on the idea that the activity of an encoding ensemble must be as information-rich as possible. This typically involves a maximisation of mutual information between the retinal input and the activity of the encoding ensemble. We used a classical ICA algorithm called *fastICA* (Hyvärinen and Oja, 2000) which achieves this through an iterative estimation of input negentropy. The pre-processing in this implementation was performed using a truncated Principal Component Analysis (PCA) transform ($\tilde{d} = 150$ components were used), leading to low-pass filtering and local decorrelation akin to centre-surround processing reported in the LGN. The model fit a total of 33,750 weights. If the input patches are denoted by the columns of a matrix (say $\mathbf{X}$), the LGN activity $\mathbf{L}$ can be written as:

$$\mathbf{L} = \widetilde{\mathbf{U}}^T \mathbf{X}_C \tag{3}$$

Here, $\mathbf{X}_C = \mathbf{X} - \langle \mathbf{X} \rangle$ and $\widetilde{\mathbf{U}}$ is a matrix composed of the first $\tilde{d}\ (= 150)$ principal components of $\mathbf{X}_C$. The activity of these LGN filters was then used to drive the ICA V1 layer consisting of 150 units, with its activity $\mathbf{\Sigma}$ being given by:

$$\Sigma = \mathbf{WL} \tag{4}$$

Here, $\mathbf{W}$ is the un-mixing matrix which is optimised during learning. The recovery of the RFs for ICA was relatively straight forward, as, in our implementation, they were assumed to be equivalent to the filters which must be applied to a given input to generate the corresponding V1 activity. The RFs (denoted as columns of a matrix, say $\boldsymbol{\xi}$) were given by:

$$\boldsymbol{\xi} = \widetilde{\mathbf{U}} \mathbf{W}^T + \langle \mathbf{X} \rangle \tag{5}$$

Spike-Timing Dependent Plasticity

Spike-timing dependent plasticity is a biologically observed, Hebbian-like learning rule which relies on local spatiotemporal patterns in the input. We used a feedforward model based on an abstract rank-based STDP rule (Chauhan et al., 2018). The pre-processing in the model consisted of half-rectified ON/OFF filtering using difference-of-Gaussian kernels based on the properties of magno-cellular LGN cells. The outputs of these filters were converted to relative first-spike latencies using a monotonically decreasing function ($1/x$ was used), and only the earliest 10% spikes were allowed to propagate to V1 (Delorme et al., 2001; Masquelier and Thorpe, 2007). For each iteration, spikes within this 10% window were used to drive an unsupervised network of 225 integrate-and-fire neurones. The membrane potential of a V1 neurone was given by:

$$u(t) = \mathrm{H}(\theta - u) \sum_{i \in LGN} w_i(t)\, \delta(t - t_i) \tag{6}$$

Here, t_i denotes the latency of the i-th pre-synaptic neuron, H is the Heavisde function, and θ is the spiking threshold. During learning, changes in the synaptic weights between LGN and V1 were governed by a rank-based, simplified version of the STDP rule proposed by Gütig et al. (2003). After each iteration, the change (Δw) in the weight (w) of a given synapse was given by:

$$\Delta w = \begin{cases} -\alpha^- \cdot (w - w_{min})^{\mu^-} \mathrm{K}(t, \tau^-),\ t \le 0 \\ \alpha^+ \cdot (w_{max} - w)^{\mu^+} \mathrm{K}(t, \tau^+),\ t > 0 \end{cases} \tag{7}$$

Here, Δt is the difference between the post- and pre-synaptic spike times, the constants $\alpha^{\pm}$ describe the learning rates for long-term potentiation (LTP) and depression (LTD), respectively, $\mu^{\pm} \in [0,1]$ characterise the non-linearity of the multiplicative updates, K is a windowing function, and $\tau^{\pm}$ are the time-scales for LTP and LTD windows. Note that w is soft-bound such that $w \in (w_{min}, w_{max})$. The model used $w_{min} = 0$ (thalamocortical connections are known to be excitatory in nature), and $w_{max} = 1$. Since the intensity-to-latency conversion operates on an arbitrary time-scale, weight updates were based on the spike-order rather than precise spike-timing (rank-based). This meant that the window for LTP (τ^+) was variable and driven by the first 10% thalamic spikes, while the window for LTD (τ^-) was theoretically infinite. During updates, the weight was increased if a presynaptic spike occurred before the postsynaptic spike (causal firing), and decreased if it occurred after the post-synaptic spike (acausal firing). The learning rates were $\alpha^+ = 5 \times 10^{-3}$ and $\alpha^- = 0.75 \times \alpha^+$, and the nonlinearities were $\mu^+ = 0.65$ and $\mu^- = 0.05$. The model has previously been shown to be robust to both internal and external noise, and the parameter values were chosen from a range which best approximates the behaviour of the model under a realistic, V1-like regime (Chauhan et al., 2018). The neural population was homogeneous, with each neuron described by the exact same set of parameters.

During each iteration of learning, the population followed a winner-take-all inhibition rule wherein the firing of one neurone reset the membrane potentials of all other neurones. A total of 50,625 weights were fit by the model. After learning, this inhibition was no longer active and multiple units were allowed to fire for each input – allowing us to measure the behaviour of the network during testing. This also renders the model feed-forward only, making it comparable to SC and ICA. The RFs of the converged neurones were recovered using a linear approximation. If w_i denotes the weight of the synapse connecting a given neurone to the ith LGN filter with RF $\boldsymbol{\psi}_{\mathbf{i}}$, the RF $\boldsymbol{\xi}$ of the neurone was given by:

$$\boldsymbol{\xi} = \sum_{i \in LGN} w_i \boldsymbol{\psi}_{\mathrm{i}} \tag{8}$$

Evaluation Metrics

Gabor Fitting

Linear approximations of RFs obtained by each encoding strategy were fitted using 2-D Gabor functions. This is motivated by the fact that all the encoding schemes considered here lead to linear, simple-cell-like RFs. In this case, the goodness-of-fit parameter (R^2) provides an intuitive measure of how Gabor-like a given RF is. The fitting was carried out using an adapted version of the code available at https://uk.mathworks.com/matlabcentral/fileexchange/60700-fit2dgabor-data-options (Ecke et al., 2021).

Frequency-Normalised Spread Vector

The shape of the RFs approximated by each encoding strategy was characterised using frequency-normalised spread vectors (FSVs) (Ringach, 2002; Chauhan et al., 2018). For a RF fitted by a Gabor-function with sinusoid carrier frequency f and envelope size $\sigma = \left[\sigma_x\ \sigma_y\right]^T$, the FSV is given by:

$$\left[n_x\ n_y\right]^T = \left[\sigma_x\ \sigma_y\right]^T f \tag{9}$$

While n_x provides an intuition of the number of cycles in the RF, n_y is a cycle-adjusted measure of the elongation of the RF perpendicular to the direction of sinusoid propagation. The FSV serves as a compact, intuitive descriptor of the RF shape-invariance to affine operations such as translation, rotation, and isotropic scaling.

Orientation Tuning

Orientation tuning curves (OTCs) were estimated by presenting each unit in each model with noisy oriented sine-wave grating (SWG) stimuli at its preferred frequency. The orientation was sampled in steps of 2° in the interval [0°, 180]°. For each orientation, the activity was averaged over phase values uniformly sampled in the interval [0°, 360°] using a step-size of 5°. The bandwidth of an OTC was taken as its half-width at $1/\sqrt{2}$ of the peak response (Ringach et al., 2002). The whole process was repeated 100 times, and a bootstrap procedure was used to determine 95% confidence intervals.

Fisher Information

The information content in the activity of the converged units was quantified by using approximations of the Fisher information (FI, denoted here by the symbol J). If $\mathbf{x} = \{x_1, x_2, x_3, , x_N\}$ is a random variable describing the activity of an ensemble of N independent units, the FI of the population with respect to a parameter θ is given by:

$$J(\theta) = \sum_{i=1}^{N} E\left[\left\{\frac{\partial}{\partial x}\ln P(x_i|\theta)\right\}^2\right]_{P(x_i|\theta)} \tag{10}$$

Here, $E[.]_{P(x_i|\theta)}$ denotes expectation value with respect to the firing-state probabilities of the ith neurone in response to the stimuli corresponding to parameter value θ. In our simulations, θ was the orientation (defined as the direction of travel) of a set of SWGs with additive Gaussian noise, and was sampled at intervals of 4° in the range [0°, 180°). The SWGs were presented at frequency of 1.25 cycles/visual degree, and the responses were calculated by averaging over 8 evenly spaced phase values in [0°, 360°). This effectively simulated a drifting grating design within the constraints of the computational models. Each simulation was repeated 100 times and a jackknife procedure was used to estimate 95% confidence intervals. Noise was added such that the signal-to-noise ratio (SNR) varied between −6 and 6 dB in steps of 1 dB.

Decoding Using a Linear Classifier

In addition to FI approximations, we also used a linear decoder on the population responses obtained in the FI simulations. The decoder was an error-correcting output codes model composed of binary linear-discriminant classifiers configured in a one-vs.-all scheme. Similar to the FI experiment, ground-truth values of the orientation at intervals of 4° in the range [0°, 180°) were used as the class labels, and the activity generated by the corresponding SWG stimuli with added Gaussian noise was used as the training/testing data. The SWGs were presented at a frequency of 1.25 cycles/visual degree, and the responses were calculated by averaging over 8 evenly spaced phase values in [0°, 360°). Each simulation was repeated 100 times, each time with five-fold validation. A jackknife procedure was used to estimate 95% confidence intervals.

Post-convergence Threshold Variation in STDP

To test how post-learning changes in the threshold affect the specificity of a converged network, we tested an STDP network trained using a threshold θ_{training} by increasing or decreasing its threshold (to say, θ_{testing}) and presenting it with SWGs (same stimuli as the ones used to calculate the FI). We report the results of seven simulations where the relative change in threshold was given by 25% increments/decrements, i.e.:

$$\frac{\theta_{\text{testing}} - \theta_{\text{training}}}{\theta_{\text{training}}} = \{0,\ \pm\ 0.25,\ \pm\ 0.50,\ \pm\ 0.75\} \tag{11}$$

Kullback–Leibler Divergence

For each model, we estimated probability density functions (pdfs) over parameters such as the FSVs and the population bandwidth. To quantify how close the model pdfs were to those estimated from the macaque data, we employed the Kullback–Leibler (KL) divergence. KL divergence is a directional measure of distance between two probability distributions. Given two distributions P and Q with corresponding probability densities p and q, the KL divergence (denoted D_{KL}) of P from Q is given by:

$$D_{KL}(P||Q) = \int_{\mathbf{\Omega}} p(\mathbf{x}) \log_2\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right) d\mathbf{x} \tag{12}$$

Here, $\mathbf{\Omega}$ is the support of the distribution Q. In our analysis, we considered the reference distribution p as a pdf estimated from the macaque data, and q as the pdf (of the same variable) estimated using ICA, SC, or STDP. In this case, KL divergence lends itself to a very intuitive interpretation: it can be considered as the additional bandwidth (in bits) which would be required if the biological variable were to be encoded using one of the three computational models. Note that P and Q may be multivariate distributions.

Sparsity: Gini Index

The sparseness of the encoding was evaluated using the Gini index (GI). GI is a measure which characterises the deviation of the population-response from a uniform distribution of activity across the samples. Formally, the GI (denoted here as Λ) is given by:

$$\Lambda(\mathbf{x}) = 1 - 2\int_0^1 L(F) dF \tag{13}$$

Here L is the Lorenz function defined on the cumulative probability distribution F of the neural activity (say, $\mathbf{x}$). GI is 0 if all units have the same response and tends to 1 as responses become sparser (being equal to 1 if only 1 unit responds, while others are silent). It is invariant to the range of the responses

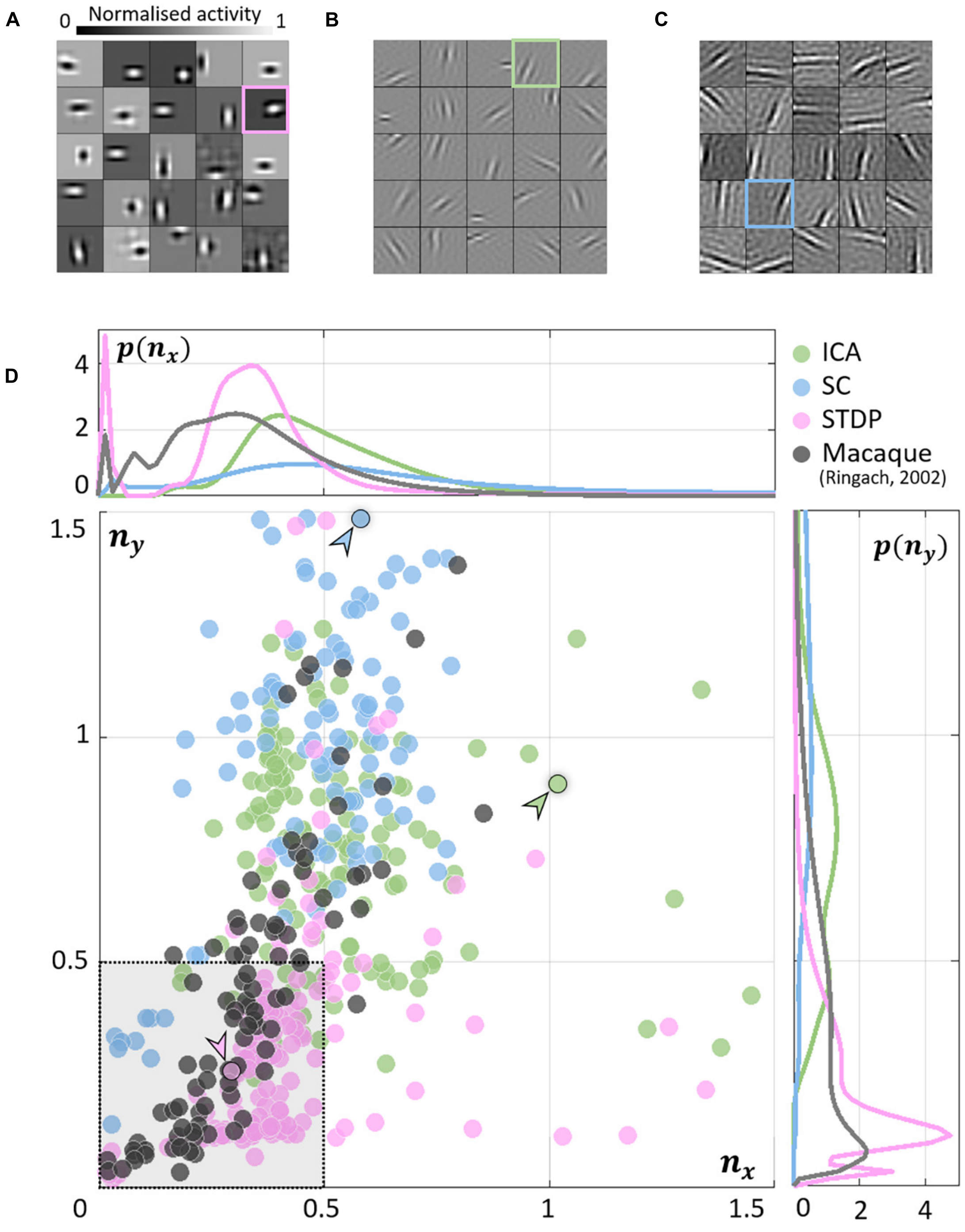


FIGURE 2 | Receptive field (RF) shape. **(A–C)** RFs of neurones randomly chosen from the three converged populations. The STDP population is shown in **(A)**, ICA in **(B)**, and SC in **(C)**. **(D)** Frequency-scaled spread vectors (FSVs). FSV is a compact metric for quantifying RF shape. n_x is proportional to the number of lobes in the RF, n_y is a measure of the elongation of the RF, and values near zero characterise symmetric, often blobby RFs. The FSVs for STDP (pink), ICA (green), and SC (blue), are shown with data from macaque V1 (black) (Ringach, 2002). Measurements in macaque simple-cells tend to fall within the square bound by 0.5 along both axes (shaded in grey, with a dotted outline). Three representative neurones are indicated by colour-coded arrows: one for each algorithm. The corresponding RFs are outlined in **(A–C)** using the corresponding colour. The STDP neurone has been chosen to illustrate a blobby RF, the ICA neurone shows a multi-lobed RF, and the SC neurone illustrates an elongated RF. Insets above and below the scatter plot show estimations of the probability density function for n_x and n_y. Both axes have been cut-off at 1.5 to facilitate comparison with biological data.

within a given sample, and robust to variations in sample-size (Hurley and Rickard, 2009). We defined two variants of the GI which measure the spatial (Λ_s) and temporal sparsity (Λ_t) of an ensemble of encoders. Given a sequence of M inputs to an ensemble of N neurones, the spatial sparsity of the ensemble response to the mth stimulus is given by:

$$\Lambda_S(m) = \Lambda\left(\left\{x_m^1, x_m^2, \ldots, x_m^N\right\}\right) \tag{14}$$

Here, x_m^n denotes the activity of the nth neurone in response to the mth input. Similarly, the temporal sparsity of the nth neurone over the entire sequence of inputs is given by:

$$\Lambda_T(n) = \Lambda\left(\left\{x_1^n, x_2^n, \ldots, x_M^n\right\}\right) \tag{15}$$

Code

The code for ICA was written in python using the sklearn library which implements the classical *fastICA* algorithm. The code for SC was based on the C++ and Matlab code shared by Prof. Bruno Olshaussen. The STDP code was based on a previously published binocular-STDP algorithm available at https://senselab.med.yale.edu/ModelDB/showmodel.cshtml?model=245409#tabs-1.

RESULTS

We used an abstract model of the early visual system with three representative stages: retinal input, LGN processing, and V1 activity (**Figure 1B**). To simulate retinal activity corresponding to natural inputs, patches of size 3° × 3° (visual angles) were sampled randomly from the Hunter–Hibbard database (Hunter and Hibbard, 2015) of natural scenes (**Figure 1A**). 10^5 patches were used to train models corresponding to three encoding schemes: ICA, SC, and STDP. Each model used a specific procedure for implementing the LGN processing and learning the synaptic weights between the LGN and V1 (see **Figure 1B** and section "Materials and Methods").

Receptive Field Symmetry

As expected, units in all models converged to oriented, edge-detector like RFs. While the RFs from ICA (**Figure 2B**) and SC (**Figure 2C**) were elongated and highly directional, STDP (**Figure 2A**) RFs were more compact and less sharply tuned. This is closer to what is observed from simple-cell recordings in the macaque (Ringach, 2002) where RFs show high circular symmetry, and do not seem to be optimally tuned for edge-detection (see Jones and Palmer, 1987 for similar data measured in the cat). To obtain a quantitative measure of this phenomenon, we fit Gabor functions to the RFs and considered the frequency-normalised spread vectors or FSVs of the fit (Eq. 9). The first component (n_x) of the FSV characterises the number of lobes in the RF, and the second component (n_y) is a measure of the elongation of the RF (perpendicular to carrier propagation). A considerable number of simple-cell RFs measured in macaque tend to fall within the square bounded by $n_x = 0.5$ and $n_y = 0.5$. The FSVs of a sample of neurones ($N = 93$) measured in the macaque V1 (Ringach, 2002) indicate that 59.1% of the neurones lay within this region (**Figure 2D**). Since they are not elongated, and contain few lobes (typically 2–3 on/off regions), they tend to be compact – making them less effective as edge-detectors compared to more crisply tuned, elongated RFs. Amongst the three encoding schemes, while a considerable number of STDP units (82.2%) tended to fall within these realistic boundaries, ICA (10.7%) and SC (4.0%) showed a distinctive shift upwards and to the right. This trend has been observed in a number of studies using models based on ICA and SC (see, e.g., Rehn and Sommer, 2007; Puertas et al., 2010; Zylberberg et al., 2011).

The inlays in **Figure 2D** provide estimations of the probability densities of two FSV parameters for the macaque data and the three models. An interesting insight into these distributions is given by the KL divergence (**Table 1**). KL divergence (Eq. 12) is a directed measure which can be interpreted as the additional number of bits required if one of the three models were used to encode data sampled from the macaque distribution. The KL divergence for the STDP model was found to be 3.0 bits indicating that, on average, it would require three extra bits to encode data sampled from the macaque distribution. In comparison, SC and ICA were found to require 8.4 and 14.6 additional bits, respectively. An examination of the KL divergence of marginal distributions of the FSV parameters showed that STDP offers excellent encoding of both the n_x (number of lobes) and the n_y (compactness) parameter. ICA does not encode either of the two parameters satisfactorily, while SC performance is closer to the STDP model (especially for parameter n_x).

Orientation Selectivity

Given this sub-optimal, symmetric nature of STDP RF shapes, we next investigated how this affected the responses of these neurones to sharp edges. In particular, we were interested in how the orientation bandwidths of the units from the three models would compare to biological data. Given the RF shape, we hypothesised that orientation selectivity would be worse for STDP compared to the ICA and SC schemes. To test this hypothesis, we simulated a typical electrophysiological experiment for estimating orientation tuning (**Figure 3A**). To each unit, we presented noisy SWGs at its preferred spatial frequency and recorded its activity as a function of the stimulus

TABLE 1 | Kullback–Leibler (KL) divergence of the distribution of macaque frequency-normalised spread vectors (FSVs) from the models.

	ICA	SC	STDP
Joint distribution			
$[n_x\, n_y]^T$	14.6	8.4	3.0
Marginal distributions			
n_x	7.6	1.4	1.3
n_y	14.0	3.8	0.4

*The receptive-field (RF) shape of the neurones from the models and measurements in macaque V1 **(Ringach, 2002)** was parametrised by estimating the frequency-normalised spread vectors (FSVs). FSVs are characterised by two parameters n_x and n_y: n_x is proportional to the number of lobes in the receptive field, and n_y is modulated by its elongation perpendicular to the direction of periodicity. The KL divergence reflects the number of additional bits required to encode the parameter(s) of interest from the macaque data using the distributions from one of the three models (ICA, SC, or STDP). All values are in bits.*

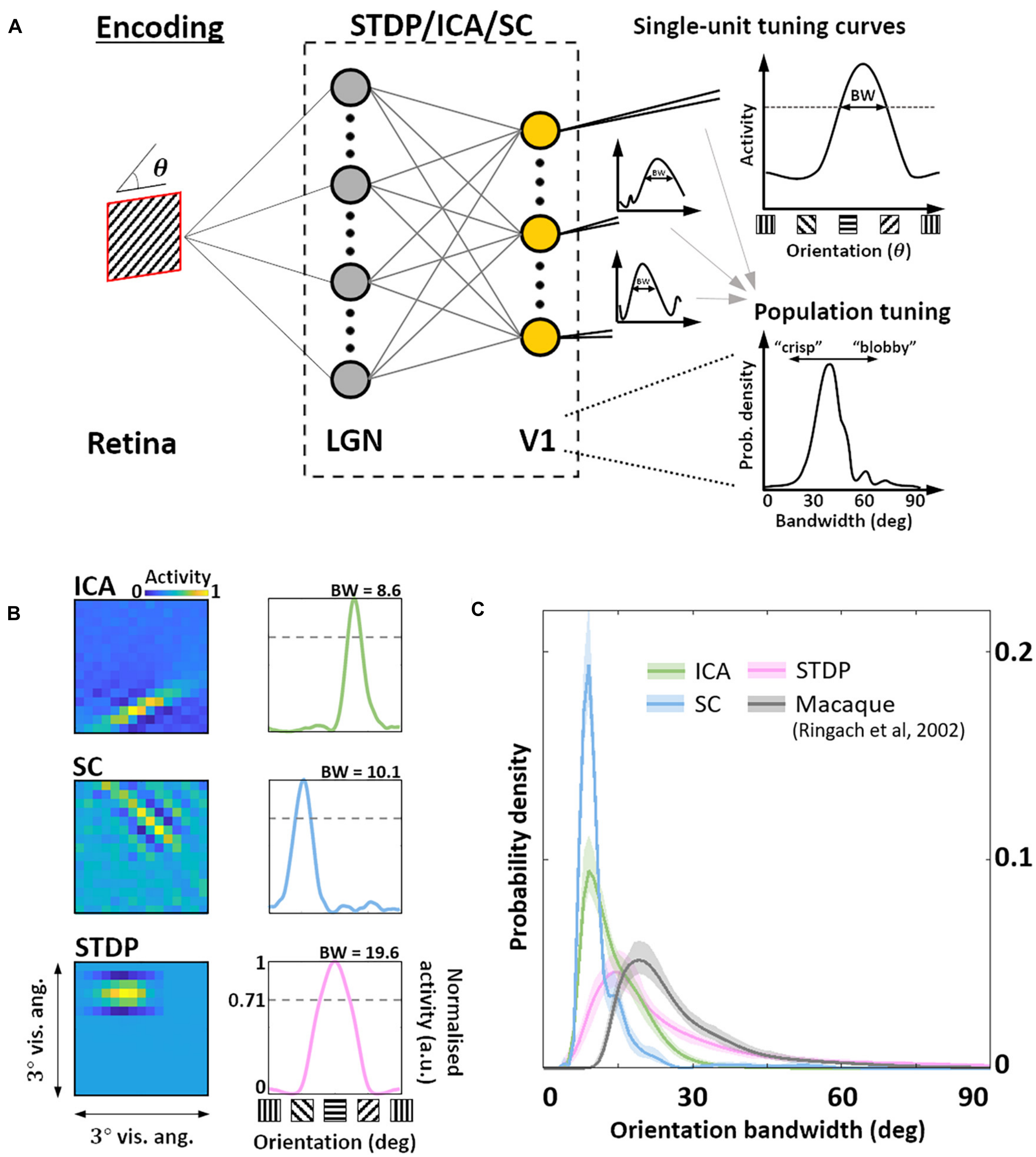


FIGURE 3 | Orientation encoding. **(A)** Orientation tuning. Sine-wave gratings with additive Gaussian noise were presented to the three models to obtain single-unit orientation tuning curves (OTCs). OTC peak identifies the preferred orientation of the unit, and OTC bandwidth (half width at $1/\sqrt{2}$ peak response) is a measure of its selectivity around the peak. Low bandwidth values are indicative of sharply tuned units while high values signal broader, less specific tuning. **(B)** Single-unit tuning curves. RF (left) and the corresponding OTC (right) for representative units from ICA (top row, green), SC (second row, blue), and STDP (bottom row, pink). The bandwidth is shown above the OTC. **(C)** Population tuning. Estimated probability density of the OTC bandwidth for the three models (same colour code as panel **B**), and data measured in macaque V1 (black) **(Ringach et al., 2002)**. Envelopes around solid lines show 95% confidence intervals estimated using a bootstrap procedure. All simulations shown here were performed at an input SNR of 0 dB.

orientation. This allowed us to plot its OTC (**Figure 3B**) and estimate the tuning bandwidth, which is a measure of the local selectivity of the unit around its peak – low values corresponding to sharply tuned neurones and higher values corresponding to broadly tuned, less selective neurones. For each of the three models, we estimated the pdf of the OTC bandwidth, and compared it to the distribution estimated over a large set of data ($N = 308$) measured in macaque V1 (Ringach et al., 2002) (**Figure 3C**). We found that ICA and SC distributions peaked at a bandwidth of about 10° (ICA: 9.1°, SC: 8.5°) while the STDP

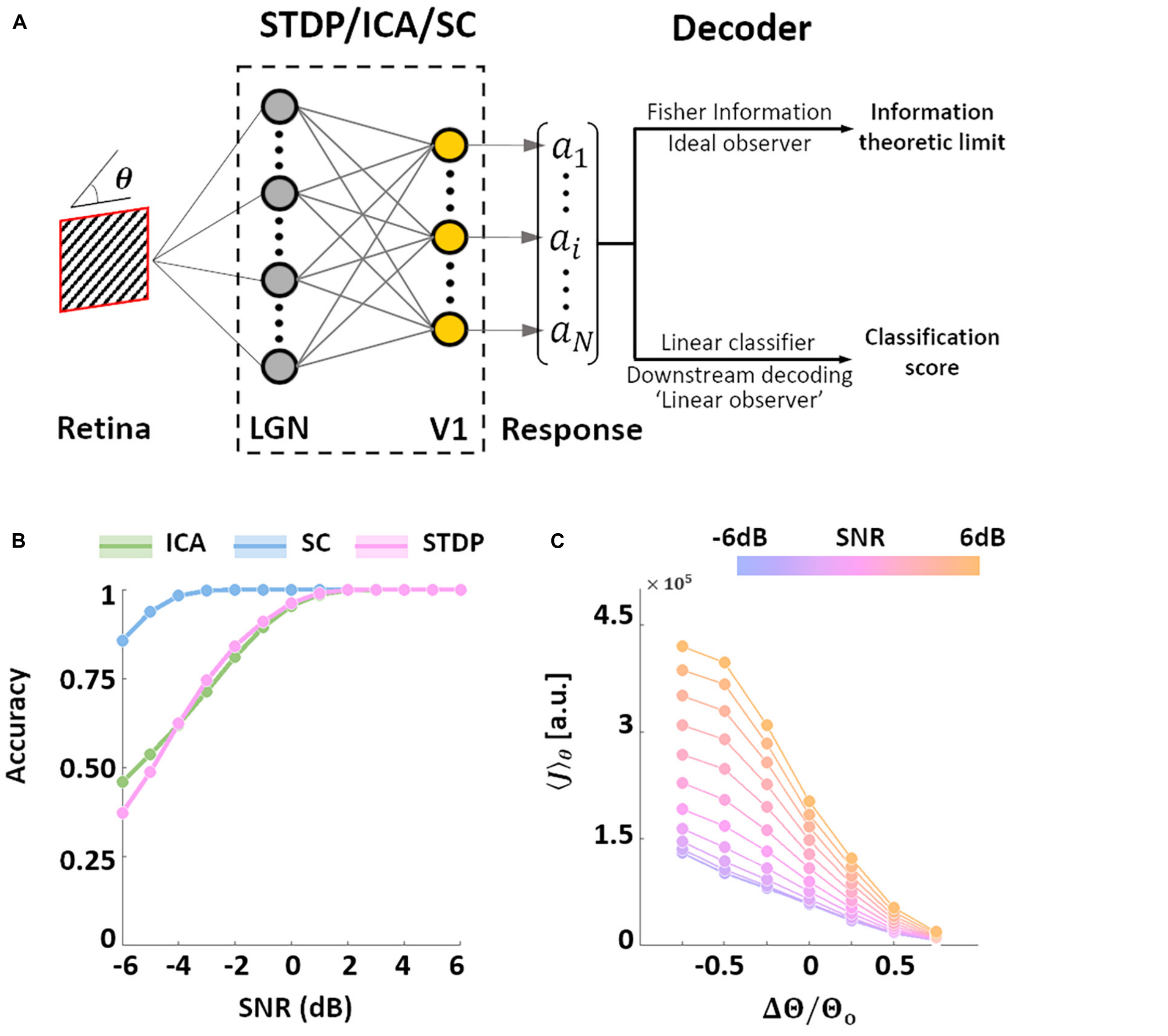


FIGURE 4 | Orientation decoding. **(A)** Retrieving encoded information. Sine-wave gratings (SWGs) with varying degrees of additive Gaussian noise were presented to the three models. The following question was then posed: how much information about the stimulus (in this case, the orientation) can be decoded from the population responses? The theoretical limit of the accuracy of such a decoder can be approximated by estimating the Fisher information (FI) in the responses. In addition, a linear decoder was also used to directly decode the population responses. This could be a downstream process which is linearly driven by the population activity, or a less-than-optimal "linear observer." **(B)** Linear decoding. The responses of each model were used to train a linear-discriminant classifier. The ordinate shows the accuracy (probability of correct classification) for each level of added noise (abscissa). Results for ICA are shown in green, SC in blue, and STDP in pink. **(C)** Post-training threshold variation in STDP. The SWG stimuli were also used to test STDP models with different values of the threshold parameter. The threshold was either increased (by 25, 50, or 75%) or decreased (by 25, 50, or 75%) with respect to the training threshold (denoted by θ_o). The abscissa denotes the relative change in threshold, and the ordinate denotes the estimated FI. The colour of the lines denotes the input SNR, which ranged from −6 dB (blue) to 6 dB (orange).

and macaque data showed much broader tunings (STDP: 15.1°, Macaque data: 19.1°). This was also reflected in the KL divergence of the macaque distribution from the three model distributions (ICA: 2.4 bits, SC: 3.5 bits, STDP: 0.29 bits). Thus, while the orientation tuning for STDP is much broader compared to ICA and SC, it is also closer to measurements in the macaque V1, indicating a better agreement with biology.

Decoding and Information Throughput

After characterising the encoding capacity of the models, we next probed the possible downstream implications of such codes. The biological goal of most neural code, in the end, is the generation of behaviour that maximises evolutionary fitness. However, due to the complicated neural apparatus that separates behaviour from early sensory processing, it is not straightforward (or at times, even possible) to analyse the interaction between the two. Bearing these limitations in mind, we employed two separate metrics to investigate this relationship. In both cases, the models were presented with oriented SWGs, followed by a decoding analysis of the resulting neural population activity (**Figure 4A**).

We examined the performance of a decoder built on linear discriminant classifiers (these classifiers assume fixed first-order correlations in the input). Such decoders can be interpreted as linearly driven feedforward populations downstream from the thalamo-recipient layer (the "V1" populations in the three models), or a simplified, "linear" observer. Not surprisingly the accuracy of the three models increases with the SNR. We found that SC was the most accurate of the three models under all tested noise-levels, while ICA and STDP showed very similar performances (**Figure 4B**). SC was also more robust to Gaussian noise compared to both ICA and STDP. A major difference between the three models tested in this study is that while ICA and SC are based on linear generative units, the STDP model has an intervening thresholding nonlinearity (Eq. 6). To test the effect of this thresholding on the information throughput of the STDP model, we ran simulations where, after training on natural images, the value of the threshold parameter in the STDP model was either increased or decreased (Eq. 11). The network was presented with SWGs (same stimuli as **Figure 4B**), and the average FI (Eq. 10) over the orientation parameter was estimated for each simulation condition. Note that in all simulations the model was first trained (i.e., synaptic learning using natural stimuli, see **Figure 1**) using the same "training" threshold, and the increase/decrease of the threshold parameter was imposed post-convergence. The FI increased for thresholds lower than the training threshold – possibly driven by an increase in the overall activity of the network. On the other hand, increasing the threshold led to lower FI due to the decreased bandwidth of neural activity. Thus, it is indeed possible to manipulate the information throughput of the spiking network by regulating the overall spiking activity in the network. This trend was found to occur robustly for all tested SNR values.

DISCUSSION

In this study, we showed that learning in a network with an abstract, rank-based STDP rule can predict biological findings at various scales. The FSVs of the converged RFs in the model show strong similarities with single-cell data measured in the macaque primary visual cortex, while the OTCs in the model closely predict measured population tuning.

Optimality in Biological Systems

In neuroscience, normative schemes are typically used to relate natural stimuli to an encoding hypothesis. Most normative encoding schemes optimise a generative reconstruction of the input by minimising an error metric (e.g., the L1 or L2 losses) over a given dataset. An alternative approach to studying stimulus encoding is through the use of process-based schemes which model known biophysical mechanisms at various levels of abstraction without making explicit assumptions about optimality. Traditionally, process-based or mechanistic schemes do not employ error metrics, and have been used to study fine-grained neuronal dynamics (Kang and Sompolinsky, 2001; Moreno-Bote et al., 2014; Harnack et al., 2015). On the other hand, normative schemes are employed to describe population-level characteristics (Olshausen and Field, 1997; van Hateren and van der Schaaf, 1998; Lee and Seung, 1999; Hoyer and Hyvärinen, 2000). In this study, we show that RFs predicted by a non-generative rank-based STDP rule are closer to electrophysiological measurements in the macaque V1 when compared to generatively optimal schemes such as ICA and SC. While this study only employs the classical variations of ICA and SC, subsequent work has demonstrated that similar suboptimalities in RF shape can also be obtained by generative models when biologically plausible nonlinearities such as thresholding operations (Rehn and Sommer, 2007; Rozell et al., 2008), or pointwise maxima operations (Puertas et al., 2010) are introduced. However, the abstract rank-based STDP model used here is free from generative optimisation and offers a much more biologically plausible, normative description of "learning"

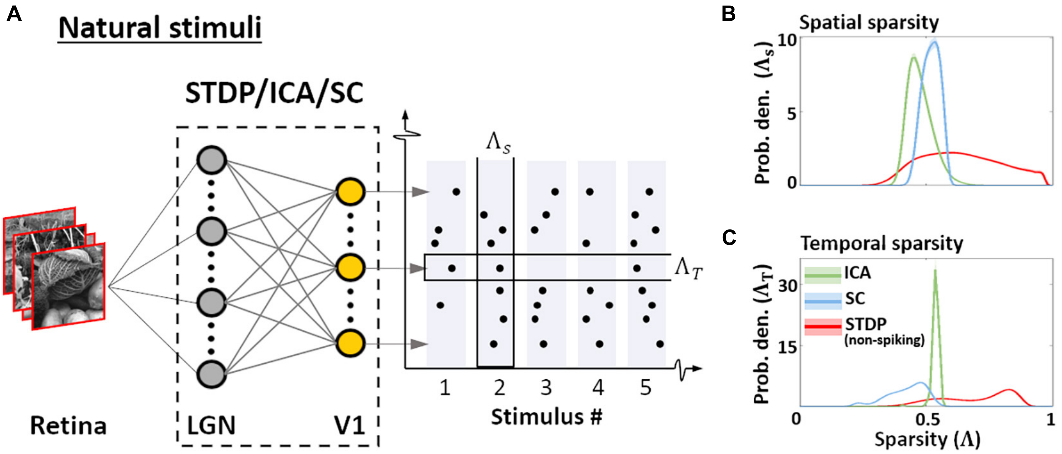


FIGURE 5 | Sparsity. **(A)** Sparsity indices. To estimate the sparsity of the non-spiking responses to natural stimuli, 10^4 patches (3° × 3° visual angle) randomly sampled from natural scenes were presented to the three models. Two measures of sparsity were defined: Spatial sparsity Index (Λ_S) was defined as the average sparsity of the activity of the entire neuronal ensemble, while Temporal sparsity Index (Λ_T) was defined as the average sparsity of the activity of single neurones to the entire input sequence. **(B)** Spatial sparsity. Estimated probability density of Λ_S for ICA (green), Sparse Coding (blue), and STDP (red). Λ_S varied between 0 (all units activate with equal intensity) and 1 (only 1 u/U activates) by definition. **(C)** Temporal sparsity. Estimated probability density of Λ_T, shown in a manner analogous to Λ_S (panel **B**). Λ_T also varied between 0 (homogeneous activity for the entire input sequence) and 1 (activity only for few inputs in the sequence).

through experience in the early visual system, where there is no sensory "ground truth" to generate errors from.

Note that while process-based models can predict suboptimalities observed in biological data, they cannot account for the theoretical insights offered by generative normative schemes. Local synaptic processes such as STDP can, in fact, be viewed as neural substrates for the overall synaptic optimisation employed by these schemes. The critique that gradient descent is inherently biologically implausible is being challenged by recent studies which frame error propagation and stochastic descent in terms of local, biologically plausible rules (see, e.g., Lillicrap et al., 2016; Melchior and Wiskott, 2019; Li, 2020). It has been demonstrated that local plasticity rules can, in fact, be adapted to describe various normative hypotheses about stimulus encoding (Savin et al., 2010; Brito and Gerstner, 2016). A growing number of insightful studies now employ hybrid encoding schemes which address multiple optimisation criteria (Perrinet and Bednar, 2015; Martinez-Garcia et al., 2017; Beyeler et al., 2019), often through local biologically realistic computation (Rozell et al., 2008; Savin et al., 2010; Zylberberg et al., 2011; Isomura and Toyoizumi, 2018).

Sparsity

Normative descriptions of the early visual system are grounded in the idea of efficiency – in terms of information transfer, and in terms of resource consumption. These assumptions, in turn, determine the behaviour of population responses to natural images. We quantified this behaviour by presenting the converged models with patches randomly sampled from the training dataset of natural images, and estimating the sparsity of the resulting activations using the Gini coefficient (Eq. 13; Hurley and Rickard, 2009). The sparsity was examined in two contexts (Barth and Poulet, 2012) as shown in **Figure 5A**. First, sparsity of the entire ensemble was estimated for each presented stimulus – this is a measure of how many neurones, on average, are employed by the ensemble to encode a given stimulus (Eq. 14). Second, the sparsity of individual neurones over the entire sequence of stimuli was estimated, allowing us to infer how frequently the features selected/encoded by the converged models occur in the sequence (Eq. 15). We denote the former as spatial sparsity (Λ_s), and the latter as temporal sparsity (Λ_t). For STDP, the indices were calculated for the membrane potential to facilitate comparison with ICA and SC activations. STDP membrane potential (red, **Figure 5B**) shows high variability in Λ_s, whereas ICA (green) and SC (blue) show much lower variance in comparison. This suggests that ICA and SC converge to features such that each image activates approximately equal number of units. On the other hand, the sparsity of the STDP neurones is more variable and stimulus-dependent, and likely driven by the relative probability of occurrence of specific features in the dataset – thus reflecting the Hebbian principal. ICA also exhibits a similar, small range for temporal sparsity Λ_t (**Figure 5C**) – suggesting that ICA encoding has uniform activation probability across its units. SC and STDP, however, show a much broader range of temporal sparsity across their units, with some units activating more frequently as compared to others.

Taken together, this suggests that the ICA encoding scheme converges to features such that the activation is distributed uniformly across the units, both for a given stimulus, and across multiple stimuli. This is likely to be driven by the objective of minimising reconstruction loss while maintaining minimal mutual information across the population. SC, on the other hand, equalises the probability of firing over the population for any given stimulus, but individual units may converge to features which occur more or less frequently. Once again, this behaviour is a consequence of the loss function which ensures that the network activity is sparse for each stimulus, but does not impose explicit constraints between the activity profile of individual units. As the STDP model is unsupervised and does not explicitly impose any generative loss function, we find high variability in both the spatial and temporal sparsity of STDP units. As shown in **Figure 4C**, this variability ensures that the information throughput of the network can be modulated by regulation of parameters such as the spiking threshold, even after the initial training.

Emerging Technologies and Process-Based Modelling in Neuroscience

Traditionally, detailed process-based models have suffered from constraints imposed by computational complexity, prohibitively long execution times which do not scale well for large networks, and hardware that is geared toward synchronous processing. On the other hand, most normative models can leverage faster computational libraries and architectures which have been developed over several decades, thereby leading to more efficient and scalable computation. However, with the growing availability of faster and more adaptable computing solutions such as neuromorphic hardware (event-based cameras, spike-based chips), and event-driven computational frameworks (e.g., Nengo: Bekolay et al., 2014; or Brian 2: Stimberg et al., 2019), implementations of such models are becoming increasingly accessible and scientifically tractable. These frameworks can be used not only to investigate detailed biophysical models or create biologically relevant machine and reinforcement learning pipelines, but to also investigate normative neuroscientific hypotheses which require unsupervised learning. In the future, we hope process-based modelling will be adopted more widely by cognitive and computational neuroscientists alike.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analysed in this study. This data can be found here: http://ringachlab.net/.

AUTHOR CONTRIBUTIONS

TC, BC, and TM conceptualised the study and interpreted the results. TC designed, programmed, and ran the simulations and wrote the original draft. All authors reviewed and revised the manuscript.

FUNDING

This work was funded by Fondation pour la Recherche Médicale (FRM: SPF20170938752, awarded to TC) and Agence Nationale de la Recherche (ANR-16-CE37-002-01, 3D3M, awarded to BC).

ACKNOWLEDGMENTS

We would like to thank Yves Trotter and Lionel Nowak for their critical feedback.

REFERENCES

Anderson, C., Van Essen, D., and Olshausen, B. (2005). "CHAPTER 3 - Directed visual attention and the dynamic control of information flow," in *Neurobiology of Attention*, eds L. Itti, G. Rees, and J. K. Tsotsos (Burlington, VT: Academic Press), 11–17. doi: 10.1016/B978-012375731-9/50007-0

Barth, A. L., and Poulet, J. F. A. (2012). Experimental evidence for sparse firing in the neocortex. *Trends Neurosci.* 35, 345–355. doi: 10.1016/j.tins.2012.03.008

Bekolay, T., Bergstra, J., Hunsberger, E., DeWolf, T., Stewart, T., Rasmussen, D., et al. (2014). Nengo: a Python tool for building large-scale functional brain models. *Front. Neuroinformatics* 7:48. doi: 10.3389/fninf.2013.00048

Bell, A., and Sejnowski, T. (1997). The "independent components" of natural scenes are edge filters. *Vision Res.* 37, 3327–3338. doi: 10.1016/S0042-6989(97)00121-1

Beyeler, M., Rounds, E., Carlson, K., Dutt, N., and Krichmar, J. (2019). Neural correlates of sparse coding and dimensionality reduction. *PLoS Comput. Biol.* 15:e1006908. doi: 10.1371/journal.pcbi.1006908

Brito, C. S. N., and Gerstner, W. (2016). Nonlinear Hebbian learning as a unifying principle in receptive field formation. *PLoS Comput. Biol.* 12:e1005070. doi: 10.1371/journal.pcbi.1005070

Bruce, N., Rahman, S., and Carrier, D. (2016). Sparse coding in early visual representation: from specific properties to general principles. *Neurocomputing* 171, 1085–1098. doi: 10.1016/j.neucom.2015.07.070

Caporale, N., and Dan, Y. (2008). Spike timing–dependent plasticity: a Hebbian learning rule. *Annu. Rev. Neurosci.* 31, 25–46. doi: 10.1146/annurev.neuro.31.060407.125639

Chauhan, T., Masquelier, T., Montlibert, A., and Cottereau, B. (2018). Emergence of binocular disparity selectivity through Hebbian learning. *J. Neurosci.* 38, 9563–9578. doi: 10.1523/JNEUROSCI.1259-18.2018

Delorme, A., Perrinet, L., and Thorpe, S. J. (2001). Networks of integrate-and-fire neurons using rank order coding B: spike timing dependent plasticity and emergence of orientation selectivity. *Neurocomputing* 38–40, 539–545. doi: 10.1016/S0925-2312(01)00403-9

Ecke, G. A., Papp, H. M., and Mallot, H. A. (2021). Exploitation of image statistics with sparse coding in the case of stereo vision. *Neural Netw.* 135, 158–176. doi: 10.1016/j.neunet.2020.12.016

Geisler, W. (2008). Visual perception and the statistical properties of natural scenes. *Annu. Rev. Psychol.* 59, 167–192. doi: 10.1146/annurev.psych.58.110405.085632

Gütig, R., Aharonov, R., Rotter, S., and Sompolinsky, H. (2003). Learning input correlations through nonlinear temporally asymmetric Hebbian plasticity. *J. Neurosci.* 23, 3697–3714.

Harnack, D., Pelko, M., Chaillet, A., Chitour, Y., and van Rossum, M. C. W. (2015). Stability of neuronal networks with homeostatic regulation. *PLoS Comput. Biol.* 11:e1004357. doi: 10.1371/journal.pcbi.1004357

Hoyer, P., and Hyvärinen, A. (2000). Independent component analysis applied to feature extraction from colour and stereo images. *Netw. Comput. Neural Syst.* 11, 191–210. doi: 10.1088/0954-898X_11_3_302

Hübener, M., and Bonhoeffer, T. (2014). Neuronal plasticity: beyond the critical period. *Cell* 159, 727–737. doi: 10.1016/j.cell.2014.10.035

Hunter, D., and Hibbard, P. (2015). Distribution of independent components of binocular natural images. *J. Vis.* 15:6. doi: 10.1167/15.13.6

Hurley, N., and Rickard, S. (2009). Comparing measures of sparsity. *IEEE Trans. Inf. Theory* 55, 4723–4741. doi: 10.1109/TIT.2009.2027527

Hyvärinen, A., and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Netw.* 13, 411–430. doi: 10.1016/S0893-6080(00)00026-5

Isomura, T., and Toyoizumi, T. (2018). Error-gated Hebbian rule: a local learning rule for principal and independent component analysis. *Sci. Rep.* 8:1835. doi: 10.1038/s41598-018-20082-0

Jones, J., and Palmer, L. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J. Neurophysiol.* 58, 1233–1258.

Kang, K., and Sompolinsky, H. (2001). Mutual Information of population codes and distance measures in probability space. *Phys. Rev. Lett.* 86, 4958–4961. doi: 10.1103/PhysRevLett.86.4958

Larsen, R. S., Rao, D., Manis, P. B., and Philpot, B. D. (2010). STDP in the developing sensory neocortex. *Front. Synaptic Neurosci.* 2:9. doi: 10.3389/fnsyn.2010.00009

Lee, D., and Seung, H. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 788–791. doi: 10.1038/44565

Li, H. L. (2020). Faster biological gradient descent learning. *ArXiv*[Preprint] ArXiv 200912745 Cs,

Lillicrap, T. P., Cownden, D., Tweed, D. B., and Akerman, C. J. (2016). Random synaptic feedback weights support error backpropagation for deep learning. *Nat. Commun.* 7:13276. doi: 10.1038/ncomms13276

Markram, H., Lübke, J., Frotscher, M., Sakmann, B., Hebb, D. O., Bliss, T. V. P., et al. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science* 275, 213–215. doi: 10.1126/science.275.5297.213

Martinez-Garcia, M., Martinez, L. M., and Malo, J. (2017). Topographic Independent Component Analysis reveals random scrambling of orientation in visual space. *PLoS One* 12:e0178345. doi: 10.1371/journal.pone.0178345

Masquelier, T. (2012). Relative spike time coding and STDP-based orientation selectivity in the early visual system in natural continuous and saccadic vision: a computational model. *J. Comput. Neurosci.* 32, 425–441. doi: 10.1007/s10827-011-0361-9

Masquelier, T., and Thorpe, S. J. (2007). Unsupervised learning of visual features through spike timing dependent plasticity. *PLoS Comput. Biol.* 3:e31. doi: 10.1371/journal.pcbi.0030031

Melchior, J., and Wiskott, L. (2019). Hebbian-Descent. *ArXiv*[Preprint] ArXiv190510585 Cs Stat. Available online at: http://arxiv.org/abs/1905.10585 [Accessed August 20, 2021],

Moreno-Bote, R., Beck, J., Kanitscheider, I., Pitkow, X., Latham, P., and Pouget, A. (2014). Information-limiting correlations. *Nat. Neurosci.* 17, 1410–1417. doi: 10.1038/nn.3807

Olshausen, B., and Field, D. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, 607–609. doi: 10.1038/381607a0

Olshausen, B., and Field, D. (1997). Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Res.* 37, 3311–3325. doi: 10.1016/S0042-6989(97)00169-7

Perrinet, L. U., and Bednar, J. A. (2015). Edge co-occurrences can account for rapid categorization of natural versus animal images. *Sci. Rep.* 5:11400. doi: 10.1038/srep11400

Puertas, J., Bornschein, J., and Lücke, J. (2010). "The maximal causes of natural scenes are edge filters," in *Advances in Neural Information Processing Systems*, eds J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta (Red Hook, NY: Curran Associates, Inc), 1939–1947.

Raichle, M. (2010). Two views of brain function. *Trends Cogn. Sci.* 14, 180–190. doi: 10.1016/j.tics.2010.01.008

Rehn, M., and Sommer, F. (2007). A network that uses few active neurones to code visual input predicts the diverse shapes of cortical receptive fields. *J. Comput. Neurosci.* 22, 135–146. doi: 10.1007/s10827-006-0003-9

Ringach, D. (2002). Spatial structure and symmetry of simple-cell receptive fields in Macaque primary visual cortex. *J. Neurophysiol.* 88, 455–463.

Ringach, D., and Shapley, R. (2004). Reverse correlation in neurophysiology. *Cogn. Sci.* 28, 147–166. doi: 10.1207/s15516709cog2802_2

Ringach, D., Shapley, R., and Hawken, M. (2002). Orientation selectivity in Macaque V1: diversity and laminar dependence. *J. Neurosci.* 22, 5639–5651. doi: 10.1523/JNEUROSCI.22-13-05639.2002

Rozell, C., Johnson, D., Baraniuk, R., and Olshausen, B. (2008). Sparse coding via thresholding and local competition in neural circuits. *Neural Comput.* 20, 2526–2563. doi: 10.1162/neco.2008.03-07-486

Savin, C., Joshi, P., and Triesch, J. (2010). Independent component analysis in spiking neurons. *PLoS Comput. Biol.* 6:e1000757. doi: 10.1371/journal.pcbi.1000757

Stimberg, M., Brette, R., and Goodman, D. F. (2019). Brian 2, an intuitive and efficient neural simulator. *eLife* 8:e47314. doi: 10.7554/eLife.47314

van Hateren, J., and van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. Biol. Sci.* 265, 359–366. doi: 10.1098/rspb.1998.0303

Wandell, B., and Smirnakis, S. (2010). Plasticity and stability of visual field maps in adult primary visual cortex. *Nat. Rev. Neurosci.* 10:873. doi: 10.1038/nrn2741

Zhaoping, L. (2006). Theoretical understanding of the early visual processes by data compression and data selection. *Netw. Comput. Neural Syst.* 17, 301–334. doi: 10.1080/09548980600931995

Zylberberg, J., Murphy, J. T., and DeWeese, M. R. (2011). A Sparse coding model with synaptically local plasticity and spiking neurons can account for the diverse shapes of V1 simple cell receptive fields. *PLoS Comput. Biol.* 7:e1002250. doi: 10.1371/journal.pcbi.1002250

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Artificial Intelligence

METHODS
published: 08 December 2021
doi: 10.3389/frai.2021.798962

Accelerating Hyperparameter Tuning in Machine Learning for Alzheimer's Disease With High Performance Computing

Fan Zhang[1,2]*, *Melissa Petersen*[1,2], *Leigh Johnson*[1,3], *James Hall*[1,3] *and Sid E. O'Bryant*[1,3]*

[1] *Institute for Translational Research, University of North Texas Health Science Center, Fort Worth, TX, United States,* [2] *Department of Family Medicine, University of North Texas Health Science Center, Fort Worth, TX, United States,* [3] *Department of Pharmacology and Neuroscience, University of North Texas Health Science Center, Fort Worth, TX, United States*

OPEN ACCESS

Edited by:
Reza Lashgari,
Shahid Beheshti University, Iran

Reviewed by:
Saman Sarraf,
Institute of Electrical and Electronics Engineers, United States
Xiaojie Zhao,
Rosalind Franklin University of Medicine and Science, United States
Daiwei Wan,
The First Affiliated Hospital of Soochow University, China

****Correspondence:***
Fan Zhang
fan.zhang@unthsc.edu
Sid E. O'Bryant
sid.obryant@unthsc.edu

Specialty section:
This article was submitted to Machine Learning and Artificial Intelligence,
a section of the journal Frontiers in Artificial Intelligence

Received: *20 October 2021*
Accepted: *15 November 2021*
Published: *08 December 2021*

Citation:
Zhang F, Petersen M, Johnson L, Hall J and O'Bryant SE (2021) Accelerating Hyperparameter Tuning in Machine Learning for Alzheimer's Disease With High Performance Computing.
Front. Artif. Intell. 4:798962.
doi: 10.3389/frai.2021.798962

Driven by massive datasets that comprise biomarkers from both blood and magnetic resonance imaging (MRI), the need for advanced learning algorithms and accelerator architectures, such as GPUs and FPGAs has increased. Machine learning (ML) methods have delivered remarkable prediction for the early diagnosis of Alzheimer's disease (AD). Although ML has improved accuracy of AD prediction, the requirement for the complexity of algorithms in ML increases, for example, hyperparameters tuning, which in turn, increases its computational complexity. Thus, accelerating high performance ML for AD is an important research challenge facing these fields. This work reports a multicore high performance support vector machine (SVM) hyperparameter tuning workflow with 100 times repeated 5-fold cross-validation for speeding up ML for AD. For demonstration and evaluation purposes, the high performance hyperparameter tuning model was applied to public MRI data for AD and included demographic factors such as age, sex and education. Results showed that computational efficiency increased by 96%, which helped to shed light on future diagnostic AD biomarker applications. The high performance hyperparameter tuning model can also be applied to other ML algorithms such as random forest, logistic regression, xgboost, etc.

Keywords: machine learning, hyperparameter tuning, alzheimer's disease, high performance computing, support vector machine

INTRODUCTION

Alzheimer's disease (AD) is the most common form of dementia. In 2020, as many as 5.8 million Americans were living with AD. This number is projected to nearly triple by 2060 (Prevention, 2021). Machine Learning (ML) methods for AD and AD Related Dementias (ADRDs) is growing faster than ever before (Waring et al., 2008; Magnin et al., 2009; O'Bryant et al., 2011a; O'Bryant et al., 2011b; O'Bryant et al., 2013; O'Bryant et al., 2014; Weiner et al., 2015; O'Bryant et al., 2016; O'Bryant et al., 2017; Grassi et al., 2018; Hampel et al., 2018; O'Bryant et al., 2018; Stamate et al., 2019; Zetterberg and Burnham, 2019; Zhang and Sejdić, 2019; Franzmeier et al., 2020; O'Bryant et al., 2020; Rodriguez et al., 2021). A PubMed search using keywords of AD and ML showed that the number of publications related to ML for AD has increased by 146 percent from just two in 2006 to 294 in 2020. For example, O'Bryant et al. developed a Support Vector Machine (SVM) model with 398 plasma samples obtained from adults with Down syndrome to predict incident mild cognitive impairment

TABLE 1 | Talon3 computer nodes.

Quanity	Memory (GB)	Cores	Description
192	64	28	Dell PowerEdge C6320 server with two 2.4 GHz Intel Xeon E5-2680 v4 14-core processors
75	32	16	Dell PowerEdge R420 server with two 2.1 GHz Intel Xeon E5-2450 eight-core processors
64	64	16	Dell PowerEdge R420 server with two 2.1 GHz Intel Xeon E5-2450 eight-core processors
8	512	32	Dell PowerEdge R720 server with four 2.4 GHz Intel Xeon E5-4640 eight-core processors
16	64	28	Dell PowerEdge R730 server with two 2.4 GHz Intel Xeon E5-2680 v4 14-core processors and two Nvidia Tesla K80 GPUS (4,992 GPU cores/card)

(MCI) (AUC = 0.92) and incident AD (AUC = 0.88) (O'Bryant et al., 2020). O'Bryant et al. also developed a precision medicine model for targeted NSAID therapy in AD based on data collected from a previously conducted clinical trial. This work included 351 patients with mild-to-moderate AD that were enrolled into one of three trial arms: 1-year exposure to rofecoxib (25 mg once daily), naproxen (220 mg twice-daily) and placebo. The SVM model yielded 98% theragnostic accuracy in the rofecoxib arm and 97% accuracy in the naproxen arm, respectively (O'Bryant et al., 2018). Magnin et al. also built a SVM model with three-dimensional T1-weighted MR images of 16 patients with AD and 22 elderly controls and obtained a 94.5% mean accuracy for AD with a mean specificity of 96.6% and mean sensitivity of 91.5% (Magnin et al., 2009).

Improving speed and capability is a huge issue in applying ML to AD. It is possible that certain ML computations can be delayed because of the large amount of time to iteration that is required, for example, time to train with hyperparameters tuning. High performance computing (HPC) can be used to help meet the increasing demands for the speed and capabilities of processing ML for AD (Eddelbuettel, 2021). With the fast processing ability of high-performance computing systems, faster results can be delivered, which in turn would not only speed up finding the optimal hyperparameters for AD with ML models but would also identify opportunities to fix issues in hyperparameter tuning for AD ML models.

In this paper, based on the multicores parallel structure of Talon3 high performance computing provided by the University of North Texas, we present a high performance computing workflow to support our parallel SVM hyperparameter tuning. We applied the multicore high performance SVM hyperparameter tuning to 100 times repeated 5-fold cross-validation model for longitudinal MRI data of 150 subjects with 64 subjects classified as demented and 86 subjects classified as nondemented. The computational time was dramatically reduced by up to 96% for the high performance SVM hyperparameter tuning model. The multicores parallel structure and the high performance SVM hyperparameter tuning model can be used for other ML applications.

MATERIALS AND METHODS

Parallel Structure

We used the Talon3 system (**Table 1**) provided by University of North Texas for this study due to its convenient computing services, which allowed us to import/export/execute large and complex parallel ML. The hardware configuration of the Talon3 contains the following: more than 8,300 CPU cores, 150,000 GPU cores, Mellanox FDR InfiniBand network, and over 1.4 Petabytes of Lustre File Storage. The amount of AD data necessary for performing ML with a PC workstation is massive. For example, in one study with 300 samples, to process just the 100 times repeated 5-fold cross-validation for hyperparameters tuning with SVM, it would require about 3 h of consecutive CPU time and 12 GB of storage with a local computer.

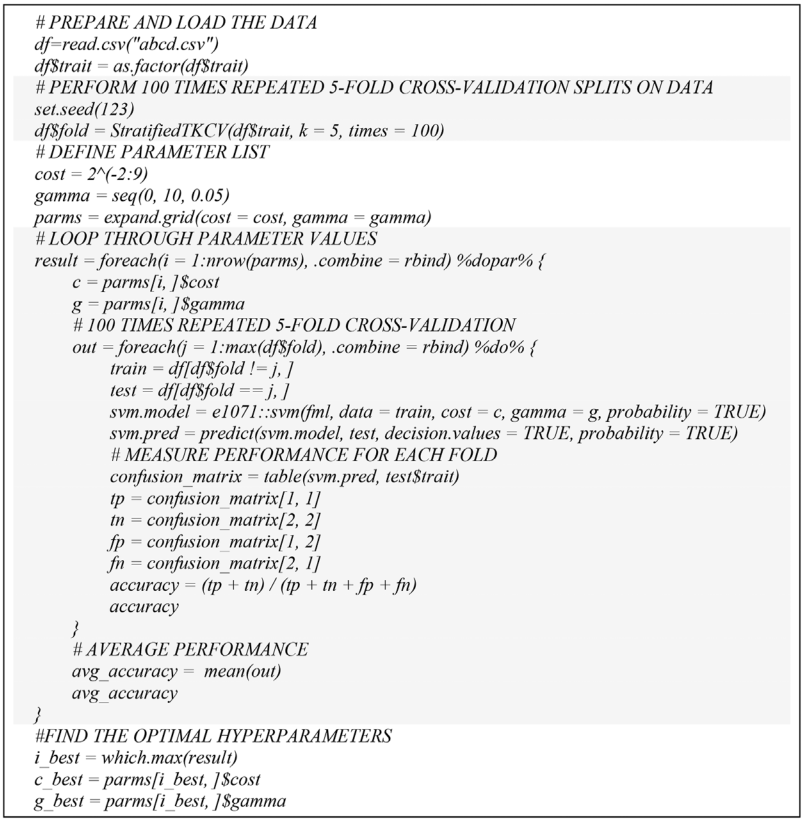
TABLE 2 | Performance for testing set after hyperparameter tuning.

	Actual demented	Actual nondemented
Predicted demented	9	1
Predicted nondemented	3	16
Precision/PPV	90.00%	
Accuracy	86.21%	
Sensitivity	75.00%	
Specificity	94.12%	
NPV	84.21%	
AUC	90.80%	
PPV12	63.49%	
NPV12	96.50%	

For parallel computing, the Talon3 provides several options including: SNOW, Rmpi, and multicore. We chose multicore because it executes parallel tasks on a single node as opposed to multiple nodes and the level of flexibility is higher than the other two options. For multicore parallel programming, submitting high performance ML includes two parts: a shell script and an R script. The shell script we submitted for multicore is for a single node with 28 cores in C6320. And for the R script high performance ML, we used doParallel (Michelle Wallig et al., 2020; Eddelbuettel, 2021) and foreach (Michelle Wallig and Steve, 2020; Eddelbuettel, 2021) packages.

Parallel SVM Hyperparameter Tuning

Based on the above parallel structure in Talon3 and doParallel package, we developed a high performance computing workflow to support our parallel SVM hyperparameter tuning (**Figure 1**). We used a grid search approach to find the best model parameters in terms of accuracy. This procedure mainly contains three steps: 1) define a grid to vary cost and gamma, 2) perform 100 times



```
# PREPARE AND LOAD THE DATA
df=read.csv("abcd.csv")
df$trait = as.factor(df$trait)
# PERFORM 100 TIMES REPEATED 5-FOLD CROSS-VALIDATION SPLITS ON DATA
set.seed(123)
df$fold = StratifiedTKCV(df$trait, k = 5, times = 100)
# DEFINE PARAMETER LIST
cost = 2^(-2:9)
gamma = seq(0, 10, 0.05)
parms = expand.grid(cost = cost, gamma = gamma)
# LOOP THROUGH PARAMETER VALUES
result = foreach(i = 1:nrow(parms), .combine = rbind) %dopar% {
    c = parms[i, ]$cost
    g = parms[i, ]$gamma
    # 100 TIMES REPEATED 5-FOLD CROSS-VALIDATION
    out = foreach(j = 1:max(df$fold), .combine = rbind) %do% {
        train = df[df$fold != j, ]
        test = df[df$fold == j, ]
        svm.model = e1071::svm(fml, data = train, cost = c, gamma = g, probability = TRUE)
        svm.pred = predict(svm.model, test, decision.values = TRUE, probability = TRUE)
        # MEASURE PERFORMANCE FOR EACH FOLD
        confusion_matrix = table(svm.pred, test$trait)
        tp = confusion_matrix[1, 1]
        tn = confusion_matrix[2, 2]
        fp = confusion_matrix[1, 2]
        fn = confusion_matrix[2, 1]
        accuracy = (tp + tn) / (tp + tn + fp + fn)
        accuracy
    }
    # AVERAGE PERFORMANCE
    avg_accuracy =  mean(out)
    avg_accuracy
}
#FIND THE OPTIMAL HYPERPARAMETERS
i_best = which.max(result)
c_best = parms[i_best, ]$cost
g_best = parms[i_best, ]$gamma
```

FIGURE 1 | Pseudo code for parallel SVM hyperparameter tuning.

repeated 5-fold cross-validation splits on training data, and 3) tune the cost and gamma of the SVM model.

100 Times Repeated 5-Fold Cross-Validation

A single run of the 5-fold cross-validation (O'Bryant et al., 2019) may result in a noisy estimate of model parameters. We adopted 100 times repeated 5-fold cross-validation (Kublanov et al., 2017) to improve the estimation of optimal parameters of the ML model. This involves simply repeating the cross-validation procedure 100 times and reporting the mean performance across all folds from all runs. This mean performance is then used for the determination of optimal parameters.

Metrics

The following eight measurements were involved in our evaluation: 1) Sensitivity (also called recall), the proportion of actual positive pairs that are correctly identified; 2) Specificity, the proportion of negative pairs that are correctly identified; 3) Precision, the probability of correct positive prediction; 4) Accuracy, the proportion of correctly predicted pairs; 5) Area Under the Curve; 6) Negative Predictive Value (NPV), the probability that subjects with a negative screening test truly don't have the disease; 7) Negative Predictive Value at base rate of 12% (NPV12); and 8) Positive Predictive Value at base rate of 12% (PPV12).

RESULTS

We downloaded open access longitudinal MRI data available on nondemented and demented older adults (Marcus et al., 2010a). The dataset consisted of longitudinal MRI data from 150 subjects aged 60 to 96. 72 of the subjects were classified as "nondemented" throughout the study. 64 of the subjects were classified as "demented" at the initial visit and remained so throughout the study. 14 subjects were classified as "nondemented" at the initial visit and were subsequently characterized as "demented"at a later study visit. For each subject, three to four individual T1-weighted magnetization prepared rapid gradient-echo (MP-RAGE) images were acquired in a single imaging session (Marcus et al., 2010b). The subject-independent model we developed for parallel hyperparameter tuning is not based on a classifier trained for each subject individually. We chose the following five imaging and clinical variables to predict the status of AD: SES (Socioeconomic Status), MMSE (Mini Mental State Examination), eTIV (Estimated Total Intracranial Volume), nWBV (Normalize Whole Brain Volume), and ASF (Atlas Scaling Factor). Measurements of these variables in this cohort

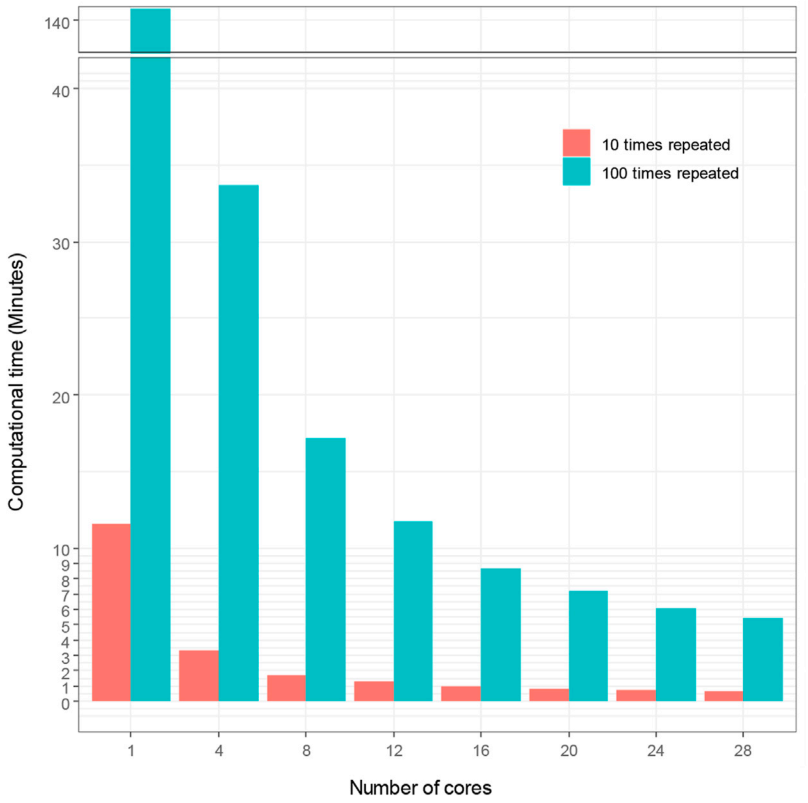


FIGURE 2 | Computational time vs. number of cores with SVM modeling.

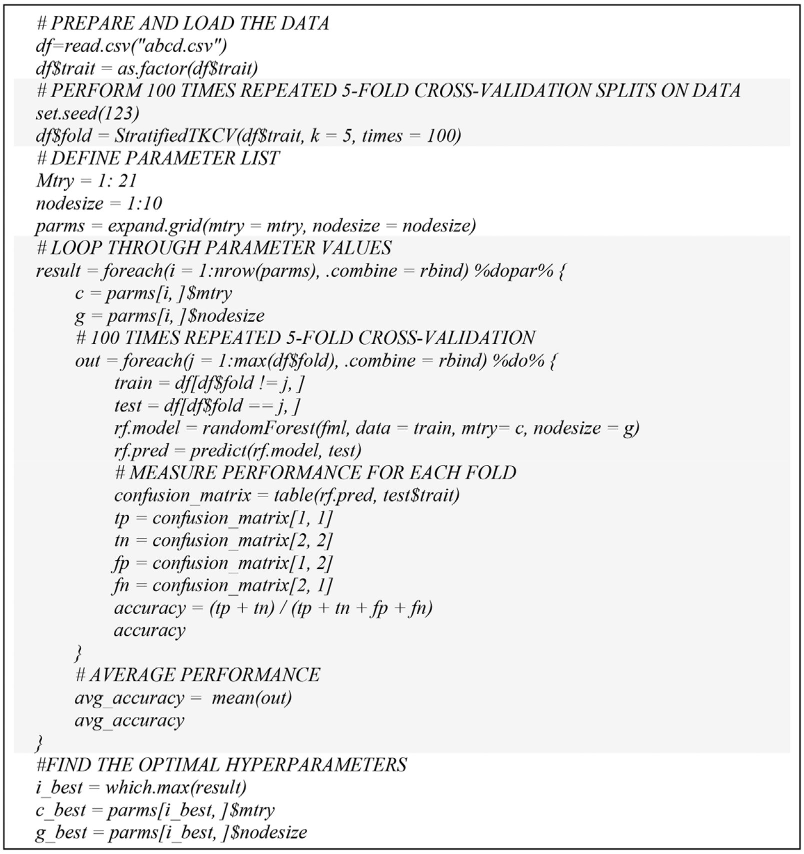
including clinical dementia rating scale (CDR), nWBV, eTIV, ASF, etc. have been previously described elsewhere (Marcus et al., 2010b). Three demographic variables: Sex, Age, and Edu (Years of education) were also added as covariates.

We used sbatch commands to submit shell scripts and R scripts for comparing computational time for hyperparameter tuning under different number of cores and repeated times of 5-fold cross-validation in Talon3. With SVM modeling, we demonstrate in **Figure 2** how the number of cores affects the computational time of hyperparameter tuning, which shows that the computational time decreases proportionally as the number of cores increases. **Figure 2** also demonstrates that the repeated times of the 5-fold cross-validation algorithm for hyperparameter affects the computational time. The computational time increased proportionally with the increasing repeated times. The time spent initially for the hyperparameter tuning without using high performance computing is very large (140.73 min for t = 100; 11.58 min for t = 10). The computational time decreased to 5.44 and 0.67 min, for t = 100 and for t = 10 respectively, when we used 28 cores to accelerate hyperparameter tuning in ML. We thereby reduced computational time by up to 96%, with high performance ML model.

The optimal hyperparameters we obtained are the same for all runs (gamma = 0.005, cost = 32). We used grid search method and set boundary for the two parameters: cost and gamma as suggested in the paper (Hsu et al., 2003) where fine grid search was on cost = (2, 32) and gamma = [2(−7), 2(−3)]. We extended the cost boundary to (0.25, 512) and the gamma boundary to (0, 10) to catch as much change as possible. Variables importance under the SVM model with the optimal hyperparameters shows that the MMSE, nWBV, and SES are leading variables in predicting dementia (AD) status. Out of the three demographic variables, education was shown to be less important for the SVM model than Age and Sex.

With the optimal hyperparameters, the average performance that the SVM model achieved for a testing set of 12 Demented and 17 Nondemented is reported on below for both 100 times repeated 5 -fold cross-validation and 10 times repeated 5 -fold cross-validation. The performance (**Table 2**) is slightly higher than previously reported at https://www.kaggle.com/hyunseokc/detecting-early-alzheimer-s, which achieved accuracy = 0.82, sensitivity = 0.70, and AUC = 0.82 for SVM. Our results show that the high performance SVM hyperparameter tuning workflow that we presented can significantly reduce computational time while maintaining the necessary accuracy.

In order to demonstrate the extensibility of our hyperparameter tuning workflow to other ML models, we also followed the SVM hyperparameter tuning workflow (**Figure 3**) and adopted random forest into our parallel hyperparameter tuning workflow (**Figure 4**). We obtained consistent results that the computation time for hyperparameter tuning of random forest was also remarkedly reduced (**Figure 5**). The computational time was reduced from 47.67 to 2.24 min by



```
# PREPARE AND LOAD THE DATA
df=read.csv("abcd.csv")
df$trait = as.factor(df$trait)
# PERFORM 100 TIMES REPEATED 5-FOLD CROSS-VALIDATION SPLITS ON DATA
set.seed(123)
df$fold = StratifiedTKCV(df$trait, k = 5, times = 100)
# DEFINE PARAMETER LIST
Mtry = 1: 21
nodesize = 1:10
parms = expand.grid(mtry = mtry, nodesize = nodesize)
# LOOP THROUGH PARAMETER VALUES
result = foreach(i = 1:nrow(parms), .combine = rbind) %dopar% {
      c = parms[i, ]$mtry
      g = parms[i, ]$nodesize
      # 100 TIMES REPEATED 5-FOLD CROSS-VALIDATION
      out = foreach(j = 1:max(df$fold), .combine = rbind) %do% {
            train = df[df$fold != j, ]
            test = df[df$fold == j, ]
            rf.model = randomForest(fml, data = train, mtry= c, nodesize = g)
            rf.pred = predict(rf.model, test)
            # MEASURE PERFORMANCE FOR EACH FOLD
            confusion_matrix = table(rf.pred, test$trait)
            tp = confusion_matrix[1, 1]
            tn = confusion_matrix[2, 2]
            fp = confusion_matrix[1, 2]
            fn = confusion_matrix[2, 1]
            accuracy = (tp + tn) / (tp + tn + fp + fn)
            accuracy
      }
      # AVERAGE PERFORMANCE
      avg_accuracy =  mean(out)
      avg_accuracy
}
#FIND THE OPTIMAL HYPERPARAMETERS
i_best = which.max(result)
c_best = parms[i_best, ]$mtry
g_best = parms[i_best, ]$nodesize
```

FIGURE 3 | Pseudo code for parallel RF hyperparameter tuning.

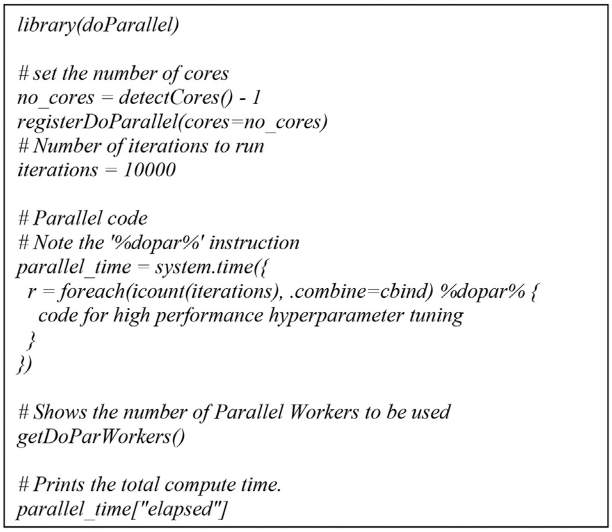
95% and from 5.25 min to 18.17 s by 94%, for $t = 100$ and $t = 10$ respectively.

We also tested the adaptability of our hyperparameter tuning workflow to the Texas Alzheimer's Research and Care Consortium (TARCC) dataset (Zhang et al., 2021). The TARCC dataset contains a total of 300 cases (150 AD cases; 150 Normal Control cases). Each subject (at one of the five participating TARCC sites) undergoes an annual standardized assessment, which includes a medical evaluation, neuropsychological testing, and a blood draw. The same blood-based biomarkers in (Zhang et al., 2021) were used as features for parallel hyperparameter tuning. Even when adopting a new TARCC dataset into our parallel hyperparameter tuning workflow (**Figure 1**), we obtained consistent results, which showed that the computation time for the hyperparameter tuning of the new TARCC dataset was also remarkedly reduced by about 96%. The computational time was reduced from 311.5 to 12.48 h by 96% and from 34.03 to 1.6 h by 95%, for $t = 100$ and $t = 10$ respectively.

DISCUSSIONS

HPC advances have successfully helped scientists and researchers to achieve various breakthrough innovations in the field of Omics-medicine, technology, retail, banking and so on (Merelli et al., 2014). For example, HPC has been applied to Next Generation Sequencing that is extremely data-intensive and needs ultra-powerful workstations to process the ever-growing data (Schmidt and Hildebrandt, 2017). Hyperparameter tuning component of ML can be a high-performance computing problem as it requires a large amount of computation and data motion. ML requires a computationally-intensive grid search and lots of computational power to help enable faster tuning cycles. Introducing HPC to ML can take advantage of high volumes of data as well as speed up the process of hyperparameter tuning.

Therefore, we presented a parallel hyperparameter tuning workflow with HPC to exploit modern parallel infrastructures to execute large-scale calculations by simultaneously using multiple compute resources. The rationales are 1) the foreach package that the workflow is based on supports parallel execution and provides a new looping construct for executing R code repeatedly. Specifically, a problem is broken into discrete parts that can be solved concurrently and an overall control/coordination mechanism is employed; 2) the foreach package can be used with a variety of different parallel computing systems, include NetWorkSpaces and snow; and 3) foreach can be used with iterators, which allows the data to be specified in a very flexible way.



```
library(doParallel)

# set the number of cores
no_cores = detectCores() - 1
registerDoParallel(cores=no_cores)
# Number of iterations to run
iterations = 10000

# Parallel code
# Note the '%dopar%' instruction
parallel_time = system.time({
  r = foreach(icount(iterations), .combine=cbind) %dopar% {
    code for high performance hyperparameter tuning
  }
})

# Shows the number of Parallel Workers to be used
getDoParWorkers()

# Prints the total compute time.
parallel_time["elapsed"]
```

FIGURE 4 | R script for parallel computing.

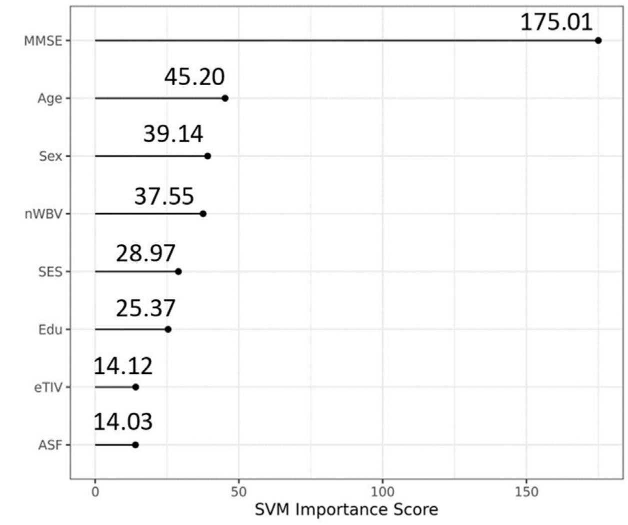
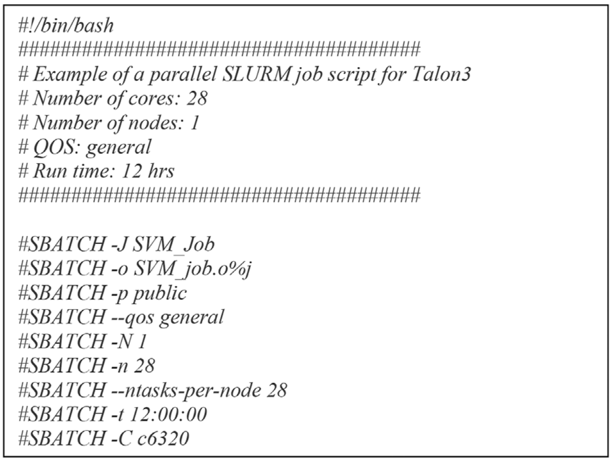


FIGURE 5 | Variable importance of the eight variables.

The multicore high performance SVM hyperparameter tuning workflow we presented is hardware-agnostic and can be used in HPCs of most U.S. universities or commerical clouds for example, Amazon AWS, Microsoft Azure, Google Cloud, etc. Before executing the multicores high performance SVM hyperparameter tuning, R package (V4.0.3, Linux) and doParallel and foreach libraries should be installed successfully, which are met for HPCs in most U.S. universities or commerical clouds. There are mainly two of the most popular job schedulers used for requesting resources allocations on a multi user cluster: 1) the Simple Linux Utility for Resource Management (Slurm) and 2) the Portable Batch System (PBS). In **Figure 6**, we described the shell script for parallel computing for the Slurm system. Similary a shell script for parallel computing for PBS system is as followed.



```
#!/bin/bash
########################################
# Example of a parallel SLURM job script for Talon3
# Number of cores: 28
# Number of nodes: 1
# QOS: general
# Run time: 12 hrs
########################################

#SBATCH -J SVM_Job
#SBATCH -o SVM_job.o%j
#SBATCH -p public
#SBATCH --qos general
#SBATCH -N 1
#SBATCH -n 28
#SBATCH --ntasks-per-node 28
#SBATCH -t 12:00:00
#SBATCH -C c6320
```

FIGURE 6 | Shell script for parallel computing.

The multicore high performance SVM hyperparameter tuning workflow significantly reduced computational time while maintaining a consistent detection accuracy. The workflow was diagrammed through a multicore computing pseudo code using the doParallel package in R for high performance hyperparameter tuning. The basic idea of multicore computing is to allow a single program, in this case R, to run multiple threads simultaneously in order to reduce the "walltime" required for completion. The doParallel package in R is one of several "parallel backends" for the foreach. It establishes communication between multiple cores, even on different physical "nodes" linked by network connections. The foreach function evaluates an expression for each value of the counter (iterator) "case". The %dopar% operator is used to execute the code in parallel. Using %do% instead would lead to sequential computation by the primary process. When parallelizing nesting for loops, there is always a question of which loop to parallelize. If the task and number of iterations vary in size, then it's really hard to know which loop to parallelize. We parallelized the outer loop in our SVM hyperparameter tuning because this would result in larger individual tasks, and larger tasks can often be performed more efficiently than smaller tasks. The hyperparameter tuning could be parallelized at the inner loop also if the outer loop doesn't have many iterations and the tasks are already large.

The multicores high performance hyperparameter tuning workflow can also be used for other ML such as random forest, logistic regression, xgboost, etc. For example, we demonstrated that a random forest model can be adopted into our parallel hyperparameter tuning model (**Figure 3**) and the results we obtained were consistent in that the computation time for hyperparameter tuning of random forest models were remarkedly reduced (**Figure 7**). In the future, we plan to use Rmpi library to create multinodes parallel computing workflow for hyperparameter tuning when Talon3 supports multinodes parallel computing to run R script.

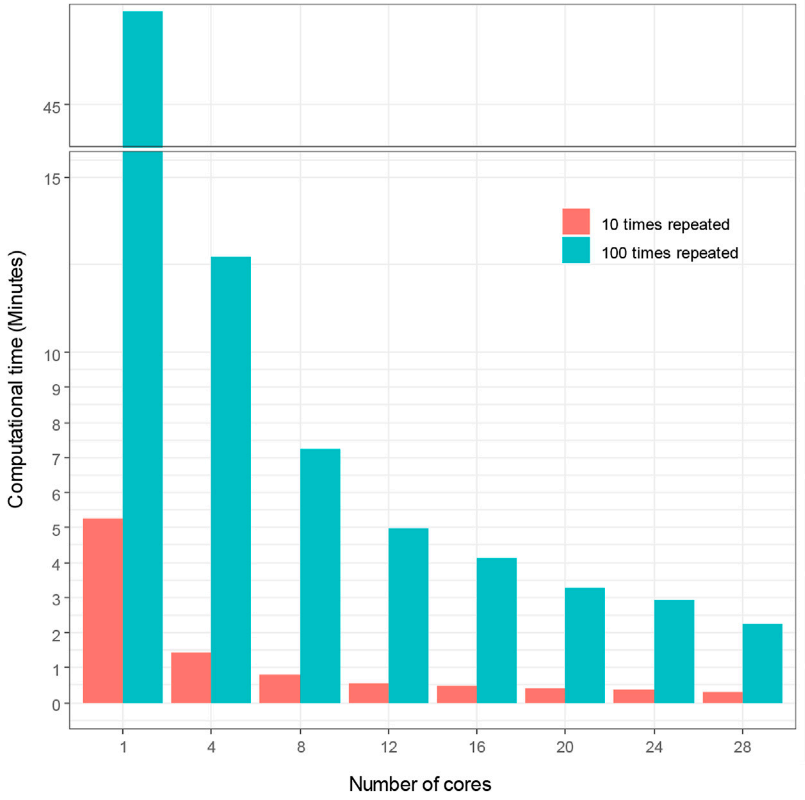


FIGURE 7 | Computational time vs. number of cores with RF modeling.

Our optimal hyperparameter model also showed that MMSE, Age, Sex, nWBV, and SES are important variables in AD diagnosis, which is consistent with previous findings. For example, Arevalo-Rodriguez et al. found that baseline MMSE scores can achieve a sensitivity of 76% and specificity of 94% for predicting conversion from MCI to dementia (in general) and a sensitivity of 89% and specificity of 90% for predicting conversion from MCI to AD dementia (Arevalo-Rodriguez et al., 2015). Advanced age and sex are two of the most prominent risk factors for dementia. Females are more likely to be susceptible for developing AD dementia than males (Podcasy and Epperson, 2016). Podcasy at Penn PROMOTES Research on Sex in Health and examined sex and gender differences in the development of dementia and suggested that researchers should consider sex as a biological variable for dementia research (Podcasy and Epperson, 2016). Rose et al. evaluated the combination of cerebrospinal fluid biomarkers with education and normalized whole-brain volume (nWBV) to predict incident cognitive impairment (Roe et al., 2011). They concluded that time to incident of cognitive impairment is moderated by education and nWBV for individuals with normal cognition had higher levels of cerebrospinal fluid tau and ptau at baseline (Roe et al., 2011). Khan et al. and Leong et al. assessed the role of various features on the prognosis of AD, and found that sex, age, MMSE, nWBV, and SES were significantly associated with and made an impact on the occurrence of AD (Leong and Abdullah, 2019; Khan and Zubair, 2020).

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.oasis-brains.org.

AUTHOR CONTRIBUTIONS

Conception and design of study: FZ and SO'B. Acquisition, and analysis of data: SO'B, FZ, and MP. Drafting manuscript or figures: SO'B, FZ, LJ, MP, and JH.

FUNDING

Research reported in this publication was supported by the National Institute on Aging of the National Institutes of Health under Award Number R01AG058537, R01AG054073, R01AG058533, and 3R01AG058533-02S1. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

ACKNOWLEDGMENTS

The authors acknowledge the North Texas Scientific Computing, a division under the office of the CIO at University of North Texas, for providing Talon computational and data analytics resources that have contributed to the research results reported within this paper. Data were provided by OASIS: Longitudinal: Principal Investigators: D. Marcus, R. Buckner, J. Csernansky, J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382.

REFERENCES

Arevalo-Rodriguez, I., Smailagic, N., Roqué I Figuls, M., Ciapponi, A., Sanchez-Perez, E., Giannakou, A., et al. (2015). Mini-Mental State Examination (MMSE) for the Detection of Alzheimer's Disease and Other Dementias in People with Mild Cognitive Impairment (MCI). *Cochrane Database Syst. Rev.* 2015, CD010783. doi:10.1002/14651858.CD010783.pub2

Eddelbuettel, D. (2021). *Parallel Computing R. A. Brief Review* 13, e1515. doi:10.1002/wics.1515

Franzmeier, N., Koutsouleris, N., Benzinger, T., Goate, A., Karch, C. M., Fagan, A. M., et al. (2020). Alzheimer's Disease Neuroimaging, I., Dominantly Inherited AlzheimerPredicting Sporadic Alzheimer's Disease Progression *via* Inherited Alzheimer's Disease-informed Machine-learning. *Alzheimer's Demen.* 16, 501–511. doi:10.1002/alz.12032

Grassi, M., Perna, G., Caldirola, D., Schruers, K., Duara, R., and Loewenstein, D. A. (2018). A Clinically-Translatable Machine Learning Algorithm for the Prediction of Alzheimer's Disease Conversion in Individuals with Mild and Premild Cognitive Impairment. *Jad* 61, 1555–1573. doi:10.3233/jad-170547

Hampel, H., O'Bryant, S. E., Molinuevo, J. L., Zetterberg, H., Masters, C. L., Lista, S., et al. (2018). Blood-based Biomarkers for Alzheimer Disease: Mapping the Road to the Clinic. *Nat. Rev. Neurol.* 14, 639–652. doi:10.1038/s41582-018-0079-7

Hsu, C.-W., Chang, C.-C., and Lin, C.-J. (2003). A Practical Guide to Support Vector Classification Chih-Wei Hsu. Chih-Chung Chang, and Chih-Jen Lin. Available at: https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

Khan, A., and Zubair, S. (2020). Longitudinal Magnetic Resonance Imaging as a Potential Correlate in the Diagnosis of Alzheimer Disease: Exploratory Data Analysis. *JMIR Biomed. Eng.* 5, e14389. doi:10.2196/14389

Kublanov, V. S., Dolganov, A. Y., Belo, D., and Gamboa, H. (2017). Comparison of Machine Learning Methods for the Arterial Hypertension Diagnostics. *Appl. Bionics Biomech.* 2017, 5985479. doi:10.1155/2017/5985479

Leong, L. K., and Abdullah, A. A. (2019). Prediction of Alzheimer's Disease (AD) Using Machine Learning Techniques with Boruta Algorithm as Feature Selection Method. *J. Phys. Conf. Ser.* 1372, 012065. doi:10.1088/1742-6596/1372/1/012065

Magnin, B., Mesrob, L., Kinkingnéhun, S., Pélégrini-Issac, M., Colliot, O., Sarazin, M., et al. (2009). Support Vector Machine-Based Classification of Alzheimer's Disease from Whole-Brain Anatomical MRI. *Neuroradiology* 51, 73–83. doi:10.1007/s00234-008-0463-x

Marcus, D. S., Fotenos, A. F., Csernansky, J. G., Morris, J. C., and Buckner, R. L. (2010a). Open Access Series of Imaging Studies: Longitudinal MRI Data in Nondemented and Demented Older Adults. *J. Cogn. Neurosci.* 22, 2677–2684. doi:10.1162/jocn.2009.21407

Merelli, I., Pérez-Sánchez, H., Gesing, S., and D'agostino, D. (2014). High-performance Computing and Big Data in Omics-Based Medicine. *Biomed. Res. Int.* 2014, 825649. doi:10.1155/2014/825649

Michelle Wallig, M. C., Steve, W., and Dan, T. (2020). doParallel: Foreach Parallel Adaptor for the 'parallel' Package. R package version 1.0.16. Available at: https://cran.r-project.org/web/packages/doParallel/index.html.

Michelle Wallig, M., and Steve, W. (2020). foreach: Provides Foreach Looping Construct R package version 1.5.1. Available at: https://cran.r-project.org/web/packages/foreach/index.html.

O'bryant, S. E., Zhang, F., Johnson, L. A., Hall, J., Edwards, M., Grammas, P., et al. (2018). A Precision Medicine Model for Targeted NSAID Therapy in Alzheimer's Disease. *J. Alzheimers Dis.* 66, 97–104. doi:10.3233/JAD-180619

O'bryant, S. E., Zhang, F., Silverman, W., Lee, J. H., Krinsky-Mchale, S. J., Pang, D., et al. (2020). Proteomic Profiles of Incident Mild Cognitive Impairment and Alzheimer's Disease Among Adults with Down Syndrome. *Alzheimers Dement (Amst)* 12, e12033. doi:10.1002/dad2.12033

O'bryant, S. E., Edwards, M., Johnson, L., Hall, J., Villarreal, A. E., Britton, G. B., et al. (2016). A Blood Screening Test for Alzheimer's Disease. *Alzheimer's Demen. Diagn. Assess. Dis. Monit.* 3, 83–90. doi:10.1016/j.dadm.2016.06.004

O'bryant, S. E., Edwards, M., Zhang, F., Johnson, L. A., Hall, J., Kuras, Y., et al. (2019). Potential Two-step Proteomic Signature for Parkinson's Disease: Pilot Analysis in the Harvard Biomarkers Study. *Alzheimer's Demen. Diagn. Assess. Dis. Monit.* 11, 374–382. doi:10.1016/j.dadm.2019.03.001

O'bryant, S. E., Mielke, M. M., Rissman, R. A., Lista, S., Vanderstichele, H., Zetterberg, H., et al. (2017). Blood-based Biomarkers in Alzheimer Disease: Current State of the Science and a Novel Collaborative Paradigm for Advancing from Discovery to clinicBlood-Based Biomarkers in Alzheimer Disease: Current State of the Science and a Novel Collaborative Paradigm for Advancing from Discovery to Clinic. *Alzheimer's Demen.* 13, 45–58. doi:10.1016/j.jalz.2016.09.014

O'bryant, S. E., Xiao, G., Barber, R., Huebinger, R., Wilhelmsen, K., Edwards, M., et al. (2011a). Texas Alzheimer's, R., Care, C., and Alzheimer's Disease Neuroimaging, IA Blood-Based Screening Tool for Alzheimer's Disease that Spans Serum and Plasma: Findings from TARC and ADNI. *PLoS One* 6, e28092. doi:10.1371/journal.pone.0028092

O'bryant, S. E., Xiao, G., Barber, R., Reisch, J., Hall, J., Cullum, C. M., et al. (2011b). Texas Alzheimer'sA Blood-Based Algorithm for the Detection of Alzheimer's Disease. *Dement Geriatr. Cogn. Disord.* 32, 55–62. doi:10.1159/000330750

O'bryant, S. E., Xiao, G., Edwards, M., Devous, M., Gupta, V. B., Martins, R., et al. (2013). Texas Alzheimer's, R., and Care, CBiomarkers of Alzheimer's Disease Among Mexican Americans. *Jad* 34, 841–849. doi:10.3233/jad-122074

O'bryant, S. E., Xiao, G., Zhang, F., Edwards, M., German, D. C., Yin, X., et al. (2014). Validation of a Serum Screen for Alzheimer's Disease across Assay Platforms, Species, and Tissues. *Jad* 42, 1325–1335. doi:10.3233/jad-141041

Podcasy, J. L., and Epperson, C. N. (2016). Considering Sex and Gender in Alzheimer Disease and Other Dementias. *Dialogues Clin. Neurosci.* 18, 437–446. doi:10.31887/DCNS.2016.18.4/cepperson

Prevention, C. F. D. C. A. (2021). Alzheimer's Disease and Healthy Aging. Available at: https://www.cdc.gov/aging/aginginfo/alzheimers.html.

Rodriguez, S., Hug, C., Todorov, P., Moret, N., Boswell, S. A., Evans, K., et al. (2021). Machine Learning Identifies Candidates for Drug Repurposing in Alzheimer's Disease. *Nat. Commun.* 12, 1033. doi:10.1038/s41467-021-21330-0

Roe, C. M., Fagan, A. M., Grant, E. A., Marcus, D. S., Benzinger, T. L., Mintun, M. A., et al. (2011). Cerebrospinal Fluid Biomarkers, Education, Brain Volume, and Future Cognition. *Arch. Neurol.* 68, 1145–1151. doi:10.1001/archneurol.2011.192

Schmidt, B., and Hildebrandt, A. (2017). Next-generation Sequencing: Big Data Meets High Performance Computing. *Drug Discov. Today* 22, 712–717. doi:10.1016/j.drudis.2017.01.014

Stamate, D., Kim, M., Proitsi, P., Westwood, S., Baird, A., Nevado-Holgado, A., et al. (2019). A Metabolite-based Machine Learning Approach to Diagnose Alzheimer-type Dementia in Blood: Results from the European Medical Information Framework for Alzheimer Disease Biomarker Discovery Cohort. *Alzheimer's Demen. Translational Res. Clin. Interventions* 5, 933–938. doi:10.1016/j.trci.2019.11.001

Waring, S., O'bryant, S., Reisch, J., Diaz-Arrastia, R., Knebl, J., and Doody, R. (2008). The Texas Alzheimer's Research Consortium Longitudinal Research Cohort: Study Design and Baseline Characteristics. *Tex. Public Health J* 60, 9–13.

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Cedarbaum, J., et al. (2015). 2014 Update of the Alzheimer's Disease Neuroimaging

Initiative: A Review of Papers Published since its Inception. *Alzheimers Dement* 11, e1–120. doi:10.1016/j.jalz.2014.11.001

Zetterberg, H., and Burnham, S. C. (2019). Blood-based Molecular Biomarkers for Alzheimer's Disease. *Mol. Brain* 12, 26. doi:10.1186/s13041-019-0448-1

Zhang, F., Petersen, M., Johnson, L., Hall, J., and O'Bryant, S. E. (2021). Recursive Support Vector Machine Biomarker Selection for Alzheimer's Disease. *Jad* 79, 1691–1700. doi:10.3233/jad-201254

Zhang, Z., and Sejdić, E. (2019). Radiological Images and Machine Learning: Trends, Perspectives, and Prospects. *Comput. Biol. Med.* 108, 354–370. doi:10.1016/j.compbiomed.2019.02.017

Conflict of Interest: SO'B has multiple pending and issued patents on blood biomarkers for detecting and precision medicine therapeutics in neurodegenerative diseases. He is a founding scientist and owns stock options in Cx Precision Medicine, Inc.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Neuroscience

ORIGINAL RESEARCH
published: 24 December 2021
doi: 10.3389/fnins.2021.743720

The Critical Modulatory Role of Spiny Stellate Cells in Seizure Onset Based on Dynamic Analysis of a Neural Mass Model

***Saba Tabatabaee*[1], *Fariba Bahrami*[1]* *and Mahyar Janahmadi*[2]**

[1] Human Motor Control and Computational Neuroscience Laboratory, School of Electrical and Computer Engineering (ECE), College of Engineering, University of Tehran, Tehran, Iran, [2] Department of Physiology, Neuroscience Research Center, School of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran

Growing evidence suggests that excitatory neurons in the brain play a significant role in seizure generation. Nonetheless, spiny stellate cells are cortical excitatory non-pyramidal neurons in the brain, whose basic role in seizure occurrence is not well understood. In the present research, we study the critical role of spiny stellate cells or the excitatory interneurons (EI), for the first time, in epileptic seizure generation using an extended neural mass model inspired by a thalamocortical model originally introduced by another research group. Applying bifurcation analysis on this modified model, we investigated the rich dynamics corresponding to the epileptic seizure onset and transition between interictal and ictal states caused by EI connectivity to other cell types. Our results indicate that the transition between interictal and ictal states (preictal signal) corresponds to a supercritical Hopf bifurcation, and thus, the extended model suggests that before seizure onset, the amplitude and frequency of neural activities gradually increase. Moreover, we showed that (1) the altered function of GABAergic and glutamatergic receptors of EI can cause seizure, and (2) the pathway between the thalamic relay nucleus and EI facilitates the transition from interictal to ictal activity by decreasing the preictal period. Thereafter, we considered both sensory and cortical periodic inputs to study model responses to various harmonic stimulations. Bifurcation analysis of the model, in this case, suggests that the initial state of the model might be the main cause for the transition between interictal and ictal states as the stimulus frequency changes. The extended thalamocortical model shows also that the amplitude jump phenomenon and non-linear resonance behavior result from the preictal state of the modified model. These results can be considered as a step forward to a deeper understanding of the mechanisms underlying the transition from normal activities to epileptic activities.

Keywords: neural mass model, bifurcation analysis, linear/non-linear resonance, excitatory interneurons, preictal state

OPEN ACCESS

Edited by:
Reza Lashgari,
Shahid Beheshti University, Iran

Reviewed by:
Qing Yun Wang,
Beihang University, China
Guosheng Yi,
Tianjin University, China
Sandesh Panthi,
University of Otago, New Zealand

****Correspondence:***
Fariba Bahrami
fbahrami@ut.ac.ir

Specialty section:
This article was submitted to Neural Technology, a section of the journal Frontiers in Neuroscience

Received: *19 July 2021*
Accepted: *22 November 2021*
Published: *24 December 2021*

Citation:
Tabatabaee S, Bahrami F and Janahmadi M (2021) The Critical Modulatory Role of Spiny Stellate Cells in Seizure Onset Based on Dynamic Analysis of a Neural Mass Model. Front. Neurosci. 15:743720. doi: 10.3389/fnins.2021.743720

INTRODUCTION

Epilepsy is one of the most common disorders of the central nervous system (CNS), which causes sudden abnormal and synchronized brain activities, resulting in seizures. Interactions between excitatory and inhibitory neurons shape mainly brain activities, and some transitory disparity in the inhibitory/excitatory balance can trigger a seizure. One-third of people with epilepsy are likely to have drug-resistant epilepsy, and treatments such as surgery or stimulation-based treatments like deep brain stimulation (DBS) can be considered for these kinds of drug-resistant patients (Shan et al., 2021). A seizure can be composed of four distinct states including preictal, ictal, interictal, and postictal. The preictal state appears before the seizure begins and indicates that seizures do not simply start out of nothing. Experimental studies (Bartolomei et al., 2004; Huberfeld et al., 2011) on human tissue and intracranial EEG signals have shown that preictal spikes are distinguishable from ictal and interictal signals. The preictal signal can be used to predict seizure occurrence. Understanding the mechanism of generation of the preictal period opens the way for a more precise seizure prediction and thereby more reliable automatic interventions to prevent the seizure occurrence (Moghim and Corne, 2014).

Different seizures can be categorized into focal and generalized onset seizures. Tonic, clonic, and absence seizures are the generalized onset types of seizures, and they occur when a widespread activity triggers in both hemispheres of the brain (Scheffer et al., 2017). Furthermore, there are two groups of seizures that have known hemispheric origins, i.e., generalized and focal onset seizures (Fisher, 2017). Both groups are categorized into two subclasses: motor and non-motor seizures. Both subclasses have several motor seizure types in common (e.g., tonic and clonic seizures). However, each type has different manifestations and symptoms. On the other hand, absence seizures are among well-known non-motor generalized onset seizure groups that are also considered as sensory and emotional seizures (Fisher, 2017). In tonic–clonic seizures, a clonic activity follows the tonic activity. Electrophysiological observations have shown the two-way transition between absence and tonic–clonic seizures (Mayville et al., 2000), and dysfunction of the cortical and/or thalamic circuitries is believed to produce the absence, clonic, and tonic epileptic activities, together with transitions between them. Recorded electroencephalogram (EEG) signals indicate that seizures can vary in frequency contents. A typical absence seizure can be considered as approximately synchronous spike and wave discharges (SWD) with a frequency of 2∼4 Hz, but atypical absence seizures (another kind of non-motor generalized onset seizures) show different frequency ranges (<2 and >4 Hz) (Velazquez et al., 2007). A tonic seizure is a fast-spiking activity (>14 Hz) with low amplitude, and a clonic seizure is slow-wave activity (approximately 3 Hz) with high amplitude.

Several studies have shown that interneurons are cell types, which play key roles in the initiation and termination of epileptic seizures (Khan et al., 2018; Tran et al., 2020). Interneurons in the cortex not only control the activity of pyramidal neurons but also receive thalamic relay nucleus inputs. Therefore, they are important factors in transferring and integrating the sensory information coming from the thalamus to the cortex (Danober et al., 1998). Interneurons are traditionally regarded as inhibitory neurons, but more precisely, there are two kinds of interneurons in the CNS, i.e., excitatory and inhibitory interneurons. Inhibitory interneurons use the neurotransmitter gamma-aminobutyric acid (GABA) or glycine, and excitatory interneurons (EI) are spiny stellate cells in the neocortex of the human brain and use glutamate as their neurotransmitters (Okhotin, 2006). A study on the human brain (Steriade and Contreras, 1998) has shown that the neurons in the neocortex play a crucial role in SWD. Another study on the juvenile mice using multi-photon imaging (Neubauer et al., 2014) suggested that the neocortex has an intrinsic predisposition for seizure generation and pathological recruitment of the thalamus into joint synchronous epileptic activities. A genetic mutation study on mice has shown that in the spiny stellate cells (in the neocortex), an alteration in the kinetic of N-methyl-D-aspartate receptors (NMDA) was sufficient to cause neuronal hyper-excitability leading to epileptic activity in the brain (Lozovaya et al., 2014). An experimental study on monkeys has shown that synchronous discharges of EI could spread the epileptic activity in the brain, and EI could play an important role in the initiation and propagation of SWD, in which the spike is because of the extracellularly synchronous and powerful depolarization of EI, and the wave is because of the inhibitory interneurons. Therefore, SWD can be generated as a result of the interactions between inhibitory and EI (Steriade, 1974). Nevertheless, it is not well understood how EI in the neocortex of the human brain can explicitly cause a seizure. Therefore, understanding the epileptogenic role of the EI in the neocortex of the human brain is crucial to deepen our knowledge about neocortical epilepsy.

On the other hand, computational modeling of epilepsy has provided dynamic insights into the mechanism underlying the transition from normal to epileptic activities. Fan et al. (2015) used a modified computational field model of a cortical microzone and showed that the two-way transitions between absence and tonic–clonic epileptic seizures are induced by disinhibition between slow and fast inhibitory interneurons. Nevertheless, they ignored the role of thalamic circuitry and its interaction with the cerebral cortex in the proposed computational model, whereas, in Zhang et al. (2015) and Law et al. (2018) it is shown that the mechanism of seizures depends on the dysfunctionality in the function of the thalamus and or cortex. Experimental evidence suggests that interactions between the thalamus and cerebral cortex influence the initiation and propagation of SWD (Pinault and O'Brien, 2005). For that reason, Taylor et al. (2014) developed a thalamocortical model to investigate the SWD generation. Liu and Wang (2017) introduced a thalamocortical model inspired by the model developed by Taylor et al. (2014) and Fan et al. (2015). In their model, they considered the dual pathways between fast and slow inhibitory interneurons in the cortex and simulated normal activity, clonic, absence, and tonic seizures. However, none of the models mentioned above did consider the EI population in the thalamocortical circuitry, and therefore, they did not study the role of the EI population in transition from the interictal to the ictal state.

In the present study, we modified and extended a thalamocortical model, originally proposed by Liu and Wang (2017). The original model describes absence, tonic, clonic, and tonic–clonic seizure generation. That is one of the reasons we chose this model and extended it to investigate the role of EI on seizure onset using dynamical analysis. For this purpose, we introduced an additional group of interneurons into the model, which was excitatory and considered their interactions with pyramidal neurons, inhibitory interneurons, and thalamic relay nucleus in the brain. As will be shown, this model not only is capable of producing different types of epileptic seizures such as absence, clonic and tonic, but also generates the preictal period. The original model of Liu and Wang (2017) does not describe the preictal period.

The organization of the paper is as follows. In section "Materials and Methods," the modified epileptic dynamical model inspired by the model proposed in Liu and Wang (2017) is introduced. Then, in section "Results," we first explore the role of impairment in the activity of the EI in interictal to ictal transition. Using the modified model, we investigate the role of GABAergic and glutamatergic receptors of EI in seizure generation and transition between interictal and ictal states. Through numerical simulations and bifurcation analysis, we show that dysfunction of excitatory and inhibitory receptors of EI leads to interictal to ictal transition. Moreover, we investigate the effect of impairments in the interactions between the thalamic relay nucleus and EI and we will investigate how they give rise to interictal to preictal transition and also facilitate the occurrence of an epileptic seizure. Then, we investigate the behavior of the extended model in a more general framework when the EI function properly, but there are dysfunctions in the thalamus, more exactly in the synaptic strength between the thalamic relay nucleus and reticular nucleus. Finally, in section "Discussion," we examine the frequency responses of the modified thalamocortical model subjected to various sensory and cortical periodic inputs to identify the effect of various stimuli on

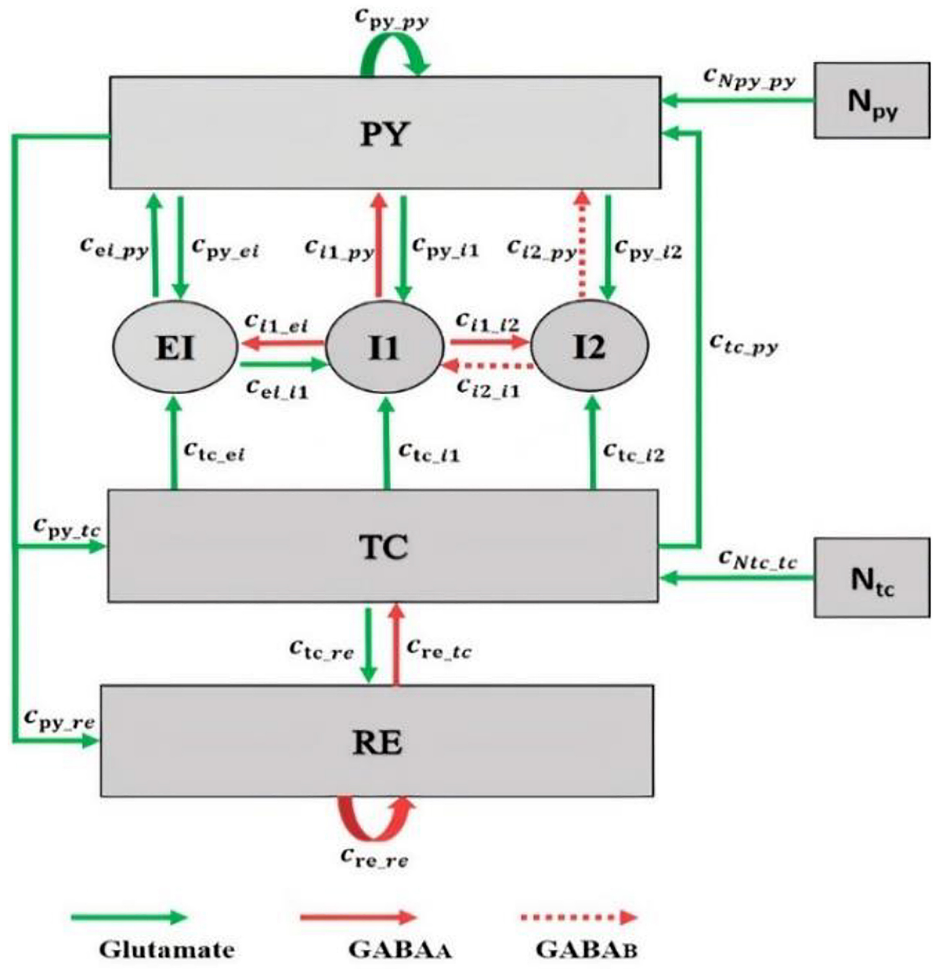


FIGURE 1 | Schematic of the thalamocortical model including the excitatory interneuron population in the cortex. The model involves cortical and thalamic subnetworks. In the cortical subnetwork, PY is the pyramidal neuronal population, EI is the excitatory interneurons population, and I1 and I2 are the fast and slow inhibitory interneurons populations, respectively. In the thalamic subnetwork, TC is the thalamic relay nucleus population and RE is the thalamic reticular nucleus population. Green arrows define the excitatory glutamatergic receptor. Red solid and dashed arrows define the $GAGA_A$ and $GABA_B$ inhibitory receptors, respectively. The value of parameters of this model are as follows: $c_{py\_py} = 1.89$, $c_{py\_i1} = 4$, $c_{i1\_py} = 1.8$, $c_{re\_re} = 0.01$, $c_{tc\_re}= 10$, $c_{re\_tc}= 1.4$, $c_{py\_tc}= 3$, $c_{py\_re} = 1.4$, $c_{tc\_py} = 1$, $c_{py\_i2} = 1.5$, $c_{tc\_i1} = 0.05$, $c_{tc\_i2} = 0.05$, $c_{ei\_i1} = 0.05$, $c_{ei\_py} = 0.442$, $c_{i2\_py} = 0.05$, $c_{i2\_i1} = 0.1$, $c_{i1\_i2} = 0.5$, $c_{Npy\_py} = 1$, $c_{Ntc\_tc} = 1$, $tau_1 = 21.5$, $tau_2 = 31.5$, $tau_3 = 0.1$, $tau_4 = 4.5$, $tau_5 = 3.8$, $tau_6 = 3.9$, $h_{py} = –0.4$, $h_{i1} = –3.4$, $h_{i2} = –2$, $h_{ei} = –1$, $h_{tc} = –2.5$, $h_{re} = –3.2$, $\varepsilon = 250{,}000$, $a_{tc} = 0.02$, $a_{py} = 0.02$, $B_{Npy} = 0.7$, and $B_{Ntc} = 0.1$. These parameters were obtained by trials and error such that the dynamic behavior of the model (i.e., the bifurcation diagrams) was the same as the original model developed by Liu and Wang (2017).

epileptic seizures. The obtained results show that the model starts its non-linear resonator behavior when the shift from normal activity to preictal state takes place. Therefore, the amplitude jump phenomenon corresponding to chaotic behavior takes place before the onset of a seizure.

MATERIALS AND METHODS

Model Structure

Recent studies Jiang et al. (2018) and Zhang et al. (2020) have shown that during generalized seizures, dysfunctionality has been identified in the thalamocortical network. Therefore, both the cortex and thalamus play an important role in seizure generation. Accordingly, we modified the thalamocortical model of Liu and Wang (2017) and as shown in **Figure 1**. This model is a neural mass model, which considers the inhibitory and excitatory neuronal populations and their connections in the brain. The cortical section of this model includes excitatory pyramidal neuron population (PY) and mutual fast and slow inhibitory interneuron populations (I1, I2) having inhibitory $GABA_A$ and $GABA_B$ receptors. Activation of ionotropic GABAA receptors causes fast inhibitory postsynaptic potentials (IPSPs) by allowing the influx of Cl^- into the postsynaptic cells, while activation of metabotropic GABAB receptors mediates a slow inhibition by inducing K^+ efflux. The thalamic subsystem includes a population of excitatory thalamic relay nucleus (TC) and the inhibitory population of neurons located in the reticular nucleus (RE). To investigate the effect of spiny stellate cells on seizure onset, we considered the EI population, EI, in our thalamocortical model (see **Figure 1**). In this new model version, we will investigate the synaptic connectivity strength of EI to explore dynamics, which lead to preictal state and dynamics during transitions between absence, tonic, and clonic seizures.

In this thalamocortical model, we analyze the interactions between neuronal populations in the network by varying the strength of the synaptic connections. These synaptic connections are associated with the type of receptors. In the brain, receptors are either excitatory (glutamatergic) or inhibitory (GABAergic). In the cortex, the pyramidal population is excitatory n and there is a mutual excitatory connection between pyramidal and EI by the glutamatergic receptors (Feldmeyer et al., 2002). Excitatory and inhibitory interneurons are locally (Sun et al., 2006), and based on this fact, we consider an excitatory and inhibitory connection between EI and fast inhibitory interneurons. The thalamus is globally connected to the EI by the thalamic relay nucleus, and the thalamic relay nucleus is regarded as an excitatory population (da Costa and Martin, 2011). Here, we consider an excitatory connection from the thalamic relay nucleus population to the EI.

The thalamic relay nucleus receives all the sensory information from different parts of the brain, and then this information is sent to the appropriate area in the cortex for further processing (Taylor et al., 2015; Castejon et al., 2016). Here, we investigate the dynamic of the extended model when it receives cortical and sensory inputs. In this regard, we consider an input (N_{tc}) to the thalamic relay nucleus by adding a periodic sensory input with frequency ftc, bias B_{Ntc}, and amplitude a_{tc}. Moreover, we consider a cortical input (N_{py}) to the pyramidal neuronal population by adding a biased sinusoidal waveform with bias of B_{Npy}, frequency of fpy, and amplitude of a_{py}.

We implement the model using equations developed by Liu and Wang (2017) and the Amari neural field equations (Amari, 1977) (for the EI population). The differential equations of the model are given below.

$$\begin{aligned}\frac{\mathrm{dPY}}{\mathrm{dt}} &= \mathrm{tau}_1\ (\mathrm{h_{py}} - \mathrm{PY} + \mathrm{c_{py\_py}}\ \mathrm{f(PY)} - \mathrm{c_{i1\_py}}\ \mathrm{f(I1)} \\ &\quad + \mathrm{c_{tc\_py}}\ \mathrm{f(TC)} - \mathrm{c_{i2\_py}}\ \mathrm{f(I2)} + \mathrm{c_{ei\_py}}\ \mathrm{f(EI)}\ ) + \mathrm{N_{py}}\end{aligned} \tag{1}$$

$$\begin{aligned}\frac{dI1}{dt} &= tau_2\ (h_{i1} - I1 + \ c_{py\_i1} f\ (PY) - c_{i2\_i1} f\ (I2) \\ &\quad + c_{tc\_i1} f\ (TC) + \ c_{ei\_i1} f\ (EI))\end{aligned} \tag{2}$$

$$\begin{aligned}\frac{dI2}{dt} &= tau_3\ (h_{i2} - I2 + \ c_{py\_i2} f\ (PY) - c_{i1\_i2} f\ (I1) \\ &\quad + c_{tc\_i2} f\ (TC))\end{aligned} \tag{3}$$

$$\begin{aligned}\frac{dEI}{dt} &= tau_4\ (h_{ei} - EI + \ c_{py\_ei} f\ (PY) - c_{i1\_ei} f\ (I1) \\ &\quad + \ c_{tc\_ei} f\ (TC))\end{aligned} \tag{4}$$

$$\begin{aligned}\frac{dTC}{dt} &= \ tau_5\ (h_{tc} - TC + \ c_{py\_tc} f\ (PY) - c_{re\_tc} f\ (RE)) \\ &\quad + N_{tc}\end{aligned} \tag{5}$$

$$\begin{aligned}\frac{dRE}{dt} &= tau_6\ (h_{re} - RE + \ c_{py\_re} f\ (PY) - c_{re\_re} f\ (RE) \\ &\quad + \ c_{tc\_re} f\ (TC))\end{aligned} \tag{6}$$

$$N_{py} = B_{Npy} + a_{py}\ sin(2\pi\ fpy\ t) \tag{7}$$

$$N_{tc} = B_{Ntc} + a_{tc}\ sin(2\pi\ ftc\ t) \tag{8}$$

$$Model\ Output = (PY + I1 + I2 + EI)/4 \tag{9}$$

Dimensionless parameters $c_{1,\ldots,16,iny,ei,in1,in2}$ are the connectivity parameters, which determine the coupling strength between the populations, and $h_{py,i1,i2,ei,tc,re}$ are input parameters, $tau_{1,2,\ldots,6}$ are time scale parameters, and $f(x) = 1/(1+\varepsilon^{-x})$ is the transfer function, where ε determines the steepness and x = PY,I1,I2,EI,TC and RE. N_{py} (cortical input) and N_{tc} (sensory input) are inputs to the pyramidal population and thalamic relay nucleus population, respectively. These parameters were obtained by trials and error such that the dynamic behavior of the model (i.e., the bifurcation diagrams) has a similar mechanism as the model developed by Liu and Wang (2017). The output of the model is taken as the mean activity of the four cortical populations.

Simulations

Simulations are performed by the standard fourth-order Runge–Kutta integration using MATLAB 9.4 and the MatCont environments, with a step size of 0.0039 s. We set the time window at 60 s and use the last 2 s of the time series to analyze the stable state of the time domain and the deterministic behavior of the model. We use the frequency domain and time domain analyses to explore the transitions from interictal to ictal states. We extract the dominant frequency (the frequency that carries the maximum energy) from the power spectral density using the fast Fourier transform, and for the bifurcation analysis, we extract the extrema (local maximum and minimum) of the mean of four cortical populations from the time series. By doing so, we can observe seven different epileptic activities such as interictal state (or the normal background activity), preictal state, which occurs before the seizure onset, slow rhythmic activity that can be observed at seizure onset or during the seizures, typical absence seizures, atypical absence seizures, and tonic and clonic seizures. For 1D bifurcation diagrams (sections "Transition Dynamics Produced by GABAergic Receptor-Mediated Inhibition in EI," "Transition Dynamics Produced by Glutamatergic Receptor-Mediated Excitation in Excitatory Interneurons," and "Transition Dynamics Produced by Glutamatergic Receptor-Mediated Excitation in Reticular Nucleus"), we obtain the maximum and minimum of model output as we change the bifurcation parameter. For hybrid bifurcation analysis (section "Hybrid Cooperation of GABAergic and Glutamatergic Receptors of EI in Epileptic Transition Dynamics") when both of the parameters of c_{py_ei} and c_{i1_ei} are changed, we computed the local maxima, local minima, and frequency of the last two seconds of the model output, and then we sorted the local maxima and minima from large to small to obtain the largest (Pmax1) and the smallest (Pmax2) maxima and the largest (Pmin1) and the smallest (Pmin2) minima. We summarized our categorization in **Table 1**.

RESULTS

In this section, the role of EI in transition from interictal to ictal and the frequency response of the thalamocortical model, when it receives sensory and cortical inputs, are examined using various bifurcation analyses and dynamical simulations. In section "Transition Dynamics," to explore the transition dynamics of the model we assume that the inputs of the model are constant, i.e., $N_{py} = B_{Npy}$ and $N_{tc} = B_{Ntc}$. In section "Frequency Analysis of the Different Initial States of the Model," for investigating the frequency response of the model, a sinusoidal waveform signal is added to both sensory and cortical inputs, that is, $N_{py} = B_{Npy} + a_{py}$ sin(2 fpy t) and $N_{tc} = B_{Ntc} + a_{tc}$ sin(2ftc t).

Transition Dynamics

EI are connected to the pyramidal neurons and fast inhibitory interneurons by glutamatergic and GABAergic receptors, respectively. The synaptic strength is not static, and the changes in the neurotransmitters released by the excitatory and inhibitory neurons in the brain can result in short-term or long-term changes in synaptic strength. Antiepileptic drugs either block glutamatergic receptors or facilitate the function of GABAergic receptors, which result in a change in the glutamatergic and GABAergic neurotransmitter release (Rogawski, 2011; Vashchinkina et al., 2014). Moreover, a ketogenic diet can bring about the altered function of receptors and change in neurotransmitter release (Zhang et al., 2018). Experimental studies have shown that changes in the function of GABAergic and glutamatergic receptors in the cortex of rats can be seen in genetic absence, tonic, and clonic seizure onset (Cortez et al., 2004; Jones et al., 2008; Errington et al., 2011). However, the transition dynamics from interictal to ictal and transitions between absence, tonic, and clonic seizures caused by the abnormalities in glutamatergic and GABAergic receptors of EI are still unclear. In this section, by using bifurcation analysis, we explore the effect of impairment in the function of GABAergic and glutamatergic receptors of the EI population on the epileptic dynamics of the model as the synaptic strengths c_{i1-ei} and c_{py-ei} change, respectively.

Transition Dynamics Produced by GABAergic Receptor-Mediated Inhibition in Excitatory Interneurons

Increasing the GABAergic inhibition is traditionally believed to suppress epileptic seizures; however, a computational study has shown that before seizure onset, the activity of GABAergic interneurons increases, leading to synchronous neuronal activity and epileptic seizures (Rich et al., 2020). Based on an experimental study on humans, antiepileptic drugs such as midazolam that increase the GABAergic neurotransmitters in the brain can trigger seizure (Montenegro et al., 2001). Abnormality in $GABA_A$ receptors brings about a shift in the chloride reversal potential in neurons, which in turn results in changing the behavior of GABA from inhibitory to excitatory behavior and cause seizure (Khalilov et al., 2003). Therefore, the role of GABAergic receptor epileptic seizures is more complex than one could assume that they have only inhibitory roles and anticonvulsant agents always inhibit their function. Here, we investigate the effect of increasing inhibitory function of GABAergic receptors of EI in the neocortex on seizure generation. We explore the epileptiform activity induced by the dysfunction of the GABAergic receptor of the EI population as the synaptic strength c_{i1-ei} changes. Results of **Figure 2** show that the model displays rich dynamics as the parameter c_{i1-ei} varies.

In **Figure 2**, the overall bifurcation is shown in detail for variations of c_{i1-ei} as the bifurcation parameter. In the bifurcation diagram from left to right, the red line corresponds to the stable fixed points, which are related to the normal background activity (**Figure 2A**), and upon increasing c_{i1-ei} to ∼0.349, a stable fixed point coalesces with a stable limit cycle and a supercritical Hopf bifurcation point (H_1) happens; with increasing of c_{i1-ei} to ∼0.355, a period doubling bifurcation (PD1) occurs, which forms the preictal spike patterns (transition from normal background activity to ictal activity, **Figure 2B**). With a further increase of the c_{i1-ei} parameter to ∼0.37, another period doubling bifurcation (PD2) happens and the transition from preictal state to clonic seizure takes place and shapes clonic seizure patterns (**Figure 2C**). In addition, as c_{i1-ei} is increased up to ∼0.458, the first fold limit cycle bifurcation (LPC1) occurs, and another fold limit cycle bifurcation (LPC2) takes place with further increase in c_{i1-ei}. With the coexistence of two stable limit cycles in the range of c_{i1-ei} bounded by twofold limit cycle bifurcations (LPC1 and LPC2), a bistable region appears

TABLE 1 | Specific characteristics of seven simulated signals.

Type of the signal	Dominant frequency range	The largest (Pmin1) and the smallest minima (Pmin2)	The largest (Pmax1) and the smallest maxima (Pmax2)
Normal background activity	DF = 0 Hz	–	–
Preictal spikes	DF < 3.5 Hz	Pmin1-Pmin2 < 0.2	0.01 < Pmax1-Pmax2 < 0.12
Slow rhythmic activity	DF < 15 Hz	–	–0.8 < Pmax1 <–0.1
Typical absence seizure	DF ∈ [2,4] Hz	0.004 < Pmin1-Pmin2	–
Atypical absence seizure	DF > 4 Hz and DF < 2 Hz	0.01 < Pmin1-Pmin2	–
Clonic seizure	DF ∈ [2,4] Hz	Pmin1-Pmin2 < 0.15	0.01 < Pmax1-Pmax2
Tonic seizure	DF ≥ 14 Hz	Pmin1-Pmin2 < 0.01	0.01 < Pmax1-Pmax2

Dominant frequency, the largest minima, the largest maxima, the largest minima, and the largest maxima are characteristics used to separate seven different mean activities of the cortical populations.

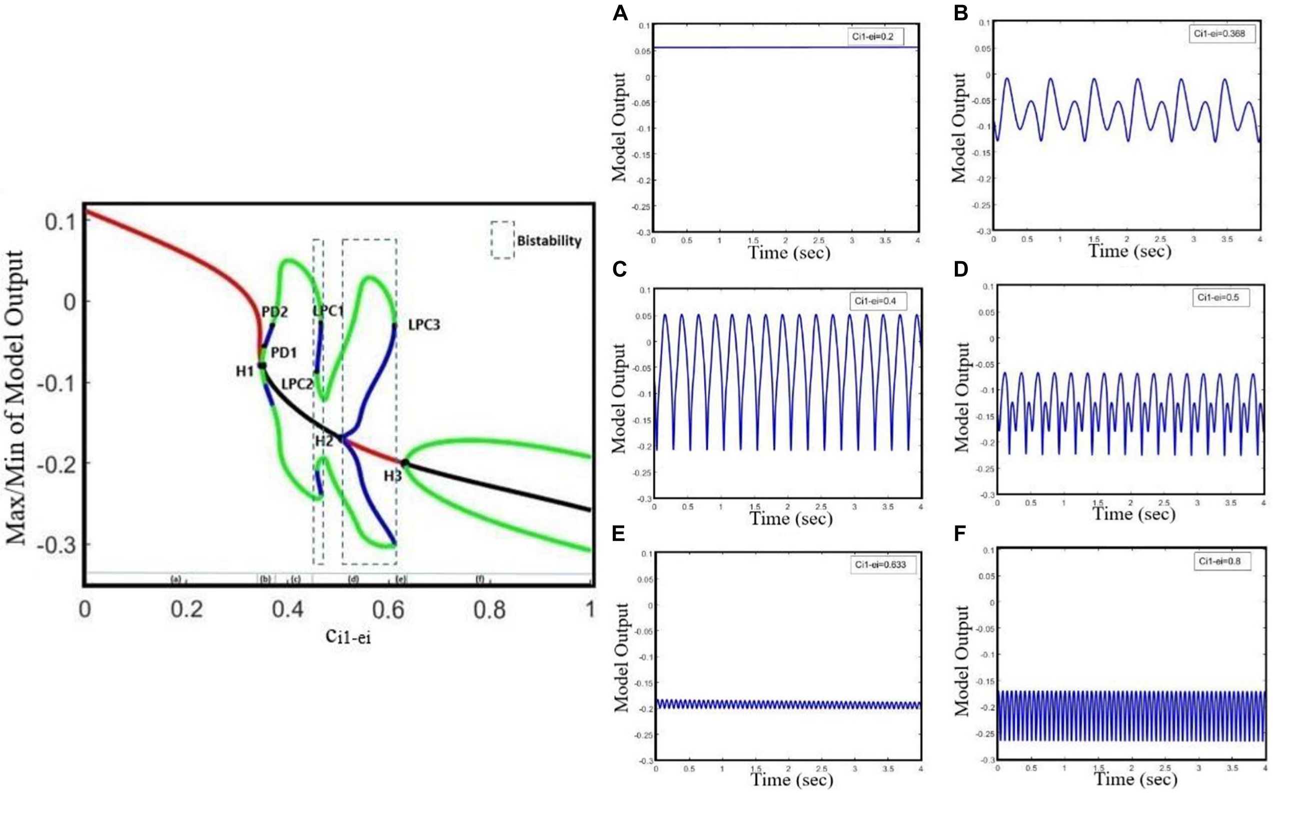


FIGURE 2 | Bifurcation diagram and corresponding time series of the model output for different values of c_{i1_ei}. The model output is defined as the mean value of the output voltage of PY, I1, I2, and EI populations. The bifurcation diagram of the model (left) is calculated and plotted for c_{i1_ei} as the bifurcation parameter and with c_{py_ei} = 0.8, **ctc_ei** = 4.5, a_{tc} = 0, and a_{py} = 0. We also set a_{tc} = 0 and a_{py} = 0. In the plot, blue and green cycles represent the unstable and stable limit cycles, respectively, and red and black lines represent the stable and unstable equilibrium points, respectively. According to the diagram as the parameter c_{i1_ei} changes, the model produces **(A)** normal background firing, **(B)** preictal spikes, **(C)** clonic discharges, **(D)** SWD, **(E)** slow rhythmic activity, and **(F)** tonic discharges. LPC1, LPC2, and LPC3 are fold limit cycle bifurcation, H1 and H3 are supercritical Hopf bifurcation, H2 is subcritical Hopf, and PD1 and PD2 are periods of doubling bifurcations.

and creates the coexistence of two different amplitude SWD patterns (**Figure 2D**). Moreover, stable points appear again at c_{i1-ei} around 0.508, where the first subcritical Hopf bifurcation (H_2) takes place, and another bistable region turns out with coexistence of stable fixed points and a stable limit cycle until the third fold limit cycle bifurcation (LPC3) at $c_{i1-ei} \sim 0.611$ takes place. Upon increasing the c_{i1-ei}, the stable focus points, which are corresponding to slow rhythmic activity, take place (**Figure 2E**), and then at $c_{i1-ei} \sim 0.634$ another supercritical Hopf bifurcation (H_3) happens and shapes the tonic seizure patterns (**Figure 2F**).

Transition Dynamics Produced by Glutamatergic Receptor-Mediated Excitation in Excitatory Interneurons

Studies have shown that antiepileptic drugs that are expected to reduce excitation in the brain, on the contrary, can have a paradoxical effect and brings about an aberrant synchronization

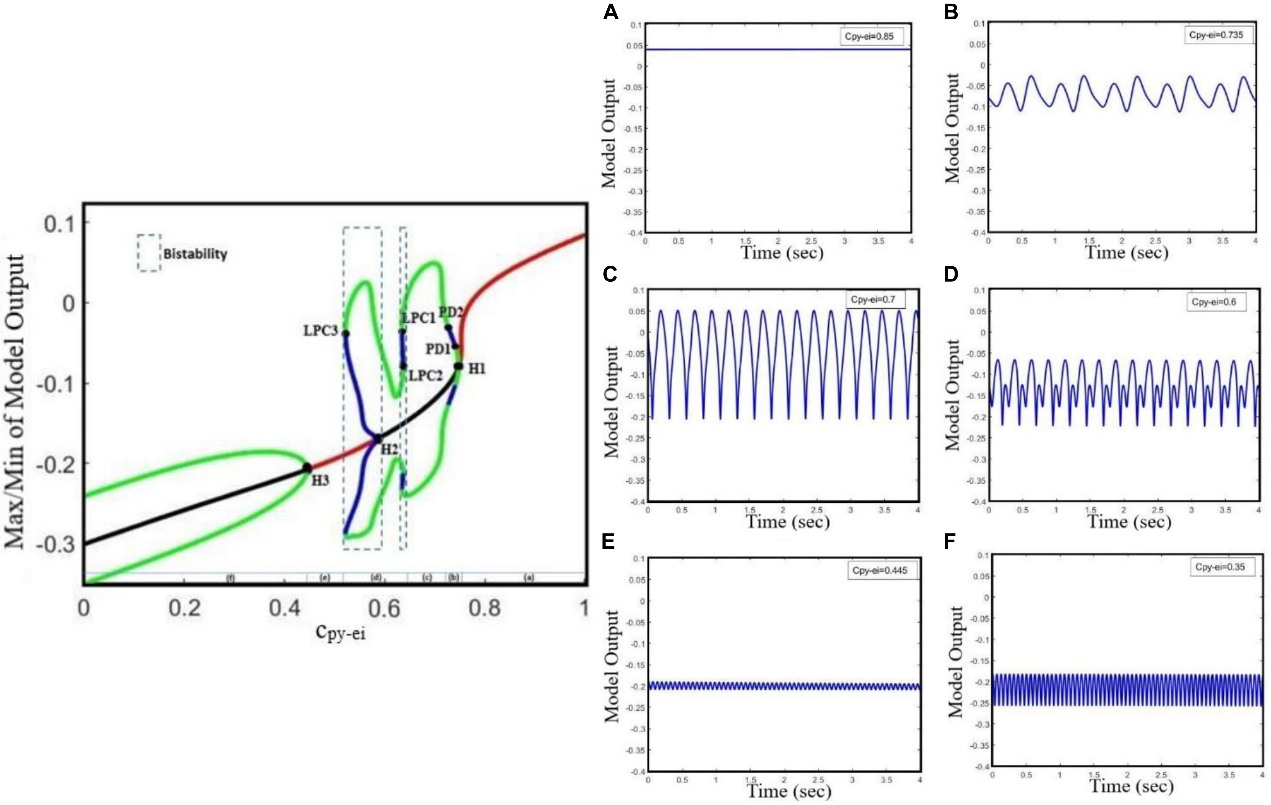


FIGURE 3 | Bifurcation diagram and corresponding time series of the model output for different values of c_{py-ei}. The bifurcation diagram of the model is calculated and plotted for c_{py-ei} as the bifurcation parameter with c_{i1-ei} = 0.3 and c_{tc-ei} = 4.5. We also set a_{tc} = 0 and a_{py} = 0. In the bifurcation plot, blue and green cycles represent the unstable and stable limit cycles, respectively, and red and black lines represent the stable and unstable points, respectively. It can be found from the bifurcation diagram that as the parameter c_{py-ei} decreases, **(A)** normal background firing (interictal), **(B)** preictal spikes, **(C)** clonic discharges, **(D)** SWD, and **(E)** slow rhythmic activity and tonic discharges can appear. **(F)** Tonic discharges. LPC1, LPC2, and LPC3 are fold limit cycle bifurcation, H1 and H3 are supercritical Hopf bifurcation, H2 is subcritical Hopf, and PD1 and PD2 are period-doubling bifurcations.

in neural activity (Chaves and Sander, 2005; Thomas et al., 2006). Based on the *in vitro* study on mice in the neocortex, impairment in glutamate release, which is either due to the decreased function of glutamate receptors or due to loss of glutamate receptors, causes the generalized absence and tonic–clonic epileptic seizures (Seal et al., 2008). A study on the neocortex of mice (Maheshwari et al., 2013) has shown that the paradoxical seizure exacerbation effect of the antiepileptic medication can be explained by the unintended suppression of inhibitory interneurons following the NMDA receptor blockade. Additionally, in an *in vivo* study on a mouse model of tuberous sclerosis complex (TSC), the mutation in either TSC1 or TSC2 can disturb the function of NMDA receptors in the EI, which in turn can cause a shift from normal activity to ictal activity (Lozovaya et al., 2014). However, the mechanisms underlying the transition from interictal to ictal activities associated with glutamatergic receptors on EI remained unclear. Here, we take c_{py-ei} as the growing bifurcation parameter to investigate the dynamics of a model to explore the epileptiform activity induced by the glutamatergic receptor function of EI. According to **Figure 3**, by gradually decreasing the EI excitability in the cortex, we can observe rich epileptiform transition dynamics.

According to the bifurcation diagram in **Figure 3**, from right to left, we can observe the normal background activity, preictal spikes, clonic seizures, absence seizures, slow rhythmic activity, and tonic seizures, respectively, combined with corresponding time series, as the parameter c_{py-ei} decreases. Hence, suppression of glutamatergic receptors on EI can lead to the transition from interictal to preictal signals, and also transition between epileptic seizures. With decreasing synaptic strength c_{py-ei}, the system encounters normal background activity (**Figure 3A**) for the normal value of c_{py-ei}. As c_{py-ei} becomes smaller, a supercritical Hopf bifurcation point (H_1) happens and the stable fixed points become unstable at c_{py-ei} ∼0.74. Upon approach of c_{py-ei} to ∼0.73, a period-doubling bifurcation (PD1) takes place and shapes the two-amplitude preictal spike pattern (**Figure 3B**). With further decrease of the c_{py-ei} parameter to ∼0.72, another period-doubling bifurcation (PD2) happens and then starts to form clonic seizure patterns (**Figure 3C**). In addition, with a further decrease in the c_{py-ei} parameter to ∼0.678, a bistable region including two stable limit cycles takes place between the first- and second-fold limit cycle bifurcations (LPC1 and LPC2) and shapes the SWD patterns (**Figure 3D**). As the c_{py-ei} parameter decreases to ∼0.591, the first subcritical Hopf bifurcation (H_2) occurs and the stable points appear again. The second bistable region consists of stable focus points and a stable limit cycle happens between the subcritical Hopf bifurcation (H_2) and the third fold limit cycle (LPC3) until c_{py-ei} is approaching to ~0.532. Upon decreasing the c_{py-ei} parameter, slow rhythmic activity (**Figure 3E**) happens, and then at c_{py-ei} ~0.451 the second supercritical Hopf bifurcation (H_3) arises from the steady state and shapes the tonic seizure patterns (**Figure 3F**).

Altogether, this model showed that impairment of GABAergic or glutamatergic receptors in the EI population causes the transition from normal background activity to seizures (preictal signal) and the transition between absence, tonic, and clonic seizures. This model shows the possible existence of supercritical Hopf bifurcation with growing amplitude oscillations when the transition from interictal to ictal states occurs. In order to have a deeper understanding of the interictal to ictal transition dynamics, we investigate the bifurcation of the cortical and sensory input frequencies in our model.

Hybrid Cooperation of GABAergic and Glutamatergic Receptors of Excitatory Interneurons in Epileptic Transition Dynamics

The genesis of generalized seizures requires interdependencies in different thalamocortical connections. Studies on genetic rat models have shown that the increasing excitatory coupling strength from the thalamus to the cortex may facilitate the maintenance and propagation of absence seizures (Sitnikova et al., 2008; Lüttjohann and Pape, 2019). However, based on the paradoxical behavior of glutamatergic and GABAergic receptors discussed in sections "Transition Dynamics Produced by GABAergic Receptor-Mediated Inhibition in Excitatory Interneurons," and ""Transition Dynamics Produced by Glutamatergic Receptor-Mediated Excitation in Excitatory Interneurons," here, we explore the effect of decreased glutamatergic synaptic strength from the thalamus to the neocortex on seizure propagation using the extended model. In this section, we investigate the effect of a dual collaboration of intra-EI GABAergic and glutamatergic synaptic strength in the cortex on the transition between different kinds of activity, either from interictal to ictal signal or from absence seizure to tonic and clonic seizures. Moreover, we explore the influence of the TC to EI pathway dominated by glutamatergic receptors on the initiation and propagation of epileptic activities based on the altered ratio of the dual cooperation of intra-EI GABAergic and glutamatergic receptors in the cortex. According to **Figure 4**, the pattern evolutions for different values of c_{tc-ei} are obtained to investigate the state transition produced by this model and their corresponding dominant frequency distributions as the parameters of c_{py-ei} and c_{i1-ei} varies in region [0.1,0.9] × [0.2,0.9].

For higher values of synaptic strength like $c_{tc-ei} = 4.5$ (**Figures 4A,B**), this model demonstrates seven different types of activities, when ci1-ei varies from 0.2 to 0.9. We denoted these seven activities as follows: type1, normal background activity (DF = 0 Hz in **Figure 4B**); type2, preictal activity (DF < 3.5 Hz in **Figure 4B**); type6, clonic seizure (DF $\leq$ 7 Hz in **Figure 4B**); type5, atypical absence seizure (DF > 4 Hz and DF < 2 Hz in **Figure 4B**); type4, typical absence seizure (2 Hz < DF < 4 Hz in **Figure 4B**); type3, slow rhythmic activity (DF = 0 Hz in **Figure 4B**); type7, tonic seizure (DF $\geq$ 14 Hz in **Figure 4B**). Upon decreasing the synaptic strength c_{tc-ei} from 4.5 to 4 (**Figures 4C,D**), one can see that the surface of the yellow region (atypical absence seizure; Type5) and red region (tonic seizure; Type7) increases with the reduction level of excitation of intra-EI. Accordingly, the regions correspond to normal background activity (Type1), preictal activity (Type2), clonic seizure (Type6), typical absence seizure (Type4), and slow rhythmic activity (Type3) decreases. Biologically, decreasing the excitatory coupling strength from TC to EI interrupts the normal activity of the thalamic relay nucleus in the thalamocortical network. This interruption facilitates the transition from interictal to ictal, i.e., with the increasing of c_{py-ei} and c_{i1-ei} from their normal state, the transition from normal background activity to ictal activity and also the transition between clonic, absence, and tonic seizures can occur more easily. Upon further decreasing of c_{tc-ei} to 3.5 (**Figures 4E,F**), the yellow and red regions correspond to atypical absence seizure (Type5) and tonic seizure (Type7) extend, respectively. However, the whole other regions decrease, which shows that the transition from normal background activity to epileptic seizure activities in this model can be influenced by the intra-EI GABAergic and glutamatergic receptors in the cortex under the impact of the thalamus to the cortex synaptic strength.

In summary, the intra-EI GABAergic and glutamatergic receptors in the cortex can elicit seven different kinds of activity in this model under the effect of excitatory synaptic strength from TC to EI. The lower excitatory synaptic strength of c_{tc-ei}, in the model causes a faster transition from interictal to ictal state. This implies the important role of the thalamus to cortex synaptic strength in the initiation and maintenance of generalized epileptic seizures.

Transition Dynamics Produced by Glutamatergic Receptor-Mediated Excitation in Reticular Nucleus

Increasing the GABAergic inhibition is predominantly believed to suppress epileptic seizures; however, studies (Klaassen et al., 2006; Cope et al., 2009; Wong, 2010) have reported that enhanced GABAergic inhibition in the brain promotes seizure as well. The reticular nucleus (RE) is mainly composed of inhibitory interneurons that their epileptogenic role in seizure is investigated by Liu and Wang (2017) in their thalamocortical model. Liu and Wang (2017) did not consider the EI population in their model when they investigated the role of TC to RE synaptic strength in epileptic activities. Therefore, here, we investigate how the strength of the excitatory synapses of TC affects the function of the inhibitory reticular nucleus population in seizure generation in the presence of EI. By increasing the strength of the excitatory synapses from TC to RE, we increase the inhibitory behavior of RE in the proposed thalamocortical model. In **Figure 5**, we show that the proposed model demonstrates a rich dynamic and that the transition from interictal to ictal occurs when we increase c_{tc-re}. In **Figure 5**, from left to right, first, we observe the normal background activity. After increasing the parameter c_{tc-re} to ~9.4, the transition from steady state to the limit cycle of preictal (supercritical Hopf bifurcation H1)

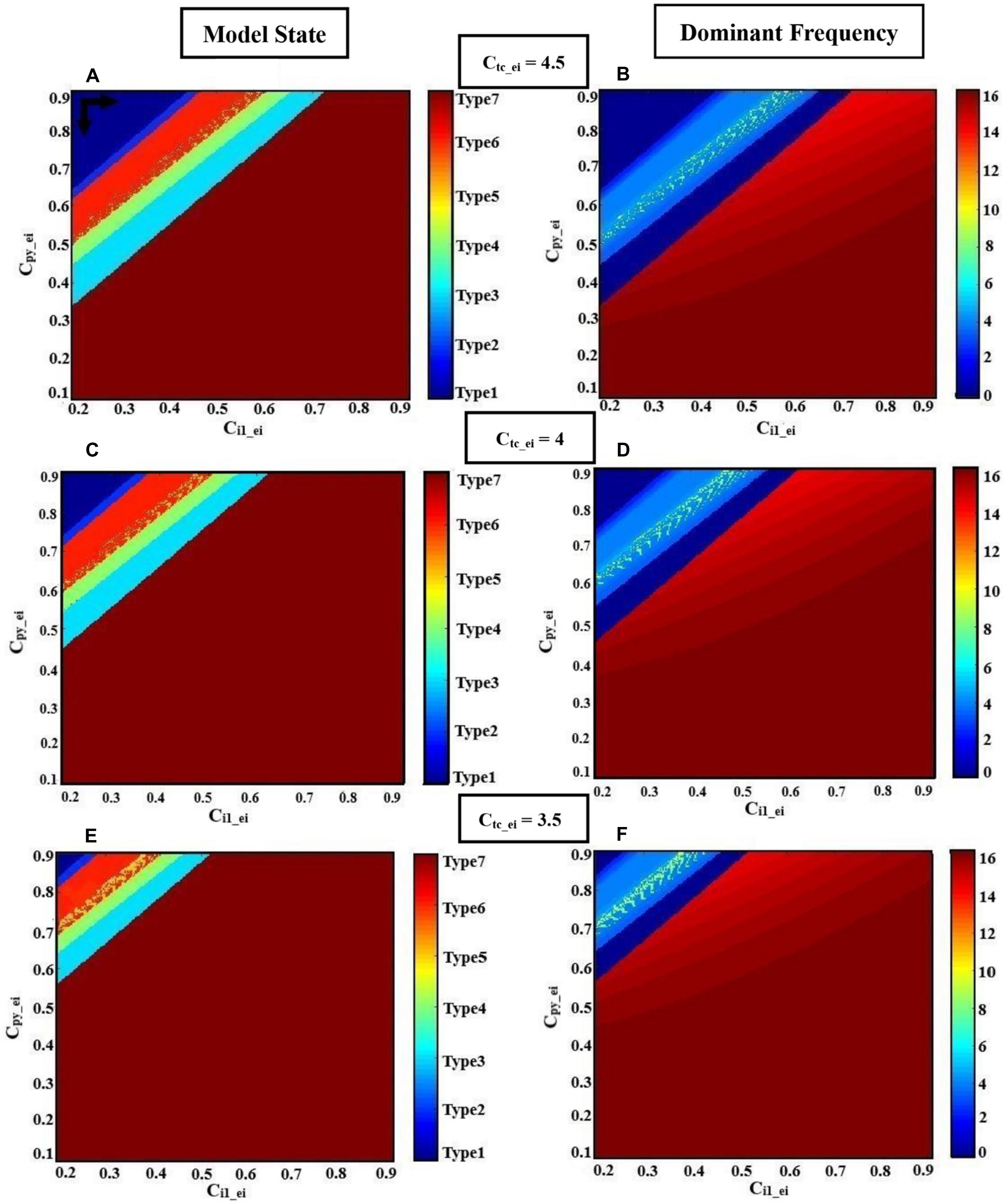


FIGURE 4 | Hybrid modulation of the model. Patterns of evolutions of the model states are shown when the glutamatergic and GABAergic synaptic strength of EI (c_{py-ei}, c_{i1-ei}) in the cortex are changed. Each row corresponds to a given TC to EI excitatory synaptic strength (c_{tc-ei}). The left column shows the states (normal or epileptic activities), and the right column shows the dominant frequencies (DF) of the model output. We set $a_{tc} = 0$ and $a_{py} = 0$. The various activities **(A,C,E)** and their corresponding dominant frequencies **(B,D,F)** with decreasing excitatory synaptic strength c_{tc-ei} from top to bottom, 4.5, 4, and 3.5, respectively, are shown. The model produces different firing states with variation of c_{py-ei} and c_{i1-ei}. These states are normal background firing (Type1; DF = 0 Hz), preictal activity (Type2; DF < 3.5 Hz), slow rhythmic activity (Type3; DF = 0 Hz), typical absence seizure (Type4; 2 Hz < DF < 4 Hz), atypical absence seizure (Type5; DF > 4 Hz and DF < 2 Hz), clonic seizure (Type6; DF ≤ 7 Hz), and tonic seizure (Type7; DF ≥ 14 Hz). The black horizontal and vertical arrows represent the transition routes from the normal state to the pathological states.

happens. Further increasing the c_{tc-re} to ∼9.58 and 9.63, the first and second period-doubling bifurcations (PD1, PD2) take place, respectively. After PD2, the limit cycles corresponding to clonic seizure take place. With increasing of c_{tc-re} to ∼10.1 the first fold limit cycle (LPC1) and then at $c_{tc-re} \approx 10.3$ the second fold limit cycles (LPC2), the model generates the first bistable

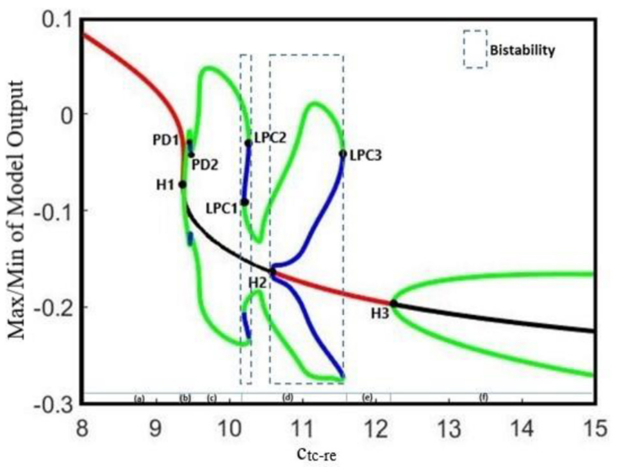


FIGURE 5 | The bifurcation diagram for the bifurcation parameter c_{tc-re}. Other parameters of the model are c_{py-ei} = 0.75, c_{i1-ei} = 0.33, and c_{tc-ei} = 4.2. We also set a_{tc} = 0 and a_{py} = 0. In the bifurcation diagram, blue and green cycles represent the unstable and stable limit cycles, respectively. The red and black lines represent the stable and unstable points, respectively. LPC1, LPC2, and LPC3 are fold limit cycle bifurcation, H1 and H3 are supercritical Hopf bifurcation, H2 is subcritical Hopf, and PD1 and PD2 are period doubling bifurcations.

region between LPC1 and LPC2, which shape the limit cycle corresponding to absence seizure activity. The second bistable region occurs between the subcritical Hopf bifurcation H2 at ∼10.6 and the third fold limit cycle LPC3 at ∼11.9. Then, at $c_{tc-re} \approx 11.8$ the transition from limit cycles of absence seizures and steady state occurs. Upon increasing of c_{tc-re} to ∼12.1, the transition from the stable equilibrium to the limit cycle of tonic seizure happens. By increasing the parameter c_{tc-re} to 9.5, we can observe the first supercritical Hopf bifurcation (H1) and the transition from interictal to preictal state. Then by further increasing c_{tc-re} and inducing more impairment in the TC to RE pathway, we can observe that limit cycles correspond to clonic and absence seizures, respectively.

Frequency Analysis of the Different Initial States of the Model

According to clinical observations, sensory (Dawson, 1947) and cortical (Bezard et al., 1999) stimulations can provoke epileptic seizures. Therefore, in this section, we examine the effect of sensory and cortical inputs in the interictal to ictal transition using the extended thalamocortical model. Studies on thalamocortical models (Haghighi and Markazi, 2017, 2019) have shown that changes in the frequencies of cortical and sensory periodic inputs can cause a transition from chaotic to periodic activities in the output of the model. Therefore, we will first examine the effect of frequencies of the cortical and sensory inputs on the transition behavior of our model, and even more importantly, we will discuss the effect of the initial state of the model on this transition.

Figure 6 shows various bifurcation diagrams for different initial states of the model, such as normal background activity (interictal), preictal activity, and clonic, absence, and tonic seizures. To change the initial state of the model, we used different values for parameter c_{py_ei}. These values were selected using the results of the bifurcation analysis shown in **Figure 3**, in which c_{py_ei} is the bifurcation parameter. In **Figures 6A,B**, we set c_{py_ei} = 0.76, which is in the range corresponding to interictal activities. In **Figures 6C,D**, by decreasing the parameter c_{py_ei} to 0.748, the initial state of the model corresponds to preictal activities. In **Figures 6E,F**, we change the initial state of the model to a clonic seizure state by decreasing the value of c_{py_ei} to 0.72. As it is shown in **Figures 6G–J**, by further decreasing c_{py_ei} to 0.58 and then to 0.4, we set the model in absence seizure and then in tonic seizure states, respectively.

On the other hand, for each state of the model, we changed the frequencies of cortical and sensory periodic inputs separately with a frequency step increment of 0.01 Hz. For the interictal initial state (**Figures 6A,B**), the model shows a linear resonant behavior with a resonant frequency of fpy ≈ 0.2 and ftc ≈ 4.7 as the cortical and sensory input frequencies increase. The model, in this case, behaves like a bandpass filter for the sensory input and a low pass filter for the cortical input. For c_{py_ei} = 0.748, when the frequency of cortical input is increased, a chaotic behavior is resulted. In **Figure 6C**, with further increase in the cortical input frequency (fpy), we notice a jump near the frequency of fpy ≈ 1.7. Moreover, by increasing the frequency of the sensory input, two jumps take place near the frequencies of ftc ≈ 1.8 and ftc ≈ 4 (**Figure 6D**). Then, upon increasing fpy and ftc, the model returns to its normal periodic behavior. This observation indicates that, when the initial state of the model changes from normal background activity to preictal activity, the behavior of the model changes from a linear resonator to a non-linear resonator. In **Figures 6E,F**, where the model is in the clonic seizure state, by increasing the frequencies of cortical and sensory inputs, we observe that the model first displays a noticeable chaotic behavior around 3 Hz, which is the main frequency of the clonic seizure, and then returns to its expected periodic behavior. In **Figures 6G,H**, when the model is in the absence seizure state, increasing the cortical and sensory input frequencies yields two chaotic behaviors with two different amplitude ranges; the reason for this is the coexistence of the two different SWD patterns with different amplitudes. This chaotic behavior can be seen around the main frequency of absence seizure (fpy ∈ [2, 4]). As we can see in **Figures 6I,J**, when the state of the model changes to tonic seizure activities, we observe a chaotic behavior around the main frequency of tonic seizure (fpy ≈ 16). As it is demonstrated in **Figures 6F,H,J**, when the model is in the seizure activity state, by increasing the frequency of sensory input, two peaks in the frequency response of the model are observed. These two peaks can be seen more clearly in tonic seizures (**Figure 6J**) and correspond to the two jumps already observed in **Figure 6D**.

Figure 7 demonstrates the effect of the initial state of the model on the transition between small-amplitude oscillation and large-amplitude seizure activities, all in the time domain. We set c_{py_ei} equal to 0.76 in order to adjust the model in its normal state, and then we change the model state from interictal (normal activity) to preictal activity by decreasing the c_{py_ei} parameter to 0.748. In **Figure 7A**, the model receives a cortical input with

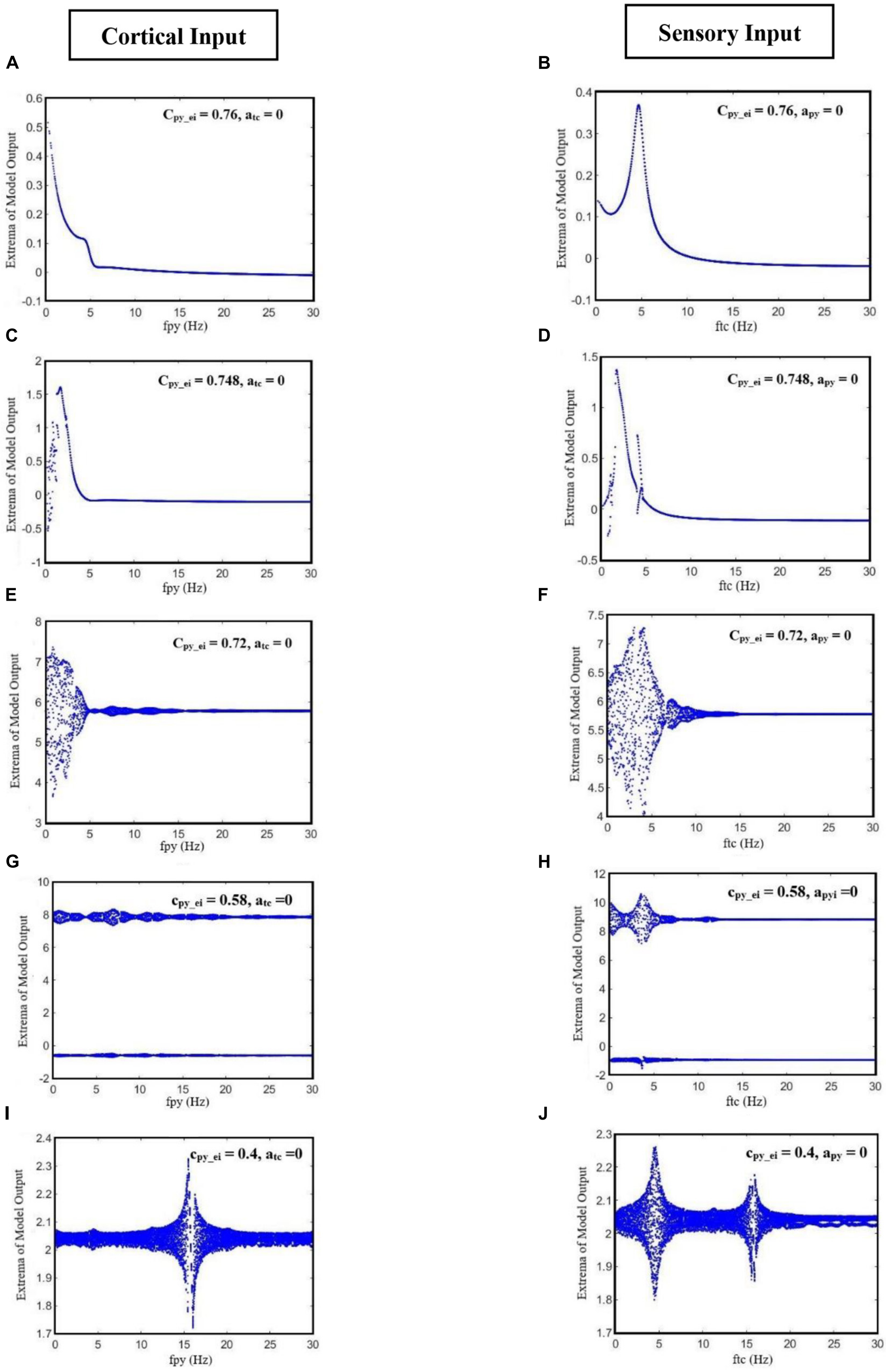


FIGURE 6 | The bifurcation diagrams of the model output for different values of c_{py-ei}, when the frequencies of the cortical input (fpy) and sensory input (ftc) change. Other parameters have the numerical values of c_{i1-ei} =0.3 and c_{tc-ei} =4.5. For the bifurcation diagrams shown in the left column **(A,C,E,G,I)**, fpy is the bifurcation parameter and a_{tc} has been set to zero; and for the bifurcation diagrams of the right column **(B,D,F,H,J)**, ftc is the bifurcation parameter and a_{py} is equal to zero.

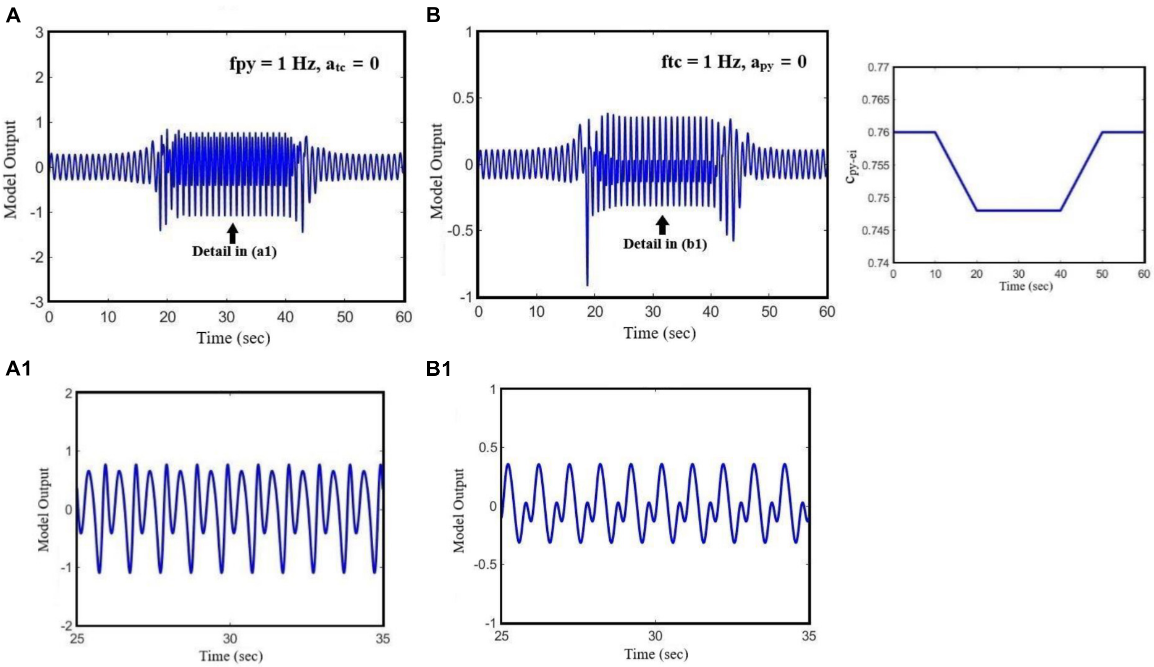


FIGURE 7 | The effect of state transition of the model on its output when the parameter of c_{py_ei} changes. The effect of the initial state of the model on the transition between periodic signal and seizure when the cortical and sensory input frequencies are constant. In **(A)**, we set the parameters of fpy = 1 Hz and atc = 0. In **(B)**, we set ftc = 1 HZ and a_{py} = 0. The pattern of glutamatergic synaptic strength, c_{py_ei}, is presented on the right-hand side diagram.

a fixed frequency, fpy = 1 Hz, in the absence of sensory input. However, in **Figure 7B**, we only consider the effect of the sensory input on the output of the model and we set its frequency ftc = 1 Hz. In **Figures 7A,B**, it is shown that based on the bifurcation diagram of **Figure 3**, by changing c_{py_ei} from 0.76 to 0.748, the state transition from normal activity to typical absence seizure activity occurs, and then it returns to the normal activity as c_{py_ei} returns to its initial value.

DISCUSSION

Traditionally, mechanisms underlying seizures have been considered to be due to increased excitation, decreased inhibition, or even both of them resulting in hyper-excitability and seizure generation. However, glutamatergic and GABAergic neurotransmitters can exert paradoxical effects and cause seizures. Antiepileptic drugs do not necessarily work by decreasing excitation and increasing the inhibition of neural activities. Studies Cossart et al. (2005), Fritschy (2008), Kaila et al. (2014), Knoflach et al. (2016), and Trevelyan (2016) have shown that some seizures occur when inhibition is enhanced in the brain. Therefore, it might be oversimplifying if one considers the role of GABAergic inhibition in the brain as the only antiepileptics. In fact, understanding the seemingly contradictory role of the excitatory and inhibitory neurons in the brain can lead to new therapies for epileptic seizures. Therefore, in the present study, we investigated the opposite effect of GABAergic and glutamatergic receptors in seizure onset through the dynamical analysis of an extended model. In this direction, we proposed an extended neural mass model, which considers the role of the spiny stellate cell population connectivity with other cortical neural populations in different states of seizure generation. The original thalamocortical model, which was developed by Liu and Wang (2017), was modified and extended and then used to simulate the preictal activity that has a prevailing role in epileptic seizure prediction. Using this model, we investigated the interactions between EI and the glutamatergic pyramidal and inhibitory interneurons in the cortex, and how they lead to epileptic seizures. This model enabled us to generate preictal activities before the clonic seizure. To make this point clearer, we calculated the bifurcation diagram of the output of the model for $\mathbf{c}_{i1-py}$, as the bifurcation parameter, in two cases: (1) without the EI population in the model and (2) at the presence of EI. As one can see in **Figures 8A,B** by increasing the inhibitory synaptic strength between PY and I1, first a clonic seizure and then a tonic seizure occur. In **Figure 8A**, when the EI population is not considered in the model, there exist no preictal activities (please notice the high-amplitude clonic activity after Hopf bifurcation H1). However, in **Figure 8B**, when the role of EI in the model is considered, the preictal activity emerges in the bifurcation diagram right after Hopf bifurcation H1 (please notice the low-amplitude preictal activity after H1). Therefore, one can conclude that although the order and list of bifurcations that occur when the inhibition between fast interneurons and pyramidal neurons increases are the same for both cases (extended model and the original model developed by Liu and Wang, 2017), the emergence of the preictal activities

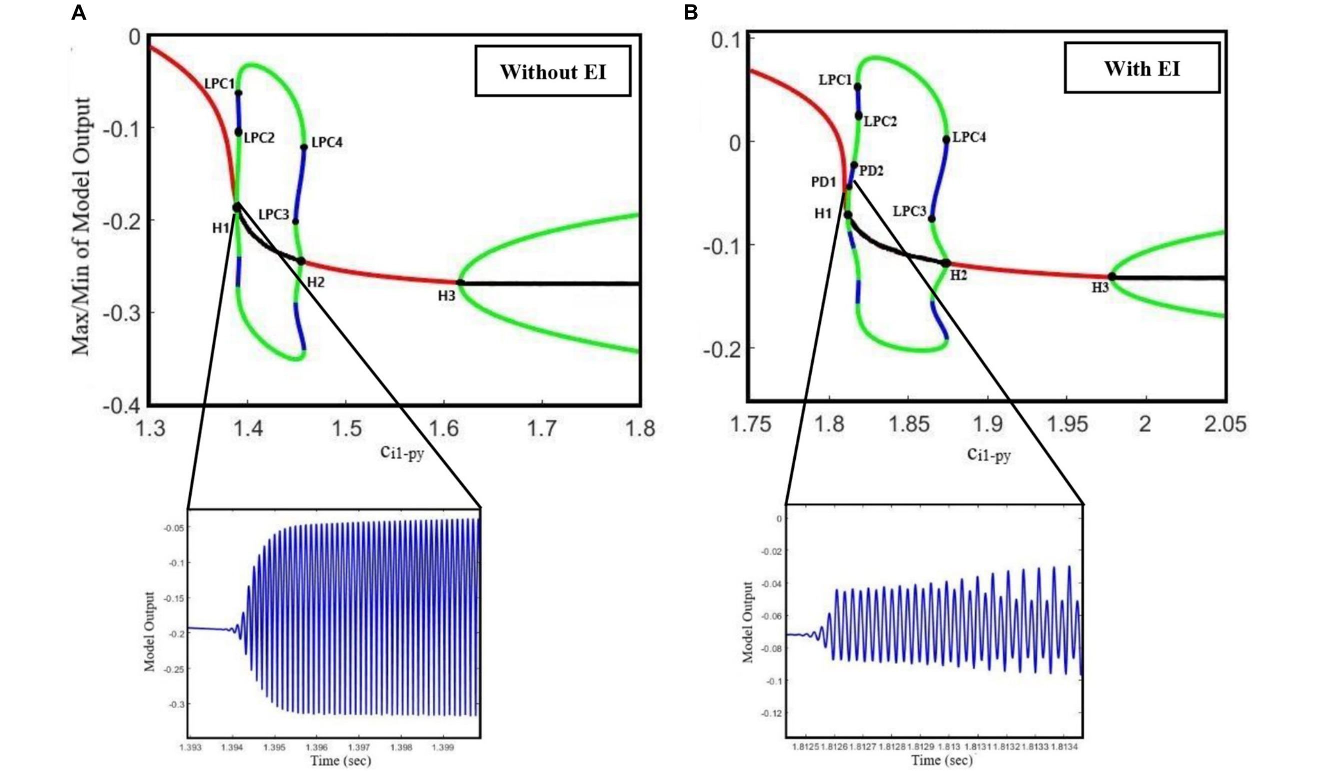


FIGURE 8 | (A) Bifurcation diagram of the output of the model when the excitatory interneurons are not considered in the model (parameters are set based on the caption of **Figure 1**). The bifurcation parameter is the inhibition strength between I1 (fast interneurons) and PY (pyramidal) population (c_{i1_py}). **(B)** Bifurcation diagram of the output of the model when the excitatory interneurons are considered in the model (c_{py_ei}= 0. 8, c_{i1-ei}= 0.3, c_{tc-ei}= 4.5). The bifurcation parameter is the inhibition strength between I1 (fast interneurons) and PY (pyramidal) population (c_{i1_py}).

in the model depends on the EI. Without these neurons (and certainly the feedback loop created by them), the model does not show any preictal activities and the seizure starts abruptly.

Our results show the richness of the dynamics that the proposed model can generate, including normal background activity, preictal spikes, slow rhythmic activity, clonic seizure, typical absence seizure, and tonic seizure, as we decreased and increased the glutamatergic and GABAergic synaptic strengths of the EI, respectively. Furthermore, bifurcation diagrams of the model were obtained by varying the coupling strength of GABAergic and glutamatergic receptors in EI. The diagrams in **Figures 2**, **3**, **5** show that the interictal to ictal transition occurs when we increased the glutamatergic excitation and GABAergic inhibition in the cortex. Based on our bifurcation diagrams, pathological transitions are consequences of supercritical and subcritical Hopf bifurcations as well as the fold limit cycle bifurcation. The onset of preictal discharges is characterized by a supercritical Hopf bifurcation. This results in a preictal activity, which is characterized by growths in amplitude and frequency of neural activities. As mentioned before, Liu and Wang (2017) developed a thalamocortical model, which is originally inspired by Taylor et al. (2014) and Fan et al. (2015) to study the epileptogenic role of the synaptic strength between TC and ER in the thalamus. Using that model, they did not consider the spiny stellate cell population and its role in epileptic seizure generation. One of the most important features lacking in their results was the preictal state of the model. In fact, the model was not able to generate the limit cycles of preictal state in their bifurcation analysis when they investigated the excitatory pathway from TC to RE. As a result, in their bifurcation diagrams there was an abrupt appearance of ictal limit cycles after the normal background activities. However, in **Figure 5** we evaluated the extended thalamocortical model in a more general framework and we showed that this model is capable of generating the preictal state when the synaptic strength from TC to RE is changed. We showed that considering the role of EI was crucial for the emergence of the preictal state. Using this model, we demonstrated that ictal discharges do not appear abruptly after a period of interictal activities. As it was shown in the bifurcation diagrams (**Figures 2**, **3**, **5**), we can simulate the gradual increase in frequency and amplitude from normal activities to ictal activities. Consequently, the model demonstrates the key role of the spiny stellate cells of the cortex in interictal to ictal transition dynamics.

In order to have a deeper insight into the dynamics of interictal to ictal transition, we also explored the effect of cortical and sensory periodic (sinusoidal) inputs on the output of the extended and modified model. Our simulation results reveal that the transition from normal activity to seizure activity occurs when a perturbation in the dynamic structure of the model relocates the model state from interictal to preictal state. Our results show that during the preictal state, the extended model is more susceptible to sensory and cortical inputs, which in turn causes typical absence seizures. Moreover, the frequency response of the model (**Figure 6**) makes it clear that model responses to the cortical and sensory input are dependent on the initial state of the model. According to the bifurcation analysis depicted in **Figure 6**, when the model is in its normal state, it behaves as a linear resonator. Therefore, when we stimulate the model with sensory and cortical stimuli, the model output does not show a chaotic behavior and comes back to its normal behavior as the input cortical and sensory frequencies are increased. On the other hand, when any impairments occur in the glutamatergic receptors of the EI, these cause the model to enter the preictal state. Then, the same sensory and cortical stimulations can cause the model to act as a non-linear resonator and we can observe chaotic behaviors and jump phenomenon. This non-linear behavior increases when the initial state of the model changes to the ictal state. In general, the cortical stimulations evoke more peak-to-peak amplitudes of cortical output in this model, in comparison to sensory stimulations, when it is in clonic and tonic seizure states. On the contrary, the peak-to-peak amplitude of the model output when it is in the seizure activity state is more intense when it receives sensory stimulations. This demonstrates that absence seizures are more sensitive to sensory stimuli than cortical stimulations. Also, we investigated the effect of the initial state of the model in absence seizure generation. As is shown in **Figure 7**, when the extended model is in its normal state, we observe the low-amplitude activities, which correspond to normal background activity. However, by changing the model state from normal to preictal state, we can observe that the absence seizures take place. According to our observation, we suggest that alteration in the initial states of the brain can be considered as one of the principal causes of the absence and photosensitive seizures.

Finally, we examined the role of the thalamus in epileptic seizure activities. We investigated the role of cooperation of glutamatergic and GABAergic receptors of EI under the effect of the thalamic relay nucleus (TC) in seizure generation and propagation. As is shown in **Figure 4**, the results have shown the dependence of cortical function on the thalamic one in the initiation, propagation, and termination of epileptic seizures. Recent studies Chang et al. (2020) and Prathaban and Balasubramanian (2020) did not consider the thalamus and its synaptic connectivity with other neuronal populations such as EI in the cortical models they worked with. However, based on the bifurcation analysis obtained from the hybrid modulation of intra-EI excitatory and inhibitory synaptic strength in the extended neural mass model, the pathway from the thalamic relay nucleus to the EI facilitates the transitions between different types of epileptic activities, either from interictal to ictal or between clonic, absence, and tonic seizures. We believe that these results provide useful insights for understanding more thoroughly the function of the EI population and dynamics caused by their connections to other populations in the thalamocortical circuitry and the transitions between interictal to preictal and ictal states. Therefore, the extended model is more suitable to be used in epileptic seizure prediction and abatement.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found in the following address: https://github.com/SabaTabatabaee/Neural-Mass-model-for-Epilepsy.

AUTHOR CONTRIBUTIONS

ST, FB, and MJ guaranteed the integrity of the entire study. ST contributed to literature research, formal analysis, and writing the original draft. FB and MJ contributed to project administration and reviewed and edited the manuscript. All authors contributed to conceptual design of the study, methodology, and validation.

REFERENCES

Amari, S. I. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biol. Cybern.* 27, 77–87. doi: 10.1007/BF00337259

Bartolomei, F., Wendling, F., Regis, J., Gavaret, M., Guye, M., and Chauvel, P. (2004). Pre-ictal synchronicity in limbic networks of mesial temporal lobe epilepsy. *Epilepsy Res.* 61, 89–104. doi: 10.1016/j.eplepsyres.2004.06.006

Bezard, E., Boraud, T., Nguyen, J. P., Velasco, F., Keravel, Y., and Gross, C. (1999). Cortical stimulation and epileptic seizure: a study of the potential risk in primates. *Neurosurgery* 45, 346–350. doi: 10.1097/00006123-199908000-00030

Castejon, C., Barros-Zulaica, N., and Nuñez, A. (2016). Control of somatosensory cortical processing by thalamic posterior medial nucleus: a new role of thalamus in cortical function. *PLoS One* 11:e0148169. doi: 10.1371/journal.pone.0148169

Chang, S., Wei, X., Su, F., Liu, C., Yi, G., Wang, J., et al. (2020). Model predictive control for seizure suppression based on nonlinear auto-regressive moving-average volterra model. *IEEE Trans. Neural Syst. Rehabil. Eng.* 28, 2173–2183. doi: 10.1109/TNSRE.2020.3014927

Chaves, J., and Sander, J. W. (2005). Seizure aggravation in idiopathic generalized epilepsies. *Epilepsia* 46, 133–139.

Cope, D. W., Di Giovanni, G., Fyson, S. J., Orbán, G., Errington, A. C., Lőrincz, M. L., et al. (2009). Enhanced tonic GABA A inhibition in typical absence epilepsy. *Nat. Med.* 15, 1392–1398. doi: 10.1038/nm.2058

Cortez, M. A., Wu, Y., Gibson, K. M., and Snead, O. C. III (2004). Absence seizures in succinic semialdehyde dehydrogenase deficient mice: a model of juvenile absence epilepsy. *Pharmacol. Biochem. Behav.* 79, 547–553. doi: 10.1016/j.pbb.2004.09.008

Cossart, R., Bernard, C., and Ben-Ari, Y. (2005). Multiple facets of GABAergic neurons and synapses: multiple fates of GABA signalling in epilepsies. *Trends Neurosci.* 28, 108–115. doi: 10.1016/j.tins.2004.11.011

da Costa, N. M., and Martin, K. A. (2011). How thalamus connects to spiny stellate cells in the cat's visual cortex. *J. Neurosci.* 31, 2925–2937. doi: 10.1523/JNEUROSCI.5961-10.2011

Danober, L., Deransart, C., Depaulis, A., Vergnes, M., and Marescaux, C. (1998). Pathophysiological mechanisms of genetic absence epilepsy in the rat. *Prog. Neurobiol.* 55, 27–57. doi: 10.1016/s0301-0082(97)00091-9

Dawson, G. D. (1947). Investigations on a patient subject to myoclonic seizures after sensory stimulation. *J. Neurol. Neurosurg. Psychiatry* 10:141. doi: 10.1136/jnnp.10.4.141

Errington, A. C., Gibson, K. M., Crunelli, V., and Cope, D. W. (2011). Aberrant GABA A receptor-mediated inhibition in cortico-thalamic networks of succinic semialdehyde dehydrogenase deficient mice. *PLoS One* 6:e19021. doi: 10.1371/journal.pone.0019021

Fan, D., Wang, Q., and Perc, M. (2015). Disinhibition-induced transitions between absence and tonic-clonic epileptic seizures. *Sci. Rep.* 5:12618. doi: 10.1038/srep12618

Feldmeyer, D., Lübke, J., Silver, R. A., and Sakmann, B. (2002). Synaptic connections between layer 4 spiny neurone-layer 2/3 pyramidal cell pairs in juvenile rat barrel cortex: physiology and anatomy of interlaminar signalling within a cortical column. *J. Physiol.* 538, 803–822. doi: 10.1113/jphysiol.2001.012959

Fisher, R. S. (2017). The new classification of seizures by the international league against epilepsy 2017. *Curr. Neurol. Neurosci. Rep.* 17:48.

Fritschy, J. M. (2008). Epilepsy, E/I balance and GABAA receptor plasticity. *Front. Mol. Neurosci.* 1:5. doi: 10.3389/neuro.02.005.2008

Haghighi, H. S., and Markazi, A. H. (2017). A new description of epileptic seizures based on dynamic analysis of a thalamocortical model. *Sci. Rep.* 7:13615. doi: 10.1038/s41598-017-13126-4

Haghighi, H. S., and Markazi, A. H. (2019). Dynamic origin of spike and wave discharges in the brain. *Neuroimage* 197, 69–79. doi: 10.1016/j.neuroimage.2019.04.047

Huberfeld, G., de La Prida, L. M., Pallud, J., Cohen, I., Le Van Quyen, M., Adam, C., et al. (2011). Glutamatergic pre-ictal discharges emerge at the transition to seizure in human epilepsy. *Nat. Neurosci.* 14, 627–634. doi: 10.1038/nn.2790

Jiang, S., Luo, C., Gong, J., Peng, R., Ma, S., Tan, S., et al. (2018). Aberrant thalamocortical connectivity in juvenile myoclonic epilepsy. *Int. J. Neural Syst.* 28:1750034. doi: 10.1142/S0129065717500344

Jones, N. C., Salzberg, M. R., Kumar, G., Couper, A., Morris, M. J., and O'Brien, T. J. (2008). Elevated anxiety and depressive-like behavior in a rat model of genetic generalized epilepsy suggesting common causation. *Exp. Neurol.* 209, 254–260. doi: 10.1016/j.expneurol.2007.09.026

Kaila, K., Ruusuvuori, E., Seja, P., Voipio, J., and Puskarjov, M. (2014). GABA actions and ionic plasticity in epilepsy. *Curr. Opin. Neurobiol.* 26, 34–41. doi: 10.1016/j.conb.2013.11.004

Khalilov, I., Holmes, G. L., and Ben-Ari, Y. (2003). In vitro formation of a secondary epileptogenic mirror focus by interhippocampal propagation of seizures. *Nat. Neurosci.* 6, 1079–1085. doi: 10.1038/nn1125

Khan, A. A., Shekh-Ahmad, T., Khalil, A., Walker, M. C., and Ali, A. B. (2018). Cannabidiol exerts antiepileptic effects by restoring hippocampal interneuron functions in a temporal lobe epilepsy model. *Br. J. Pharmacol.* 175, 2097–2115. doi: 10.1111/bph.14202

Klaassen, A., Glykys, J., Maguire, J., Labarca, C., Mody, I., and Boulter, J. (2006). Seizures and enhanced cortical GABAergic inhibition in two mouse models of human autosomal dominant nocturnal frontal lobe epilepsy. *Proc. Natl. Acad. Sci. U.S.A.* 103, 19152–19157. doi: 10.1073/pnas.0608215103

Knoflach, F., Hernandez, M. C., and Bertrand, D. (2016). GABAA receptor-mediated neurotransmission: not so simple after all. *Biochem. Pharmacol.* 115, 10–17. doi: 10.1016/j.bcp.2016.03.014

Law, N., Smith, M. L., and Widjaja, E. (2018). Thalamocortical connections and executive function in pediatric temporal and frontal lobe epilepsy. *Am. J. Neuroradiol.* 39, 1523–1529. doi: 10.3174/ajnr.A5691

Liu, S., and Wang, Q. (2017). Transition dynamics of generalized multiple epileptic seizures associated with thalamic reticular nucleus excitability: a computational study. *Commun. Nonlinear Sci. Numer. Simul.* 52, 203–213.

Lozovaya, N., Gataullina, S., Tsintsadze, T., Tsintsadze, V., Pallesi-Pocachard, E., Minlebaev, M., et al. (2014). Selective suppression of excessive GluN2C expression rescues early epilepsy in a tuberous sclerosis murine model. *Nat. Commun.* 5:4563. doi: 10.1038/ncomms5563

Lüttjohann, A., and Pape, H. C. (2019). Regional specificity of cortico-thalamic coupling strength and directionality during waxing and waning of spike and wave discharges. *Sci. Rep.* 9:2100. doi: 10.1038/s41598-018-37985-7

Maheshwari, A., Nahm, W., and Noebels, J. (2013). Paradoxical proepileptic response to NMDA receptor blockade linked to cortical interneuron defect in stargazer mice. *Front. Cell. Neurosci.* 7:156. doi: 10.3389/fncel.2013.00156

Mayville, C., Fakhoury, T., and Abou-Khalil, B. (2000). Absence seizures with evolution into generalized tonic-clonic activity: clinical and EEG features. *Epilepsia* 41, 391–394. doi: 10.1111/j.1528-1157.2000.tb00178.x

Moghim, N., and Corne, D. W. (2014). Predicting epileptic seizures in advance. *PLoS One* 9:e99334. doi: 10.1371/journal.pone.0099334

Montenegro, M. A., Guerreiro, M. M., Caldas, J. P., Moura-Ribeiro, M. V., and Guerreiro, C. A. (2001). Epileptic manifestations induced by midazolam in the neonatal period. *Arq. NeuroPsiquiatr.* 59, 242–243. doi: 10.1590/s0004-282x2001000200018

Neubauer, F. B., Sederberg, A., and MacLean, J. N. (2014). Local changes in neocortical circuit dynamics coincide with the spread of seizures to thalamus in a model of epilepsy. *Front. Neural Circuits* 8:101. doi: 10.3389/fncir.2014.00101

Okhotin, V. E. (2006). Cytophysiology of spiny stellate cells in the striate cortex and their role in the excitatory mechanisms of intracortical synaptic circulation. *Neurosci. Behav. Physiol.* 36, 825–836.

Pinault, D., and O'Brien, T. J. (2005). Cellular and network mechanisms of genetically-determined absence seizures. *Thalamus Relat. Syst.* 3:181. doi: 10.1017/S1472928807000209

Prathaban, B. P., and Balasubramanian, R. (2020). Prediction of epileptic seizures using Grey Wolf optimized model driven mathematical approach. *Microprocess. Microsyst.* 103370. doi: 10.1016/j.micpro.2020.103370

Rich, S., Chameh, H. M., Rafiee, M., Ferguson, K., Skinner, F. K., and Valiante, T. A. (2020). Inhibitory network bistability explains increased interneuronal activity prior to seizure onset. *Front. Neural Circuits* 13:81. doi: 10.3389/fncir.2019.00081

Rogawski, M. A. (2011). Revisiting AMPA receptors as an antiepileptic drug target: revisiting AMPA receptors as an antiepileptic drug target. *Epilepsy Curr.* 11, 56–63. doi: 10.5698/1535-7511-11.2.56

Scheffer, I. E., Berkovic, S., Capovilla, G., Connolly, M. B., French, J., Guilhoto, L., et al. (2017). ILAE classification of the epilepsies: position paper of the ILAE commission for classification and terminology. *Epilepsia* 58, 512–521.

Seal, R. P., Akil, O., Yi, E., Weber, C. M., Grant, L., Yoo, J., et al. (2008). Sensorineural deafness and seizures in mice lacking vesicular glutamate transporter 3. *Neuron* 57, 263–275. doi: 10.1016/j.neuron.2007.11.032

Shan, W., Mao, X., Wang, X., Hogan, R. E., and Wang, Q. (2021). Potential surgical therapies for drug-resistant focal epilepsy. *CNS Neurosci. Ther.* 27, 994–1011. doi: 10.1111/cns.13690

Sitnikova, E., Dikanev, T., Smirnov, D., Bezruchko, B., and Van Luijtelaar, G. (2008). Granger causality: cortico-thalamic interdependencies during absence seizures in WAG/Rij rats. *J. Neurosci. Methods* 170, 245–254. doi: 10.1016/j.jneumeth.2008.01.017

Steriade, M. (1974). Interneuronal epileptic discharges related to spike-and-wave cortical seizures in behaving monkeys. *Electroencephalogr. Clin. Neurophysiol.* 37, 247–263. doi: 10.1016/0013-4694(74)90028-5

Steriade, M., and Contreras, D. (1998). Spike-wave complexes and fast components of cortically generated seizures. I. Role of neocortex and thalamus. *J. Neurophysiol.* 80, 1439–1455. doi: 10.1152/jn.1998.80.3.1439

Sun, Q. Q., Huguenard, J. R., and Prince, D. A. (2006). Barrel cortex microcircuits: thalamocortical feedforward inhibition in spiny stellate cells is mediated by a small number of fast-spiking interneurons. *J. Neurosci.* 26, 1219–1230. doi: 10.1523/JNEUROSCI.4727-04.2006

Taylor, P. N., Thomas, J., Sinha, N., Dauwels, J., Kaiser, M., Thesen, T., et al. (2015). Optimal control based seizure abatement using patient derived connectivity. *Front. Neurosci.* 9:202. doi: 10.3389/fnins.2015.00202

Taylor, P. N., Wang, Y., Goodfellow, M., Dauwels, J., Moeller, F., Stephani, U., et al. (2014). A computational study of stimulus driven epileptic seizure abatement. *PLoS One* 9:e114316. doi: 10.1371/journal.pone.0114316

Thomas, P., Valton, L., and Genton, P. (2006). Absence and myoclonic status epilepticus precipitated by antiepileptic drugs in idiopathic generalized epilepsy. *Brain* 129, 1281–1292. doi: 10.1093/brain/awl047

Tran, C. H., Vaiana, M., Nakuci, J., Somarowthu, A., Goff, K. M., Goldstein, N., et al. (2020). Interneuron desynchronization precedes seizures in a mouse model of Dravet syndrome. *J. Neurosci.* 40, 2764–2775. doi: 10.1523/JNEUROSCI.2370-19.2020

Trevelyan, A. J. (2016). Do cortical circuits need protecting from themselves? *Trends Neurosci.* 39, 502–511. doi: 10.1016/j.tins.2016.06.002

Vashchinkina, E., Panhelainen, A., Aitta-Aho, T., and Korpi, E. R. (2014). GABAA receptor drugs and neuronal plasticity in reward and aversion: focus on the ventral tegmental area. *Front. Pharmacol.* 5:256. doi: 10.3389/fphar.2014.00256

Velazquez, J. L., Huo, J. Z., Dominguez, L. G., Leshchenko, Y., and Snead, I. I. I. O. C. (2007). Typical versus atypical absence seizures: network mechanisms of the spread of paroxysms. *Epilepsia* 48, 1585–1593. doi: 10.1111/j.1528-1167.2007.01120.x

Wong, M. (2010). Too much inhibition leads to excitation in absence epilepsy. *Epilepsy Curr.* 10, 131–133. doi: 10.1111/j.1535-7511.2010.01379.x

Zhang, C. H., Sha, Z., Mundahl, J., Liu, S., Lu, Y., Henry, T. R., et al. (2015). Thalamocortical relationship in epileptic patients with generalized spike and wave discharges—a multimodal neuroimaging study. *NeuroImage* 9, 117–127. doi: 10.1016/j.nicl.2015.07.014

Zhang, Y., Jiang, L., Zhang, D., Wang, L., Fei, X., Liu, X., et al. (2020). Thalamocortical structural connectivity abnormalities in drug-resistant generalized epilepsy: a diffusion tensor imaging study. *Brain Res.* 1727:146558. doi: 10.1016/j.brainres.2019.146558

Zhang, Y., Xu, J., Zhang, K., Yang, W., and Li, B. (2018). The anticonvulsant effects of ketogenic diet on epileptic seizures and potential mechanisms. *Curr. Neuropharmacol.* 16, 66–70. doi: 10.2174/1570159X15666170517153509

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 03 January 2022
doi: 10.3389/fnins.2021.744190

Toward a Multimodal Computer-Aided Diagnostic Tool for Alzheimer's Disease Conversion

Danilo Pena[1], Jessika Suescun[2], Mya Schiess[2], Timothy M. Ellmore[3], Luca Giancardo[1] and the Alzheimer's Disease Neuroimaging Initiative*

[1] Center for Precision Health, School of Biomedical Informatics, University of Texas Health Science Center, Houston, TX, United States, [2] Department of Neurology, McGovern Medical School, University of Texas Health Science Center, Houston, TX, United States, [3] Department of Psychology, The City College of New York, New York, NY, United States

Alzheimer's disease (AD) is a progressive neurodegenerative disorder. It is one of the leading sources of morbidity and mortality in the aging population AD cardinal symptoms include memory and executive function impairment that profoundly alters a patient's ability to perform activities of daily living. People with mild cognitive impairment (MCI) exhibit many of the early clinical symptoms of patients with AD and have a high chance of converting to AD in their lifetime. Diagnostic criteria rely on clinical assessment and brain magnetic resonance imaging (MRI). Many groups are working to help automate this process to improve the clinical workflow. Current computational approaches are focused on predicting whether or not a subject with MCI will convert to AD in the future. To our knowledge, limited attention has been given to the development of automated computer-assisted diagnosis (CAD) systems able to provide an AD conversion diagnosis in MCI patient cohorts followed longitudinally. This is important as these CAD systems could be used by primary care providers to monitor patients with MCI. The method outlined in this paper addresses this gap and presents a computationally efficient pre-processing and prediction pipeline, and is designed for recognizing patterns associated with AD conversion. We propose a new approach that leverages longitudinal data that can be easily acquired in a clinical setting (e.g., T1-weighted magnetic resonance images, cognitive tests, and demographic information) to identify the AD conversion point in MCI subjects with AUC = 84.7. In contrast, cognitive tests and demographics alone achieved AUC = 80.6, a statistically significant difference ($n = 669$, $p < 0.05$). We designed a convolutional neural network that is computationally efficient and requires only linear registration between imaging time points. The model architecture combines Attention and Inception architectures while utilizing both cross-sectional and longitudinal imaging and clinical information. Additionally, the top brain regions and clinical features that drove the model's decision were investigated. These included the thalamus, caudate, planum temporale, and the Rey Auditory Verbal Learning Test. We believe our method could be easily translated into the healthcare setting as an objective AD diagnostic tool for patients with MCI.

Keywords: mild cognitive impairment, ADNI, longitudinal, deep learning, neuroimaging, clinical features, multimodal

OPEN ACCESS

Edited by:
Ali Ghazizadeh,
Sharif University of Technology, Iran

Reviewed by:
Seyed-Mahdi Khaligh-Razavi,
Massachusetts Institute of Technology, United States
Seyed Amir Hossein Batouli,
Tehran University of Medical Sciences, Iran

****Correspondence:***
Luca Giancardo
luca.giancardo@uth.tmc.edu

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *19 July 2021*
Accepted: *09 November 2021*
Published: *03 January 2022*

Citation:
Pena D, Suescun J, Schiess M, Ellmore TM, Giancardo L and the Alzheimer's Disease Neuroimaging Initiative (2022) Toward a Multimodal Computer-Aided Diagnostic Tool for Alzheimer's Disease Conversion. Front. Neurosci. 15:744190. doi: 10.3389/fnins.2021.744190

INTRODUCTION

Alzheimer's disease (AD) is a progressive cognitive decline that severely disrupts activities of daily living. It is estimated that the number of people affected by AD will triple to over 120 million people by 2050, costing the United States alone billions of dollars in healthcare expenses (Lane et al., 2018). Further, no medications are currently available that can either reverse or stop the cognitive decline in subjects with AD. There is a clear need to develop novel treatments for those with AD. To accomplish this, early detection and identification of AD will facilitate the development of biomarkers and support the discovery of novel molecules by providing the right population for clinical trials.

Early dementia detection is paramount to decrease the chance of further comorbidities and mortality (Ahmed et al., 2019). This is especially relevant in clinical environments outside large academic centers, such as community hospitals, where resources are limited. Subjects with mild cognitive impairment (MCI) have many of the neurological deficits found in AD subjects. Additionally, about 10–15% of subjects with MCI will progress to AD every year (Plassman et al., 2008). This estimate is variable, with higher rates in clinical centers and some treatment trials; and lower numbers in population-based studies. Hence, subjects with MCI represent the perfect prodromal population for the exploration of conversion biomarkers, which has been one focus of the neurocognitive field (Desikan et al., 2010; Jack et al., 2010; Landau et al., 2010; Young et al., 2014; Liu et al., 2017; Ottoy et al., 2019; Giorgio et al., 2020). Creating a computer-assisted diagnosis (CAD) tool would provide an objective instrument for early AD diagnosis in patients with MCI. The vast majority of community hospitals can perform basic neuropsychological assessments and T1 magnetic resonance imaging (MRI); as such, we propose a multi-modal approach that combines both data sources to objectively and efficiently confirm the AD diagnosis in patients with MCI (which are at high risk of conversion).

For years, researchers have been investigating neuroimaging-based biomarkers in conjunction with computational tools to find early signs of AD within MCI subjects. Studies have looked at the differences between all of the combinations of healthy controls (CN), AD subjects, MCI subjects who have converted to AD (cMCI), and MCI subjects who have stayed stable (sMCI) (Mateos-Pérez et al., 2018). To determine the crucial features of an MCI subject which eventually converts to AD, we decided to focus on a cMCI vs. sMCI comparison. Current works have combined many types of data and a host of machine learning techniques. Recent papers have used T1-weighted MRI images and linear support vector machines (Sun et al., 2017; Tong et al., 2017), positron emission tomography (PET) and random forests (Nozadi et al., 2018), clinical information/neuropsychological measurements with ensemble learning (Grassi et al., 2019), and T1-weighted and diffusion MRI with linear models (Xu et al., 2019) to predict MCI conversion. However, many of these techniques require dimensionality reduction techniques, feature selection, lengthy image pre-processing pipelines, and other tabular data transformations that all require *a priori* hypotheses and increase the model and hyperparameter search space (Moradi et al., 2015; Ahmed et al., 2019).

Thus, scientists have turned to deep learning methods to abstract some of these steps that may incur bias throughout the pipeline. This class of models allows the incorporation of different types of data that form complex, non-linear relationships that could potentially provide more information about the conversion risk of an MCI subject. Some of the recent deep learning techniques for MCI classification use multimodal data types. These include T1-weighted MRI imaging with clinical variables (Spasov et al., 2019), cerebrospinal fluid imaging and longitudinal brain volumetric features (Lee et al., 2019a), T1-weighted and hippocampal imaging (Li et al., 2019), and a recurrent neural network (RNN) structure that uses cerebrospinal fluid, cognitive, and imaging biomarkers (Lee et al., 2019b). Using an array of data has been shown to have additive effects over using one data type alone for MCI classification. Researchers are also interested in developing a better understanding of the disease progression. Groups have predicted MCI clinical trajectories through a longitudinal feature framework (Bhagwat et al., 2018) and have used gray matter density maps at multiple time points as inputs to an RNN (Cui et al., 2019). This extension of data through time within one subject's trajectory has proven a complicated but necessary problem to be able to incorporate all potential clinically available data (Lawrence et al., 2017). This is a non-exhaustive list of neuroimaging deep learning models for AD/MCI detection and prediction, and we refer to recent comprehensive reviews (Rathore et al., 2017; Ansart et al., 2021) for a complete list. This body of work focuses on the prediction of future AD in MCI subjects or diagnosis of AD using cohorts of subjects included in studies after their AD diagnosis, and therefore likely to have the disease for many years. To our knowledge, limited to no attention has been given to the development of automated CAD systems able to diagnose the conversion from MCI to AD, in patient cohorts followed longitudinally. This is important as these CAD systems could be used by neurologists and non-specialized physicians to monitor their MCI patients.

In our work, we propose to fill in this gap with a model that combines multi-modal longitudinal data that can be easily acquired in the vast majority of clinical settings in the industrialized world (e.g., T1-weighted magnetic resonance images, cognitive tests, and demographic information). This model is based on a compact convolutional neural network architecture that combines Attention and Inception modules which is computationally efficient and requires only linear registration between imaging time points. We test the conversion diagnosis performance of our model in a cohort of subjects that received a confirmed AD diagnosis after having MCI in a previous visit (cMCI) and subjects that remained with a stable MCI diagnosis (sMCI). Our dataset has a relatively large sample size (440 sMCI vs. 229 cMCI) compared to related methodological studies, which has been a common criticism (Mateos-Pérez et al., 2018).

MATERIALS AND METHODS

Data

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database[1] in October 2019. The ADNI was launched in 2003 as a public–private partnership led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD.

Demographic information used in this study is displayed in **Table 1**. The majority of subjects were categorized as white (>93%) and non-Hispanic (>97%). Differences between sex counts were tested using Fisher's exact test, and differences in baseline age, time between sessions, and years of education were evaluated with Wilcoxon rank-sum tests. *p*-Values of less than 0.05 were considered statistically significant.

For each subject, T1-weighted structural magnetic resonance images (MRI) were taken at two different time points in addition to clinical and demographic variables (age, sex, and years of education) available from ADNI. The clinical variables included APOe4 genotypes, neuropsychological cognitive tests like Montreal Cognitive Assessment (MoCA), Mini-Mental State Exam (MMSE), and the Dementia Rating Scale (CDRSB), the AD Assessment Scale (ADAS13, ADAS11, and ADASQ4), memory evaluations from the Rey Auditory Verbal Learning Test (RAVLT), and the functional activities questionnaire (FAQ). Additionally, we used AD and CN subjects to pre-train the model, and these subjects' demographics are in **Supplementary Material**.

The time points used for cMCI subjects were chosen by selecting the session when the subjects were diagnosed with AD (session two) and the previous session where the subjects were still not converted (session one). Sessions for the subjects in the other cohorts (sMCI, AD, and CN) were chosen by selecting two consecutive sessions where both imaging and clinical evaluation were present. The current dataset did not allow a design to match the time between sessions for the whole cohort, this potential confounder is accounted for in our analysis.

Mild cognitive impairment conversion was clinically adjudicated by trained clinicians as described in the ADNI protocol. Any subject who converted back from AD to MCI was excluded from the study. Any subject included in the sMCI cohort remained stable for all sessions present in the ADNI dataset. For the subjects who converted (cMCI), the MRI images selected were based on the closest imaging session to the conversion adjudication; as such, we assumed that the T1 brain image would be representative of the status of the subject at the time of conversion as it is unlikely to significantly change in this time period. The average elapsed time between the time of conversion and the second imaging session was -0.7 ± 1.4 years.

The information on the conversion date can be found in the DXSYM_PDXCONV_ADNIALL.csv file from the ADNI database.

Image Preprocessing Pipeline

As shown in **Figure 1**, the T1-weighted MRI images were pre-processed according to the steps outlined in our previous work (Pena et al., 2019). In summary, the two images at two time points were normalized and aligned to each other first and then registered to a common space using a linear registration algorithm. The normalization involved motion correction, non-uniform intensity normalization, and skull strip as implemented in the first pre-processing stages of the Freesurfer 6.0 pipeline. The final common interpatient space was derived from 2 mm MNI T1 template which was cropped of the background space to reduce the computational complexity of the network for a final resolution of 64 × 80 × 64. This pipeline was shown to drastically decrease the pre-processing time compared with conventional image processing pipelines such as the wull FreeSurfer-based ones (Pena et al., 2019). These steps were extended to the full MCI cohort used in this study.

Nine clinical variables were used at two different imaging sessions (e.g., cross-sectional variables). In addition, the longitudinal signed differences for each of these variables. Note that while the APOe4 genotype is not expected to change between sessions, it has followed the same processing for consistency and simplifying the evaluation of the feature importance. Age, sex, and years of education were also concatenated in the final feature vector used in the model. These clinical variables were all normalized by their mean value.

Deep Learning Pipeline

Experimental Design

A 10-fold stratified cross-validation procedure was employed for model training and evaluation. Each fold was split into training, validation, and test sets with proportions of 80, 10, and 10%, respectively. Each fold maintained the distribution of sMCI/cMCI. Binary cross-entropy and the Adam were the loss function and optimizers used, respectively (Kingma and Ba, 2015). Each of the 10-folds had 75 epochs, and an early stopping condition of 10 epochs was implemented based on the model's validation loss. Cyclical learning rates were used to dynamically change the learning rate throughout the training process (Smith, 2017). This method has been shown to potentially allow the model to "jump" out of local minima to subsequently find a lower minimum to reduce the overall loss. The upper and lower bounds for the learning rates were 1e−5 to 1e−8. A batch size of 4 was used in the experiments. The area under the receiver operating curves (AUC) and balanced accuracy were the experimental

TABLE 1 | Demographics and time between imaging sessions of MCI subjects used in this study (1 SD).

	sMCI	cMCI	*p*-Value
Number of subjects	440	229	
Baseline age, years [mean (SD)]	73.4 (7.7)	74.2 (7.1)	0.167
Time between sessions, years [mean (SD)]	2.9 (2.2)	3.7 (1.9)	<0.0001
Years of education [mean (SD)]	15.8 (2.9)	15.8 (2.7)	0.831
Sex [male, *n* (%)]	260 (59.1)	134 (58.5)	0.934

[1] adni.loni.usc.edu

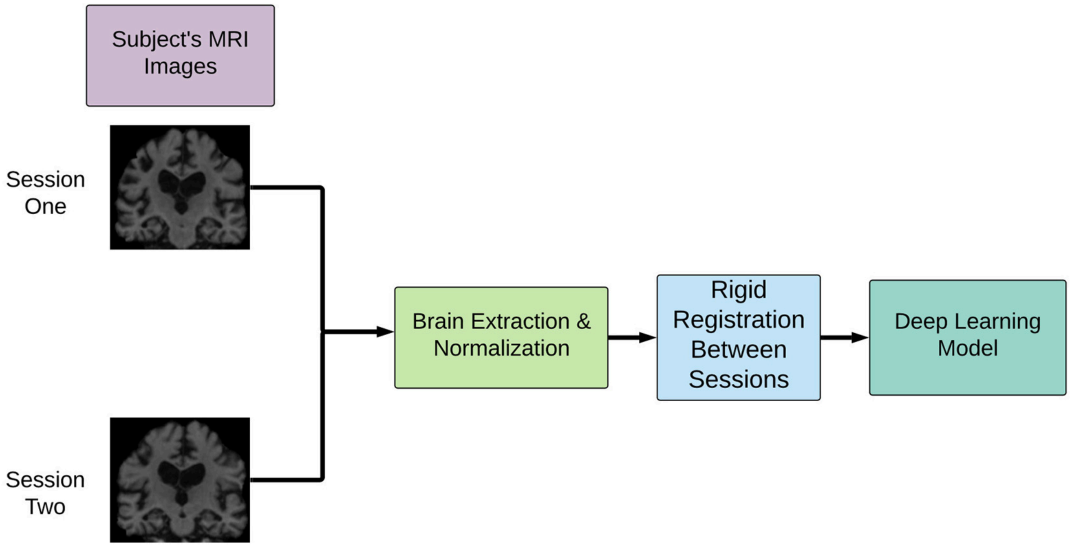

Subject's MRI Images
Session One
Session Two
Brain Extraction & Normalization
Rigid Registration Between Sessions
Deep Learning Model

FIGURE 1 | Overview of image pre-processing pipeline implementation. This pipeline involves rigid registration to align the patient's brains intra-patient first and then inter-patient to a common space. cMCI subjects' session one was before conversion, and session two was the imaging session at the clinically deemed conversion date. sMCI subjects, by definition, are not diagnosed with an AD conversion at any point.

evaluation metrics. The DeLong's test for statistical significance was used to test differences between AUC curves (DeLong et al., 1988). AUC curves' 95% confidence intervals were calculated using a Monte Carlo resampling simulation with 1,000 iterations, and in each iteration, 80% of the total subjects' probabilities were randomly chosen.

Two of the experiments used a transfer learning approach where an additional set of 190 AD and 243 CN subjects were first used to pre-train the network aimed at a simpler task first. None of the 433 AD/CN subjects in this pretraining step were used for the cross-validation, this avoided any risk of data leakage. The pretraining step followed all image pre-processing, hyperparameters, and initializations as stated in the text above, except for the cross-validation procedure. Finally, the pre-trained model was then used as the starting point for the weights used in the MCI classification task.

A fully connected network using only clinical variables was tested to obtain baseline comparison with the multi-modal network. This model is effectively equivalent to a logistic regression trained using the same optimization technique and validation approach as the multi-modal network; as such, it will allow for a fair evaluation of the relative improvement of adding brain imaging to the clinical data. The input feature vectors were the clinical variables, and the outputs were the same as the multi-modal network.

The experiments outlined were completed using Python 3.7, Keras version 2.2, and TensorFlow 1.14. The graphical processing units used were GeForce RTX 2080 Ti with 11 GB RAM. The training times varied between 30 and 90 s per epoch, depending on the architecture and experimental setup. The computational performance at inference time, which is more relevant to evaluate the ease of deployment of the model in a clinical environment, is discussed in section "Results."

Deep Learning Architecture

The network architecture employed was inspired by a model that learned from spatial symmetry between brain hemispheres in the stroke detection task (Barman et al., 2019; Sheth et al., 2019). Our previous work extended this model in the AD-progression and time domain (Pena et al., 2019). This study has implemented a new network that combines cross-sectional and longitudinal imaging data with clinical features, which can be trained end-to-end on the MCI conversion classification task. Further, we focused our efforts on a less parameterized network to improve computational efficiency, as we are primarily concerned with the clinical application of this class of methods. This was possible through residual attention-based modules (Wang et al., 2017), allowing the network to focus on specific areas of the image with an Inception-based network, which leads to learning convolutional filters at different scales.

From a high level (**Figure 2**), the model can learn a complex representation of two images at different time points through the two subnetworks (Attention and Inception) in addition to the temporal differences of the two brains through the subtraction layer. This subtraction layer is sensitive to changes and is referred to as the "longitudinal" portion of the network. In the attention module, cross-sectional information is added through skip connections. The output from these two subnetworks is combined with clinical variables in a final dense layer for prediction.

This model has several benefits:

- It has the potential to identify structural changes in T1-weighted MRI scans over time, which is vital for determining MCI conversion while utilizing commonly available clinical information.

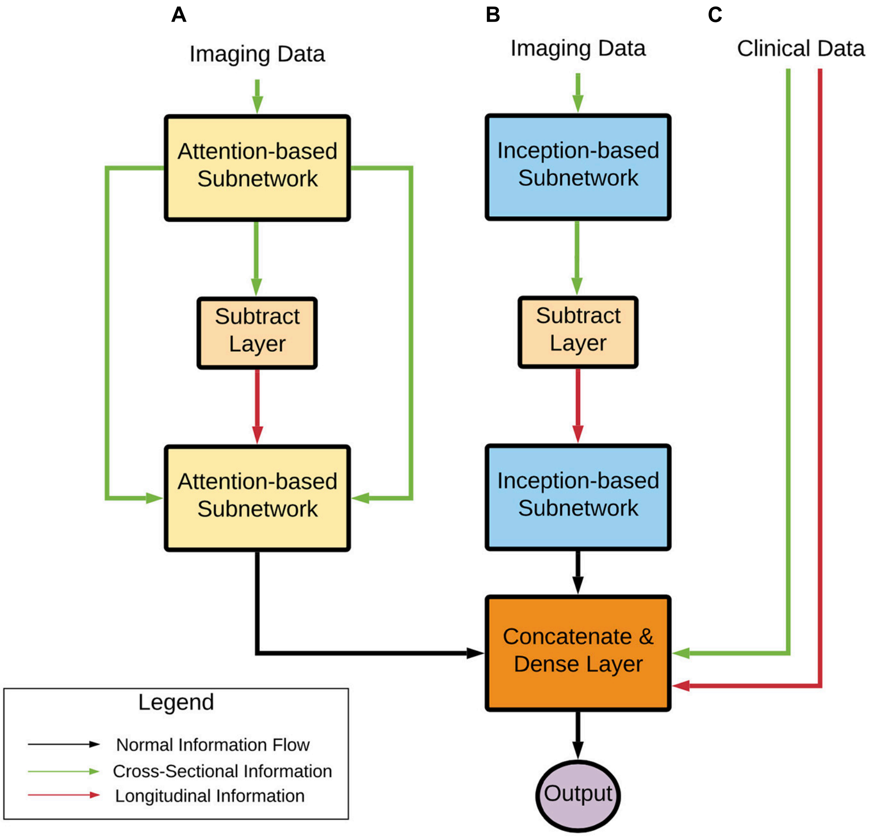

A
Imaging Data
Attention-based Subnetwork
Subtract Layer
Attention-based Subnetwork
B
Imaging Data
Inception-based Subnetwork
Subtract Layer
Inception-based Subnetwork
C
Clinical Data
Concatenate & Dense Layer
Output
Legend
Normal Information Flow
Cross-Sectional Information
Longitudinal Information

FIGURE 2 | Deep learning architecture high-level overview. Imaging data pass through **(A)** an Attention-based subnetwork and **(B)** an Inception-based subnetwork. The Attention-based network includes skip connections that concatenate cross-sectional information to the processed longitudinal information. The Inception-based network only contains longitudinal information. These two subnetworks' outputs are combined with clinical variables **(C)** that contain both cross-sectional and computed longitudinal differences. Finally, these subnetworks are combined and input into a prediction layer. Note that transfer learning approaches were used with AD and CN data, as stated earlier.

- It uses attention-based networks and deliberately leverages a less parameterized network that inherently regularizes the weights to only focus on important information.
- It incorporates an inception-based network that allows the model to use multi-resolution to represent the images at different scales in a non-sparse fashion.

Residual Attention Modules

Wang et al. (2017) extended the previously studied attention mechanism and applied it to their approach for image-level classification. Their overall network was composed of blocks named the residual attention module. These modules combined normal convolutional blocks (e.g., convolution, back normalization, and max pooling) with a U-Net inspired structure (Ronneberger et al., 2015) through a multiplication operator. The U-Net subunit allows the model to learn important information representing the input image through an encoder-decoder-like structure. The convolutional blocks allow the model to pay "attention" to these critical parts of the image through multiplication. This output then goes through another series of convolutional layers for further learning. Wang et al. (2017) stacked these residual attention modules to create a deep structure with complex attention mechanisms at different scales of the images. However, to create a less parameterized network, we limited the proposed network to just one residual attention block. For additional details about these modules, we refer the readers to the original publications.

Inception Modules

The inception modules used in this paper were inspired by the work done by Szegedy et al. (2015) and were extended to the 3-dimensional space (3D). The inception modules used were a combination of multi-resolution 3D convolutional layers. These layers were composed of three parallel operations: 1 × 1 × 1, 3 × 3 × 3, and 5 × 5 × 5 convolutions with two filters. Previous work has shown that this module can produce meaningful results in neuroimaging applications (Barman et al., 2019, 2020; Pena et al., 2019). This style of operation allows the model to view an image or input at different scales to learn different types of spatial information. The outputs were then concatenated and served as input to the next network layer. For additional details about these modules, we refer the readers to the original publications.

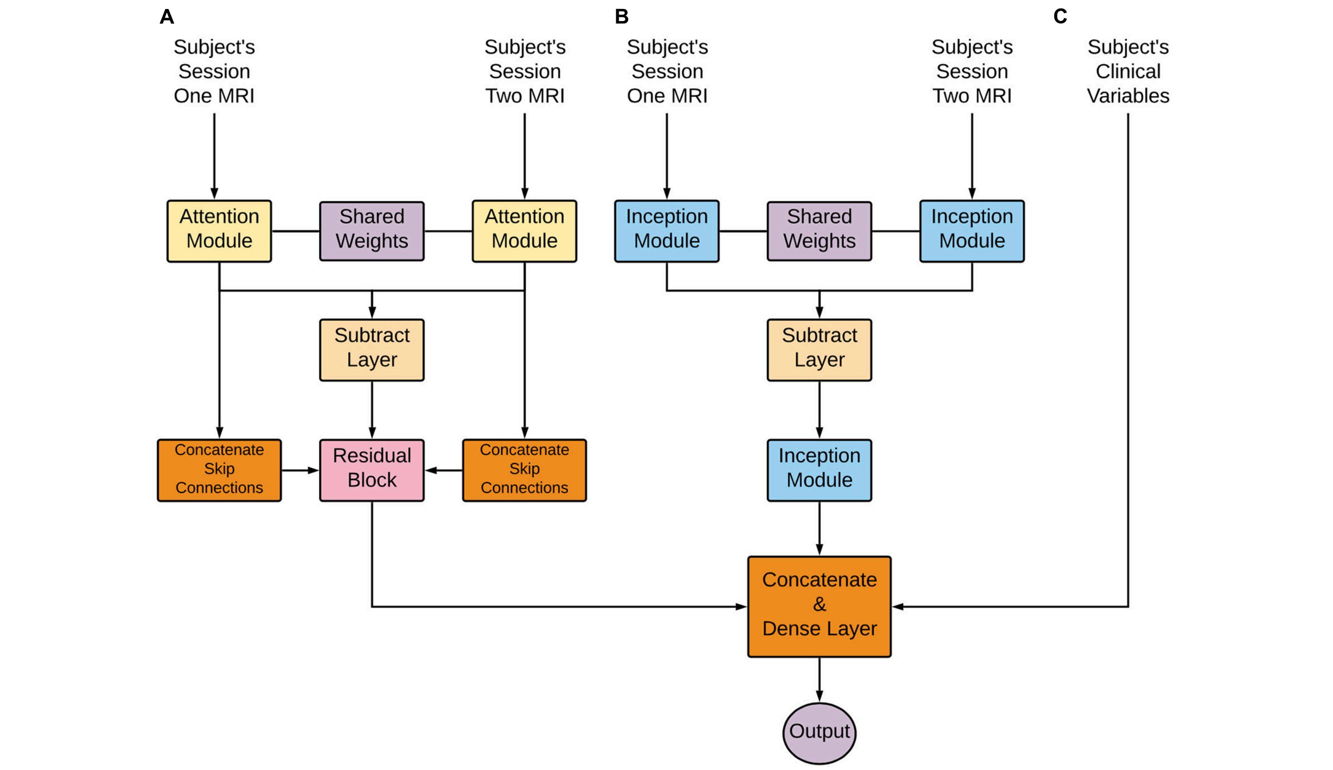

A
Subject's Session One MRI
Subject's Session Two MRI
Attention Module
Shared Weights
Attention Module
Subtract Layer
Concatenate Skip Connections
Residual Block
Concatenate Skip Connections
B
Subject's Session One MRI
Subject's Session Two MRI
Inception Module
Shared Weights
Inception Module
Subtract Layer
Inception Module
C
Subject's Clinical Variables
Concatenate & Dense Layer
Output

FIGURE 3 | Deep learning architecture detailed overview. Longitudinal images pass through **(A)** an Attention-based network and **(B)** an Inception-based network. These subnetworks are composed of an initially shared weight representation, a subtraction layer, and a subsequent flatten layer. These two subnetworks' outputs are combined with clinical variables **(C)**. Lastly, this concatenation is put through a dense layer for the final prediction.

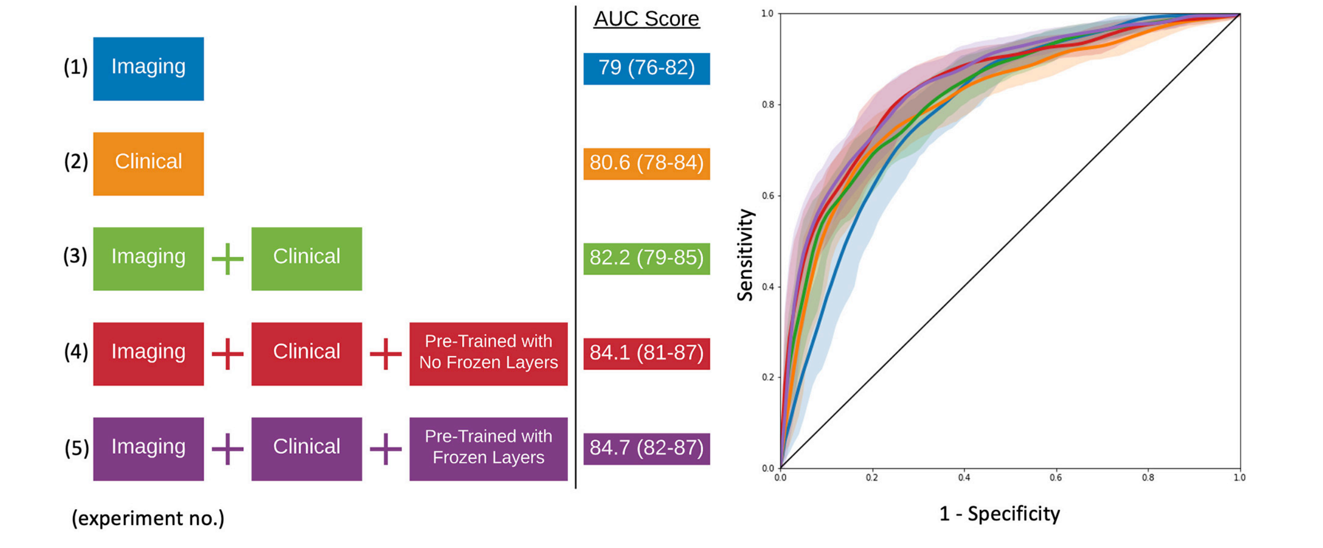


FIGURE 4 | Model experiments and associated AUC scores by varying input data and the use of transfer learning ***(left)***. ROC curves for comparing model performance from the experiments conducted ***(right)***.

Layers

From the overall network perspective (**Figure 3** below), the first module layers learned a representation shared between the first and second imaging time points. This representation proceeded to a subtraction layer that took the difference between the two sessions, and this difference was the input to another module.

These longitudinal outputs then went through another module to further learn from the differences between the two sessions. Note that the attention subnetwork incorporated longitudinal and cross-sectional information through the addition of skip connections, as seen in the figure below. Next, these outputs were flattened and concatenated with each other to form an imaging and clinical feature vector. Finally, the prediction layer was used for prediction utilizing the SoftMax activation function.

The code repository for this publication can be found at https://gitlab.com/lgianca/deepsymnet-att.

Confounding Variable Adjustment

A logistic regression model was fit with the deep neural network's probability output, baseline age, time between imaging sessions, and sex to adjust for any potential confounders inherent in the data chosen for the model. The logistic regression model coefficients, the 95% confidence intervals, and corresponding *p*-values were reported.

Feature Importance

To develop an intuition about which voxels from the T1-weighted MRI images and features from the clinical variables, we employed the epsilon layer-wise relevance propagation (e-LRP) method (Bach et al., 2015). The e-LRP method starts from the prediction layer and works its way backward through the network. Layer-by-layer, the relevance of each of the previous layer's nodes is computed until the operation reaches the input data layer. Each feature in the input data is assigned a final relevance score that describes how important that feature was for the final prediction. The codebase used in our experiments follows the implementation of DeepExplain (Ancona et al., 2018).

As stated above, relevance scores are projected onto the input data, which, in this study, are the two T1-weighted images and the subject's clinical feature vector. We used a global and regional method to compute the magnitude of relative importance for the voxels in the MRI images. For the global method, each subject's MRI relevance map was added for both sessions, and absolute values were used to remove the risk of canceling out relevance scores. Then, a heatmap allowed for the visualization of this global method.

For the regional method, the cortical and subcortical regions were segmented for each subject *via* the Harvard–Oxford atlas (Caviness et al., 1996). Then, for each subject and session, the summation of all the voxels' magnitude in each region was calculated. This value was then divided by the volume of that particular region, resulting in a normalized relevance magnitude for a particular brain region. This final value allowed the different regions to be compared to one another on a similar scale. A similar approach was used to find the relative importance of the clinical features. The unsigned value for a clinical feature was added for each subject and then ranked in order of importance based on the magnitude of the total value. The overall method is described in greater detail in our previous work (Pena et al., 2019).

RESULTS

This study aims to (1) evaluate the use of different deep learning architectures, input data modalities, and transfer learning for MCI conversion classification using a computationally efficient architecture and to (2) investigate the important imaging and clinical features that drove the model's decision based on the e-LRP method.

Model Evaluation

As seen in **Figure 4** and **Table 2**, the model that used imaging and clinical input data was pre-trained using AD, and CN subjects with frozen weights had the highest AUC score (Experiment 5). This model was considered the "best" performing model in this paper. The pre-trained model where all of the weights could be fine-tuned had the best-balanced accuracy. **Table 2** also shows that the improvement between solely using clinical variables (Experiment 1) and the best model that combined clinical and T1 imaging was statistically significant. Further, our best model was the only one significantly greater than the model that used clinical variables only. The average time taken to pre-process an image and for the model to make a prediction was 129.7 ± 19.8 and 0.12 ± 0.05 s, respectively.

We evaluate the individual importance of the Inception and Attention subnetworks with ablation studies. We use as a base model and training strategy what has been described in Experiment 5. In order to account for the artificial advantage that the architectures might have solely on the basis of having more parameters, we increased the number of convolutional filters in each of the independent subnetworks to make them comparable with the full network. In **Table 3**, we show that the Inception-based subnetwork overperforms the Attention-based subnetworks. However, their combination (with the addition of the clinical data) outperformed the two architectures individually, even if the number of parameters was comparable.

In order to evaluate the computational efficiency of the model, we evaluated the time required to generate a prediction at inference time (i.e., after model training) on an off-the-shelf laptop without using any GPUs. We repeated this 100 times and achieved an average execution time of 1.56 s (0.10 std).

TABLE 2 | Model experiments' metric comparison for balanced accuracy, AUC score, and testing for significant differences between AUC curves.

Experiment	Balanced accuracy	AUC score	
(1) Clinical variables only	75.5	80.6	
(2) T1-weighted MRI only	73.2	79	p=0.51
(3) Clinical variables + T1-weighted MRI	75.4	82.2	p=0.41
(4) Clinical variables + T1-weighted MRI (pre-trained network: no frozen layers)	78.2	84.1	p=0.07
(5) Clinical variables + T1-weighted MRI (pre-trained network: frozen layers)	77.8	84.7	**p<0.05**

p-Values were computed from the DeLong test for correlated ROC curves to reject the null hypothesis that there is no statistical difference between the AUCs.

TABLE 3 | Ablation studies indicate that the combination of the two Attention and Inception-based subnetworks overperform the two individual subnetworks.

Experiment	Number of parameters	Balanced accuracy	AUC score
Attention-based subnetwork (imaging + pre-training with frozen layers)	374,398	65.7	68.6
Inception-based subnetwork (imaging + pre-training with frozen layers)	333,352	73.6	77.6
Best performing network (imaging + clinical + pre-training with frozen layers)	355,572	77.8	84.7

Note that the number of convolutional filters in the subnetworks were increased to have comparable parameters with the entire network (i.e., within 10%).

This does not take into account the file conversion, initial brain extraction and linear registration steps required, which can take from tens of seconds to a few minutes, depending on the software used. This compares favorably to the "*de facto*" Freesurfer-based longitudinal pipeline that can take an average of 17 h per subject (Pena et al., 2019) or methods relying on non-linear registration and extraction of the warp field, taking each image into template space. For example Spasov et al. (2019) report approximately 19,200 h of CPU time on a high-performance parallel computing cluster to non-linearly register the images, which is ∼19 h per subject.

The Network as a Clinical Decision Support Tool

With the final model, we investigated the strength of the signal (deep network output probability) between MCI subjects who eventually converted to AD and those who stayed stable, as seen in **Figure 5**. The starting point for the cMCI subjects is higher than the sMCI subjects since there was some indication of AD conversion-like progression using MRI imaging time points before the actual conversion. However, this signal strengthens when an imaging time point around AD conversion is included, shown by the tendency toward higher probabilities on the right side of the figure. The sMCI group has a smaller slope with respect to time as there is no indication of AD conversion. This makes for a clear, qualitative difference between the two groups. The network derives a much stronger signal at the conversion point, indicating its ability to recognize patterns distinctly associated with AD conversion.

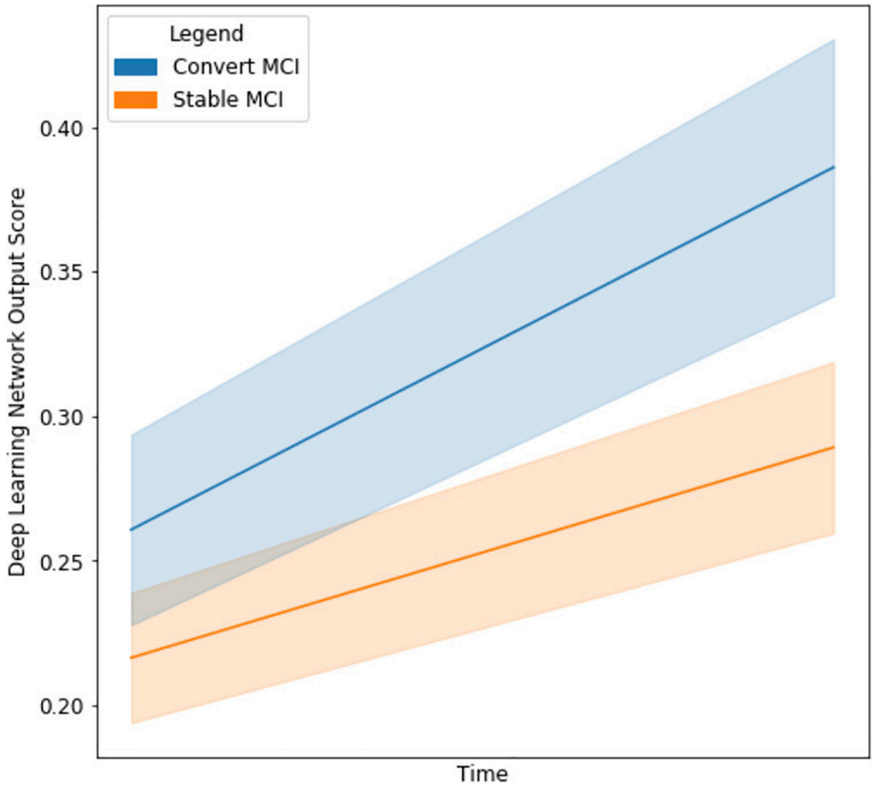


FIGURE 5 | Line graph visualizing the difference in output network scores between cMCI (*blue*) and sMCI subjects (*orange*) with 95% confidence intervals in the shaded regions. The darker lines represent the mean trajectory based on the distribution of scores of the respective groups. Note that the cMCI subjects' starting score is computed using the network and both imaging time points before conversion. The ending score includes the second time point when the AD conversion was diagnosed. The sMCI subjects' starting scores are taken using the baseline and time point near the baseline date, and their ending score is using the baseline and a later date.

TABLE 4 | Summary of the logistic regression coefficients, confidence intervals, and *p*-values for model output probability and confounding variables.

Baseline age	Time between sessions	Years of education	Sex	Deep learning output probability
0.0140 (0.001–0.027)*	0.0916 (0.010–0.174)	0.0361 (−0.024 to 0.096)	0.0521 (−0.358 to 0.462)	−4.0714 (−4.700 to −3.443)***

**p* < 0.05.*
****p* < 0.0001.*

Confounding Variable Adjustment

Further, the deep learning output probabilities were assessed for statistical significance with a logistic regression model and potential confounding variables. As seen in **Table 4**, the output probability remained statistically significant ($p < 0.0001$). Interestingly, though there were group differences between the imaging variable, as seen in **Table 1**, these differences were not significant when combined with the output probabilities.

Feature Importance

Next, model feature importance was evaluated for both the imaging and clinical inputs. The imaging feature importance was completed on both a voxel-wise level and a brain regional level (subcortical vs. cortical regions), as seen in **Figure 6**. These saliency maps are smoother than our previous work (Pena et al., 2019), and we attributed this improvement to the use of attention-based modules and a less parameterized network. These model characteristics perform significant regularization, highlighting only the most informative regions for the given task.

Further, the top five regions from the cortical and subcortical regions were plotted in **Figures 6B,C**. After volume normalization, the thalamus, caudate, pallidum, and lateral ventricle subcortical regions contained the highest overall contribution to the model's decision. For the cortical regions, the planum temporale and parietal operculum cortex had the highest contributions. Both the cortical and subcortical regions had similar contribution magnitudes, as seen in the *x*-axis of **Figure 7**.

The clinical variables' contributions are shown in descending order in **Figure 8**. The RAVLT score from the second session was the most important clinical feature, with ADAS11 from the second session and MoCA scores from the first session following.

DISCUSSION

This study employed a deep learning model to enable a CAD system able to provide an AD conversion diagnosis in an MCI cohort followed longitudinally. The model combined both Attention and Inception modules and was designed to be less parameterized to form a sparse yet rich representation of the input imaging and clinical features. The experiments performed demonstrated that the combination of imaging and clinical features produced a better model than using either type of data alone. Also, a model pre-trained on AD and CN subjects that served as a baseline for MCI classification was a better starting point for subsequent model fine-tuning than random weight initialization. Further, the brain regions that drove the model's decision were visualized and quantified through the e-LRP method. The clinical features included in the model were also ranked and analyzed for relevance.

One of the main contributions of this network architecture is the combination of longitudinal and cross-sectional information. The subtraction operation was used between the two imaging and clinical time points and their respective features; thus, the network could learn from the differences over time. Further, this information was preserved throughout the training process by keeping the raw signal from the individual time points (e.g., cross-sectional data). The imaging-focused part of the network was divided into the Attention and Inception-based mechanisms. The attention module extended the residual attention used in computer vision, allowing the model to introduce sparsity into the network parameters. This allows the model to focus on certain parts of the brain input data related to MCI conversion to AD. The inception modules used 3D convolutional filters to find information at different spatial scales and granularity. We empirically show that using a combination of these modules, both the balanced accuracy and AUC were higher than using these modules individually in a network.

Further, we showed that the network improved AUC performance by incorporating more information in the time domain (cross-sectional and longitudinal) and in data modality (T1-weighted MRI and clinical features). This has been shown to be the case in related MCI and AD research (Goryawala et al., 2015; Spasov et al., 2019). The best model was also pre-trained on a cohort of AD and CN subjects. This model's only trainable layer was the dense layer right before the prediction layer. Exclusively fine-tuning of the penultimate layer allowed the model to focus on changing a smaller number of weights compared to the entire model. This transfer learning setup also assumed that the brain representation from the AD and CN subjects was a good representation for an MCI application, making intuitive sense since this is modeling a progression pattern in subjects at high risk of developing AD in their lifetime. Finally, after controlling for several potential confounding factors, the network output probabilities remained statistically significant. Once the model is trained, the whole model can run in ∼1.5 s plus the time required to perform basic pre-processing involving file conversion, skull stripping, and linear registration (typically tens of seconds to a few minutes) on an off-the-shelf laptop without GPU. This would enable a neurologist to use this system as a computer-aided diagnostic tool during the office visit once the required imaging and/or clinical variables are acquired.

Once the experiments were completed, the crucial features for driving the model's decision were investigated. To narrow down the imaging analysis for interpretation, we focused on the subcortical and cortical regions. The thalamus, caudate, pallidum, and lateral ventricle had the highest overall activation magnitude for the subcortical regions. Cholinergic synapses have a high density in several parts of the brain, including the thalamus, and have played a central role in research in aging and

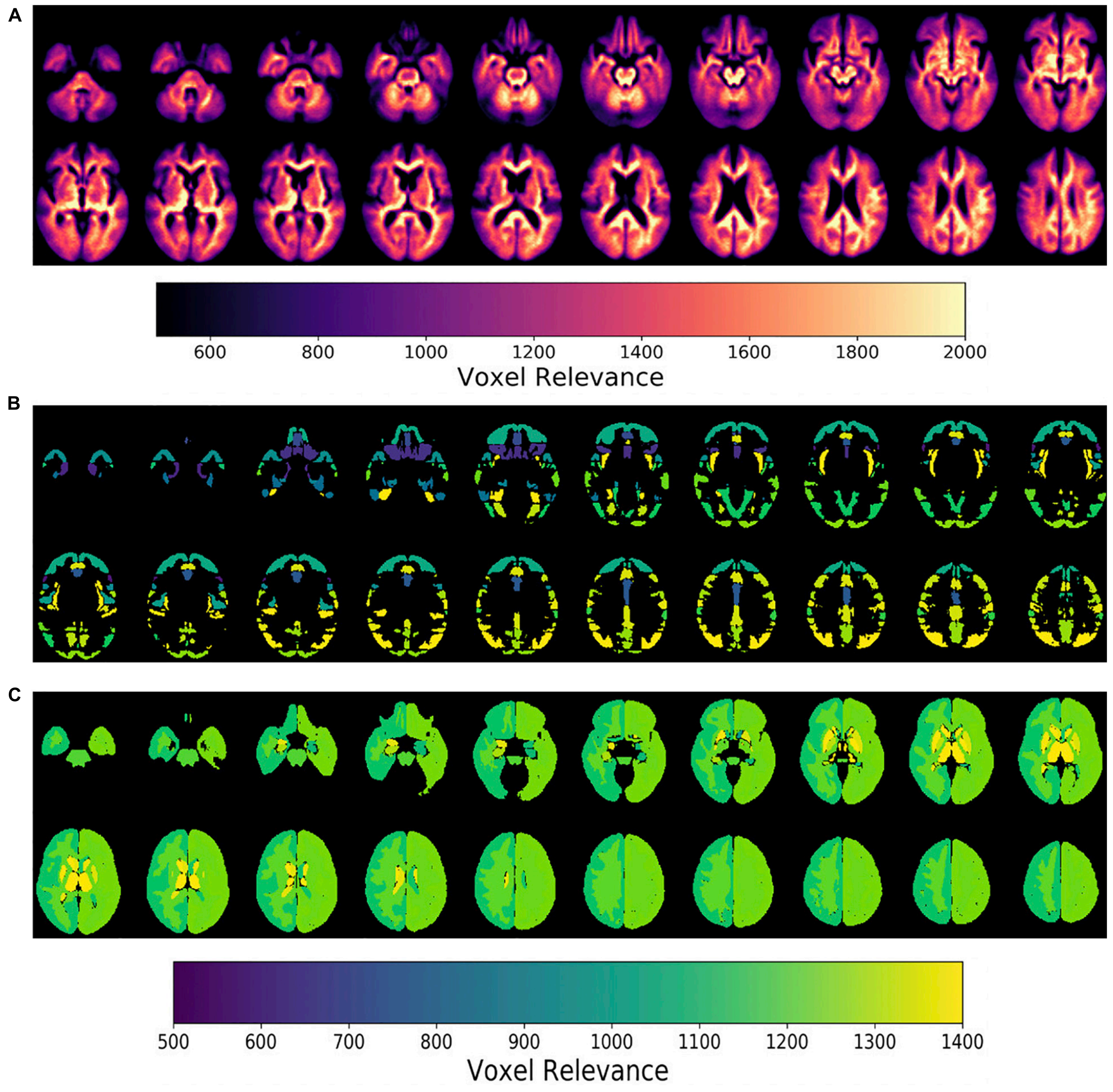


FIGURE 6 | Epsilon layer-wise relevance propagation relevance maps displaying voxel-level and region-level contribution to model output probability at different brain slices. The maps are at the **(A)** voxel-level, **(B)** cortical, and **(C)** subcortical levels. Maps **(B,C)** are normalized to the brain region volume. The scales indicate the degree of voxel contribution magnitude.

cognitive decline. Cholinesterase inhibitors are considered first-line treatments for mild and moderate AD (Hampel et al., 2018). Likewise, an *in vivo* imaging study found reduced serotonin transporter availability in MCI subjects in the thalamus compared to controls (Smith et al., 2017). Qing et al. (2017) found that impairment of spatial navigation skills, a clinical feature of AD found in MCI subjects, was significantly correlated to neuroimaging variable changes in the pallidum and thalamus.

Similarly, Fischer et al. (2017) investigated mobility changes in subjects with MCI. They found that decreased gray matter volume in the caudate nucleus was associated with a lower speed in functional mobility tasks. Crocco et al. (2018) applied a cognitive stress test to AD and MCI subjects and showed that negative clinical results were related to dilation of the lateral ventricle, among other regions. Yi et al. (2016) found that gray matter volumes in subcortical regions, including, but not limited to, the thalamus, caudate, and pallidum, were significantly reduced in MCI subjects when compared to controls. Additionally, many of these subcortical volumes were correlated with cognitive function.

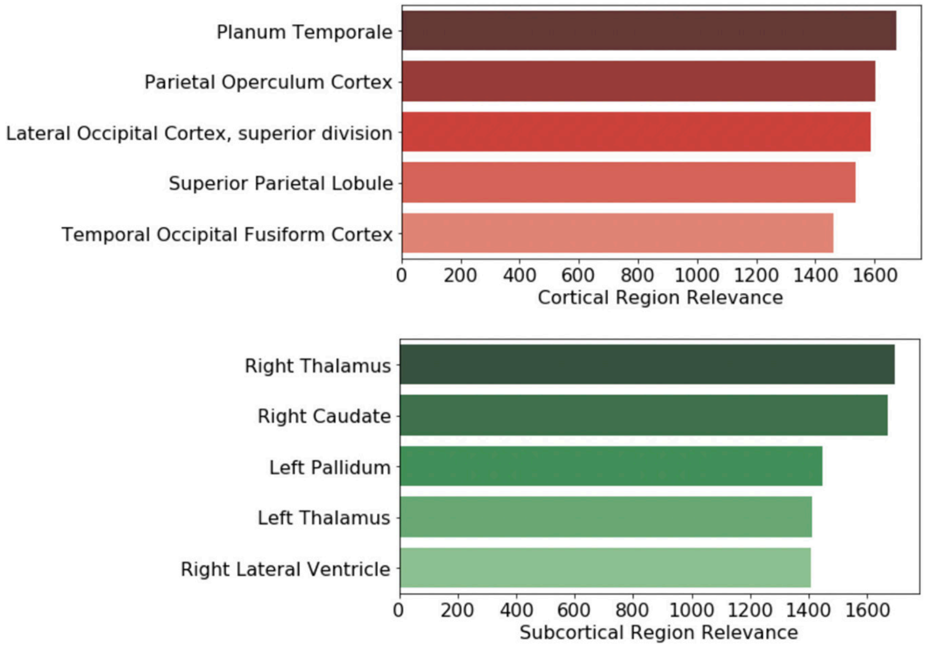


FIGURE 7 | Brain regional contribution magnitude of the top five (*top*) cortical and (*bottom*) subcortical regions. The contributions were calculated by summating all the magnitudes within the brain region and then normalized to the brain region volume.

For the cortical regions, the planum temporale, operculum cortex, and occipital cortex were some of the top regions with associated findings in AD and MCI literature. Researchers using several independent AD datasets found that anatomical changes in the planum temporale and thalamus were among the top features for their predictive model (Giraldo et al., 2018; Li et al., 2020). Others have found that changes in cortical minicolumn organization and premortem cognitive scores were significantly related in the planum temporale, potentially reflecting a phenomenon in brain atrophy in AD subjects (Chance et al., 2011). Alternatively, this might indicate the importance of auditory processing in MCI to AD progression as planum temporale is involved in auditory processing. A clinical study that used functional connectivity imaging and associated metrics found decreased intrinsic connectivity in the operculum cortex among MCI and AD subjects (Xie et al., 2012). Finally, a PET study demonstrated a significant and high overlap in hypoperfusion and hypometabolism in AD subjects in the occipital cortex (Riederer et al., 2018).

From the clinical features, ADAS, MoCA, and MMSE scores are among the top five variables with the most relevance for the model's decision. This is unsurprising as MoCA and MMSE are the most widely used screening tools in clinical practice. ADAS is frequently used as a progression measurement in both clinical settings and clinical trials.

Interestingly, the RAVLT, a recent memory test, was the variable with the most relevance for the model's decision for the first session, and it was ranked as one of the top five variables for the second session. Memory for recent events is distinctively impaired in AD and is served by the hippocampus, entorhinal cortex, and related structures in the medial temporal lobe.

This could indicate that the RAVLT provides more complementary information that is harder to directly learn from the imaging alone. Multiple clinical and neuroimaging studies have shown the importance of this variable in AD and MCI research; one of the earliest was performed by Estévez-González et al. (2003). More recently, Eliassen et al. (2017) used PET imaging and clinical scores to show that RAVLT were significant predictors in changes in cortical thickness between MCI and CN participants. A neuroimaging study conducted by Moradi et al. (2017) found that the MRI-based volumetric features were suitable variables for predicted parts of the RAVLT tool using an elastic net-based linear regression model. Russo et al. (2017) found that parts of the RAVLT assessment can have differences in discrimination accuracy and response bias between MCI and AD subjects, indicating there could be diagnostic specificity if using different test portions.

Our study has some limitations. First, the absolute classification performance of our method was lower than some found in the literature (Liu et al., 2017; Tong et al., 2017; Spasov et al., 2019) that report AUC scores around 90%. However, these models focus on the prediction of future AD rather than an actual diagnosis of the AD conversion, and they typically involve a very long pre-processing pipeline that would be hard to use in clinical settings. The use of longitudinal data to output an AD diagnosis can also be considered a limitation as it requires data from two-timepoints. The dataset used did include a majority of Caucasian non-Hispanic population, as such, the generalizability of the algorithm needs to be further

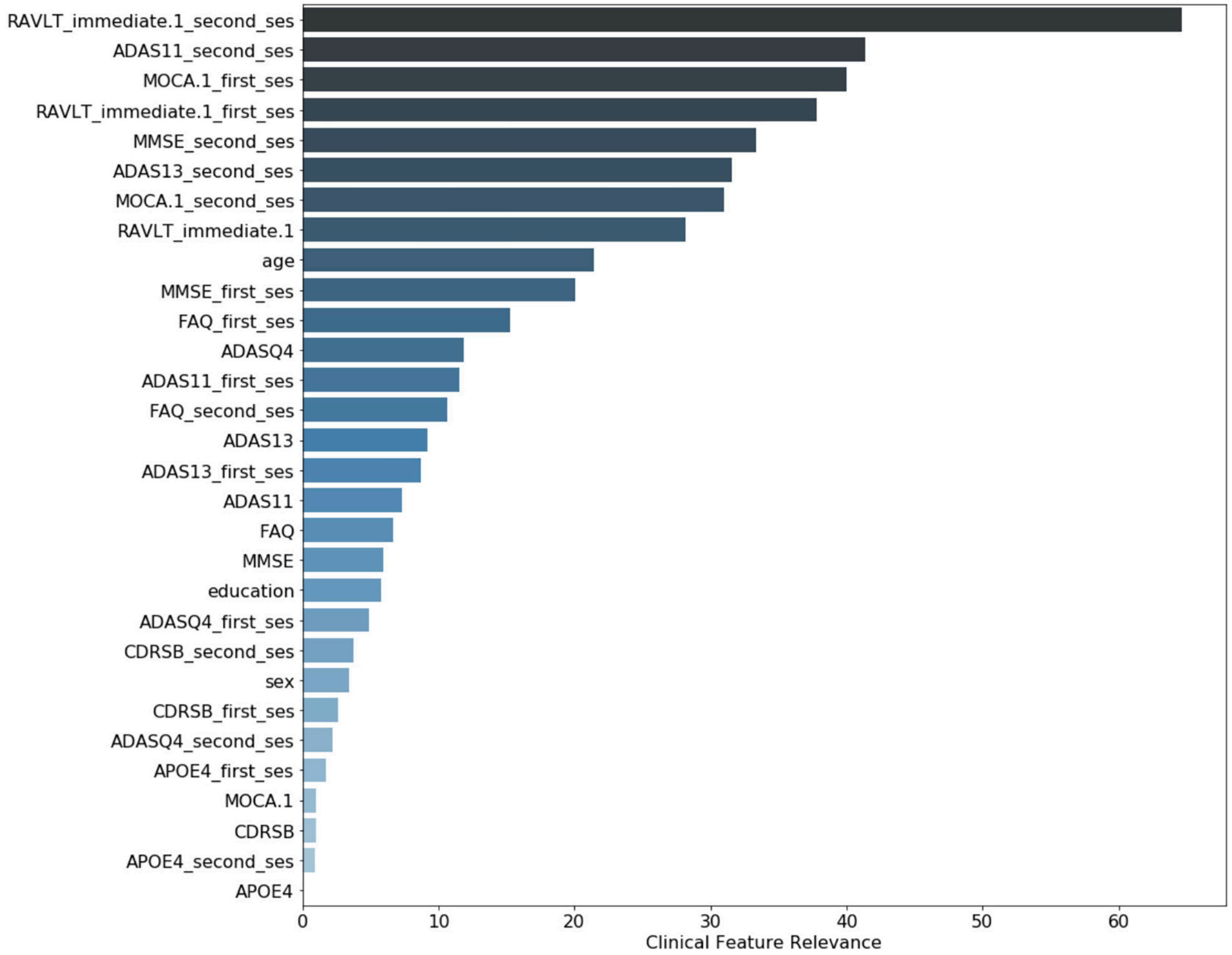


FIGURE 8 | Clinical variable-level contribution magnitude. These values were calculated by summation of the contribution across all subjects. Neurological clinical variables denoted with "first_ses" and "second_ses" correspond to the subjects' first and second clinical sessions, respectively. The clinical variables without a suffix represent the longitudinal change in that particular variable over time. Demographic variables included were age, sex, and years of education.

confirmed on an entirely external dataset including a more diverse population. Finally, while the model can handle missing imaging or clinical data (as a whole), it currently cannot leverage clinical data with missing variables (unless imputation is used).

Future work could extend the ADNI dataset to incorporate multiple sources. This would increase the model's generalizability to bias and errors that are inherent to different datasets. Also, adding more time points by extending this model using recursive neural networks or Gaussian processes algorithms could give a more nuanced trajectory signal that may unearth a strong signal for MCI progression and conversion to AD.

CONCLUSION

In this paper, we introduce a novel method that utilizes T1-weighted MRI and clinical data at two-time points to diagnose AD in patients with MCI. At a high level, the model is a deep learning framework that combines residual Attention and Inception modules while taking advantage of cross-sectional and longitudinal data. The epsilon layer-wise propagation method allowed the interpretation of essential brain regions and clinical features that drove the model's output. Some of the top subcortical and cortical regions included the thalamus, caudate, planum temporale, and operculum cortex. Further, RAVLT was the clinical feature that had the highest contribution to the final prediction. This method could easily be translated to the healthcare environment because it integrates variables commonly used in a clinical setting and has a fast image processing and prediction pipeline. This instrument could potentially be used as an objective and efficient diagnostic tool for patients at high risk of AD conversion.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: adni.loni.usc.edu.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

DP, LG, MS, and JS contributed to conception and design of the study. DP performed the experiments and analysis, and wrote the first draft of the manuscript. DP, TE, and LG interpreted the results. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI was funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd. and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development, LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research was providing funds to support ADNI clinical sites in Canada. Private sector contributions were facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization was the Northern California Institute for Research and Education, and the study was coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data were disseminated by the Laboratory for Neuro Imaging at the University of Southern California. LG was supported by a Learning Healthcare Award funded by the UTHealth Center for Clinical and Translational Science (CCTS), the Translational Research Institute through NASA Cooperative Agreement NNX16AO69A, NIH grants UL1TR003167 and R01NS121154, and a Cancer Prevention and Research Institute of Texas grant (RP 170668).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2021.744190/full#supplementary-material

REFERENCES

Ahmed, M. R., Zhang, Y., Feng, Z., Lo, B., Inan, O. T., and Liao, H. (2019). Neuroimaging and machine learning for dementia diagnosis: recent advancements and future prospects. *IEEE Rev. Biomed. Eng.* 12, 19–33. doi: 10.1109/RBME.2018.2886237

Ancona, M., Ceolini, E., Oztireli, C., and Gross, M. (2018). Towards a better understanding of gradient-based attribution methods for deep neural networks. *ICLR* 2018, 1–16.

Ansart, M., Epelbaum, S., Bassignana, G., Bône, A., Bottani, S., Cattai, T., et al. (2021). Predicting the progression of mild cognitive impairment using machine learning: a systematic, quantitative and critical review. *Med. Image Anal.* 67:101848. doi: 10.1016/j.media.2020.101848

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One* 10:e0130140. doi: 10.1371/journal.pone.0130140

Barman, A., Inam, M. E., Lee, S., Savitz, S., Sheth, S., and Giancardo, L. (2019). "Determining ischemic stroke from CT-Angiography imaging using symmetry-sensitive convolutional networks," in *Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, (Piscataway, NJ: IEEE), doi: 10.1109/ISBI.2019.8759475

Barman, A., Lopez-Rivera, V., Lee, S., Vahidy, F. S., Fan, J. Z., Savitz, S. I., et al. (2020). "Combining symmetric and standard deep convolutional representations for detecting brain hemorrhage," in *Proceedings of the Medical Imaging 2020: Computer-Aided Diagnosis*, (Bellingham, DC), doi: 10.1117/12.2549384

Bhagwat, N., Viviano, J. D., Voineskos, A. N., Chakravarty, M. M., and Initiative, A. D. N. (2018). Modeling and prediction of clinical symptom trajectories in Alzheimer's disease using longitudinal data. *PLoS Comput. Biol.* 14:e1006376. doi: 10.1371/journal.pcbi.1006376

Caviness, V. S., Meyer, J., Makris, N., and Kennedy, D. N. (1996). MRI-Based topographic parcellation of human neocortex: an anatomically specified method with estimate of reliability. *J. Cogn. Neurosci.* 8, 566–587. doi: 10.1162/jocn.1996.8.6.566

Chance, S. A., Clover, L., Cousijn, H., Currah, L., Pettingill, R., and Esiri, M. M. (2011). Microanatomical correlates of cognitive ability and decline: normal ageing, MCI, and Alzheimer's disease. *Cereb. Cortex* 21, 1870–1878. doi: 10.1093/cercor/bhq264

Crocco, E. A., Loewenstein, D. A., Curiel, R. E., Alperin, N., Sara, J., Harvey, P. D., et al. (2018). A novel cognitive assessment paradigm to detect Pre-mild cognitive impairment (PreMCI) and the relationship to biological markers of Alzheimer's disease. *J. Psychiatr. Res.* 96, 33–38. doi: 10.1016/j.jpsychires.2017.08.015.A

Cui, R., Liu, M., Alzheimer's Disease Neuroimaging Initiative (2019). RNN-based longitudinal analysis for diagnosis of Alzheimer's disease. *Comput. Med. Imaging Graph.* 73, 1–10. doi: 10.1016/J.COMPMEDIMAG.2019.01.005

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametetric approach. *Biometrics* 44, 837–845. doi: 10.2307/2531595

Desikan, R. S., Cabral, H. J., Settecase, F., Hess, C. P., Dillon, W. P., Glastonbury, C. M., et al. (2010). Automated MRI measures predict progression to Alzheimer's disease. *Neurobiol. Aging* 31, 1364–1374. doi: 10.1016/j.neurobiolaging.2010.04.023

Eliassen, C. F., Reinvang, I., Selnes, P., Fladby, T., and Hessen, E. (2017). Convergent results from neuropsychology and from neuroimaging in patients with mild cognitive impairment. *Dement. Geriatr. Cogn. Disord.* 43, 144–154. doi: 10.1159/000455832

Estévez-González, A., Kulisevsky, J., Boltes, A., Otermín, P., and García-Sánchez, C. (2003). Rey verbal learning test is a useful tool for differential diagnosis in the preclinical phase of Alzheimer's disease: comparison with mild cognitive impairment and normal aging. *Int. J. Geriatr. Psychiatry* 18, 1021–1028. doi: 10.1002/gps.1010

Fischer, B. L., Bacher, R., Bendlin, B. B., Birdsill, A. C., Ly, M., Hoscheidt, S. M., et al. (2017). An examination of brain abnormalities and mobility in individuals with mild cognitive impairment and Alzheimer's disease. *Front. Aging Neurosci.* 9:86. doi: 10.3389/fnagi.2017.00086

Giorgio, J., Landau, S. M., Jagust, W. J., Tino, P., Kourtzi, Z., and Alzheimer's Disease Neuroimaging Initiative (2020). Modelling prognostic trajectories of cognitive decline due to Alzheimer's disease. *NeuroImage Clin.* 26:102199. doi: 10.1016/j.nicl.2020.102199

Giraldo, D. L., García-Arteaga, J. D., Cárdenas-Robledo, S., and Romero, E. (2018). Characterization of brain anatomical patterns by comparing region intensity distributions: applications to the description of Alzheimer's disease. *Brain Behav.* 8:e00942. doi: 10.1002/brb3.942

Goryawala, M., Zhou, Q., Barker, W., Loewenstein, D. A., Duara, R., and Adjouadi, M. (2015). Inclusion of neuropsychological scores in atrophy models improves diagnostic classification of alzheimer's disease and mild cognitive impairment. *Comput. Intell. Neurosci.* 2015:865265. doi: 10.1155/2015/865265

Grassi, M., Rouleaux, N., Caldirola, D., Loewenstein, D., Schruers, K., Perna, G., et al. (2019). A novel ensemble-based machine learning algorithm to predict the conversion from mild cognitive impairment to Alzheimer's disease using socio-demographic characteristics, clinical information, and neuropsychological measures. *Front. Neurol.* 10:756. doi: 10.3389/fneur.2019.00756

Hampel, H., Mesulam, M. M., Cuello, A. C., Farlow, M. R., Giacobini, E., Grossberg, G. T., et al. (2018). The cholinergic system in the pathophysiology and treatment of Alzheimer's disease. *Brain* 141, 1917–1933. doi: 10.1093/brain/awy132

Jack, C. R., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., et al. (2010). Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurol.* 9:119. doi: 10.1016/S1474-4422(09)70299-6

Kingma, D. P., and Ba, J. (2015). Adam: a method for stochastic optimization. *arXiv [preprint]*. doi: 10.1063/1.4902458

Landau, S. M., Harvey, D., Madison, C. M., Reiman, E. M., Foster, N. L., Aisen, P. S., et al. (2010). Comparing predictors of conversion and decline in mild cognitive impairment(Podcast)(e–Pub ahead of print). *Neurology* 75, 230–238. doi: 10.1212/WNL.0b013e3181e8e8b8

Lane, C. A., Hardy, J., and Schott, J. M. (2018). Alzheimer's disease. *Eur. J. Neurol.* 25, 59–70. doi: 10.1111/ene.13439

Lawrence, E., Vegvari, C., Ower, A., Hadjichrysanthou, C., De Wolf, F., and Anderson, R. M. (2017). A systematic review of longitudinal studies which measure Alzheimer's disease biomarkers. *J. Alzheimers Dis.* 59, 1359–1379. doi: 10.3233/JAD-170261

Lee, G., Kang, B., Nho, K., Sohn, K. A., and Kim, D. (2019a). MildInt: deep learning-based multimodal longitudinal data integration framework. *Front. Genet.* 10:617. doi: 10.3389/fgene.2019.00617

Lee, G., Nho, K., Kang, B., Sohn, K. A., Kim, D., and the Alzheimer's Disease Neuroimaging Initiative (2019b). Predicting Alzheimer's disease progression using multi-modal deep learning approach. *Nat. Sci. Rep.* 9:1952. doi: 10.1038/s41598-018-37769-z

Li, F., Liu, M., and the Alzheimer's Disease Neuroimaging Initiative (2019). A hybrid convolutional and recurrent neural network for hippocampus analysis in Alzheimer's disease. *J. Neurosci. Methods* 323, 108–118. doi: 10.1016/J.JNEUMETH.2019.05.006

Li, X., Xia, J., Ma, C., Chen, K., Xu, K., Zhang, J., et al. (2020). Accelerating structural degeneration in temporal regions and their effects on cognition in aging of MCI patients. *Cereb. Cortex* 30, 326–338. doi: 10.1093/cercor/bhz090

Liu, K., Chen, K., Yao, L., Guo, X., and the Alzheimer's Disease Neuroimaging Initiative (2017). Prediction of mild cognitive impairment conversion using a combination of independent component analysis and the cox model. *Front. Hum. Neurosci.* 11:33. doi: 10.3389/fnhum.2017.00033

Mateos-Pérez, J. M., Dadar, M., Lacalle-Aurioles, M., Iturria-Medina, Y., Zeighami, Y., and Evans, A. C. (2018). Structural neuroimaging as clinical predictor: a review of machine learning applications. *NeuroImage Clin.* 20, 506–522. doi: 10.1016/j.nicl.2018.08.019

Moradi, E., Hallikainen, I., Hänninen, T., Tohka, J., and the Alzheimer's Disease Neuroimaging Initiative (2017). Rey's auditory verbal learning test scores can be predicted from whole brain MRI in alzheimer's disease. *NeuroImage Clin.* 13, 415–427. doi: 10.1016/j.nicl.2016.12.011

Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J., and the Alzheimer's Disease Neuroimaging Initiative (2015). Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *NeuroImage* 104, 398–412. doi: 10.1016/j.neuroimage.2014.10.002

Nozadi, S. H., Kadoury, S., and the Alzheimer's Disease Neuroimaging Initiative (2018). Classification of Alzheimer's and MCI patients from semantically parcelled PET images: a comparison between AV45 and FDG-PET. *Int. J. Biomed. Imaging* 2018:1247430. doi: 10.1155/2018/1247430

Ottoy, J., Niemantsverdriet, E., Verhaeghe, J., De Roeck, E., Struyfs, H., Somers, C., et al. (2019). Association of short-term cognitive decline and MCI-to-AD dementia conversion with CSF, MRI, amyloid- and 18F-FDG-PET imaging. *NeuroImage Clin.* 22:101771. doi: 10.1016/j.nicl.2019.101771

Pena, D., Barman, A., Suescun, J., Jiang, X., Schiess, M. C., Giancardo, L., et al. (2019). Quantifying neurodegenerative progression with DeepSymNet, an end-to-end data-driven approach. *Front. Neurosci.* 13:1053. doi: 10.3389/fnins.2019.01053

Plassman, B. L., Langa, K. M., Fisher, G. G., Heeringa, S. G., Weir, D. R., Ofstedal, M. B., et al. (2008). Prevalence of cognitive impairment without dementia in the United States. *Ann. Intern. Med.* 148, 427–434.

Qing, Z., Li, W., Nedelska, Z., Wu, W., Wang, F., Liu, R., et al. (2017). Spatial navigation impairment is associated with alterations in subcortical intrinsic activity in mild cognitive impairment: a resting-state fMRI study. *Behav. Neurol.* 2017:e6364314. doi: 10.1155/2017/6364314

Rathore, S., Habes, M., Iftikhar, M. A., Shacklett, A., and Davatzikos, C. (2017). A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *NeuroImage* 155, 530–548. doi: 10.1016/j.neuroimage.2017.03.057

Riederer, I., Bohn, K. P., Preibisch, C., Wiedemann, E., Zimmer, C., Alexopoulos, P., et al. (2018). Alzheimer disease and mild cognitive impairment: integrated pulsed arterial spin-labeling MRI and 18F-FDG PET. *Radiology* 288, 198–206. doi: 10.1148/radiol.2018170575

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science*, eds N. Navab, J. Hornegger, W. Wells, and A. Frangi (Cham: Springer), doi: 10.1007/978-3-319-24574-4_28

Russo, M. J., Cohen, G., Campos, J., Martin, M. E., Clarens, M. F., Sabe, L., et al. (2017). Usefulness of discriminability and response bias indices for the evaluation of recognition memory in mild cognitive impairment and alzheimer disease. *Dement. Geriatr. Cogn. Disord.* 43, 1–14. doi: 10.1159/000452255

Sheth, S. A., Lopez-Rivera, V., Barman, A., Grotta, J. C., Yoo, A. J., Lee, S., et al. (2019). Machine learning-enabled automated determination of acute ischemic core from computed tomography angiography. *Stroke* 50, 3093–3100. doi: 10.1161/STROKEAHA.119.026189

Smith, G., Barret, F., Joo, J. H., Nassery, N., Savonenko, A., Sodums, D., et al. (2017). Molecular imaging of serotonin degeneration in mild cognitive impairment. *Neurobiol. Disord.* 105, 31–41. doi: 10.1016/j.physbeh.2017.03.040

Smith, L. N. (2017). "Cyclical learning rates for training neural networks," in *Procrrdings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, (Piscataway, NJ: IEEE).

Spasov, S., Passamonti, L., Duggento, A., Liò, P., and Toschi, N. (2019). A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease. *NeuroImage* 189, 276–287. doi: 10.1016/j.neuroimage.2019.01.031

Sun, Z., van de Giessen, M., Lelieveldt, B. P. F., and Staring, M. (2017). Detection of conversion from mild cognitive impairment to Alzheimer's disease using longitudinal brain MRI. *Front. Neuroinformatics* 11:16. doi: 10.3389/fninf.2017.00016

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). "Going deeper with convolutions," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Piscataway, NJ: IEEE), doi: 10.1109/CVPR.2015.7298594

Tong, T., Gao, Q., Guerrero, R., Ledig, C., Chen, L., Rueckert, D., et al. (2017). A novel grading biomarker for the prediction of conversion from mild cognitive impairment to Alzheimer's disease. *IEEE Trans. Biomed. Eng.* 64, 155–165. doi: 10.1109/TBME.2016.2549363

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., et al. (2017). "Residual attention network for image classification," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Honolulu, HI: IEEE), doi: 10.1109/CVPR.2017.683

Xie, C., Bai, F., Yu, H., Shi, Y., Yuan, Y., Gang, C., et al. (2012). Abnormal insula functional network is associated with episodic memory decline in amnestic mild cognitive impairment. *NeuroImage* 63, 320–327.

Xu, L., Yao, Z., Li, J., Lv, C., Zhang, H., and Hu, B. (2019). Sparse feature learning with label information for alzheimer's disease classification based on magnetic resonance imaging. *IEEE Spec. Sect. Emerg. Trends Issues Chall. Array Signal Process. Appl. Smart City* 7, 26157–26167. doi: 10.1109/ACCESS.2019.2894530

Yi, H. A., Möller, C., Dieleman, N., Bouwman, F. H., Barkhof, F., Scheltens, P., et al. (2016). Relation between subcortical grey matter atrophy and conversion from mild cognitive impairment to Alzheimer's disease. *J. Neurol. Neurosurg. Psychiatry* 87, 425–432. doi: 10.1136/jnnp-2014-309105

Young, A. L., Oxtoby, N. P., Daga, P., Cash, D. M., Fox, N. C., Ourselin, S., et al. (2014). A data-driven model of biomarker changes in sporadic Alzheimer's disease. *Brain* 137, 2564–2577. doi: 10.1093/brain/awu176

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Neuroscience

ORIGINAL RESEARCH
published: 07 January 2022
doi: 10.3389/fnins.2021.780344

Are Brain–Computer Interfaces Feasible With Integrated Photonic Chips?

Vahid Salari[1,2], Serafim Rodrigues[1]*, Erhan Saglamyurek[3,4,5], Christoph Simon[3,5,6] and Daniel Oblak[3,5]**

[1] Basque Center for Applied Mathematics (BCAM), Bilbao, Spain, [2] Quantum Biology Laboratory, Howard University, Washington, DC, United States, [3] Department of Physics and Astronomy, University of Calgary, Calgary, AB, Canada, [4] Department of Physics, University of Alberta, Edmonton, AB, Canada, [5] Institute for Quantum Science and Technology, University of Calgary, Calgary, AB, Canada, [6] Hotchkiss Brain Institute, University of Calgary, Calgary, AB, Canada

OPEN ACCESS

Edited by:
Ali Ghazizadeh,
Sharif University of Technology, Iran

Reviewed by:
Dimiter Prodanov,
Interuniversity Microelectronics Centre (IMEC), Belgium
Vassiliy Tsytsarev,
University of Maryland, College Park, United States

***Correspondence:**
Vahid Salari
vsalari@bcamath.org
Serafim Rodrigues
srodrigues@bcamath.org
Daniel Oblak
doblak@ucalgary.ca

Specialty section:
This article was submitted to Neural Technology, a section of the journal Frontiers in Neuroscience

Received: *20 September 2021*
Accepted: *22 November 2021*
Published: *07 January 2022*

Citation:
Salari V, Rodrigues S, Saglamyurek E, Simon C and Oblak D (2022) Are Brain–Computer Interfaces Feasible With Integrated Photonic Chips? Front. Neurosci. 15:780344. doi: 10.3389/fnins.2021.780344

The present paper examines the viability of a radically novel idea for brain–computer interface (BCI), which could lead to novel technological, experimental, and clinical applications. BCIs are computer-based systems that enable either one-way or two-way communication between a living brain and an external machine. BCIs read-out brain signals and transduce them into task commands, which are performed by a machine. In closed loop, the machine can stimulate the brain with appropriate signals. In recent years, it has been shown that there is some ultraweak light emission from neurons within or close to the visible and near-infrared parts of the optical spectrum. Such ultraweak photon emission (UPE) reflects the cellular (and body) oxidative status, and compelling pieces of evidence are beginning to emerge that UPE may well play an informational role in neuronal functions. In fact, several experiments point to a direct correlation between UPE intensity and neural activity, oxidative reactions, EEG activity, cerebral blood flow, cerebral energy metabolism, and release of glutamate. Therefore, we propose a novel skull implant BCI that uses UPE. We suggest that a photonic integrated chip installed on the interior surface of the skull may enable a new form of extraction of the relevant features from the UPE signals. In the current technology landscape, photonic technologies are advancing rapidly and poised to overtake many electrical technologies, due to their unique advantages, such as miniaturization, high speed, low thermal effects, and large integration capacity that allow for high yield, volume manufacturing, and lower cost. For our proposed BCI, we are making some very major conjectures, which need to be experimentally verified, and therefore we discuss the controversial parts, feasibility of technology and limitations, and potential impact of this envisaged technology if successfully implemented in the future.

Keywords: ultraweak photon emission, brain-computer interface, photonic interferometry, pattern recognition, integrated photonic circuit, on-chip photon detection, quantum technology

1. INTRODUCTION

Brain–computer interface (BCI), or generally brain–machine interface (BMI), is a computer (machine)-based system that maps brain signals into computer (machine) commands or actions. This mapping may involve intermediate analysis and processing. Moreover, a closed-loop BCI is also possible, whereby the brain is stimulated via relevant neuro-bio-signals. The most common brain signals used in BCIs are electromagnetic, that is, of classical/non-quantum origin. Herein, we turn attention to an exciting and emergent literature that reveals the brain also emits "photons," which are quanta of electromagnetic waves. The intensity of these emissions varies from a few photons to several hundred photons per second per square centimeter, mainly with spectral range of 200–800 nm (Salari et al., 2015). A caveat is that most single-photon sensitive detectors used in the experiments were only sensitive up to about 900 nm. Hence, observations with detector platforms that are sensitive in the 900–1,600 nm range, such as superconducting nano-wire single-photon detectors (SNSPDs) (Marsili et al., 2013), which also can be shaped as arrays (Wollman et al., 2019), may reveal hidden obscured about the UPE light.

The body of evidence for ultraweak photon emission (UPE) is fast growing and is being independently observed by different scientific communities/labs. Due to infancy of the research field, many different terms are used to describe this phenomenon, including biophotons, ultraweak photon emission, ultraweak bioluminescence, self-bioluminescent emission, photoluminescence, delayed luminescence, ultraweak luminescence, spontaneous chemiluminescence, ultraweak glow, biochemiluminescence, metabolic chemiluminescence, dark photobiochemistry, etc. (Salari et al., 2017; Esmaeilpour et al., 2020). In this report, we will henceforth adopt the term UPE. It has been evidenced that neurons and other living cells (e.g., in plants, animals, and humans) have spontaneous UPE (Cifra and Pospisil, 2014; Pospisil et al., 2014) mediated via their metabolic reactions associated with physiological conditions. In 1967, it was first reported that electric pulses in neurons can induce weak photon emission (in the visible region of the EM spectrum) due to chemical reactions accompanying pulses, while a dead-neuron does not exhibit any photon emission (Artem'ev et al., 1967). In 1984 (Imaizumi et al., 1984) and 1985 (Suzuki et al., 1985), it was demonstrated experimentally that after the induction of hypoxia states in a rat brain, UPE increases. Isojima et al. (1995) showed that there is a correlation between the intensity of UPE and neural metabolic activity in the rat hippocampal slice. In 1997, Zhang et al. (1997) revealed that the intensity of UPE from intact brains isolated from chick embryos was higher than the medium in which the brain was immersed. In 1999, Kobayashi et al. (1999b) detected spontaneous UPE in the rat's cortex *in vivo* without adding any chemical agent or employing external excitation and found that the UPE correlates with the electroencephalography (EEG) activity, cerebral blood flow, and hyperoxia, and the addition of glutamate increases UPE, which is mainly originated from the energy metabolism of the inner mitochondrial respiratory chain through the production of reactive oxygen species (ROS). Kataoka et al. (2001) detected spontaneous UPE from cultured rat cerebellar granule neurons in the visible range and demonstrated that the UPE depends on the neuronal activity and cellular metabolism. Then, a fascinating experimental discovery by Sun et al. revealed that photons can be conducted along neuronal fibers. In 2011, Wang et al. (2011) show-cased *in vitro* experimental evidence of spontaneous UPE and visible light-induced UPE (delayed luminescence) from freshly isolated rat's whole eye, lens, vitreous humor, and retina. Subsequently, in 2014 (Tang and Dai, 2014) Tang and Dai provided experimental evidence that the glutamate-induced UPE can be transmitted along the axons and in neuronal circuits in mouse.

These observations raise the following intriguing question: what are the underlying physiological processes that underpin UPE? Specifically, in the brain what are the associated neurophysiological processes? Although a complete picture has not been provided, it has been shown that the origin of UPE is in direct connection with the ROS. Moreover, its intensity has a direct correlation with thermal, chemical, and mechanical stress, the mitochondrial respiratory chain, cell cycle, neural activity, EEG activity, cerebral blood flow, cerebral energy metabolism, and release of glutamate. Experiments also show that cells can absorb photons by photochemical processes and slowly release these photons as delayed luminescence (Scordino et al., 2014). Interestingly, it has been shown that delayed luminescence emitted from the biological samples provide valid and predictive information about the functional status of biological systems (Musumeci et al., 2005; Niggli et al., 2005, 2008). All this opens novel exciting mathematical and physical questions at the interface of quantum biology. For example, if we consider UPE in the context of metabolism, then there has been efforts to propose quantum-metabolism (Demetrius, 2003). As it is well-known, biological systems are essentially isothermal and as such energy flow in living organisms is mediated by differences in the turnover time of various metabolic processes in the cell, which occur cyclically. The mean cycle time (τ) of these metabolic processes (turnover of essentially redox reactions) are related to the metabolic rate (g), that is, the rate at which the organism transforms the free energy of nutrients into metabolic work. This is related to two coupled chains (electron-proton transport) of the ATP system in the mitochondria. In quantum-metabolism the main variables are metabolic rate, the entropy production rate, and the mean cycle time. Then the fundamental unit of energy is given by $E(\tau) = g\tau$, where g is related to the electron–proton transport. Noteworthy, this is in contrast, but has some correspondence to quantum thermodynamics, where the thermal energy per molecule is given by $E = K_bT$, which relates specific heat, Gibbs–Boltzmann entropy, and absolute temperature T. The difference is that biological systems work far from thermodynamic equilibrium, hence in quantum-metabolism the variables depend on fluxes (rates of change of energetic values). On top of this, Albrecht-Buehler (1995) hypothesized that the electron–proton transport releases photons ($E = h\nu$, where E is the photon energy, h is plank constant, and ν photons frequency). Other researchers have contemplated at why UPE displays wide variety of frequencies, with Popp suggesting that these are coherent and mediated by DNA, thus it may regulate life processes of an organism (Popp et al., 1984, 1988). However, the coherence idea of UPE is still under debate (Salari and Brouder, 2011) and it is yet unclear if UPE is just a by-product in biological metabolism or it has some informational or functional role.

So far, UPE signals have only been studied in the context of basic science and has not been considered for experimental and clinical applications or novel technologies such as BCIs. The present article takes that first step forward and propose an implant BCI chip based on UPE. Since UPE is correlated to several sub-cellular, cellular, and neural tissue processes, there is also the potential that it can be used as a novel technological probe/bio-marker for both normal brain function and pathological conditions. In the subsequent sections, we will first briefly review the traditional classical methods in BCI and then we will focus our discussion toward UPE detection and pattern recognition for the development of a novel UPE-based skull implant BCI.

2. CLASSICAL BRAIN–COMPUTER INTERFACE TECHNOLOGY

In traditional BCI techniques, different types of signal acquisition may be used, depending on the application. In the following, we briefly review four types of brain signals, their properties, and the suitable machine interfaces.

- **Electroencephalography (EEG) signals**

  EEG is the most employed method to detect electrical activity of the brain by use of small electrodes attached to the scalp (Niedermeyer and da Silva, 2004). These signals are recorded by a machine for tracing both normal brain function and diagnosing pathological conditions (e.g., epilepsy). In stimulus (e.g., visual cue) induced EEG, there is positive deflection of voltage with a latency (delay between stimulus and response) of roughly 250–500 ms, which is called event-related potentials (ERP). Examples of such ERP is the so-called P300 formed at time 300 ms, which is related to decision making. Indeed, cognitive impairment is often correlated with modifications in the P300 (Polich, 2007). It is considered an endogenous potential, as its occurrence links not to a stimulus' physical attributes, but a person's reaction to it. More specifically, the P300 is thought to reflect processes involved in stimulus evaluation or categorization. The presence, magnitude, topography, and timing of this signal are often used as metrics of cognitive function in decision-making processes and hence used in BCIs. The P300 has several desirable qualities for pattern recognition. First, the waveform is consistently detectable and is elicited in response to precise stimuli. The P300 waveform can also be evoked in nearly all subjects with little variation in measurement techniques, which help simplify interface designs and permit greater usability. The speed at which an interface can operate depends on how detectable the signal is despite "noise." One negative characteristic of the P300 is that the waveform's amplitude requires averaging multiple recordings to isolate the signal. This and other post-recording processing steps determine the overall speed of a BCI interface (Donchin et al., 2000).

- **Magnetoencephalography (MEG) signals**

  MEG is a functional neuroimaging technique monitoring brain activity via magnetic fields of electrical currents in the brain, using SQUIDs (superconducting quantum interference devices), which are very sensitive magnetometers operated in a cryogenic environment. Another type of magnetometer is spin exchange relaxation-free (SERF) magnetometer (Hämäläinen et al., 1993), which can increase portability of MEG scanners, while it features sensitivity equivalent to that of SQUIDs. A typical SERF magnetometer is relatively small and does not require bulky cooling system to operate. It has been demonstrated that MEG could work with a type of SERF, i.e., chip-scale atomic magnetometer (CSAM) (Sander et al., 2012), where its development can be used efficiently for BCI. Basically, MEG may provide signals with higher spatiotemporal resolution than EEG, and therefore useful for an increased BCI communication speed.

- **Electrocorticography (ECoG) signals**

  ECoG uses electrodes placed directly on the surface of the brain to record electrical activity from the cerebral cortex, i.e., an invasive technology that involves removing a part of the skull to expose the brain surface to enable the implant of an electrode grid on the surface of the brain, i.e., called craniotomy, which is a surgical procedure performed either under general anesthesia or under local anesthesia if patient interaction is required for functional cortical mapping. The spatial and temporal resolution of the resulting signal is higher and the signal to noise ratio (SNR) superior to those of EEG due to the closer proximity to neural activity. Thus, ECoG is a promising recording technique for use in BCI, especially for decoding imagined speech or music, in which users simply imagine words, sentences, or music that the BCI can directly interpret (Shenoy et al., 2007).

- **Functional near-infrared spectroscopy (fNIRS) signals**

  fNIRS is a non-invasive optical imaging technique that measures changes in hemoglobin (Hb) concentrations in the brain by means of the characteristic absorption spectra of Hb in the near-infrared (NIR) range (Scholkmann et al., 2014). fNIRS tomography makes use of the fact that light penetrates up to several centimeters into biological tissue, i.e., a safe technique that is minimally invasive and which relies on small, relatively inexpensive easy-to handle technology, and provides relatively low spatial resolution. The penetration range of light in tissue limits the size of the target tissue volume. fNIRS can be used in BCI for the restoration of movement capability for people with motor disabilities. fNIRS cannot afford high error rates for safety purposes, and must be fast enough to provide real-time control. Several fNIRS-BCI studies have tried to improve classification accuracies and information transfer rates (Naseer and Hong, 2015).

3. POTENTIAL APPLICATION OF UPE IN BCI

UPE is largely mediated by cellular metabolism and it is presently believed that it is merely a by-product (i.e., epiphenomenon). A tempting question is whether it is possible (or not) to retrieve information from stochastic emission of UPE? In previous sections, we already saw that there are different experimental reports on significant correlations between UPE emission and neuronal activity and associated metabolic processes (Isojima et al., 1995; Kobayashi et al., 1999b; Kataoka et al., 2001; Tang and Dai, 2013). Therefore, even if UPE is an epiphenomenon, its intensity can be a proxy for tracking the underlying neural information that dynamically changes under various conditions. Indeed, UPE seem to include information for monitoring physiological variations in a neuronal tissue. Note that for EEG signals we have a similar scenario. Indeed, EEG signals do not provide specific information about single neurons. Rather, it reflects a non-trivial summation of the synchronous activity of thousands of neurons and not that of a single neuron or dendrite. Thus, retrieving patterns as information from EEG is a data-science activity typically involving statistical comparisons between different brain states (e.g., normal and abnormal brain states).

Scholkmann (2015, 2016) hypothesized that UPE may be is used by neurosystems as an additional signal enabling cell-to-cell communication and coupling. Indeed, Sun et al. (2010) found that UPE can conduct along the neural fibers. It has been hypothesized based on numerical simulations that neurons (or myelinated axons) may act as optical fibers and, hence, may conduct light associated with UPE (Kumar et al., 2016), and through these waveguides UPE may even mediate long-range quantum entanglement in the brain (Kumar et al., 2016; Zarkeshian et al., 2018; Simon, 2019). These myelinated axons are tightly wrapped by the myelin sheath, which has a higher refractive index (Antonov et al., 1983) than the inside of the axon and the interstitial fluid outside. Myelin is an insulating layer (sheath) around nerves, which is formed by two types of specialized glial cells, oligodendrocytes in the central nervous system (CNS) and Schwann cells in the peripheral nervous system (Simons and Trajkovic, 2006). Muller glia cells have also been suggested to guide photons within mammalian eyes (Franze et al., 2007; Agte et al., 2011; Reichenbach and Bringmann, 2013). These observations suggest that UPE and bioelectronic activities are not independent biological phenomena in the nervous system, and their synergistic action may take on considerable function in neural (quantum) signal and information processes (Salari et al., 2016b; Wang et al., 2016).

3.1. UPE Intensity From the Surface of the Human Brain

The UPE observed to date has been extremely weak. However, the true UPE intensity within neurons can be significantly higher than the one expected from the UPE measured a short distance away from the brain, as was done in all previous observations. Since photons are strongly scattered and absorbed in cellular or neural systems, the corresponding intensity of UPE within the organism or brain can even be two orders of magnitude higher (Slawinski, 1988; Chwirot, 1992). Based on the data from experiments with rat brain—employing a 2D photon-counting tube with a photocathode featuring a minimum detectable radiant flux density of $9.9 \times 10^{-17} W/\mathrm{cm}^2$ under 1-s observation time—the intensity of UPE has approximately 100 $\frac{\mathrm{counts}}{\mathrm{sec.cm}^2}$ from the cortex surface (Imaizumi et al., 1984; Adamo et al., 1989; Kobayashi et al., 1999a,b). Moreover, the limited quantum efficiency (QE) of the detector may impede the detection of UPE due to the limited SNR. Regarding the human brain, the neuronal density in V1 in visual cortex is $60 \times 10^6 \frac{\mathrm{Neurons}}{\mathrm{cm}^3}$ in postmortem human brains (Pakkenberg and Gundersen, 1987), It should be noted that postmortem studies use fixatives, which lead to shrinking of the tissue. The result is that the cell density is overestimated, while the volume of the extracellular space is underestimated. The reported number can be used only as an absolute best-case scenario for the interface. The V1 thickness is about 0.2 cm, and V1 surface area of one hemisphere is about 26 cm^2 in adult humans. At least, 10^6 neurons in object-related areas and 30×10^6 neurons in the entire visual cortex are activated by a single-object image (Levy et al., 2004). Based on a rough estimation, about 10^6 free radicals can be produced by each brain cell per second (Bokkon et al., 2010), which yields $10^6 \times 10^6 = 10^{12}$ free radicals produced by human visual neurons per second in V1 of one hemisphere during perception of a single-object image. Since UPE mainly originates from free radicals, the actual UPE intensity inside neuronal cells is expected to be considerably higher than the intensity measured by a detector outside [e.g., 100 counts/(s.cm^2)]. If the QE of an ideal photodetector is close to 100%, we conjecture that it may measure the UPE intensity at the cortex surface at least on the order of 1,000 counts/(sec.cm^2) for an object visualization.

4. SKULL-IMPLANT SETUP FOR THE UPE-BASED BCI

We now provide the complete design specification of a radically novel skull-implant that can facilitate a UPE-based BCI (see **Figure 1**). The envisaged BCI is not aimed for deep brain implants (although possible) but rather for intracranial brain surface implant (i.e., minimally invasive). The environment of a closed skull (after surgical implantation) is sufficiently dark and, therefore, a suitable environment for the detection of UPE signals. Once the UPE signals are detected, they are wirelessly relayed to a machine, computer, or smartphone. We also envisage alternative designs with closed-loop signals (photons) for modulating the metabolic processes of a neural tissue. However, herein we will only consider the read-out of UPE signals. The center-piece of the envisaged technology is the UPE-based integrated chip, which we will discuss at length in the subsequent sections. The integrated photonic chip is assembled from different component parts; specifically, a receiver optical plane (ROP), optical fibers, a photonic interferometery circuit, a complementary metal-oxide-semiconductor (CMOS) detector array, a battery, and a wireless system (see **Figure 2**). The use of the implantable CMOS image sensor has been described in recent years especially for optogenetic imaging (Tokuda et al., 2021).

The UPE photons first enter an ROP on the chip, which is essentially a photo-receiver array made up of optical fibers of size of $N \times N$, where N is the number of pixels (or fibers) in each row or column and each pixel is indeed an optical fiber

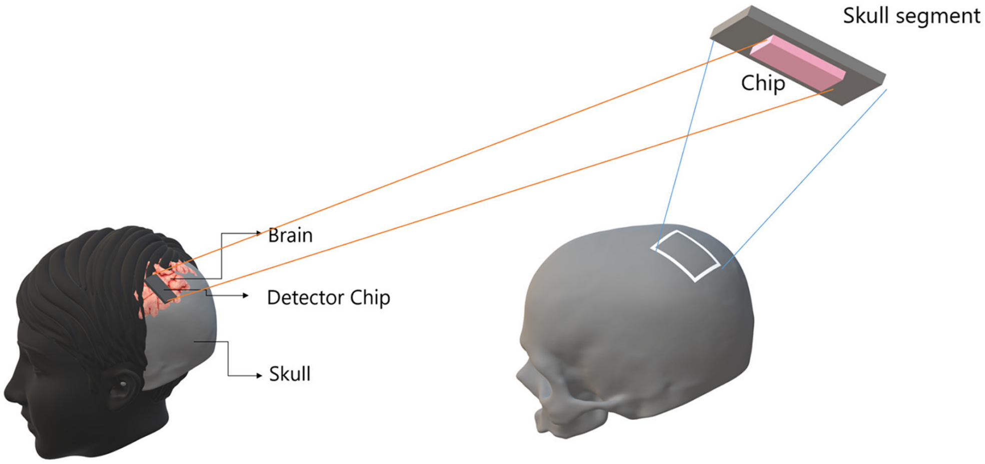


FIGURE 1 | A detector chip can be installed on the interior surface of the skull without touching the brain tissue, i.e., non-invasive. The environment of the closed skull in the head is sufficiently dark and therefore it is a suitable environment for the detection of UPE with the installed chip. The intensity of UPE is stronger close to the surface of the brain, which can be captured by chip on the skull.

that couples into a waveguide on the chip, using grating couplers (Cheng et al., 2020). Alternatively, the UPE light can be directly coupled to waveguides created by femto-second laser-writing and since these can be patterned at different depths in the chip (Nolte et al., 2003), and they can directly facilitate serialization step. Subsequently, the $N \times N$ pixels are serialized into a 1D vector (where $N' = N \times N$ is the number of optical fibers connected to the waveguides in the optical interferometer with N' input ports in a series and linear 1D form, and therefore N' CMOS pixels in a single row as the output port on the PIC). In fact, the received photons on ROP are guided to the optical interferometer via optical fibers. The advantage of an optical interferometer is that it may discriminate the emission patterns of photons. We estimate that UPE intensity ranges 10–1,000 counts per second per each cm^2 of the whole array, depending on how active a neuron or neural tissue is at a given time instant. In fact, we expect that similar and non-similar UPE emissions (in wavelength) generate different detection distributions, where interference will occur between photons with similar wavelength (i.e., emanating from the same-type neural processes). Thus, the detection distributions for similar-wavelength photons will be closer to an optical interference pattern, which is uniquely determined by the wavelength of these interfering photons. In this regard, one of the concerns may arise from the fact that UPE emission over a broad range of wavelengths can lead to the observation of different patterns at the same time, rendering an ambiguous combination of several independent patterns. Such complexity may bring disadvantages over the direct detection (i.e., no interferometer), or even could cause wrong interpretations. This potential problem can be alleviated by classifying those different wavelength patterns, again with pattern recognition techniques in machine learning, such as (PCA) (Jolliffe, 2002), which allows distinguishing the differences in an ensemble of patterns, and identifying each pattern according to the respective wavelength, after many sets of training data. The optical interferometer photons are then converted into electrical signal via the CMOS array (see **Figure 6** for details). Finally, these signals are wirelessly linked to a smartphone or computer for pattern recognition/extraction. Noteworthy, since the number of detected photons is relatively low and because the data acquisition is in real time, the recognition of patterns should be done via machine learning protocols, e.g., convolutional neural networks (CCN), which is a powerful tool for 2D pattern recognition. We subsequently discuss in more detail each component part of the UPE-based electronic chip.

4.1. On-Chip Photonic Integrated Circuits

We base our proposed technology on photonic integrated circuits (PICs) (Coldren et al., 2012). These are chip that contains photonic components that operate with light (photons), where photons pass through optical components such as waveguides (equivalent to a resistor or electrical wire in an electronic chip). With electronic integrated circuits arriving at the end of their integration capacity, PICs have the potential to be the preferred technology. Nowadays, photonic platforms present several advantages for quantum information protocols enabling long coherence times, full connectivity, scalability, and operation in room temperature. Different photonic degrees of freedom including polarization, spectral, spatial, and temporal modes can be used to encode information, providing different experimental resources for a wide variety of quantum information tasks.

For our application, we consider a PIC containing an optical interferometer. A linear interferometer can be fabricated through silica-on-silicon or laser-written integrated interferometers, or electrically and optically interfaced optical chips (Szameit et al., 2007; Spring et al., 2012; Carolan et al., 2015), which makes a simple processor reducing the amount of physical resources needed for implementation.

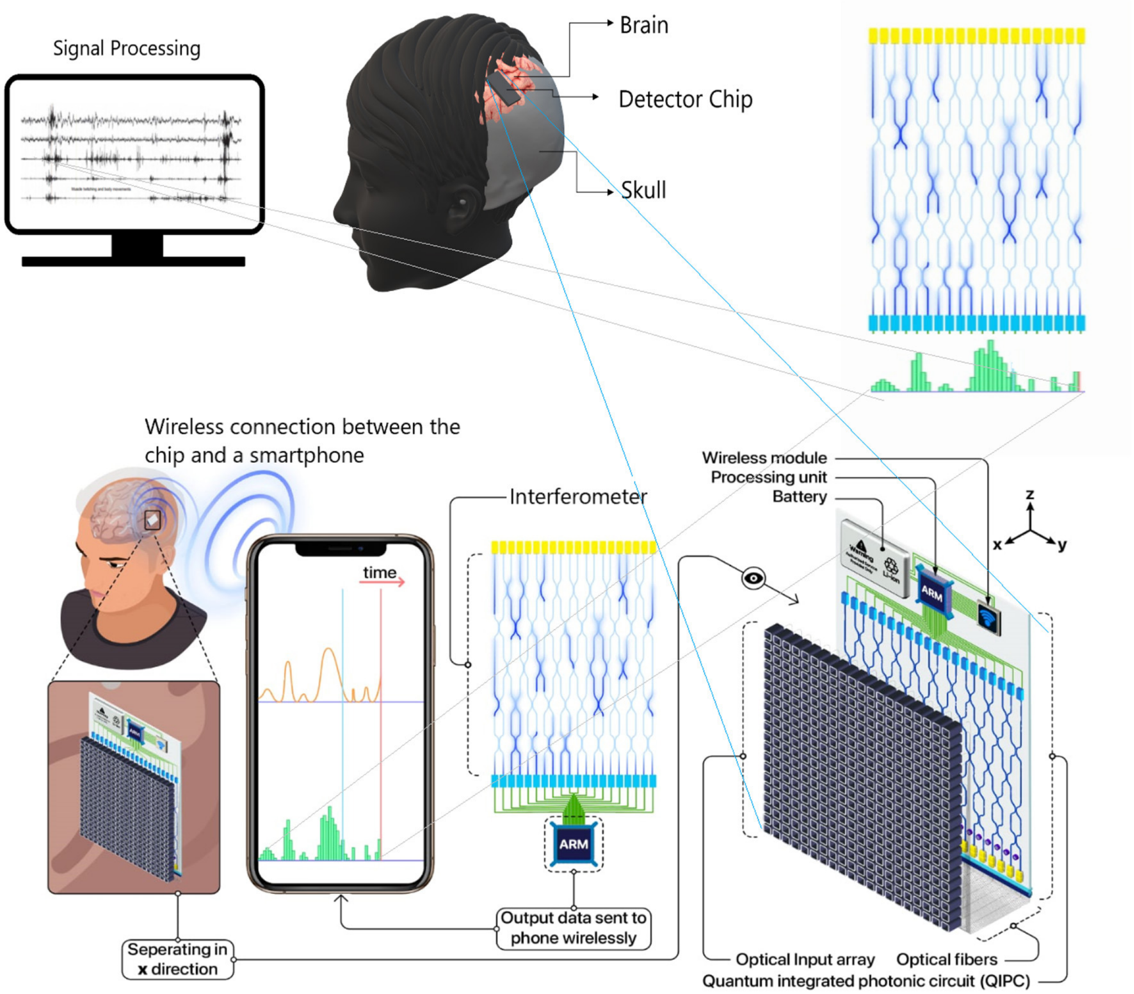

Signal Processing
Brain
Detector Chip
Skull
Wireless connection between the chip and a smartphone
Interferometer
time
ARM
Wireless module
Processing unit
Battery
z
x
y
Output data sent to phone wirelessly
Seperating in x direction
Optical Input array
Optical fibers
Quantum integrated photonic circuit (QIPC)

FIGURE 2 | In a brain–computer interface (BCI) proposal, an optical chip is implanted on the interior surface of the skull. A few number of ultraweak photon emission (UPE) photons interfere in a photonic chip and the results are detected as different single photon distributions at detectors vs. time. This results are communicated via wireless signals from the detector part of the chip to a receiver (e.g., smartphone or a computer).

4.2. Photons Statistics and Distributions

In the context of optics, coherence is a property of light. In a simplified picture, coherence is the ability of light to make interference, e.g., in the double-slit interference experiment light can create interference patterns (bright and dark bands) for both a wave (classical) and photon (quantum) picture. Thus, coherence of light can be both of a classical and quantum character. For example, thermal states of light can be described in the classical and the quantum framework, while other states, such as squeezed states, can only be described in the quantum framework. One of the essential conditions to show the coherence property of light is for its intensity/photon-number distribution to be a Poisson distribution. However, this condition is not sufficient to conclude that the light is certainly coherent. Other types of sources may yield a Poisson distribution, e.g., shot noise and dark noise. In the following paragraphs, we will introduce a couple of photon-number distributions in order to demonstrate how this measure provides insight into the nature of the UPE light being emitted.

The photocount statistics of coherent light is a Poisson distribution (Cifra and Brouder, 2015).

$$P_n(t,T) = \frac{\langle n\rangle^n}{n!} e^{-\langle n\rangle} \qquad (1)$$

where $\langle n\rangle$ is the average number of photons measured between time t and time $t + T$. The variance of Poisson distribution is equal to its mean $\langle(\Delta n)^2\rangle = \langle n\rangle$. The deviation of the photon-number distribution from the Poisson distribution is measured by the Fano factor F such that $\langle(\Delta n)^2\rangle = \langle n\rangle F$, or by the Mandel parameter $Q = F - 1$. A photocount statistics is said to be super-Poissonian if $F > 1$ and $Q > 0$, and sub-Poissonian (and therefore non-classical) if $F < 1$ and $Q < 0$. Hence, the shift from a Poisson distribution is a sign of non-classical (quantum)

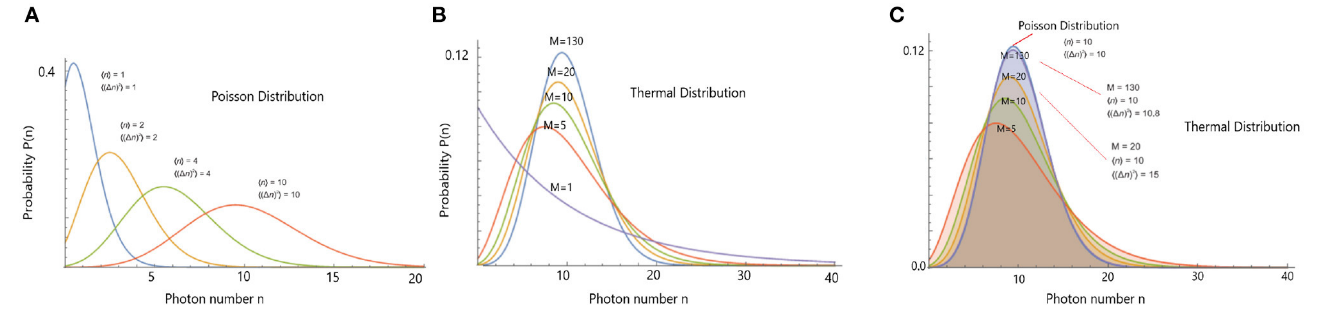


FIGURE 3 | (A) Poisson distribution for four different average values of photon counts $\langle n \rangle$. **(B)** Demonstration of thermal field photocount distribution for different number of thermal modes for the average number of 10 photons. **(C)** Thermal field photocount distribution (with similar $\langle n \rangle$) approaches Poisson distribution for a large number of modes *M*.

characteristics of the light (Cifra and Brouder, 2015) while a Poisson distribution is a sign of classicality.

The photocount statistics of a thermal source with M modes is approximated by the expression

$$P_n(t, T, M) = \frac{(n + M - 1)!}{n!(M - 1)!})(1 + \frac{M}{\langle n \rangle})^{-n}(1 + \frac{\langle n \rangle}{M})^{-M} \qquad (2)$$

where $\langle n \rangle$ is the average number of photons and M is the number of field modes (Cifra and Brouder, 2015). An important characteristic of these states is the relation between the variance and the mean $\langle (\Delta n)^2 \rangle = \langle n \rangle + \frac{\langle n \rangle^2}{M}$. The coefficient M is generally very large for chaotic sources. So that the relation between the variance and the mean is close to that of a coherent state, i.e., for large M, $P_n(t, T, M)$ approaches a Poisson distribution (see **Figure 3**). In relation to UPE, it is important to know whether the photocount statistics can distinguish between the coherent and thermal emissions, because photocount statistics of thermal light becomes equal to that of a coherent state when the number of modes M is large. Since the photocount statistics are not able to discriminate between a coherent and a thermal state with many modes.

Another type of emission is super-radiance, which is the coherent emission of light by several sources, and its main characteristic is the fact that the intensity of the emitted light can vary with the square of the number of sources because they can emit in phase. The photocount statistics of super-radiant emission is sub-Poissonian (Cifra and Brouder, 2015), and the photon state of a super-radiant system is generally not a coherent state.

4.2.1. Photon Detection With Interference

The photons collected onto our chip will then be propagated through a PIC featuring several interference paths and other components. The model of the effect of the PIC on the incident photons aims to predict the probability distribution of photons at the detector following their propagation and interference in a linear interferometer. The experimental setup only requires photodetectors and linear optical elements, i.e., beam splitters and phase shifters. Suppose the chip is injected with an input state of single photons of UPE, $|S\rangle = |s_1, s_2, ..., s'_N\rangle$ where s_k are the number of UPE emitted photons in the kth mode and injected into the chip. The output state of the chip can be written as $|O\rangle = |x_1, x_2, ..., x'_N\rangle$. For the sake of simplicity, suppose that there are four outputs on the chip. Therefore, probabilities of output detection for $N = 1$ input photons in case there is no dissipation in the circuit are $P_{|1000\rangle}$, $P_{|0100\rangle}$, $P_{|0010\rangle}$, $P_{|0001\rangle}$, and for $N = 2$ input photons the probabilities at the output are $P_{|1001\rangle}$, $P_{|1010\rangle}$, $P_{|1100\rangle}$, $P_{|0110\rangle}$, $P_{|0011\rangle}$, $P_{|0101\rangle}$, $P_{|2000\rangle}$, $P_{|0200\rangle}$, $P_{|0020\rangle}$, $P_{|0002\rangle}$. Now, we consider a general case for N' outputs. The signal processing and the interpretation of the signals require machine learning techniques. As the signal acquisition is performed through an interferometer, different interference patterns may form. We suggest a pattern recognition approach via convolutional neural networks (CNNs) (Fukushima, 1980) for an efficient interpretation of output signals on the photonic interferometer chip. Here, the conjecture is that a synchronous activity in a specific region of cortex makes synchronous similar metabolism with similar chemical reactions producing similar ROS by-products simultaneously, and therefore the probability of detection of similar photons (even with a low probability of interference in the interferometer) during a specific brain activity is higher than the normal state with stochastic photon emissions. Discrimination between the interference pattern of active and normal states will be non-trivial but tractable via machine learning. This conjecture is expected to be reasonable based on highly synchronized brain activities for different specific cognitive tasks. In fact, the photonic chip continuously produces data under normal and active states of the brain. The patterns can be recognized by studying the data and classifications via discrimination between the signals of normal and active states. In such a state, both supervised and unsupervised learning can be performed on software. This can be an advantage of the method.

The idea of using UPE signals for BCI applications still remains at the level of conjecture, relying on a mere fact that UPE shows correlations with some brain activities. Therefore, from a BCI point of view, such correlations are very important because for almost all types of brain signals for BCI applications, it is hardly possible to extract specific information from the signals directly. With an analysis of signals over thousands of

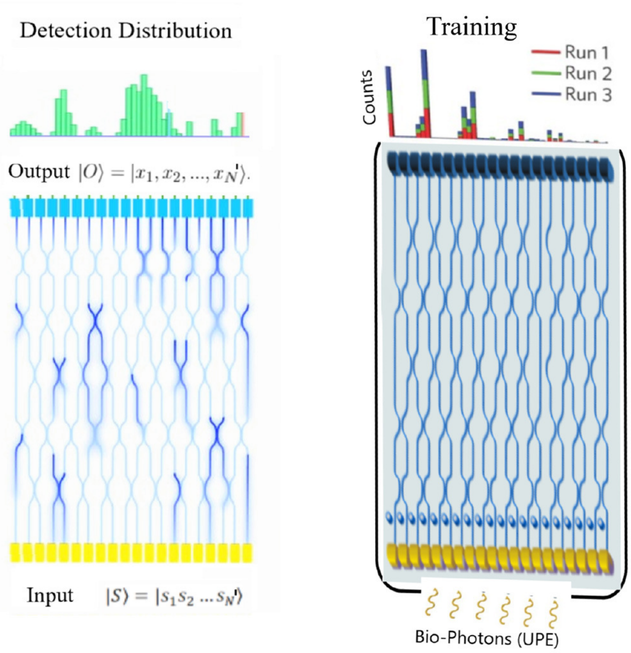

Detection Distribution
Output $|O\rangle = |x_1, x_2, ..., x_{N'}\rangle$.
Input $|S\rangle = |s_1 s_2 ... s_{N'}\rangle$
Training
Counts
Run 1
Run 2
Run 3
Bio-Photons (UPE)

FIGURE 4 | On-chip optical interferometer with N' inputs and N' outputs. The output patterns can be processed for feature extraction via machine learning techniques. It is expected that for each cognitive task or decision making, a similar pattern (in average) forms after many runs under training for specific tasks. The features of the average pattern can be recognized by deep learning methods, or specifically by convolutional neural networks (CNNs) on a software.

training trials, it will be possible to obtain an average pattern with specific features (for feature extraction) that finally make it easy for a specific algorithm to recognize the pattern in the next acquisition signals directly. Here, we suggest using a machine-learning algorithm to discriminate variations and extraction of features by enhancement of training data. A deep-learning algorithm becomes stronger in learning with increasing the training data to a specific level. This is a benefit for an implanted chip since it is always creating thousands of patterns easily to be processed by software on a computer or a smartphone. There is no need to perform separate experiments each time for training. Therefore, a deep-learning algorithm can learn how to understand features from UPE signals and interpret them according to the relevant cognitive task. Thus, data analysis of the output UPE signals of the chip can be performed via machine learning in general and deep learning specifically. For instance, a possibility is via deep learning method called CNN technique, which enables high-resolution pattern recognition. Since CNNs are ideal for 2D imaging processing, then the UPE signals detected at the receiver optical plane pixel-array can be readily adapted for CNN (see **Figure 4**). The pattern analysis can be enhanced depending on the details our architecture. CNN error minimization methods are used to optimize convolutional networks in order to implement quite powerful pattern transformations. This is very useful when the input is spatially or temporally distributed. The first layer of a CNN generally implements non-linear template-matching at a relatively fine spatial resolution, extracting basic features of the data. Subsequent layers learn to recognize particular spatial combinations of previous features, generating “patterns of patterns” in a hierarchical manner. If down-sampling is implemented, subsequent layers perform pattern recognition at progressively larger spatial scales, with lower resolution. A CNN with several down-sampling layers enables processing of large spatial arrays, with relatively few free weights. As we discussed before, an ensemble of wavelengths may make different patterns at the same time and obscure the interference patterns, where a PCA algorithm (Jolliffe, 2002) can find the differences between different patterns in the overlapped patterns, and classify each pattern for the relevant wavelength after many sets of training data.

4.3. Implementation Feasibility

We now discuss the feasibility of fabricating all elements of our envisaged skull-implant UPE-based BCI (to be followed with **Figure 5**).

4.3.1. Chip Ingredients

The design and fabrication of PICs is a mature technology, which is realized on a variety of material platforms, which are tailored to the needs and requirements of the application at hand. Available platforms for lithography-based fabrication include silicon photonics [Silicon on Insulator (220 nm and 3 μm SOI), Si-based silica on silicon (SiO_2, also known as PLC), and silicon nitride (SiN and TriPleX)], III-V photonics such as indium phosphide (InP), gallium arsenide (GaAs) and derivatives, and finally lithium niobate ($LiNbO_3$) and other more exotic materials (Liang and Bowers, 2009; Washburna and Bailey, 2011; Fang and Zhao, 2012; Arakawa et al., 2013; Chrostowski and Hochberg, 2015; Muñoz et al., 2017; Boes et al., 2018; Zhu et al., 2021). It should be noted that the SIO platform is not a suitable candidate for the UPE in the visible spectrum as the relatively small band-gap of silicon renders it completely opaque below a wavelength of about 1,000 nm. SiN, which, on the other hand, is transparent in the visible wavelength-range and features compatibility with CMOS technology (Romero-Garcia et al., 2013), appears to be a strong candidate as a PIC platform for our proposed BCI. As an alternative to the lithography-based PIC, femto-second laser-written waveguides (FLWs) in SiO_2 (glass) have in recent years been used to successfully implement advanced PICs (Davis et al., 1996; Marshall et al., 2009). The unique advantage of FLWs is that the ability to define waveguides in three dimensions, i.e., including at different depths in the chip. This allows more complex routing, such as the crossing of waveguides (Marshall et al., 2009).

Choosing the right technology will be the starting point for having a successful integrated chip. By integrating all devices into a single chip, complex assembly, alignment, and stabilization processes are avoided, and packaging and testing are greatly simplified. Moreover, it is the only way to scale up complexity when moving over 20–30 components into a single package.

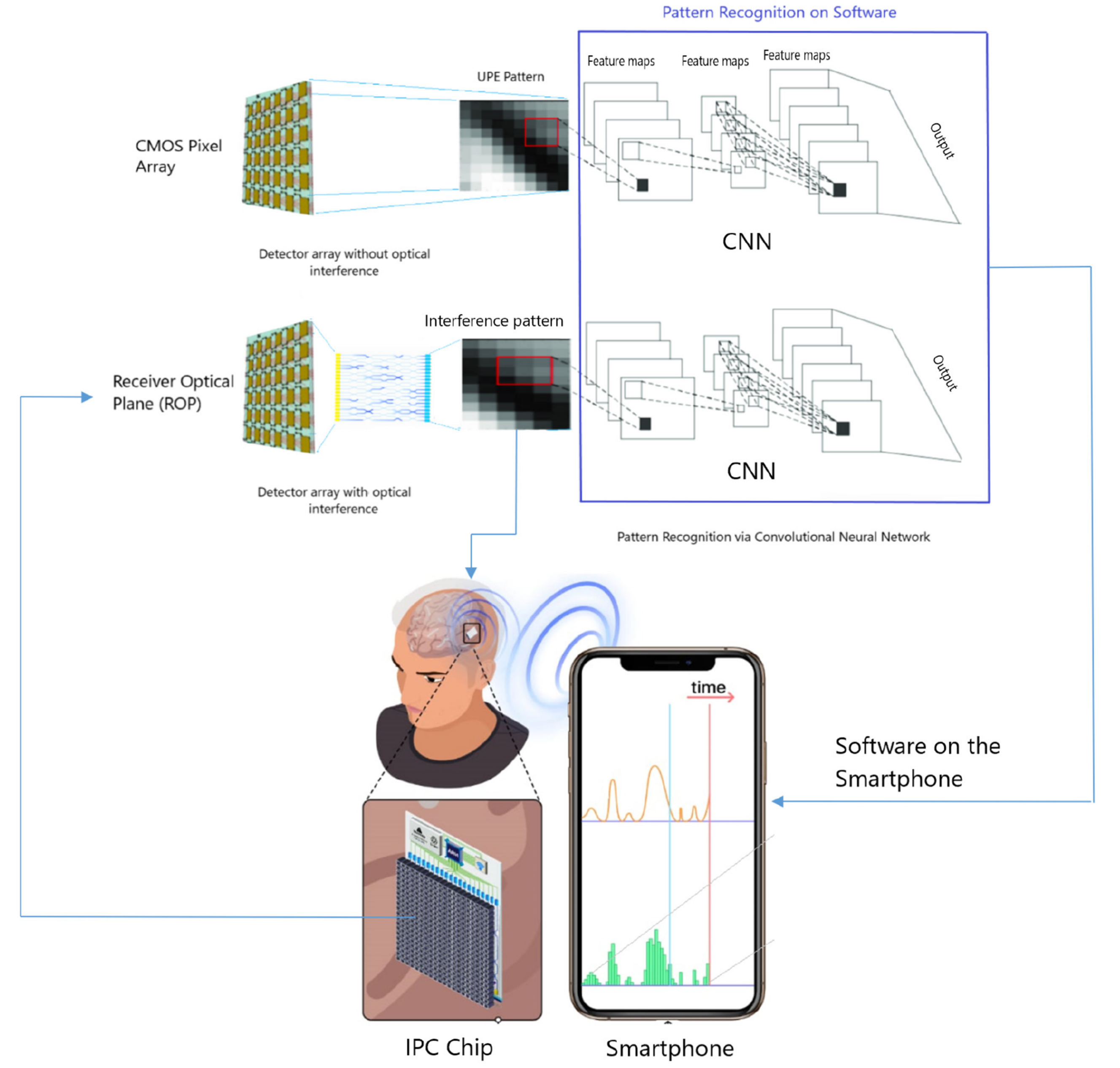

Pattern Recognition on Software
Feature maps
Feature maps
Feature maps
UPE Pattern
CMOS Pixel Array
Output
CNN
Detector array without optical interference
Interference pattern
Receiver Optical Plane (ROP)
Output
CNN
Detector array with optical interference
Pattern Recognition via Convolutional Neural Network
time
Software on the Smartphone
IPC Chip
Smartphone

FIGURE 5 | Feature extraction and pattern recognition of detected ultraweak photon emission (UPE) by a chip composed of complementary metal-oxide-semiconductor (CMOS) array via convolutional neural network (CNN) on a software installed on a computer, machine, or smartphone; **(top)** direct UPE detection without optical interferometer, and **(bottom)** UPE detection after the interferometer. The existence of optical interferometer is to discriminate UPE wavelengths, since interference of similar photons (in wavelength) make a different pattern with non-similar photons. One of the advantages of such an interferometer is to have a simple "spectrometry" over similar wavelengths. However, an ensemble of wavelengths may make different patterns at the same time and obscure the interference patterns which may not make advantage over a direct detection, but one can classify those ensemble patterns with pattern recognition techniques such as PCA, which can find the differences between different patterns in the overlapped patterns, and classify each pattern for the relevant wavelength after many sets of training data. The direct detection of UPE by CMOS array and indirect detection after an optical interferometer both can be used for UPE data acquisition.

The selection of the integration material will then determine the capabilities and limitations for the technology platform, making some of them more appropriate for certain applications than others. This is thus a critical choice and needs to be carefully evaluated.

4.4. Noise and Loss in the PIC

Design of an PIC, testing and packaging from the beginning should be done carefully. The steps are device level (optical, thermal, and material simulations), circuit level (virtual lab to test performance), system level (PIC connected to a CMOS array),

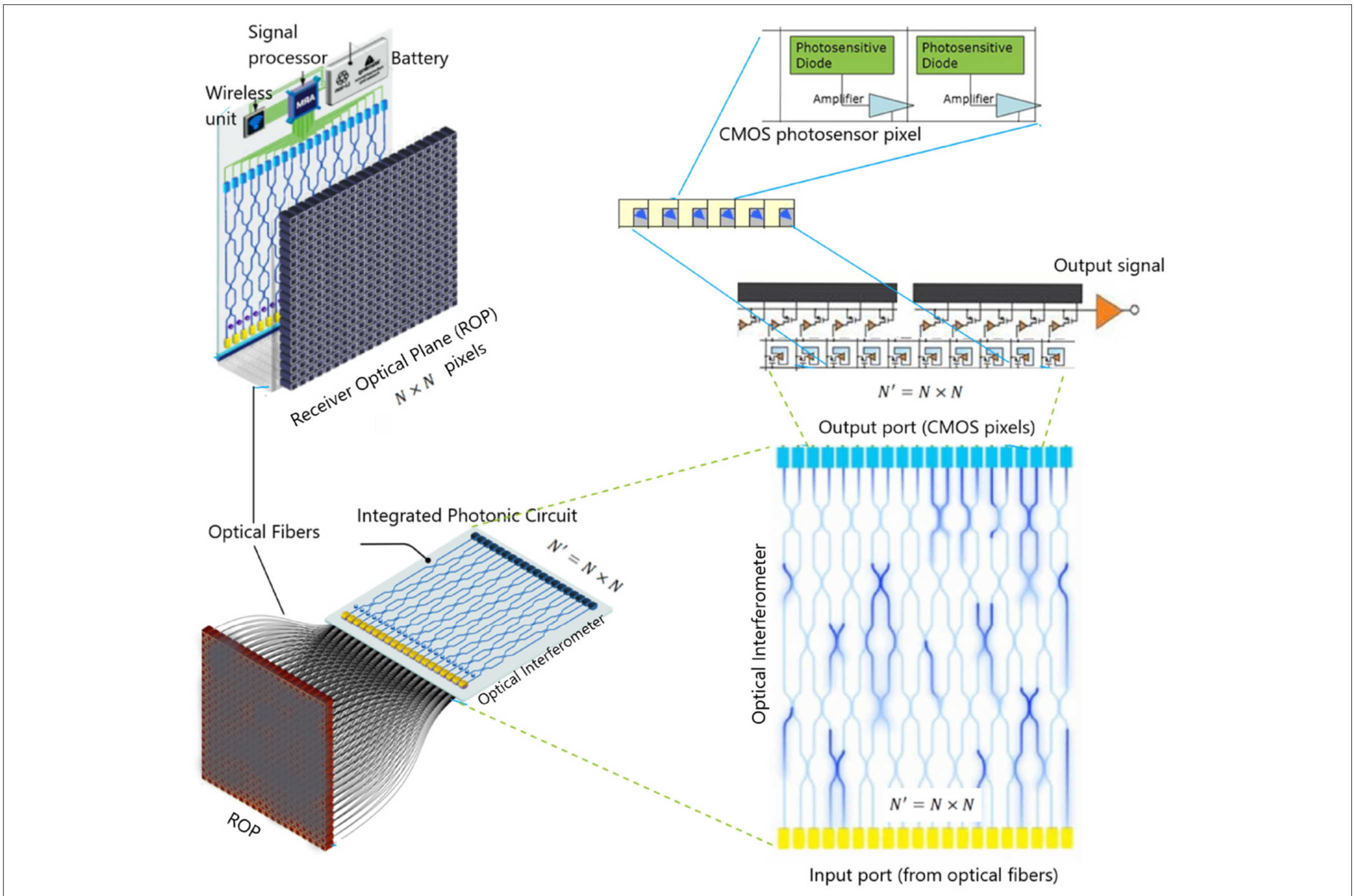


FIGURE 6 | A typical on-chip ultraweak photon emission (UPE) detector can be built from an array of optical fibers connected to an integrated photonic circuit, which has an output gate composed of complementary metal-oxide-semiconductor (CMOS) photosensor array.

layout level (generate the design intent), verification, simulation of each process step, fabrication, and finally packaging. Moreover, a software should be designed to process the detected signals. Here, we would like to estimate the noise magnitude in the optical section of the PIC. The optical section is composed of receiver optical plane (ROP), optical fibers (OF), and optical interferometer (OI).

4.4.1. Noise and Loss in the Receiver Optical Plane

First, we note that blackbody radiation is not a significant source of photons in the visible wavelength range at body temperature. On our BCI, photons are directly coupled to the ROP's fibers that are very close (approximately in contact) with the cortex, thereby leading to a minimal coupling loss. In terms of noise, shot noise [also known as "quantum noise" (Gardiner and Zoller, 2004) or "photon noise"] is the most important contribution in the ROP. It describes the fluctuations of the number of photons received due to their occurrence independent of each other. Optical detection is said to be "photon noise limited" as only the shot noise remains. Just as with other forms of shot noise, the fluctuations in a photo-current due to shot noise scale as the square-root of the average intensity:

$$\mathrm{SN} := \mid \sqrt{(n - \langle n \rangle)^2} \mid$$

4.4.2. Loss in Optical Fibers

The intensity of photons will become lower when traveling through the core of fiber optic. Thus, the signal strength becomes weaker. This loss of light power is generally called fiber optic loss or attenuation. This decrease in power level is described in dB. There are two types of loss in optical fibers known as intrinsic fiber core attenuation (mainly due to light absorption and scattering) and extrinsic fiber attenuation due to bending loss as well as splicing (or coupling) loss between the fibers and chip. Given that the length of the fibers are to be in centimeter scale, the former will be negligible. However, bending and splicing/coupling loss can be significant depending on the process of binding the fibers to the photonic chip. For example, based on subwavelength gratings, it has been shown that it is possible to couple broadband light with very low coupling losses. Guiding of visible light in the wavelength range of 550–650 nm with losses down to 6 dB/cm is feasible using silicon gratings (having absorption of 13,000 dB/cm at this wavelength), which are fabricated with standard silicon photonics technology. This approach allows one to overcome traditional limits of the various established photonics technology platforms with respect to their suitable spectral range (Urbonas et al., 2021).

4.4.3. Noise and Loss in Optical Interferometer

The main elements of an optical interferometer on a photonic chip are couplers and optical modulators, as illustrated in **Figure 7**. There are different types of optical modulators such as MEMS, liquid crystal on silicon (LCOS), electro-optic $LiNbO_3$ waveguide, III-IV semiconductor optical amplifier (SOA), Mach-Zehnder interferometer (MZI), and micro-ring resonator (MRR) (Stefanov et al., 2020). Compared with the above technologies, the silicon photonic modulators based on silicon-on-insulator (SOI) platform attract more attention because of high device density, whose volume is 1/1,000 of silicon dioxide devices, functional integration with active photonic devices and complementary metal oxide semiconductor (CMOS) circuit, and fabrication process compatible with a mature CMOS manufacturing technology. One of the state of art of the silicon photonic modulator engine that is very useful for quantum interference is MZI. A typical 2 × 2 MZI modulator cell consists of two 3 dB coupler and a dual-waveguide arm between them. One of the arms has a phase shifter based on the change of refractive index. Since the silicon has both strong thermo-optic (T-O) effect ($1.86{\times}10^{-4}$ K^{-1}) (Stefanov et al., 2020), the phase shifter can be categorized as T-O switch with a heater and electro-optic (E-O) switch with a p-i-n junction diode. The T-O switch has a response time of microsecond-scale to millisecond-scale, while the E-O switches have a response time of nanosecond-scale.

The loss in on-chip optical interferometers arise from non-unity coupling from fiber to the input ports of the chips as well as attenuation through the waveguides patterned on the chip. As discussed above, the coupling loss can be significantly less than 1 dB through the advanced coupling methods. However, the waveguide propagation loss is given by the chip platform. Depending on the wavelength, this loss can vary substantially, in particular in the wavelength range of 300–700 nm, as shown in **Table 1**.

4.5. Noise and Loss in the CMOS Sensor Array

Noise can be produced by fluctuations in signal that makes uncertainty in detection. Essentially, the signal-to-noise ratio (SNR) is the ratio of pattern signal to the total noise. For larger SNR, it is easier to distinguish pattern from noise, which makes a higher confidence in measurements.

CMOS (Complementary metal–oxide–semiconductor) primary noise sources are shot (photon) noise (i.e., SN), dark noise (i.e., DN), and read noise (i.e., RN). Shot noise is due to physical property of light, regardless of sensor, and it is $\mathrm{SN} = \sqrt{\mathrm{Signal}}$. Dark noise is temperature dependent and higher for global shutter and its magnitude is obtained as $\mathrm{DN} = \sqrt{\mathrm{Dark\ Current}}$. Read noise includes Random Telegraph Noise (RTN), which is non-Gaussian, and depends on multiply column and pixel amplifiers, RN = Read Noise. RTN is the most significant component of CMOS noise. The SNR for CMOS is obtained as follows:

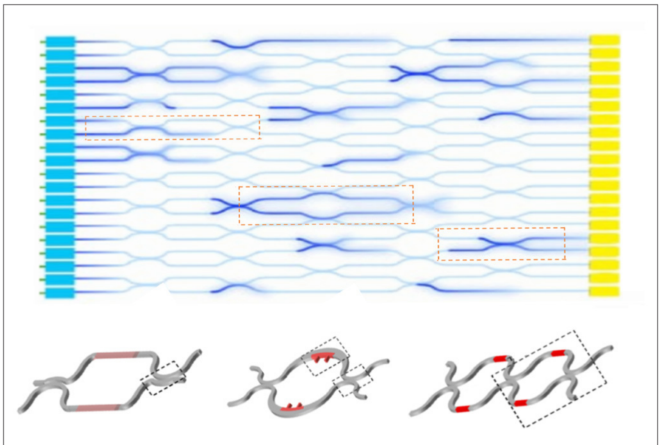


FIGURE 7 | Schematic of various of Mach-Zehnder interferometer (MZI) modulator cells in an optical interferometer. The undesirable attenuation of light in the waveguides and modulators depends on material of the chip platform as well as the dimension and structure of modulators (Stefanov et al., 2020), which determine bending and scattering loss.

$$\mathrm{SNR} = \frac{S}{\sqrt{\mathrm{SN}^2 + \mathrm{DN}^2 + \mathrm{RN}^2}} \tag{3}$$

where S is Signal=Photon flux × time × QE (Dragulinescu, 2012).

Scientific CMOS (sCMOS) sensor is a novel technology with room to grow, which allows for higher speed operation with larger pixel arrays than EMCCD and CCDs with similar noise performance to conventional CCDs.

4.5.1. Quantum Efficiency

QE is defined as

$$\mathrm{QE} = \frac{\text{Converted photons to electrons}}{\text{Total incident photons}}$$

which is a measurement of sensitivity to light. As a ratio, QE is dimensionless, but it is closely related to the responsivity, which is expressed in amps (A) per watt (W). Since the energy of a photon is inversely proportional to its wavelength, QE is often measured over a range of different wavelengths to characterize a detector efficiency at each photon energy level.

The photodetector matrix consist of CMOS-compatible photodiodes (formed between drain diffusion and p-well) with associated readout and sensor selection circuits. The spectral measurements of the photodiode have exhibited a QE better than 60% at 650 nm, and better than 40% between 500 and 850 nm (Dragulinescu, 2012).

A chip design for UPE detection can be inspired by retina implants, but with bigger array size and significantly higher QE. The irradiance on the retina even under a bright daytime illumination does not exceed 1 μW mm^{-2}. At such illumination a 20 μm diameter photodiode (having even 100% QE) can provide only 40 pA of current (Palanker et al., 2005). Basically,

TABLE 1 | Data adapted from APL Photonics 5, 020903 (2020); Optica 6(3), 380-384 (2019); and Optics Express 14 (11), 4826-4834 (2006).

	Loss vs. wavelength for various chip platforms			
Loss (dB/cm)	**300–400 nm**	**400–500 nm**	**500–600 nm**	**600–700 nm**
Aluminum nitride (AlN)	40–50	40–50	30–40	20–30
Alumina (Al_2O_3)	∼ 3	2	1	< 1
Tantalum pentoxide (Ta_2O_3)	N/A	∼ 4	∼ 2	< 1
Silicon-nitride (Si_3N_4)	N/A	5–20	< 1	< 1
Lithium niobate ($LiNbO_3$)	N/A	N/A	N/A	∼ 0.06
Femto-second laser-written waveguides in glass (SiO_2)	N/A	N/A	N/A	∼ 0.2

each photoreceptor cell can produce 1 pA with a single photon absorption (Salari et al., 2016a). To provide stimulating current on the order of 1–2 μA, which would be minimal for physiological stimulation, current amplification by a factor of about 1,000 is required. Suitable current levels would require photodiodes more than 600 μm in diameter, so that ambient light cannot be used to power more than a token number of electrodes on a retinal chip. An additional source of power will be needed for any practical chip (Palanker et al., 2005). The stimulation current for an electrode of 10 μm in diameter is on the order of 1 μA. The photodiode converts photons into electric current with efficiency of up to 0.6 AW^{-1}, thus 1.7 μW of light power will be required for activation of one pixel. If light pulses are applied for 1 ms at 50 Hz, the average power will be reduced to 83 nW/pixel. With 18,000 pixels on the chip, the total light power irradiating an implant will be 1.5 mW (Palanker et al., 2005). In the case of skull-implant PIC chip, the main difference with the retina implant is that the retina implant should activate neurons with the currents produced by external light, which needs a relatively high intensity of light, while for the PIC chip there is no need to activate neurons, and a low light intensity even with a few numbers of photons is sufficient for the CMOS pixels activation to be reported to the software. In silicon, a single-photon with a wavelength between 300 and 1,100 nm can generate only one electron–hole pair. Therefore, for visible and near-infrared light, the task of single-photon detection becomes a task of single-electron (or hole) detection. This is not easy due to the unavoidable readout noise of the sensor, which is usually too high for the reliable detection of a single electron. Another difficulty for room temperature applications are the thermal dark currents, because they are indistinguishable from photogenerated signals.

4.5.2. Chip Battery and Wireless Sectors

In order to have a dynamic chip for monitoring signals of the brain continuously, the chip requires a long lifetime battery. The size and lifetime of the battery is one of the major challenges in design of an implant chip for biomedical applications. As an alternative, replacing the battery with a miniaturized and integrated wireless power harvester aid the design of sustainable biomedical implants in smaller volumes (Masius and Wong, 2020). Currently, implanted batteries provide the energy for implantable biomedical devices. However, batteries have fixed energy density, limited lifetime, chemical side effects, and large size. Thus, researchers have developed several methods to harvest energy for implantable devices. Devices powered by harvested energy have longer lifetime and provide more comfort and safety than conventional devices. A solution to energy problems in wireless sensors is to scavenge energy from the ambient environment. Energies that may be scavenged include infrared radiant energy, wireless transfer energy, and RF radiation energy (inductive and capacitive coupling) (Hannan et al., 2014). Recently, a chip has been developed that is powered wirelessly and can be surgically implanted to read neural signals and stimulate the brain with both light and electrical current. The technology has been demonstrated successfully in rats and is designed for use as a research tool. The chip is capable of 16-ch neural recording, 8-ch electrical stimulation, and 16-ch optical stimulation, all integrated on a 5 × 3 mm^2 chip fabricated in 0.35-μm standard CMOS process. The trimodal SoC is designed to be inductively powered and communicated (Jia et al., 2020).

4.6. Biocompatibility of the Chip

Brain implants may induce side effects; for instance they may interact acutely and chronically with the brain tissue possibly causing blood–brain barrier (BBB) breach, vascular damage, micromotions, diffusion, etc. (Prodanov and Delbeke, 2016). The advantage of our suggested photonic chip is that it is minimally invasive compared to invasive implants (e.g., ECoG) since it does not need to penetrate the brain tissue.

Some of the key fundamental questions associated to brain implants are related to how long an implant can record useful neuronal signals and what degree of acquisition and decoding reliably can be achieved if the tissue is affected by chip implant. Functional neural tissue survival, distance from the chip contact to target and long-term stability are essential parameters to be considered (Prodanov and Delbeke, 2016).

In the case of photonic chip, it should be installed on the inner surface of the skull and not to be implanted directly in the brain tissue. However, there is still the possibility of a close contact with the brain meninges (i.e., layered membranes that protect the brain and spine) due to the mechanical or volume changes of the brain. In this case, it has been shown that Silicone causes the least amount of inflammation relative to other materials tested at all sacrifice points, which makes it the leading standard neurosurgical implant material and an appropriate control for studies of brain biocompatibility (Mofid et al., 1997). Thus, we envisage to adopt silicone chips but we also expect that research in biocompatibility will provide alternative and advanced materials. However, since the photonic chip can be implanted in between the meninges and the skull, there can be concerns about the limitation of brain UPE detection due to the existence of meninges. The meninges layers of the human brain are composed of three main layers: dura, arachnoid, and falx. The key question is whether light can pass through these layers and if it does, then what are the scattering and absorption effects of photons? For instance, to have a reasonable data acquisition should the dura be open? The optical properties of the human brain and its meninges have been investigated decades ago. It has been shown that meninges is approximately transparent for the near-IR range, but almost half of emissions will not pass through it in the visible range, and less than 40% of emissions can pass through the meninges in the UV range (200–400 nm) (Eggert and Blazek, 1987). As a result, based on the high efficiency of the photonic chip in the near-IR range, the existence of meninges reduces the intensity of UPE but it does not lead to a significant limitation.

Additionally, because of the aqueous and biochemically aggressive nature of the body, the lifetime of brain implants strongly depends on packaging. There are different methods for packaging, which may be especially important for the case of traditional electric chips with wireless neuromodulatory implants with increasing electrode count to have an *in vivo* lifetime comparable to a sizable fraction of a healthy patient's lifetime (>10–20 years) (Shen and Maharbiz, 2021). For our suggested photonic chip, the situation is considerably better because the chip does not have electrodes in the wet biological tissue nor contact with that, and the environment between the meninges and skull is not aqueous, and therefore the probability of water leakage in the photonic chip is minimal. If there will be an injury in the meninges layers due to some impact or accident, then the aqueous leakage may occur, where the photonic chip should be investigated for packaging based on the materials used.

5. DISCUSSION AND CONCLUSION

We propose a radically novel BCI that is based on UPE from the brain. We describe its feasibility of fabrication based on integrated photonic circuits that be readily implemented in a lab. The envisaged BCI chip can be implanted on the interior surface of the skull to monitor in real-time UPE signals emanating from the cortex surface. The proposed chip is not only useful for BCI technology but also it can be used as a photonic sensor for imaging, spectroscopy, and sensitive measurements at low light levels in several applications from biological UPE to quantum optical processing (Salari et al., 2021). Although our proposed technology is, admittedly, at the level of conjecture, requiring comprehensive tests and investigations for verification, we still envision complementary features as well as certain advantages over established technologies, including ECoG. The inherent advantage of our proposed technology is that it is minimally invasive when compared to ECoG. Furthermore, there are certain side effects that may affect the quality of data acquisition over time in ECoG, whereas we expect a relatively stable long-term data acquisition in our proposed approach. In addition, if our suggested photonic chip-technology reaches a satisfactory detection performance based on our estimations, we anticipate that it can feature some other advantages. For example, it may provide additional information about brain functioning, such as an approximately real-time imaging (in slightly longer timescales, e.g., each 15, 20, 30, 60 min, or so) and open the door to studying metabolism variations, variation of ROS production, delayed luminescence but also undertake novel and complementary studies on object visualization studies, sleep studies, and neurodegenerative diseases (Breakspear et al., 2006; Fülöp et al., 2021). Indeed, the emphasis of our conjectural paper is to develop a novel technology and methods that could provide complementary information to improve our understanding of brain activity with potential applications for BCI technologies.

Now, we would like to discuss the advantages and limitations of our proposed technology vs. the current BCI methods. On-chip PICs offer advantages such as miniaturization, higher speed, low thermal effects, large integration capacity, and compatibility with existing processing flows that allow for high yield, volume manufacturing, and lower prices. In the case of UPE detection, there is no need for on-chip single-photon sources, which is one of the most difficult challenges in PICs for quantum computation and communication. In the suggested chip, single photons are produced naturally by metabolism in neurons and therefore a lower power with battery is needed for energy consumption on an implant PIC. Loss is low in NIR range (e.g., 2×10^{-6} dB/cm). In addition, photons are bosons, which do not interact and crosstalk is minimal. A PIC for optical interferometery is efficient for the wavelengths typically in the near infrared range, 800–1,650 nm. This makes a limitation for detection of UPE photons that are in the visible range and the overlapped part to NIR, 400–800 nm. For example, loss is high for the visible range (e.g., 0.6 dB/cm at 600 nm).

Moreover, it may look that the single-photon detections on a CMOS array have a low QE besides the dark current in room temperature, which may lose considerable amounts of UPE. Another concern may be that the output of CMOS is electrons, which are charged particles and fermions, and therefore electronic crosstalk is inherent. In fact, the CMOS QE is about 75%, which is about three times higher than the photo-multiplier tubes (PMTs) with QE about 25%. The SNR of a PMT at room temperature to detect UPE photons is about 1–2, thus a cooling system is required to cool down the PMT sensor to enhance the SNR to reach 3 and higher. Obviously, there is no cooling system on an PIC chip, but in this case, the QE of the CMOS sensor can compensate the lack of a cooling system. For a simple estimation, assuming a 1 × 1 cm chip and considering the length of each CMOS pixel is 4 μm, it is possible to have 2,500 CMOS pixels as the output port on the chip, including 50×50 pixels on the ROP. According the estimations in the main text, the amount of total photon loss from the receiver optical plane (ROP) to the output of the optical interferometer (OI) is about 50%, and the QE of CMOS at the output of the OI is estimated to be 25% in body temperature under the implant conditions on the skull to have a final SNR from 1 to 2. Consequently, it is estimated that only 10% of incident photons can be safely recognized in the output and reported wirelessly to the software on a computer or smartphone. Considering 10–1,000 incident photons per second received in the ROP under a cognitive task (e.g., an object visualization), there can be 1–100 photons per second efficiently detected in the output port, which are enough to have a relatively successful implant PIC chip for an acceptable pattern for UPE processing, where the size of the machine learning program is $N \times N$ sparse matrix, which is not a difficult task for a chip size number of pixels. To conclude, in this paper, we advance major conjectures regarding the relevance of UPE patterns and decision making as well as the feature extractions from UPE signals, which need to be experimentally verified. However, despite some probable limitations in chip fabrication and efficiency, it may be used for wireless BCI signal acquisition with several advantages vs. traditional counterparts such as speed, size, minimally invasive, cheap, scalability, etc. This can be a potential step forward for real-time brain imaging and biological information processing.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

All authors contributed to the idea development and writing the manuscript.

FUNDING

VS and SR are grateful for the financial support by the Spanish State Research Agency through BCAM Severo Ochoa excellence accreditation SEV-2017-0718 and BERC 2018-2021 program. SR was also grateful for project RTI2018-093860B-C21 funded by (AEI/FEDER, UE) with acronym MathNEURO. CS acknowledges NSERC Discovery Grant RGPIN-2020-03945.

ACKNOWLEDGMENTS

VS and SR are very thankful for several helpful discussions with Jean-Bernard Bru at BCAM. VS also thanks M. Aslani for helping to find setup and illustration.

REFERENCES

Adamo, A. M., Llesuy, S. F., Pasquini, J. M., and Boveris, A. (1989). Brain chemiluminescence and oxidative stress in hyperthyroid rats. *Biochem. J.* 263, 273–277. doi: 10.1042/bj2630273

Agte, S., Junek, S., Matthias, S., Ulbricht, E., Erdmann, I., Wurm, A., et al. (2011). Muller glial cell-provided cellular light guidance through the vital Guinea-pig retina. *Biophys. J.* 101, 2611–2619. doi: 10.1016/j.bpj.2011.09.062

Albrecht-Buehler, G. (1995). Changes of cell behavior by near-infrared signals. *Cell Motil. Cytoskeleton*. 32, 299–304. doi: 10.1002/cm.970320406

Antonov, I. P., Goroshkov, A. V., Kalyunov, V. N., Markhvida, I. V., Rubanov, A. S., and Tanin, L. V. (1983). Measurement of the radial distribution of the refractive index of the Schwann's sheath and the axon of a myelinated nerve fiber *in vivo. J. Appl. Spectrosc*. 39, 822–824. doi: 10.1007/BF00662830

Arakawa, Y., Nakamura, T., Urino, Y., and Fujita, T. (2013). Silicon photonics for next generation system integration platform. *IEEE Commun. Mag.* 51, 72–77. doi: 10.1109/MCOM.2013.6476868

Artem'ev, V. V., Goldobin, A. S., and Gus'kov, L. N. (1967). Recording of light emission from a nerve. *Biofzika* 12, 1111–1113.

Boes, A., Corcoran, B., Chang, L., Bowers, J. E., and Mitchell, A. (2018). Status and potential of lithium niobate on insulator (LNOI) for photonic integrated circuits. *Laser Photon. Rev*. 12:1700256. doi: 10.1002/lpor.2017 00256

Bokkon, I., Salari, V., Tuszynski, J. A., and Antal, I. (2010). Estimation of the number of biophotons involved in the visual perception of a single-object image: biophoton intensity can be considerably higher inside cells than outside. *J. Photochem. Photobiol. B* 100, 160–166. doi: 10.1016/j.jphotobiol.2010. 06.001

Breakspear, M., Roberts, J. A., Terry, J. R., Rodrigues, S., Mahant, N., and Robinson, P. A. (2006). A unifying explanation of primary generalized seizures through nonlinear brain modeling and bifurcation analysis. *Cereb. Cortex* 16, 1296–1313. doi: 10.1093/cercor/bhj072

Carolan, J., Harrold, C., Sparrow, C., Martín-López, E., Russell, N. J., Silverstone, J. W., et al. (2015). Universal linear optics. *Science* 349, 711–716. doi: 10.1126/science.aab3642

Cheng, L., Mao, S., Li, Z., Han, Y., and Fu, H. Y. (2020). Grating couplers on silicon photonics: design principles, emerging trends and practical issues. *Micromachines* 11:666. doi: 10.3390/mi11070666

Chrostowski, L., and Hochberg, M. (2015). *Silicon Photonics Design*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781316084168

Chwirot, B. W. (1992). "Ultraweak luminescence studies of microsporogenesis in Larch," in *Recent Advances in Biophoton Research and Its Applications*, eds F. A. Popp, K. H. Li, and Q. Gu (Singapore: World Scientific Publishing Company), 259–285. doi: 10.1142/9789814439671_0010

Cifra, M., Brouder, C., Nerudova, M., and Kucera, O. (2015). Biophotons, coherence and photocount statistics: a critical review. *J. Luminesc*. 164, 38–51. doi: 10.1016/j.jlumin.2015.03.020

Cifra, M., and Pospisil, P. (2014). Ultra-weak photon emission from biological samples: defnition, mechanisms, properties, detection and applications. *J. Photochem. Photobiol. B. Biol*. 139, 2–10. doi: 10.1016/j.jphotobiol.2014.02.009

Coldren, L., Corzine, S., and Mashanovitch, M. (2012). *Diode Lasers and Photonic Integrated Circuits, 2nd Edn*. Hoboken, NJ: John Wiley and Sons. doi: 10.1002/9781118148167

Davis, K. M., Miura, K., Sugimoto, N., and Hirao, K. (1996). Writing waveguides in glass with a femtosecond laser. *Opt. Lett*. 21, 1729–1731. doi: 10.1364/OL.21.001729

Demetrius, L. (2003). Quantum statistics and alometric scaling of organisms. *Phys. A* 322, 477–490. doi: 10.1016/S0378-4371(03)00013-X

Donchin, E., Spencer, K. M., and Wijesinghe, R. (2000). The mental prosthesis: assessing the speed of a P300-based brain-computer interface. *IEEE Trans. Rehabil. Eng*. 8, 174–179. doi: 10.1109/86.847808

Dragulinescu, A. (2012). "Comparison of various structures of CMOS photodiodes in terms of dark current, photocurrent, and quantum efficiency," in *Proc. SPIE 8411, Advanced Topics in Optoelectronics, Microelectronics, and Nanotechnologies VI* (Bellingham, WA). doi: 10.1117/12.966388

Eggert, H., and Blazek, V. (1987). Optical properties of human brain tissue, meninges, and brain tumors in the spectral range of 200 to 900 nm. *Neurosurgery* 21, 459–464. doi: 10.1227/00006123-198710000-00003

Esmaeilpour, T., Fereydouni, E., Dehghani, F., Bókkon, I., Panjehshahin, M. R., Császár-Nagy, N., et al. (2020). An experimental investigation of ultraweak photon emission from adult murine neural stem cells. *Sci. Rep*. 10, 463. doi: 10.1038/s41598-019-57352-4

Fang, Z., and Zhao, C. Z. (2012). Recent progress in silicon photonics: a review. *ISRN Optics* 2012:428690. doi: 10.5402/2012/428690

Franze, K., Grosche, J., Skatchkov, S. N., Schinkinger, S., Foja, C., Schild, D., et al. (2007). Muller cells are living optical fibers in the vertebrate retina. *Proc. Natl. Acad. Sci. U.S.A*. 104, 8287–8292. doi: 10.1073/pnas.0611180104

Fukushima, K. (1980). Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybernet*. 36, 193–202. doi: 10.1007/BF00344251

Fulop, T., Tripathi, S., Rodrigues, S., Desroches, M., Bunt, T., Eiser, A., et al. (2021). Targeting impaired antimicrobial immunity in the brain for the treatment of Alzheimer's Disease. *Neuropsychiatr. Dis. Treat*. 17, 1311–1339. doi: 10.2147/NDT.S264910

Gardiner, C. W., and Zoller, P. (2004). *Quantum Noise*. New York, NY: Springer-Verlag.

Hämäläinen, M., Hari, R., Ilmoniemi, R. J., Knuutila, J., and Lounasmaa O. V. (1993). Magnetoencephalography theory, instrumentation, and applications to noninvasive studies of the working human brain. *Rev. Mod. Phys*. 65, 413–497. doi: 10.1103/RevModPhys.65.413

Hannan, M. A., Mutashar, S., Samad, S. A., and Hussain, A. (2014). Energy harvesting for the implantable biomedical devices: issues and challenges. *Biomed. Eng*. 13:79. doi: 10.1186/1475-925X-13-79

Imaizumi, S., Kayama, T., and Suzuki, J. (1984). Chemiluminescence in hypoxic brain? The first report. Correlation between energy metabolism and free radical reaction. *Stroke* 15, 1061–1065. doi: 10.1161/01.STR.15.6.1061

Isojima, Y., Isoshima, T., Nagai, K., Kikuchi, K., and Nakagawa, H. (1995). Ultraweak biochemiluminescence detected from rat hippocampal slices. *Neuroreport* 6, 658–660. doi: 10.1097/00001756-199503000-00018

Jia, Y., Guler, U., Lai, Y.P., Gong, Y., Weber, A., Li, W., et al. (2020). A trimodal wireless implantable neural interface system-on-chip. *IEEE Trans. Biomed. Circ. Syst*. 14, 1207–1217. doi: 10.1109/TBCAS.2020.3037452

Jolliffe, I. T. (2002). *Principal Component Analysis*. New York, NY: Springer-Verlag.

Kataoka, Y., Cui, Y., Yamagata, A., Niigaki, M., Hirohata, T., Oishi, N., et al. (2001). Activity-dependent neural tissue oxidation emits intrinsic ultraweak photons. *Biochem. Biophys. Res. Commun*. 285, 1007–1011. doi: 10.1006/bbrc.2001.5285

Kobayashi, M., Takeda, M., Ito, K., Kato, H., and Inaba, H. (1999a). Two-dimensional photon counting imaging and spatiotemporal characterization of ultraweak photon emission from a rat's brain *in vivo*. *J. Neurosci. Methods* 93, 163–168. doi: 10.1016/S0165-0270(99)00140-5

Kobayashi, M., Takeda, M., Sato, T., Yamazaki, Y., Kaneko, K., Ito, K., et al. (1999b). In vivo imaging of spontaneous ultraweak photon emission from a rat's brain correlated with cerebral energy metabolism and oxidative stress. *Neurosci. Res*. 34, 103–113. doi: 10.1016/S0168-0102(99)00040-1

Kumar, S., Boone, K., Tuszynski, J., Barclay, P., and Simon, C. (2016). Possible existence of optical communication channels in the brain. *Sci Rep*. 7;6:36508. doi: 10.1038/srep36508

Levy, I., Hasson, U., and Malach, R. (2004). One picture is worth at least a million neurons. *Curr. Biol*. 14, 996–1001. doi: 10.1016/j.cub.2004.05.045

Liang, D., and Bowers, J. E. (2009). Photonic integration: Si or InP substrates? *Electron. Lett*. 45, 578–581. doi: 10.1049/el.2009.1279

Marshall, G. D., Politi, A., Matthews, J. C. F., Dekker, P., Ams, M., Withford, M. J., et al. (2009). (2009). Laser written waveguide photonic quantum circuits. *Opt. Express* 17, 12546–12554 doi: 10.1364/OE.17.012546

Marsili, F., Verma, V. B., Stern, J. A., Harrington, S., Lita, A. E., Gerrits, T., et al. (2013). Detecting single infrared photons with 93% system efficiency. *Nat. Photon* 7, 210–214. doi: 10.1038/nphoton.2013.13

Masius, A. A., and Wong, Y. C. (2020). On-chip miniaturized antenna in CMOS technology for biomedical implant. *Int. J. Electron. Commun*. 115:153025. doi: 10.1016/j.aeue.2019.153025

Mofid, M. M., Thompson, R. C., Pardo, C. A., Manson, P. N., and Vander Kolk, C. A. (1997). Biocompatibility of fixation materials in the brain. *Plast. Reconstr. Surg*. 100,14–20. doi: 10.1097/00006534-199707000-00003

Muñoz, P., Micó, G., Bru, L. A., Pastor, D., Perez, D., Domenech, J. D., Fernandez, J., et al. (2017). Silicon nitride photonic integration platforms for visible, near-infrared and mid-infrared applications. *Sensors* 17:2088. doi: 10.3390/s17092088

Musumeci, F., Privitera, G., Scordino, A., Tudisco, S., and Lo Presti, C. (2005). Discrimination between normal and cancer cells by using analysis of delayed luminescence. *Appl. Phys. Lett*. 86, 153902–153901. doi: 10.1063/1.1900317

Naseer, N., and Hong, K. S. (2015). fNIRS-based brain-computer interfaces: a review. *Front. Hum. Neurosci*. 9:3. doi: 10.3389/fnhum.2015.00003

Niedermeyer, E., and da Silva, F. L. (2004). *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Philadelphia, PA: Lippincott Williams & Wilkins.

Niggli, H. J., Tudisco, S., Lanzanò, L., Applegate, L. A., Scordino, A., and Musumeci, F. (2008). Laser-ultraviolet-A induced ultra weak photon emission in human skin cells: a biophotonic comparison between keratinocytes and fbroblasts. *Indian J. Exp. Biol*. 46, 358–363. Available online at: http://nopr.niscair.res.in/bitstream/123456789/4472/1/IJEB%2046%285%29%20358-363.pdf

Niggli, H. J., Tudisco, S., Privitera, G., Applegate, L. A., Scordino, A., and Musumeci, F. (2005). Laser-ultraviolet-A-induced ultraweak photon emission in mammalian cells. *J. Biomed. Opt*. 10:024006. doi: 10.1117/1.1899185

Nolte, S., Will, M., Burghoff, J., and Tuennermann, A. (2003). Femtosecond waveguide writing: a new avenue to three-dimensional integrated optics. *Appl. Phys. A* 77, 109–111. doi: 10.1007/s00339-003-2088-6

Pakkenberg, B., and Gundersen, H. J. (1987). Neocortical neuron number in humans: effect of sex and age. *J. Comp. Neurol*. 384, 312–320. doi: 10.1002/(SICI)1096-9861(19970728)384:2<312::AID-CNE10>3.0.CO;2-K

Palanker, D., Vankov, A., Huie, P., and Baccus, S. (2005). Design of a high-resolution optoelectronic retinal prosthesis. *J. Neural Eng*. 2, S105–S120. doi: 10.1117/12.590964

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clin. Neurophysiol*. 118, 2128–2148. doi: 10.1016/j.clinph.2007.04.019

Popp, F. A., Li, K. H., Mei, W. P., Galle, M., and Neurohr, R. (1988). Physical aspects of biophotons. *Experientia* 44, 576–585. doi: 10.1007/BF01953305

Popp, F. A., Nagl, W., Li, K. H., Scholz, W., Weingärtner, O., Wolf, R. (1984). Biophoton emission. New evidence for coherence and DNA as source. *Cell Biophys*. 6, 33–51. doi: 10.1007/BF02788579

Pospisil, P., Prasad, A., and Rac, M. (2014). Role of reactive oxygen species in ultra-weak photon emission in biological systems. *J. Photochem. Photobiol. B Biol*. 139, 11–23. doi: 10.1016/j.jphotobiol.2014.02.008

Prodanov, D., and Delbeke, J. (2016). Mechanical and biological interactions of implants with the brain and their impact on implant design. *Front. Neurosci*. 10:11. doi: 10.3389/fnins.2016.00011

Reichenbach, A., and Bringmann, A. (2013). New functions of Muller cells. *Glia* 61, 651–678. doi: 10.1002/glia.22477

Romero-Garcia, S., Merget, F., Zhong, F., Finkelstein, H., and Witzens, J. (2013). Silicon nitride CMOS-compatible platform for integrated photonics applications at visible wavelengths. *Opt. Express* 21, 14036–14046. doi: 10.1364/OE.21.014036

Salari, V., Bokkon, I., Ghobadi, R., Scholkmann, F., and Tuszynski, J. A. (2016b). Relationship between intelligence and spectral characteristics of brain biophoton emission: correlation does not automatically imply causation. *Proc. Natl. Acad. Sci. U.S.A*. 113, E5540–E5541. doi: 10.1073/pnas.1612646113

Salari, V., and Brouder, C. (2011). Comment on Delayed luminescence of biological systems in terms of coherent states. *Phys. Lett. A* 375, 2531–2532. doi: 10.1016/j.physleta.2011.05.017

Salari, V., Paneru, D., Saglamyurek, E., and Ghadimi, M. (2021). Quantum face recognition protocol with ghost imaging. *arXiv preprint arXiv:2110.10088*.

Salari, V., Scholkmann, F., Bokkon, I., Shahbazi, F., and Tuszynski, J. (2016a). The physical mechanism for retinal discrete dark noise: thermal activation or cellular ultraweak photon emission? *PLoS ONE* 11:e0148336. doi: 10.1371/journal.pone.0148336

Salari, V., Scholkmann, F., Vimal, L. P., Császár, N., Aslani, M., and Bókkon, I. (2017). Phosphenes, retinal discrete dark noise, negative aferimages and retinogeniculate projections: a new explanatory framework based on endogenous ocular luminescence. *Prog. Ret. Eye Res*. 60, 101–119. doi: 10.1016/j.preteyeres.2017.07.001

Salari, V., Valian, H., Bassereh, H., Bokkon, I., and Barkhordari, A. (2015). Ultraweak photon emission in the brain. *J. Integ. Neurosci*. 14, 419–429. doi: 10.1142/S0219635215300012

Sander, T. H., Preusser, J., Mhaskar, R., Kitching, J., Trahms, L., and Knappe, S. (2012). Magnetoencephalography with a chip-scale atomic magnetometer. *Biomed. Opt. Express* 3, 981–990. doi: 10.1364/BOE.3.000981

Scholkmann, F. (2015). Two emerging topics regarding long-range physical signaling in neurosystems: membrane nanotubes and electromagnetic fields. *J. Integr. Neurosci*. 14, 135–153. doi: 10.1142/S0219635215300115

Scholkmann, F. (2016). Long range physical cell-to-cell signalling via mitochondria inside membrane nanotubes: a hypothesis. *Theor. Biol. Med. Model*. 13:1. doi: 10.1186/s12976-016-0042-5

Scholkmann, F., Kleiser, S., Metz, A. J., Zimmermann, R., Pavia, J. M., Wolf, U., and Wolf, M. (2014). A review on continuous wave functional near-infrared spectroscopy and imaging instrumentation and methodology. *Neuroimage* 85, 6–27. doi: 10.1016/j.neuroimage.2013.05.004

Scordino, A., Baran, I., Gulino, M., Ganea, C., Grasso, R., Niggli, J. H., and Musumeci, F. (2014). Ultra-weak delayed luminescence in cancer research: a review of the results by the ARETUSA equipment. *J. Photochem. Photobiol. B* 5, 76–84. doi: 10.1016/j.jphotobiol.2014.03.027

Shen, K., and Maharbiz, M. M. (2021). Ceramic packaging in neural implants. *J. Neural Eng*. 18:025002. doi: 10.1088/1741-2552/abd683

Shenoy, P., Miller, K. J., Ojemann, J. G., and Rao, R. P. N. (2007). Generalized features for electrocorticographic BCIs. *IEEE Trans. Biomed. Eng*. 55, 273–280. doi: 10.1109/TBME.2007.903528

Simon, C. (2019). Can quantum physics help solve the hard problem of consciousness? *J. Conscious. Stud*. 26, 204–218.

Simons, M., and Trajkovic, K. (2006). Neuron-glia communication in the control of oligodendrocyte function and myelin biogenesis. *J. Cell Sci*. 119(Pt 21), 4381–4389. doi: 10.1242/jcs.03242

Slawinski, J. (1988). Luminescence research and its relation to ultraweak cell radiation. *Experientia* 44, 559–571. doi: 10.1007/BF01953303

Spring, J. B., Metcalf, B. J., Humphreys, P. C., Kolthammer, W. S., Jin, X. M., Barbieri, M., et al. (2012). Boson sampling on a photonic chip. *Science* 339, 798–801. doi: 10.1126/science.1231692

Stefanov, K. D., Prest, M. J., Downing, M., George, E., Bezawada, N., and Holland, A. D. (2020). Simulations and design of a single-photon CMOS imaging pixel using multiple non-destructive signal sampling. *Sensors* 20, 2031. doi: 10.3390/s20072031

Sun, Y., Wang, C., and Dai, J. (2010). Biophotons as neural communication signals demonstrated by in situ biophoton autography. *Photochem. Photobiol. Sci*. 9, 315–322. doi: 10.1039/b9pp00125e

Suzuki, J., Imaizumi, S., Kayama, T., and Yoshimoto, T. (1985). Chemiluminescence in hypoxic brain? The second report: cerebral protective effect of mannitol, vitamin E and glucocorticoid. *Stroke* 16, 695–700. doi: 10.1161/01.STR.16.4.695

Szameit, A., Dreisow, F., Pertsch, T., and Nolte, S. (2007). Tunnermann, Andreas Control of directional evanescent coupling in fs laser written waveguides. *Opt. Express* 15, 1579–1587. doi: 10.1364/OE.15.001579

Tang, R., and Dai, J. (2013). Biophoton signal transmission and processing in the brain. *J. Photochem. Photobiol. B. Biol*. 139, 71–75. doi: 10.1016/j.jphotobiol.2013.12.008

Tang, R., and Dai, J. (2014). Spatiotemporal imaging of glutamate-induced biophotonic activities and transmission in neural circuits. *PLoS ONE* 9:e85643. doi: 10.1371/journal.pone.0085643

Tokuda, T., Haruta, M., Sasagawa, K., and Ohta, J. (2021). CMOS-based neural interface device for optogenetics. *Adv. Exp. Med. Biol*. 1293, 585–600. doi: 10.1007/978-981-15-8763-4_41

Urbonas, D., Mahrt, R. F., and Stöferle, T. (2021). Low-loss optical waveguides made with a high-loss material. *Light Sci. Appl*. 10:15. doi: 10.1038/s41377-020-00454-w

Wang, C., Bokkon, I., Dai, J., and Antal, I. (2011). Spontaneous and visible light-induced ultraweak photon emission from rat eyes. *Brain Res*. 1369, 1–9. doi: 10.1016/j.brainres.2010.10.077

Wang, Z., Wang, N., Li, Z., Xiao, F., and Dai, J. (2016). Human high intelligence is involved in spectral redshif of biophotonic activities in the brain. *Proc. Natl. Acad. Sci. U.S.A*. 113, 8753–8758. doi: 10.1073/pnas.1604855113

Washburna, A. L., and Bailey, R. C. (2011). Photonics-on-a-chip: recent advances in integrated waveguides as enabling detection elements for real-world, lab-on-a-chip biosensing applications. *Analyst* 136, 227–236. doi: 10.1039/C0AN00449A

Wollman, E. E., Verma, V. B., Lita, A. E., Farr, W. H., Shaw, M. D., Mirin, R. P., et al. (2019). Kilopixel array of superconducting nanowire single-photon detectors. *Opt. Express* 27, 35279–35289. doi: 10.1364/OE.27.035279

Zarkeshian, P., Kumar, S., Tuszynski, J., Barclay, P., and Simon, C. (2018). Are there optical communication channels in the brain? *Front. Biosci*. 23, 1407–1421. doi: 10.2741/4652

Zhang, J., Yu, W., Sun, T., and Popp, F. A. (1997). Spontaneous and light-induced photon emission from intact brains of chick embryos. *Sci. China C Life Sci*. 40, 43–51. doi: 10.1007/BF02879106

Zhu, D., Shao, L., Yu, M., Cheng, R., Desiatov, B., Xin, C., et al. (2021). Integrated photonics on thin-film lithium niobate. *Adv. Opt. Photon*. 13, 242–352. doi: 10.1364/AOP.411024

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 02 March 2022
doi: 10.3389/fnins.2022.825746

When the Whole Is Less Than the Sum of Its Parts: Maximum Object Category Information and Behavioral Prediction in Multiscale Activation Patterns

Hamid Karimi-Rouzbahani[1,2,3]* *and Alexandra Woolgar*[1,2]

[1] *Medical Research Council Cognition and Brain Sciences Unit, University of Cambridge, Cambridge, United Kingdom,* [2] *Department of Cognitive Science, Perception in Action Research Centre, Macquarie University, Sydney, NSW, Australia,* [3] *Department of Computing, Macquarie University, Sydney, NSW, Australia*

Neural codes are reflected in complex neural activation patterns. Conventional electroencephalography (EEG) decoding analyses summarize activations by averaging/down-sampling signals within the analysis window. This diminishes informative fine-grained patterns. While previous studies have proposed distinct statistical features capable of capturing variability-dependent neural codes, it has been suggested that the brain could use a combination of encoding protocols not reflected in any one mathematical feature alone. To check, we combined 30 features using state-of-the-art supervised and unsupervised feature selection procedures (n = 17). Across three datasets, we compared decoding of visual object category between these 17 sets of combined features, and between combined and individual features. Object category could be robustly decoded using the combined features from all of the 17 algorithms. However, the combination of features, which were equalized in dimension to the individual features, were outperformed across most of the time points by the multiscale feature of Wavelet coefficients. Moreover, the Wavelet coefficients also explained the behavioral performance more accurately than the combined features. These results suggest that a single but multiscale encoding protocol may capture the EEG neural codes better than any combination of protocols. Our findings put new constraints on the models of neural information encoding in EEG.

Keywords: neural encoding, multivariate pattern decoding, EEG, feature extraction, feature selection

OPEN ACCESS

Edited by:
Reza Lashgari,
Shahid Beheshti University, Iran

Reviewed by:
Abed Khorasani,
Kerman Medical University, Iran
Mohsen Bafandehkar,
Macquarie University, Australia

****Correspondence:***
Hamid Karimi-Rouzbahani
hamid.karimi-rouzbahani@mrc-cbu.cam.ac.uk

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *30 November 2021*
Accepted: *24 January 2022*
Published: *02 March 2022*

Citation:
Karimi-Rouzbahani H and Woolgar A (2022) When the Whole Is Less Than the Sum of Its Parts: Maximum Object Category Information and Behavioral Prediction in Multiscale Activation Patterns. Front. Neurosci. 16:825746. doi: 10.3389/fnins.2022.825746

INTRODUCTION

How is information about the world encoded by the human brain? Researchers have tried to answer this question using variety of brain imaging techniques across all sensory modalities. In vision, people have used invasive (Hung et al., 2005; Liu et al., 2009; Majima et al., 2014; Watrous et al., 2015; Rupp et al., 2017; Miyakawa et al., 2018) and non-invasive (EEG and MEG; Simanova et al., 2010; Carlson et al., 2013; Cichy et al., 2014; Kaneshiro et al., 2015; Contini et al., 2017) brain imaging modalities to decode object category information from variety of features of the recorded neural activations. While majority of EEG and MEG decoding studies still rely on the within-trial "mean" of activity (average of activation level within the sliding analysis window) as the main source of information (Grootswagers et al., 2017; Karimi-Rouzbahani et al., 2017b), recent theoretical and experimental studies have shown evidence that temporal variabilities of neural activity (sample to sample changes in the level of activity) form an additional channel of information encoding (Orbán et al., 2016). For example, these temporal variabilities have provided information about the "complexity," "uncertainty," and the "variance" of the visual stimulus, which correlated with the semantic category of the presented image (Hermundstad et al., 2014; Orbán et al., 2016; Garrett et al., 2020). Specifically, object categories which show a wider variability in their exemplars (e.g., houses) evoke more variable neural activation than categories which have lower variability (e.g., faces; Garrett et al., 2020). Accordingly, it is now clear that neural variabilities carry significant amounts of information about different aspects of sensory processing and may also play a major role in determining behavior (Waschke et al., 2021).

Despite the richness of information in neural variabilities, there is no consensus yet about how to quantify informative neural variabilities. Specifically, neural variabilities have been quantified using three classes of mathematical features: variance-, frequency-, and information theory-based features, each detecting specific, but potentially overlapping aspects of the neural variabilities (Waschke et al., 2021). Accordingly, previous studies have decoded object category information from EEG using variance-based (Wong et al., 2006; Mazaheri and Jensen, 2008; Alimardani et al., 2018; Joshi et al., 2018), frequency-based (Taghizadeh-Sarabi et al., 2015; Watrous et al., 2015; Jadidi et al., 2016; Wang et al., 2018; Voloh et al., 2020) and information theory-based (Richman and Moorman, 2000; Shourie et al., 2014; Torabi et al., 2017; Ahmadi-Pajouh et al., 2018) features. However, these previous studies remained silent about the temporal dynamics of category encoding as they performed the analyses (i.e., feature extraction and decoding) on the whole-trial data to maximize the decoding accuracy. On the other hand, time-resolved decoding analyses studied the temporal dynamics of category information encoding (Kaneshiro et al., 2015; Grootswagers et al., 2017; Karimi-Rouzbahani, 2018). However, few time-resolved studies have extracted any features other than the instantaneous activity at each time point, or the mean of activity across a short sliding window (e.g., by down-sampling the data), to incorporate the information contained in neural variabilities (Majima et al., 2014; Karimi-Rouzbahani et al., 2017a). Therefore, previous studies either did not focus on the temporal dynamics of information processing or did not include the contents of neural variabilities in time-resolved decoding.

Critically, as opposed to the Brain-Computer Interface (BCI) community, where the goal of feature extraction is to maximize the decoding accuracy, in cognitive neuroscience the goal is to find better neural correlates for the behavioral effect under study (Williams et al., 2007; Jacobs et al., 2009; Hebart and Baker, 2018; Woolgar et al., 2019; Karimi-Rouzbahani et al., 2021a,b). Specifically, a given feature is arguably only informative if it predicts behavior. Therefore, behavior is a key benchmark for evaluating the information content of any features including those which quantify neural variabilities. Interestingly, almost none of the above-mentioned decoding studies focused on evaluating the predictive power of their suggested informative features about behavior. Therefore, it remains unclear if the additional information they obtained from features of neural variabilities was task-relevant or epiphenomenal to the experimental conditions.

To overcome these issues, we proposed a new approach using medium-sized (50 ms) sliding windows at each time step (5 ms apart). The 50 ms time window makes a compromise between concatenating the whole time window, which in theory allows any feature to be used at the expense of temporal resolution, and decoding in a time resolved fashion at each time point separately, which might lose temporal patterns of activity (Karimi-Rouzbahani et al., 2021b). Within each window, we quantify multiple different mathematical features of the continuous data. This allows us to be sensitive to any information carried in local temporal variability in the EEG response, while also maintaining reasonable temporal resolution in the analysis. In a recent study, we extracted a large set of such features and quantified the information contained in each using multivariate classification (Karimi-Rouzbahani et al., 2021b). We balanced the number of extracted values across features using Principal Component Analysis (PCA). Across three datasets, we found that that the incorporation of temporal patterns of activity in decoding, through the extraction of spatiotemporal "Wavelet coefficients" or even using the informative "original magnitude data (i.e., no feature extraction)," provided higher decoding performance than the more conventional average of activity within each window ("mean"). Importantly, we also observed that for our Active dataset where participants categorized objects, the decoding results obtained from the same two features (i.e., Wavelet coefficients and original magnitude data) could predict/explain the participants' reaction time in categorization significantly better than the "mean" of activity in each window (Wavelet outperformed original magnitude data). We further observed that more effective decoding of the neural codes, through the extraction of more informative features, corresponded to better prediction of behavioral performance. We concluded that the incorporation of temporal variabilities in decoding can provide additional category information and improved prediction of behavior compared to the conventional "mean" of activity.

One critical open question, however, is whether we should expect the brain to encode the information via each of these features individually, or whether it may instead use combinations of these features. In other words, while each of feature may potentially capture a specific and limited aspect of the generated neural codes, the brain may recruit multiple neural encoding protocols at the same time point or in succession within the same trial. Specifically, an encoding protocol might be active only for a limited time window or for specific aspects of the visual input (Gawne et al., 1996; Wark et al., 2009). For example, it has been shown in auditory cortex that two distinct encoding protocols (millisecond-order codes and phase coding) are simultaneously informative (Kayser et al., 2009).

Another study showed that spike *rates* on 5–10 ms timescales carried complementary information to that in the *phase* of firing relative to low-frequency (1–8 Hz) local field potentials (LFPs) about which epoch of naturalistic movie was being shown (Montemurro et al., 2008). These examples suggest that two very distinct encoding protocols (rate vs. phase coding) might be at work simultaneously to provide information about distinct aspects of the same sensory input. Therefore, it might be the case that multiple neural encoding protocols contribute to the encoding of information. Alternatively, the brain may implement one general multiscale encoding protocol [e.g., multiplexing strategy which combines same-structure neural codes at different time scales (Panzeri et al., 2010)], which allows different aspects of information to be represented within a more flexible encoding protocol. More specifically, the brain might implement a general platform, which allows the representation of information at different temporal and spatial scales. For example, in visual stimulus processing, one study found that stimulus contrast was represented by latency coding at a temporal precision of ∼10 ms, whereas stimulus orientation and its spatial frequency were encoded at a coarser temporal precision (30 and 100 ms, respectively; Victor, 2000). This multiplexed encoding protocol has been suggested to provide several computational benefits to fixed encoding protocol including enhancing the coding capacity of the system (Schaefer et al., 2006; Kayser et al., 2009), reducing the ambiguity inherent to single-scale codes (Schaefer et al., 2006; Schroeder and Lakatos, 2009) and improving the robustness of neural representations to environmental noise (Kayser et al., 2009).

To see if EEG activations reflect the neural codes using several encoding protocols simultaneously, we created combinations from the large set of distinct mathematical features in our previous study (Karimi-Rouzbahani et al., 2021b). We asked whether their combination recovers more of the underlying neural code, leading to additional object category information and increased accuracy in predicting behavior, compared to the best performing individual feature from the previous study (i.e., Wavelet). Specifically, we used the same three datasets, extracted the same features from neural activity, selected the most informative features at each sliding time window and evaluated their information about object categories. We also evaluated how well each combined feature set explained behavioral recognition performance. Our prediction was that as targeted combinations of informative features provide more flexibility in detecting subtle differences, which might be ignored when using each individual feature, we should see both a higher decoding accuracy and predictive power for behavior compared to when using individual features. However, our results show that, the most informative individual feature (the Wavelet transform) outperformed all of the feature combinations (combined using 17 different feature selection algorithms). Similarly, Wavelet coefficients outperformed all combinations of features in predicting behavioral performance. Therefore, while the relationship between neuron-level encoding of information and EEG signals remains to be investigated in the future, these results provide evidence for a general multiscale encoding protocol (i.e., captured by Wavelet coefficients) rather than a combination of several protocols for category encoding in the EEG data.

MATERIALS AND METHODS

As this study uses the same set of datasets and features used in our previous study, we only briefly explain the datasets and the features. The readers are referred to our previous manuscript (Karimi-Rouzbahani et al., 2021b) as well as the original manuscripts (cited below) for more detailed explanation of the datasets and features. The datasets used in this study and the code are available online at https://osf.io/wbvpn/. The EEG and behavioral data are available in Matlab ".mat" format and the code in Matlab ".m" format.

All the open-source scripts used in this study for feature extraction were compared/validated against other implementations of identical algorithms in simulations and used only if they produced identical results. All open-source scripts of similar algorithms produced identical results in our validations. To validate the scripts, we used 1,000 random (normally distributed with unit variance and zero mean) time series each including 1,000 samples.

Overview of Datasets

We selected three highly varied previously published EEG datasets (**Table 1**) for this study to be able to evaluate the generalizability of our results and conclusions. Specifically, the datasets differed in a wide range of aspects including the recording set-up (e.g., amplifier, number of electrodes, preprocessing steps, etc.), properties of the image-set (e.g., number of categories and exemplars within each category, colorfulness of images, etc.), paradigm and task (e.g., presentation length, order and the participants' task). The EEG datasets were collected while the participants were presented with images of objects, animals, face, etc. Participants' task in Dataset 1 was irrelevant to the identity of the presented objects; they reported if the color of fixation changed from the first stimulus to the second in pairs of stimuli. Participants' task for Dataset 2 was to respond/withhold response to indicate if the presented object belonged to the category (e.g., animal) cued at the beginning of the block. Participants had no explicit active task except for keeping fixation on the center of the screen for Dataset 3. To obtain relatively high signal to noise ratios for the analyses, each unique stimulus was presented to the participants 3, 6, and 12 times in datasets 1–3, respectively. The three datasets previously successfully provided object category information using multivariate decoding methods. For more details about the datasets see the original manuscripts cited in **Table 1**.

Preprocessing

The datasets were collected at a sampling rate of 1,000 Hz. Each dataset consisted of data from 10 participants. Each object category in each dataset included 12 exemplars. For datasets 1 and 2, only the trials with correct responses were used in the analyses (dataset 3 had no task). To make the three datasets as consistent as possible, we pre-processed them differently from their original

TABLE 1 | Details of the three datasets used in the study.

Dataset	# and type of electrodes	Band-pass filtering	Notch filtering	# object categories	# stimulus repetition	Stimulus presentation time	Stimulus size (periphery)	Task	Participants' accuracy	Participants' Age (median)	Participants' gender
1 Karimi-Rouzbahani et al., 2017a	31 (Passive-10–20 system)	0.03–200 Hz	50 Hz	4	3	50 ms	2°–13.5° (0.7°–8.8°)	Color matching (passive)	%94.68	22.1	Seven male Three female
2 Karimi-Rouzbahani et al., 2019	31 (Passive-10–20 system)	0.03–200 Hz	50 Hz	4	6	900 ms	8° × 8° (0)	Object category detection (active)	%94.65	26.4	Six male Four female
3 Kaneshiro et al., 2015	128 (Passive high-density HCGSN 128)	0.03–50 Hz	No	6	12	500 ms	7.0° × 6.5° (0)	No task (fixation)	N/A	30.5	Seven male Three female

manuscripts. We performed notch-filtering on datasets 1 and 2 at 50 Hz. Datasets 1 and 2 were also band-pass-filtered in the range from 0.03 to 200 Hz. The band-pass filtering range of dataset 3 was 0.03–50 Hz, as we did not have access to the raw data to increase the upper bound to 200 Hz. Despite potential muscle artifacts in higher frequency bands of EEG (e.g., >30 Hz; da Silva, 2013; Muthukumaraswamy, 2013), the upper limit of the frequency band was selected liberally (200 Hz) to avoid missing any potential information which might be contained in high-frequency components (gamma band) of the signals (Watrous et al., 2015). As sporadic artefacts (including muscle activity, eye and movement artifacts) do not generally consistently differ across conditions (here categories), they will only minimally affect multivariate decoding analyses (Grootswagers et al., 2017; Karimi-Rouzbahani et al., 2021c). For the same reason, we did not remove the artifacts. We used finite-impulse-response filters with 12 dB roll-off per octave for band-pass filtering of datasets 1 and 2. The filtering was applied on the data before they were epoched relative to the trial onset times. Data were epoched from 200 ms before to 1,000 ms after the stimulus onset to cover most of the range of event-related neural activations. The average pre-stimulus (−200 to 0 ms relative to the stimulus onset) signal amplitude was removed from each trial of the data. For more information about each dataset see the references cited in **Table 1**.

Features

We briefly explain the 26 mathematically distinct features used in this study below. Note that 4 of the features, which were event-related potentials, were excluded from this study as they could not be defined across time. For more details about their algorithms, their plausibility and possible neural underpinnings please see Karimi-Rouzbahani et al. (2021b). Each feature was calculated for each EEG electrode and each participant separately. The following features were extracted after the raw data was filtered, epoched and baselined as explained. Each of the features was extracted from the 50 samples contained in 50 ms sliding time windows at a step size of 5 ms along each trial. The sampling rate of the data remained at 1,000 Hz and the features were extracted from the 1,000-Hz data but only calculated every 5 ms to decrease the computational load. Note that the width of the sliding analysis window needs special attention as it involves a trade-off between noise and potential information (about conditions and behavior) in EEG signals. Specifically, very short windows may lose potentially informative longer patterns, whereas very long windows might lose shorter patterns as they might be dominated by slow fluctuations. In the original work (Karimi-Rouzbahani et al., 2021b) we tested window widths between 5 and 100 ms and found 50 ms to be the most informative range for decoding, so that is the value we use here.

Mean, Variance, Skewness, and Kurtosis

These are the standard 1st to 4th moments of EEG time series. To calculate these features, we simply calculated the mean, variance, skewness and variance of EEG signals *over the samples within each sliding analysis window within each trial* (50 samples). Please note that this differs from averaging over trials, which is sometimes used to increase signal to noise ratio (Hebart and Baker, 2018). "Mean" of activity is by far the most common feature of EEG signal used in time-resolved decoding (Grootswagers et al., 2017). Specifically, in time-resolved decoding, generally the samples within each sliding time window are averaged and used as the input for the classification algorithm. People sometimes perform down-sampling of EEG time series, which either performs simple averaging or retains the selected samples every few samples. Variance (Wong et al., 2006), Skewness (Mazaheri and Jensen, 2008), and Kurtosis (Pouryazdian and Erfanian, 2009; Alimardani et al., 2018) have shown success in providing information about different conditions of visually evoked potentials.

Median

We also calculated signal's median as it is less affected by spurious values compared to the signal mean providing less noisy representations of the neural processes.

While the moment features above provide valuable information about the content of EEG evoked potentials, many distinct time series could lead to similar moment features. In order to be sensitive to this potentially informative differences nonlinear features can be used which, roughly speaking, are sensitive to nonlinear and complex patterns in time series. Below we define the most common nonlinear features of EEG time series analysis, which we used in this study.

Lempel-Ziv Complexity

We calculated the Lempel-Ziv (LZ) complexity as an index of signal complexity. This measure counts the number of unique sub-sequences within the analysis window (50 time samples), after turning the time samples into a binary sequence. To generate the binary sequence, we used the signal median, within the same analysis window, as the threshold. Accordingly, the LZ complexity of a time series grows with the length of the signal and its irregularity over time. See Lempel and Ziv (1976) for more details. This measure has previously provided information about neural responses in primary visual cortices (Szczepański et al., 2003). We used the script by Quang Thai[1] implemented based on "exhaustive complexity" which is considered to provide the lower limit of the complexity as explained by Lempel and Ziv (1976).

Higuchi and Katz Fractal Dimensions

Fractal is an indexing technique which provides statistical information determining the complexity of how data are organized within time series. Accordingly, higher fractal values, suggest more complexity and vice versa. In this study, we calculated the complexity of the signals using two methods of Higuchi and Katz, as used previously for categorizing object categories (Torabi et al., 2017; Ahmadi-Pajouh et al., 2018; Namazi et al., 2018). We used the implementations by Jesús Monge Álvarez[2] after verifying it against other implementations.

Hurst Exponent

This measure quantifies the long-term "memory" in a time series. Basically, it calculates the degree of dependence among consecutive samples of time series and functions similarly to the autocorrelation function (Racine, 2011; Torabi et al., 2017). Hurst values between 0.5 and 1 suggest consecutive appearance of high signal values on large time scales while values between 0 and 0.5 suggest frequent switching between high and low signal values. Values around 0.5 suggest no specific patterns among samples of a time series.

Sample and Approximate Entropy

Entropy measures the level of perturbation in time series. As the precise calculation of entropy needs large sample sizes and is also noise-sensitive, we calculated it using two of the most common approaches: sample entropy and approximate entropy. Sample entropy is not as sensitive to the sample size and simpler to implement compared to approximate entropy. Sample entropy, however, does not take into account self-similar patterns in the time series (Richman and Moorman, 2000). We used an open-source code[3] for calculating approximate entropy.

Autocorrelation

This index quantifies the self-similarity of a time series at specific time lags. Accordingly, if a time series has a repeating pattern at the rate of F hertz, an autocorrelation measure with a lag of 1/F will provide a value of 1. However, it would return −1 at the lag of 1/2F. It would provide values between −1 and 1 for other lags. More complex signals would provide values close to 0. A previous study has been able to decode neural information about motor imagery using the autocorrelation function from EEG signals (Wairagkar et al., 2016).

Hjorth Complexity and Mobility

These parameters measure the variation in the signals' characteristics. The complexity measure calculates the variation in a signal's dominant frequency, and the mobility measures the width of the signal's power spectrum [how widely the frequencies are scattered in the power spectrum of the signal (Joshi et al., 2018)].

Mean, Median, and Average Frequency

These measures calculate the central frequency of the signal in different ways. Mean frequency is the average of all frequency components available in a signal. Median frequency is the median normalized frequency of the power spectrum of the signal and the average frequency is the number of times the signal time series crosses zero. They have shown information about visual categories in previous studies (Jadidi et al., 2016; Iranmanesh and Rodriguez-Villegas, 2017; Joshi et al., 2018).

Spectral Edge Frequency (95%)

Spectral edge frequency (SEF) indicates the high frequency below which x percent of the signal's power spectrum exists. X was set to 95% in this study. Therefore, SEF reflects the upper-bound of frequency in the power spectrum.

Signal Power, Power, and Phase at Median Frequency

Power spectrum density (PSD) represents the intensity or the distribution of the signal power into its constituent frequency components. Signal power was used as a feature here as in previous studies (Majima et al., 2014; Rupp et al., 2017), where it showed associations between aspects of visual perception and power in certain frequency bands. Signal power is the frequency-domain representation of temporal neural variability (Waschke et al., 2021). We also extracted signal power and phase at median frequency which have previously shown to be informative about object categories (Jadidi et al., 2016; Rupp et al., 2017).

For the following features we had more than one value per trial and sliding time window. We extracted all these features but later down-sampled the values to *one* per trial using the (first) PCA procedure explained below (**Figure 1**) before using them in the feature combination procedure.

Cross-Correlation

This refers to the inter-electrode correlation of EEG time series. It simply quantifies the similarity of activations between pairs of EEG electrodes. Therefore, for each electrode, we had e-1 cross-correlation values with e referring to the number of electrodes. This measure has been shown to contain information about visual object categories before (Majima et al., 2014; Karimi-Rouzbahani et al., 2017a).

[1] https://www.mathworks.com/matlabcentral/fileexchange/38211-calc_lz_complexity

[2] https://ww2.mathworks.cn/matlabcentral/fileexchange/50290-higuchi-and-katz-fractal-dimension-measures

[3] https://www.mathworks.com/matlabcentral/fileexchange/32427-fast-approximate-entropy

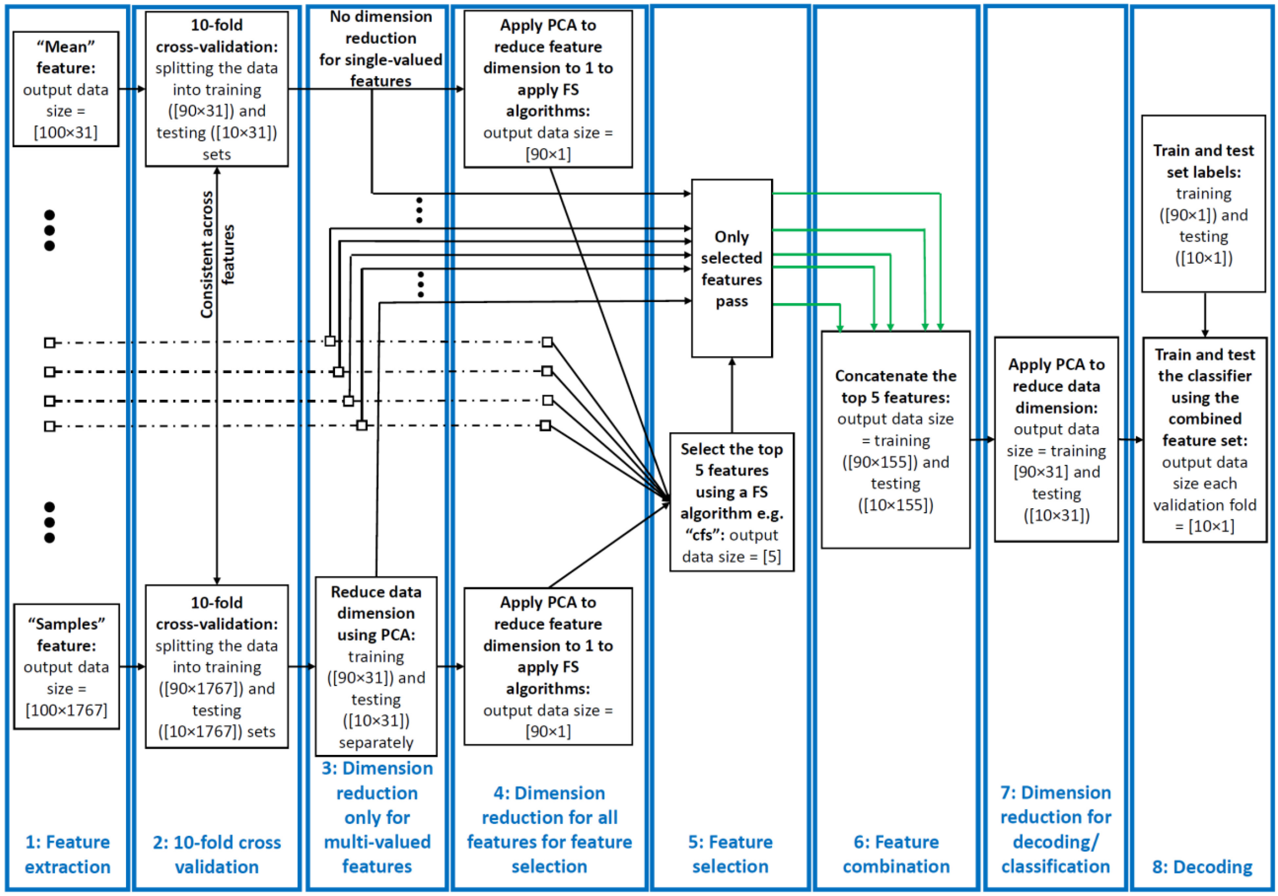

"Mean" feature: output data size = [100×31]
10-fold cross-validation: splitting the data into training ([90×31]) and testing ([10×31]) sets
No dimension reduction for single-valued features
Apply PCA to reduce feature dimension to 1 to apply FS algorithms: output data size = [90×1]
Consistent across features
Only selected features pass
Train and test set labels: training ([90×1]) and testing ([10×1])
Concatenate the top 5 features: output data size = training ([90×155]) and testing ([10×155])
Apply PCA to reduce data dimension: output data size = training [90×31] and testing ([10×31])
Train and test the classifier using the combined feature set: output data size each validation fold = [10×1]
Select the top 5 features using a FS algorithm e.g. "cfs": output data size = [5]
"Samples" feature: output data size = [100×1767]
10-fold cross-validation: splitting the data into training ([90×1767]) and testing ([10×1767]) sets
Reduce data dimension using PCA: training ([90×31]) and testing ([10×31]) separately
Apply PCA to reduce feature dimension to 1 to apply FS algorithms: output data size = [90×1]
1: Feature extraction
2: 10-fold cross validation
3: Dimension reduction only for multi-valued features
4: Dimension reduction for all features for feature selection
5: Feature selection
6: Feature combination
7: Dimension reduction for decoding/ classification
8: Decoding

FIGURE 1 | Decoding pipeline. From left to right: successive stages shown for a sample dataset comprising 100 trials of data from two categories recorded using a 31-electrode EEG amplifier. (1) Features are extracted from each trial and time window of the data. The features can be single- or multi-valued resulting in different number of values per trial and analysis time window. (2) We split the trials into training and testing sets and use the training sets in PCA and training the classifiers throughout the pipeline. (3) We used a PCA-based dimension reduction to reduce the number of values of only the multi-valued features to one equalizing them with single-valued features. (4) We used a second PCA to project all values of each feature to one dimension to be able to feed to the feature selection (FS) algorithms. (5) We selected the five most informative features using the FS algorithms. (6) We combined these features using concatenation of the selected features in their original size received from stage 4. (7) We reduced the dimension of the concatenated feature set to equalize it with the single-valued individual features from the previous study so that they could be compared. (8) We decoded/classified all pair-wise categories using the final dataset in each fold. This figure shows the procedure for a single cross-validation fold at one time point and was repeated for all the folds and time points. To avoid circularity, PCA was only ever applied on the training set and the parameters (mean and eigen vectors) used to derive the principal component of both the training and testing sets. The green arrows indicate example selected feature sets sent for combination.

Wavelet Coefficients

Considering the time- and frequency-dependent nature of ERPs, Wavelet transform seems to be a very reasonable choice as it provides a time-frequency representation of signal components. It determines the primary frequency components and their temporal position in time series. The transformation passes the signal time series through digital filters (Guo et al., 2009), each of which adjusted to extract a specific frequency (scale) at a specific time. This filtering procedure is repeated for several rounds (levels) filtering low- (approximations) and high-frequency (details) components of the signal to provide more fine-grained information about the constituent components of the signal. This can lead to coefficients which can potentially discriminate signals evoked by different conditions. Following up on a previous study (Taghizadeh-Sarabi et al., 2015), and to make the number of Wavelet features comparable in number to signal samples, we used detail coefficients at five levels D1,. . .,D5 as well as the approximate coefficients at level 5, A5. This led to 57 features in the 50 ms sliding time windows. We used the "Symlet2" basis function for our Wavelet transformations as implemented in Matlab. The multistage, variable-sized filtering procedure implemented in Wavelet coefficients, make them ideal for detecting multiscale patterns of neural activity, which has been suggested to be produced by the brain for information encoding (Panzeri et al., 2010).

Hilbert Amplitude and Phase

This transformation is a mapping function that takes a function x(t) of a real variable, and using convolution with the function, $1/\pi t$, produces another function of a real variable H(u) (t).

This technique provides amplitude and phase information of the signal in the transformed space allowing us to tease them apart and evaluate their information content about visual categories (Wang et al., 2018).

Original Magnitude Data (Samples)

We also used the post-stimulus signal samples (i.e., 50 samples in each sliding analysis window) to decode object category information without any feature extraction. This allowed us to compare the information content of the extracted features with the original signal samples to see if the former provided any extra information. Note that, this is different from averaging/down-sampling of magnitude data within the analysis windows conventionally used in multivariate decoding (Karimi-Rouzbahani et al., 2017a).

Feature Selection Algorithms

We set out to test whether neural information about object categories might be captured by combinations of the above features, better than by any one feature individually. For this, we combined the 26 extracted features using Feature Selection Library (FSLib, version 6.2.1; Roffo, 2016). Feature selection (FS), which refers to selecting a subset of features from a larger set, is generally used (for example, in machine learning) to reduce the dimensionality of the data by removing the less informative features from the dataset. FS algorithms can be categorized as supervised or unsupervised (Dash and Liu, 1997). The supervised methods receive, as input, the labels of trials for each condition (i.e., object categories here), and try to maximize the distance between conditions. We used eight different supervised FS algorithms. The unsupervised methods, on the other hand, incorporate different criteria for FS such as selecting features that provide maximum distance (i.e., *unfol*) or minimum correlation (i.e., *cfs*). The FSLib implements 19 different feature selection algorithms. As it is not yet known how the brain might recruit different encoding protocols or a potential combination of them, we used all the FS algorithms available by the FSLib to combine the features in this study, except two (rfe-SVM and L0) which we were not able to implement. Although there are other feature selection algorithms in the literature, we believe that using these 17 methods, we capture a decent range of different approaches. We set the number of selected features to 5, which was chosen to balance between including too many features, which could obscure interpretability, and including too few, which risks missing informative but lower-ranked features. Below we briefly explain the eight supervised and nine unsupervised feature selection algorithms. Readers are referred to the original manuscripts for more detail about each feature selection method as reviewed (Roffo, 2016).

Among supervised algorithms, *Relief* is a randomized and iterative algorithm that evaluates the quality of the features based on how well their values discriminate data samples from opposing conditions. This algorithm can be sensitive when used on small data samples. *Fisher* evaluates the information of features as the ratio of inter-class to intra-class distances. *Mutual Information* (mutinffs) measures the association between the data samples (observations) within each feature and their class labels. *Max-Relevance, Min-Redundancy* (mrmr) method, which is an extension of the mutual information method, is designed to follow two basic rules when selecting the features: to select the features which are mutually far away from each other while still having "high" correlation to the classification labels. As opposed to the above methods, which rank and select the features according to their specific criteria, the *Infinite latent* (ILFS) method, selects the most informative features based on the importance of their neighboring features in a graph-based algorithm. It is a supervised probabilistic approach that models the features "relevancy" in a generative process and derives the graph of features which allows the evaluation of each feature based on its neighbors. Similarly, the method of *Eigenvector Centrality* (ECFS), generates a graph of features with features as nodes and evaluates the importance of each node through an indicator of centrality, i.e., eigen vector centrality. The ranking of central nodes determines the most informative features. *LASSO* algorithm works based on error minimization in predicting the class labels using the features as regression variables. The algorithm penalizes the coefficients of the regression variables while setting the less relevant to zero to follow the minimal sum constraint. The selected features are those which have non-zero coefficients in this process. *Concave Minimization* (fsv) uses a linear programming technique to inject the feature selection process into the training of a support vector machine (SVM).

Among unsupervised FS algorithms, *Infinite FS* (InfFS), is similar to the graph-based supervised methods in which each feature is a node in a graph. Here, however, a path on a graph is a subset of features and the importance of each feature is measured by evaluating all possible paths on the graph as feature subsets in a cross-validation procedure. *Laplacian Score* (laplacian), evaluates the information content of each feature by its ability of locality preserving. To model the local geometry of the features space, this method generates a graph based on nearest neighbor and selects the features which respect this graph structure. *Dependence Guided* (dgufs) method evaluates the relationship between the original data, cluster labels and selected features. This algorithm tries to achieve two goals: to increase the dependence on the original data, and to maximize the dependence of the selected features on cluster labels. *Adaptive Structure Learning* (fsasl), which learns the structure of the data and FS at the same time is based on linear regression. *Ordinal Locality* (ufsol) is a clustering-based method which achieves distance-based clustering by preserving the relative neighborhood proximities. *Multi-Cluster* (mcfs) method is based on manifold learning and L1-regularized models for subset selection. This method selects the features such that the multi-cluster structure of the data can be best preserved. As opposed to most of the unsupervised methods which try to select the features which preserve the structure of the data, e.g., manifold learning, *L2,1-norm Regularized* (UDFS) method assumes that the class label of data can be predicted using a linear classifier and incorporates discriminative analysis and L2,1-norm minimization into a joint framework for feature selection. *Local Learning-Based* (llcfs) method is designed to work with high-dimensional manifold data. This method associates weights to features which are incorporated into the regularization procedure to evaluate their relevance for the clustering. The weights are optimized iteratively during clustering which leads to the selection of the most informative features in an unsupervised fashion. *Correlation-Based* (cfs) method simply ranks the features based on how uncorrelated they are to the other features in the feature set. Therefore, the selected features are those which are most distinct from others.

Decoding Pipeline

The pipeline used in this study for feature extraction, dimensionality reduction, feature selection, feature combination and decoding had eight stages and is summarized in **Figure 1**. Below we explain each stage of the pipeline for a simple sample dataset with 100 trials collected using a 31-electrode EEG setup. Our actual datasets, however, had varied number of trials and electrodes as explained in **Table 1**. Note that the data from all electrodes were included in the analysis and could have affected the final decoding results equally.

Feature Extraction

We extracted the set of 26 above-mentioned features from the dataset. This included features which provided one value for each sliding time window per trial (single-valued) and more than one value (multi-valued). For the sample dataset, this resulted in data matrices with 100 rows (trials) and 31 columns (electrodes) for the single-valued datasets and $31 \times e$ columns for multi-valued features, where e refers to the number of values extracted for each trial and time window.

Cross Validation

After extracting the features, we split the data into 10 folds, used 9 folds for dimension reductions and training the classifiers and the left-out fold for testing the classifiers. Therefore, we used a 10-fold cross-validation procedure in which we trained the classifier on 90% of the data and tested it on the left-out 10% of the data, repeating the procedure 10 times until all trials from the pair of categories participate once in the training and once in the testing of the classifiers. The same trials were chosen for all features in each cross-validation fold.

Dimensionality Reduction 1: Only for Multi-Valued Features

The multi-valued features explained above resulted in more than a single feature value per trial per sliding time window (e.g., cross-correlation, wavelet, Hilbert amplitude, and phase and signal samples). This could lead to the domination of the multi-valued over single-valued features in feature selection and combination. To avoid that, we used principle component analysis (PCA) to reduce the number of values in the multi-valued features to one per electrode per time window, which was the number of values for all single-valued features. Specifically, the data matrix before dimension reduction, had a dimension of n rows by $e \times f$ columns where n, e, and f were the number of trials in the dataset (consisting of all trials from all categories), the number of electrodes and the number of values obtained from a given feature (concatenated in columns), respectively. *Therefore, the columns of multi-valued features included both the spatial (electrodes) and temporal (elements of each feature) patterns of activity from which the information was obtained*. This is different from single-valued features where the columns of their data matrix only included spatial patterns of activity. As $f = 1$ for the single-valued features, for the multi-valued features, we only retained the e most informative columns that corresponded to the e eigen values with highest variance and removed the other columns using PCA. Therefore, we reduced the dimension of the data matrix to $n \times e$ which was the same for single- and multi-valued features and used the resulting data matrix for decoding. This means that, for the multi-valued features, in every analysis window, we only retained the most informative value of the extracted feature elements and electrodes (i.e., the one with the most variance in PCA). Accordingly, multi-valued features had the advantage over single-valued features as the former utilized both the *spatial* and *temporal* patterns of activity in each sliding time window, while the latter only had access to the *spatial* patterns.

Dimensionality Reduction 2: For Feature Selection

For feature selection, each feature should have a dimension of 1 to go into the FS algorithm. However, our features had as many dimensions as the number of electrodes (i.e., e). Therefore, we further reduced the dimension of each feature from e to 1 to be able to feed them to the FS algorithms, compare them and select the most informative features. This allowed us to know the general amount of information that each feature rather than each of its elements/dimensions (e.g., electrodes in single-valued features) had about object categories. Please note that, however, after finding the most informative features, we used the selected features in their original size which was e (output of step 3 goes to stage 6).

Feature Selection

Feature selection was done using 17 distinct algorithms (above) to find the five most informative features in every sliding time window. This stage only provided indices of the selected features for combination in the next stage. To avoid any circularity (Pulini et al., 2019), we applied the FS algorithms only on the training data (folds) and used the selected features in both training and testing in each cross-validation run. Please note that feature selection was performed in every analysis window across the trial. In other words, different sets of five features could be selected for each individual analysis window. This allowed multiple features to contribute at each time point (multiple codes to be in use at the same time) and for different features to be selected at different time points (different codes used at different points in the trial).

Feature Combination

We only concatenated the five selected features into a new data matrix. At this stage, we received five feature data matrices which had a dimension of $n \times e$ with n referring to the number of trials and e referring to the number of values per trial, which were 100×31 for the sample dataset explained in **Figure 1**. The combination procedure led to a concatenated data matrix of 100×155 ($n \times 5e$).

Dimensionality Reduction 3: Equalizing the Dimensions of Combined and Individual Feature Spaces

We used another round of PCA to simultaneously combine and reduce the dimensionality of each data matrix (feature space) to equalize it with the feature space of the individual features. This made the combined and individual features directly comparable, so that we could test whether a combination of the most informative features could provide additional category-related information, over and above the information decodable from individual features. Had we not controlled for the dimension of the data matrix, superior decoding for the combined features could arise trivially (due to having more predictors). Note that, whereas we knew the features which were selected on stage 5, as a result of this PCA transformation, we did not know which features contributed to the final decoding result. Therefore, in the worst case scenario, the final feature set might have only contained one of the five selected features. However, this seems unlikely to be the case as generally all inputs contribute to the distributions of the data in the PCA space. To avoid circularity (Pulini et al., 2019), we again applied the PCA algorithms on the training data (folds) only and used the training PCA parameters (i.e., eigen values and means) for both training and testing (fold) sets for dimension reduction, carrying this out in each cross-validation run separately.

Multivariate Decoding

Finally we used time-resolved multivariate decoding to test for information about object categories in the features and combinations of features. We used linear discriminant analysis (LDA) classifiers to measure the information content across all possible pairs of conditions (i.e., object categories) in each dataset. We repeated the decoding across all possible pairs of categories within each dataset, which were 6, 6 and 15 pairs for datasets 1–3, which consisted of 4, 4 and 6 object categories, respectively. Finally, we averaged the results across all combinations and reported them as the average decoding for each participant. We extracted the features from 50 ms sliding time windows in steps of 5 ms across the time course of the trial (−200 to 1,000 ms relative to the stimulus onset time). Therefore, the decoding results at each time point reflect the data for the 50 ms window around the time point, from −25 to +24 ms relative to the time point.

Decoding-Behavior Correlation

We evaluated the correlation between neural representations of object categories and the reaction time of participants in discriminating them. To that end, we generated a 10-dimensional vector of neural decoding accuracies (averaged over all pairwise category decoding accuracies obtained from each participant) at every time point and a 10-dimensional vector which contained the behavioral reaction times (averaged over all categories obtained from each participant) for the same group of 10 participants. Then we correlated the two vectors at each time point using Spearman's rank-order correlation (Cichy et al., 2014; Ritchie et al., 2015). This resulted in a single correlation value for each time point for the group of 10 participants.

Parameters of Decoding Curves

To quantitatively evaluate the patterns of decoding curves and decoding-behavior correlations, we extracted four distinct parameters from the decoding curves and one parameter from the correlation to behavior curves. All parameters were calculated in the post-stimulus time span. The "average correlation to behavior" was calculated by averaging the level of across-subject correlation to behavior. The parameters of "average decoding" and "maximum decoding" were calculated for each participant simply by calculating the average and maximum of the decoding curves. The "time of maximum decoding" and "time of first above-chance decoding" were also calculated for each participant relative to the time of the stimulus onset.

Statistical Analyses

Bayes Factor Analysis

First we asked whether we could decode object category from the combined features returned by each of the 17 FS methods. To determine the evidence for the null and the alternative hypotheses, we used Bayes analyses as implemented by Bart Krekelberg[4] based on Rouder et al. (2012). We used standard rules of thumb for interpreting levels of evidence (Lee and Wagenmakers, 2005; Dienes, 2014): Bayes factors of >10 and <1/10 were interpreted as strong evidence for the alternative and null hypotheses, respectively, and >3 and <1/3 were interpreted as moderate evidence for the alternative and null hypotheses, respectively. We considered the Bayes factors which fell between 3 and 1/3 as suggesting insufficient evidence either way.

To evaluate the evidence for the null and alternative hypotheses of at-chance and above-chance decoding, respectively, we compared the decoding accuracies obtained from all participants in the post-stimulus onset time against the decoding accuracies obtained from the same participants averaged in the pre-stimulus onset time (−200 to 0 ms). We also asked whether there was a difference between the decoding values obtained from all possible pairs of FS methods. Accordingly, we performed the Bayes factor unpaired *t-test* and calculated the Bayes factor as the probability of the data under alternative (i.e., difference; H1) relative to the null (i.e., no difference; H0) hypothesis between all possible pairs of FS methods for each dataset separately. The same procedure was used to evaluate evidence for difference (i.e., alternative hypothesis) or no difference (i.e., null hypothesis) in the maximum and average decoding accuracies, the time of maximum and above-chance decoding accuracies across FS methods for each dataset separately. To evaluate the evidence for the null or alternative hypotheses of lack of or the existence of difference between the decoding accuracies obtained from FS algorithm and the Wavelet feature, we calculated the Bayes factor between the distribution of the two distributions of decoding accuracies on every time point and for dataset separately.

Priors for the Bayes analysis can be selected based on previous work or can be estimated based on predetermined Cauchy distribution according to common effect sizes. We opted to use

[4] https://klabhub.github.io/bayesFactor/

default priors. This choice was motivated by the absence of identical studies to ours available from which we could accurately estimate priors and the awareness that publication biases in any case will tend to exaggerate effect sizes. The priors for all Bayes factor analyses were determined based on Jeffrey-Zellner-Siow priors (Zellner and Siow, 1980; Jeffreys, 1998) which are from the Cauchy distribution based on the effect size that is initially calculated in the algorithm using t-test (Rouder et al., 2012). The priors are data-driven and have been shown to be invariant with respect to linear transformations of measurement units (Rouder et al., 2012), which reduces the chance of being biased toward the null or alternative hypotheses. We did not perform correction for multiple comparisons when using Bayes factors as they are much more conservative than frequentist analysis in providing false claims with confidence (Gelman and Tuerlinckx, 2000; Gelman et al., 2012). The reason for this is that properly chosen priors [here using the data-driven approach developed by Rouder et al. (2012)], reduce the chance of making type I (false positive) errors (Gelman and Tuerlinckx, 2000; Gelman et al., 2012).

Random Permutation Testing

To evaluate the significance of correlations between decoding accuracies and behavioral reaction times, we calculated the percentage of the actual correlations that were higher (if positive) or lower (if negative) than a set of 1,000 randomly generated correlations. These random correlations were obtained by randomizing the order of participants' data in the behavioral reaction time vector (null distribution) on every time point, for each feature separately. The correlation was considered significant if surpassed 95% of the randomly generated correlations in the null distribution in either positive or negative directions ($p < 0.05$) and the p-values were corrected for multiple comparisons across time using Matlab mafdr function, where the algorithm fixes the rejection region and then estimates its corresponding error rate resulting in increased accuracy and power (Storey, 2002).

RESULTS

Do Different Ways of Combining Individual Features Affect the Level and Temporal Dynamics of Information Decoding?

As an initial step, we evaluated the level of information which can be obtained from the combination of features, each potentially capturing different aspects of the neural codes. To be as confident as possible, we used a large set of 17 distinct supervised and unsupervised FS methods to select and combine the top 5 most informative features at every time point in the time-resolved decoding procedure. The information content of features were determined based on either how much they could contribute to discriminating the target object categories (supervised) or some predefined criteria which could implicitly suggest more separation between object categories (unsupervised). We split the FS algorithms into three arbitrary groups for the sake of clearer presentation of the results (**Figure 2**).

All FS algorithms for the three datasets showed strong (BF > 10) evidence for difference from chance-level decoding at some time points/windows after the stimulus onset (**Figure 2**). This means that, any of the FS algorithms could combine the features in a way that they could decode object category information from brain signals. As expected from the difference in their mathematical formulations, however, no pairs of FS algorithms provided identical patterns of decoding in any of the three datasets. Consistently across the three datasets there was moderate (3 < BF < 10) or strong (BF > 10) evidence for continuous above-chance decoding from around 80 ms post stimulus onset for all FS algorithms. While the decoding showed evidence for above-chance accuracy (BF > 3) up until 550 ms (dataset 2) or even later than 800 ms (dataset 3) for the best FS algorithms such as UDFS, lasso and ufsol, all curves converged back to the chance-level earlier than 500 ms for dataset 1. This difference may reflect the longer stimulus presentation time for datasets 2 and 3 vs. dataset 1, which may have provided stronger sensory input for neural processing of category information, as we saw previously when evaluating individual features alone (Karimi-Rouzbahani et al., 2021b).

In order to quantitatively compare the decoding curves for the different FS algorithms, we extracted four different amplitude and timing parameters from their decoding curves as in previous studies (Isik et al., 2014): maximum and average decoding accuracies (in the post-stimulus time window), time of maximum decoding, and time of first above-chance decoding relative to stimulus onset (**Supplementary Figure 1**). Results showed that ILFS, relief and llcfs were the worst performing FS algorithms with the lowest maximum and average decoding accuracy (**Supplementary Figures 1A,B**; red boxes). UDFS, lasso and ufsol were the best performing FS algorithms leading to the highest maximum and average decoding accuracies (**Supplementary Figures 1A,B**; black boxes). Dataset 2 tended to yield higher decoding accuracies compared to the other datasets, which might be attributed to the longer presentation time of the stimuli and the active task of the participants (Roth et al., 2020; Karimi-Rouzbahani et al., 2021a,c). UDFS, ufsol and relief were among the earliest FS algorithms to reach their first above-chance and maximum decoding accuracies (**Supplementary Figures 1C,D**). However, there was not a consistent pattern of temporal precedence for any FS algorithms across the datasets.

Which Individual Features Are Selected by the Most Successful Algorithms?

The difference in the decoding patterns for different FS algorithms suggest that they used different sets of features in decoding. To see what features were selected by different FS algorithms, and whether the informative individual features were selected, we calculated the merit of each of the individual features in each FS algorithm across the time course of the trial (**Supplementary Figure 2**). Here, merit refers to the frequency of a feature being selected by the FS algorithm for decoding. We calculated the merit as the ratio of the number of times the feature

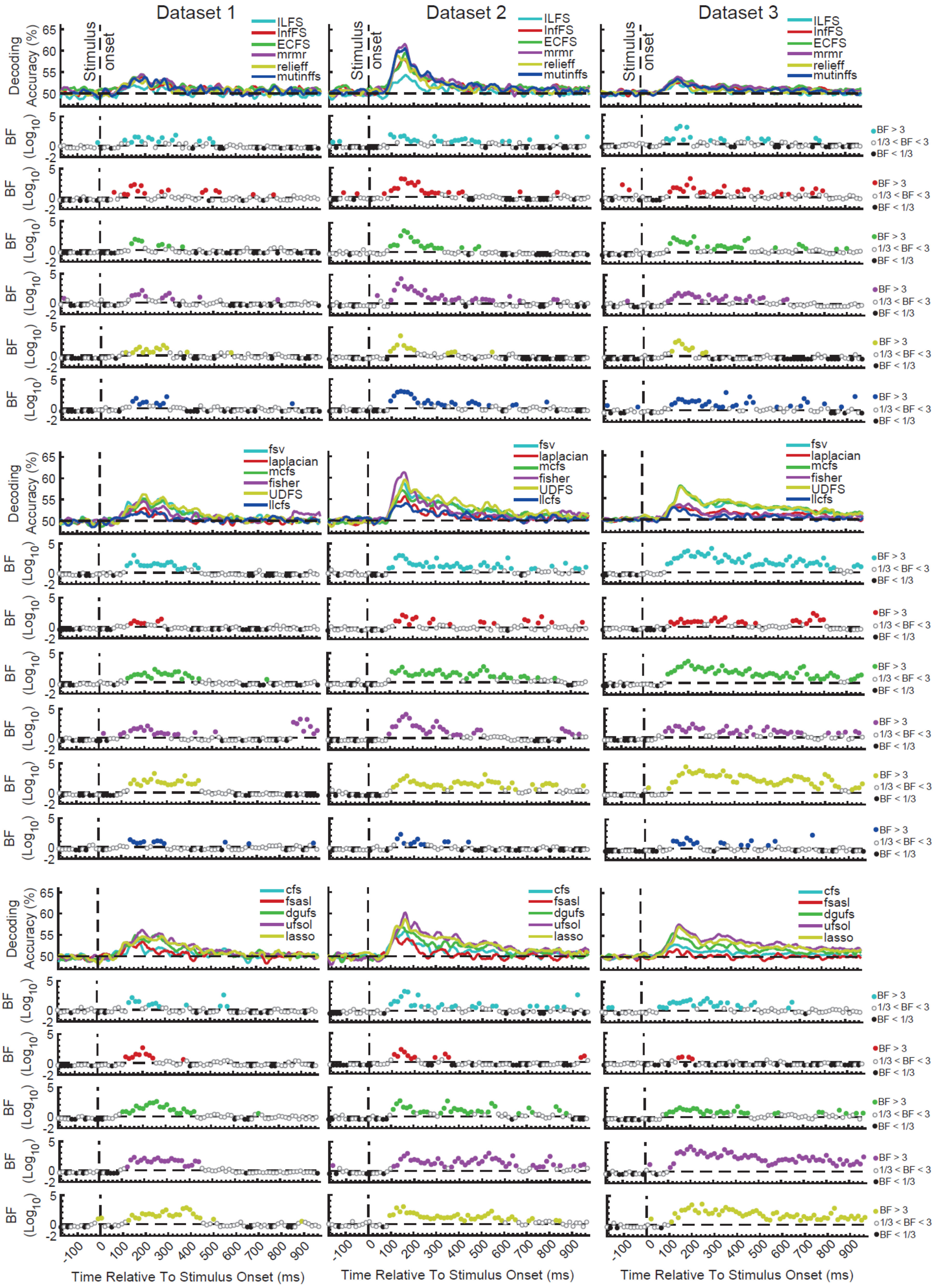


FIGURE 2 | Time-resolved decoding of object categories from the three datasets using the 17 FS methods. We split the FS algorithms into three arbitrary groups (rows) for each dataset for the sake of clearer presentation. Each column shows the results for one dataset. The top section in each of the nine panels shows the decoding accuracies across time and the bottom panels show the Bayes factor evidence for decoding to be different (H1) or not different (H0) from chance-level. The horizontal dashed lines refer to chance-level decoding, the vertical dashed lines indicates time of stimulus onset. Non-black colored filled circles in the Bayes Factors show moderate (BF > 3) or strong (BF > 10) evidence for difference from chance-level decoding, black filled circles show moderate (BF > 3) or strong (BF > 10) evidence for no difference from chance-level decoding and empty circles indicate insufficient evidence (1/3 < BF < 3) for either hypotheses.

was among the top selected five features to the number of times the decoding was performed on every time point (i.e., all possible combination of category pairs).

Visual inspection of the results suggests that each FS algorithm seemed to rely on consistent sets of features across the three datasets, which are generally different between FS algorithms. This reflects that different FS algorithms have different levels of sensitivity and distinct selection criteria. Results also showed that the merit of different features varied across the time course of trials based on their information content about object categories relative to other features (**Supplementary Figure 2**). Therefore, the recruitment of features varied across the time course of the trial: while some features were only temporarily selected (e.g., Average and Mean frequency in the laplacian method from ∼200 to 600 post-stimulus onset), there were features which were constantly used for decoding even before the stimulus onset (e.g., Cros Cor in the fsasl method), although they did not lead to any information decoding in the pre-stimulus time span (**Figure 2**). This might again be explained by the different levels of sensitivity and distinct selection criteria implemented by different FS algorithms. Importantly, the FS algorithms that provided the highest level of decoding (i.e., ufsol, lasso, and UDFS) showed the highest merits for the features of Mean, Median, Samples, and Wavelet which were among the most informative features when evaluated individually across the three datasets (Karimi-Rouzbahani et al., 2021b). On the other hand, the FS algorithms that performed most poorly (ILFS, relief, and llcfs) either used scattered sets of features (ILFS) or did not use the informative features of Mean, Median, Samples and Wavelet (llcfs and relief). Therefore, the FS algorithms that used the informative individual features outperformed other FS algorithms which did not.

Are the Neural Codes Better Captured by a Combinatorial Encoding Protocol or by a General Multiscale Encoding Protocol?

The main question of this study was to see whether the flexibility obtained by the combination of features provides any additional information about object categories compared to the best-performing individual features by detecting the neural codes more completely. In other words, we wanted to test the hypothesis that the brain uses a combination of different neural encoding protocols simultaneously as opposed to using a general multiscale encoding protocol (such as reflected in the Wavelet transform). To test this hypothesis, we directly compared the decoding accuracy obtained from the top performing individual feature from the original study (Wavelet; Karimi-Rouzbahani et al., 2021b), which is able to detect multiscale spatiotemporal patterns of information, with the decoding accuracy obtained from the top performing FS algorithm, which used a set of combined features (ufsol; **Figure 3**). Results showed consistent patterns across the three datasets with the Wavelet feature outperforming the decoding accuracies obtained by the ufsol FS algorithm across most time points. Maximum continuous evidence for difference (BF > 10) occurred between 80 and 320, 75–180, and 85–325 ms for datasets 1–3, respectively. Therefore, it seems that, at least for object categories, the coding scheme in the brain is best captured by a general multiscale encoding protocol (implemented here by the Wavelet coefficients), rather than a combination of distinct encoding protocols (captured here by different features).

Can a Combinatorial Encoding Protocol Predict Behavioral Accuracy Better Than a General Multiscale Encoding Protocol?

Our final hypothesis was that a combinatorial encoding protocol might predict the behavioral performance more accurately than a general multiscale encoding protocol as the former can potentially detect more distinctly encoded neural codes from brain activation. We could test this hypothesis only for Dataset 2 where the task was active and we had the participants' reaction times (i.e., time to categorize objects) to work with. We calculated the (Spearman's rank) correlation between the decoding accuracies and the behavioral reaction time across participants, to see whether, at each time point, participants with higher decoding values were those with the fastest reaction times. We expected to observe negative correlations between the decoding accuracies and the participants' reaction times in the post-stimulus span (Ritchie et al., 2015). Note that since correlation normalizes the absolute level of the input variables, the higher level of decoding for the individual (Wavelet) feature vs. the combined features (ufsol; **Figure 3**) does not necessarily predict a higher correlation for the individual feature of Wavelet.

Results showed significant negative correlations appearing after the stimulus onset for most FS algorithms (except dgufs) especially the laplacian algorithm which showed the most negative peak (**Figure 4A**). This confirms that the distances between object categories in neural representations have inverse relationship to behavioral reaction times (Ritchie et al., 2015). We previously observed that the individual features which provided the highest decoding accuracies could also predict the behavior most accurately (Karimi-Rouzbahani et al., 2021b). Therefore, we asked if the FS algorithms which provided the highest levels of decoding could also predict the behavior more accurately than the less informative algorithms. The rationale behind this hypothesis was that, more effective decoding of neural codes, as measured by higher "average decoding" and "maximum decoding" accuracies (**Figure 2**), should facilitate the prediction of behavior by detecting subtle but overlooked behavior-related neural codes. To test this hypothesis, we evaluated the correlation between the parameters of "maximum decoding" and "average decoding" accuracies (extracted from the decoding curve of each feature in **Figure 4A**) and the "average correlation to behavior" (calculated simply by averaging the correlation to behavior in the post-stimulus time span for each FS algorithm in **Figure 4A**). We also calculated the correlation between the "time of maximum decoding" and "time of first above-chance decoding" as control variables, which we did not expect to correlate with behavior (as in Karimi-Rouzbahani et al., 2021b). Results showed no significant correlations between any of the four parameters of decoding curves and the level of prediction of behavior (**Figure 4B**). Therefore, more efficient combinations

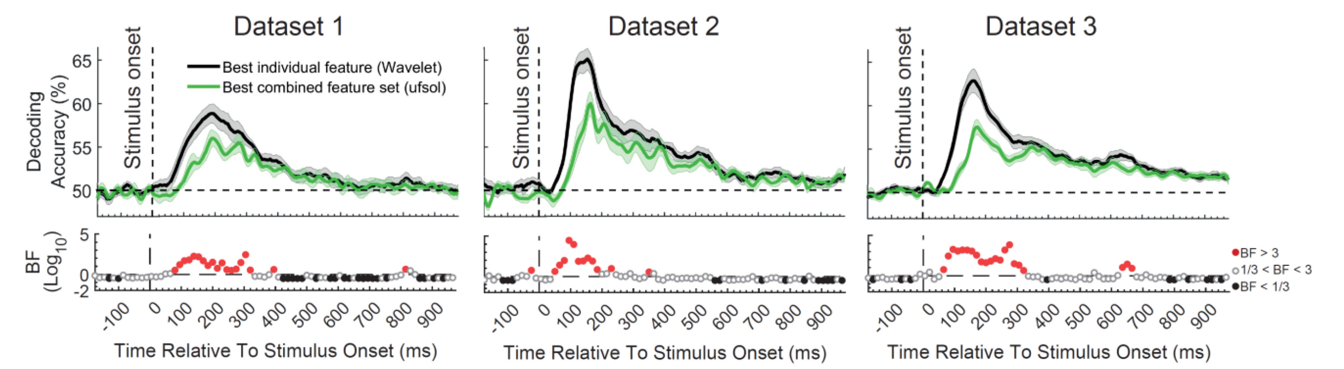


FIGURE 3 | Comparison of decoding accuracies between the most informative individual feature [Wavelet; from Karimi-Rouzbahani et al. (2021b)] and combined feature set (obtained using ufsol algorithm) from the three datasets and Bayesian evidence for a difference between them. Each column shows the results for one dataset. Thick lines show the average decoding accuracy across participants (error bars show Standard Error across participants). Top section in each panel shows the decoding accuracies across time and the bottom section shows the Bayes factor evidence for the difference between the decoding curves. The horizontal dashed lines on the top panels refer to chance-level decoding. Red filled circles in the Bayes Factors show moderate (BF > 3) or strong (BF > 10) evidence for difference between decoding curves, black filled circles show moderate (BF > 3) or strong (BF > 10) evidence for no difference and empty circles indicate insufficient evidence (1/3 < BF < 3) for either hypotheses.

of features (as measured by higher decoding accuracies) did not correspond to more accurate prediction of behavior.

To visually compare the behavioral prediction power of the top-performing individual and combined features we plotted their correlation-to-behavior results on the same figure (**Figure 4C**). For this we selected Wavelet and laplacian FS, based on them being the single feature and FS algorithm with the largest negative peak. We used this, rather than selecting based on average correlation with behavior because the temporal position of the peak can also provide some temporal indication about the timing of the decision, which if reasonable [e.g., after 200 ms post-stimulus and before the median reaction times of participants: 1,146 ms (Karimi-Rouzbahani et al., 2019)], can be more assuring about the existence of true correlation to behavior. The combined features (laplacian) did not provide a negative peak as large as the Wavelet feature, and tended to underperform Wavelet throughout the time course (**Figure 4C**). Therefore, in contradiction to our hypothesis, the combined features did not provide additional prediction of behavior compared to the individual feature of Wavelet.

DISCUSSION

Abstract models of feed-forward visual processing suggest that visual sensory information enters the brain through retina, reaches the lateral geniculate nucleus in thalamus and continues to early visual cortices before moving forward (along the ventral visual stream) to reach the anterior parts of the inferior temporal cortices where semantic information (e.g., about the category of the presented object) is extracted from the visual inputs (DiCarlo et al., 2012). However, two outstanding questions are how neurons along the way encode the information and how this information is reflected in invasively (e.g., LFPs) and non-invasively collected (e.g., EEG) neural data. While in invasively recorded data, researchers have found significant information about visual information in low-frequency power of LFPs (Belitski et al., 2008) or phase-amplitude coupling of electrocorticography (ECoG), there is no reason for these to directly imprint on EEG. In fact, there is evidence that EEG activations represent the information in a feature different [e.g., phase rather than the amplitude of slow (theta band) oscillations] from the invasive neural data such as spiking activity (Ng et al., 2013). Therefore, more detailed investigation of neural coding in EEG seems necessary.

To gain a better understanding of EEG, previous studies have extracted a wide variety of features of neural activations to extract information about visual object categories. However, they have generally used whole-trial analyses, which hide the temporal dynamics of information processing, or time-resolved decoding analyses, or considered the response at each time point separately, ignoring potentially informative temporal features of the time series data. To fill this gap, our previous study extracted and compared a large set of features from EEG in time-resolved analysis (Karimi-Rouzbahani et al., 2021b). However, an outstanding question in the literature was whether the neural code might be best captured by combinations of these features, i.e., if the brain uses a combinatorial encoding protocol to encode different aspects of the sensory input using distinct encoding protocols on the same trial (Gawne et al., 1996; Montemurro et al., 2008). Alternatively, previous invasive neural recording studies have suggested a general multiscale encoding procedure that allows the generation of all the information within the same platform (Victor, 2000; Kayser et al., 2009; Panzeri et al., 2010). To address this question we combined a large set of distinct mathematical features ($n = 26$) of the EEG time series data from three datasets, and combined them using a large set of FS algorithms ($n = 17$), each having different criteria for selection. We compared the performance of different FS algorithms using multivariate decoding of category information. Our results showed that, no matter how we combined the informative features, their combined

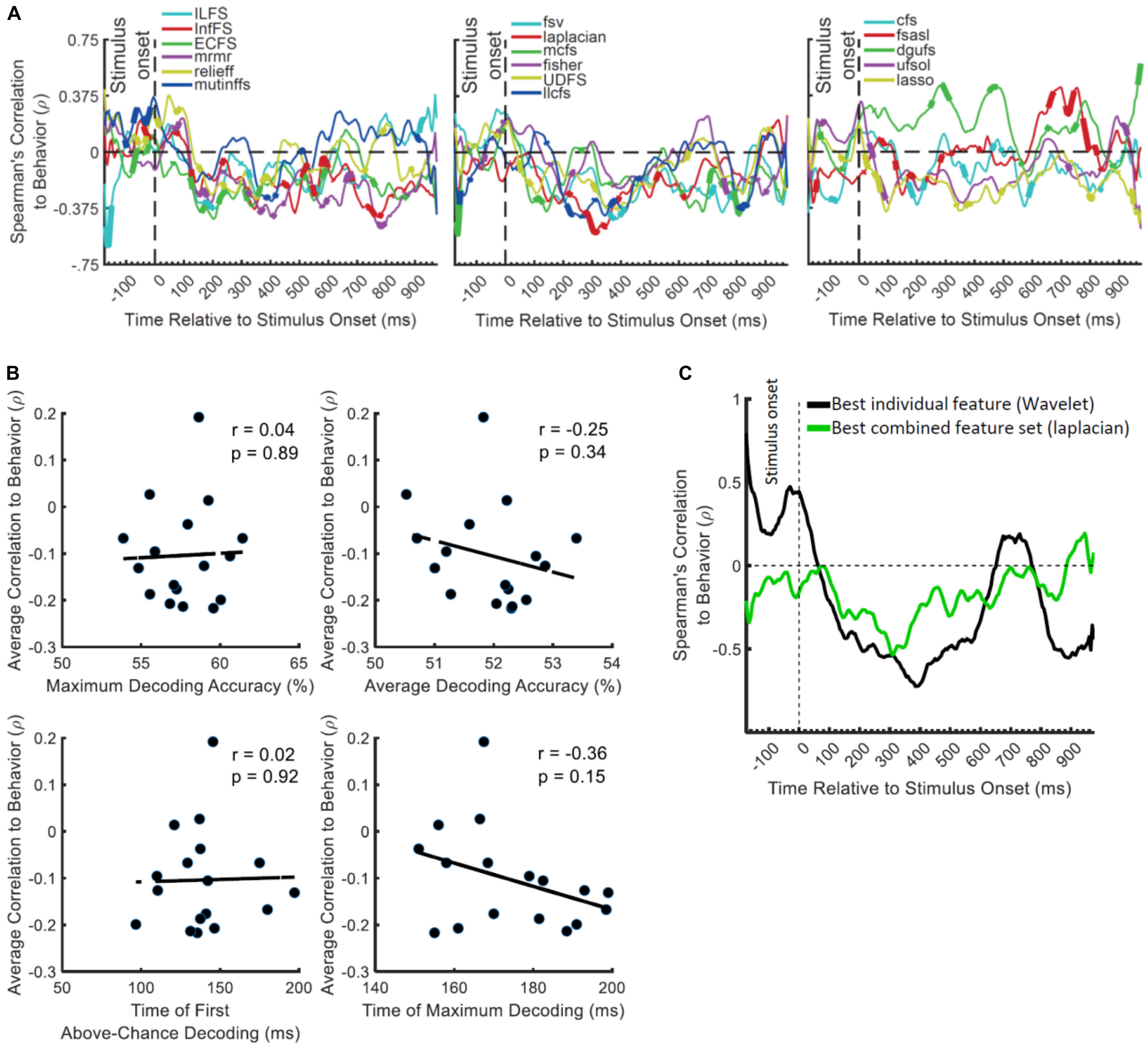


FIGURE 4 | Correlation between the decoding accuracies obtained using 17 FS algorithms and behavioral reaction time of Dataset 2. **(A)** Top section in each panel shows the (Spearman's) correlation coefficient obtained from correlating the decoding values and the reaction times for each feature separately. Thickened time points on the curves indicate time points of positively or negatively significant ($P < 0.05$; corrected for multiple comparisons) correlations as evaluated by random permutation of the variables in correlation. **(B)** Correlation between each of the amplitude and timing parameters of time-resolved decoding (i.e., maximum and average decoding accuracy and time of first and maximum decoding) with the average time-resolved correlations calculated from panel **(A)** for the set of $N = 17$ FS algorithms. The slant line shows the best linear fit to the distribution of the correlation data. **(C)** Correlation between the decoding accuracies obtained from the feature which showed the highest maximum correlation from individual features (Wavelet) and from the combined features (laplacian).

decodable information about object categories, and their power in predicting behavioral performance, was outperformed by the most informative individual feature (i.e., Wavelet), which was sensitive to multi-scale codes from the analysis time window and across electrodes (i.e., spatiotemporal specificity).

The main question of this study was whether the brain recruits and combines a number of different protocols to encode different aspects of cognitive processes involved in object category recognition ranging from sensory information to behavioral response. For example, the brain may use one encoding protocol for the encoding of feed-forward visual information processing, e.g., theta-band power, which would later in the trial be dominated by alpha/beta-band feedback information flow involved in semantic object categorization (Bastos et al., 2015). The brain may also use different encoding protocols to process different aspects of the same stimulus [e.g., contrast or the orientation of visual stimulus (Gawne et al., 1996)]. Alternatively, the brain may implement a single but multiscale protocol [e.g., multiplexing strategy which combines the codes at different time scales (Panzeri et al., 2010)] which allows different aspects of information to be represented within the same encoding protocol. Our results provide support for the latter by showing that spatiotemporally sensitive features, which can detect patterns across multiple scales (e.g., Wavelet coefficients) best capture variance in the EEG responses evoked by different categories of visual objects. Therefore, rather than a combinatorial and switching encoding protocol, the brain may instead encode object category information through a single but multiscale encoding protocol.

This study does not provide the first evidence showing that temporal patterns of activity provide information about different aspects of visual sensory input. The richness of information in the temporal patterns of activity has been previously observed in light encoding (Gollisch and Meister, 2008), co-occurrences of visual edges (Eckhorn et al., 1988), orientations in primary visual cortex (Celebrini et al., 1993) as well as object category information in the temporal cortex (Majima et al., 2014). While we do not claim that this EEG study provides direct evidence about processing of information at the level of single neurons, our findings are consistent with the above invasively-recorded neural data and provide evidence for information content in neural variability of EEG data. Our study also aligns with the recent move toward incorporating within- and across-trial temporal variability in the decoding of information from neural time series such as MEG (Vidaurre et al., 2019), EEG (Majima et al., 2014), invasive electrophysiological (Orbán et al., 2016) and even fMRI (Garrett et al., 2020) data. On the other hand, this current study contrasts with the conventional time-resolved decoding analyses which merely consider amplitude at each time point (Grootswagers et al., 2017), overlooking informative multi-scale temporal codes.

The field of Brain-Computer Interface (BCI) has already achieved great success in decoding visually evoked information from EEG representations in the past two decades, mainly through the use of rigorous supervised learning algorithms [e.g., Voltage Topographies (Tzovara et al., 2012), Independent Component Analysis (Stewart et al., 2014), Common Spatial Patterns (Murphy et al., 2011), and Convolutional Neural Networks (Seeliger et al., 2018)] or by combining multiple features (Chan et al., 2011; Wang et al., 2012; Qin et al., 2016; Torabi et al., 2017). However, the predictive power of a feature about behavior might not be as important for BCI where the goal is to maximize the accuracy of the commands sent to a computer or an actuator. In contrast, one of the most critical questions in cognitive neuroscience to understand whether the neural signatures that we observe are meaningful in bringing about behavior, as opposed to being epiphenomenal to our experimental setup (e.g., Williams et al., 2007; Jacobs et al., 2009; Ritchie et al., 2015; Hebart and Baker, 2018; Woolgar et al., 2019; Karimi-Rouzbahani et al., 2021a,b). To address this point, we evaluated whether our extracted features and their combinations were behaviorally relevant, by correlating our decoding patterns with the behavioral object recognition performance (reaction times in Dataset 2). Moreover, to directly compare the information content of the combined feature sets with the individual features, we equalized the dimensions of the data matrix for the FS algorithm to that obtained for individual features. This avoided artefactualy improving behavioral predictive power with higher dimensionality. Contrary to what we predicted, however, we observed that even the laplacian FS algorithm, which provided the best peak prediction for the behavioral performance, was outperformed by the individual Wavelet feature at most time points. Therefore, the multiscale feature of Wavelet not only provides the most decodable information, but seems to most closely reflect the neural processes involved in generating participant behavior.

One unique property of our decoding pipeline, which we believe led to the enhanced information encoding for the Wavelet feature relative to other individual features (Karimi-Rouzbahani et al., 2021b), is the incorporation of *spatiotemporal* codes in decoding in each 50 ms analysis window. The neural code can be represented in either time (across the analysis time window), space (across electrodes in EEG) or a combination of both (Panzeri et al., 2010). Specifically, most of the previous studies have evaluated the neural codes in either time, being limited by the nature of their invasive recording modality (Houweling and Brecht, 2008; Benucci et al., 2009), or space by averaging/down-sampling of data within the analysis window. However, our spatiotemporal concatenation of EEG activity across both time and electrodes (i.e., performed at the first PCA stage for individual features and at the third PCA stage for the combined features in **Figure 1**), allows the neural codes to be detected from both spatially and temporally informative patterns. The 50 ms time window chosen here makes a compromise between concatenating and decoding the whole time window in one shot, which loses the temporal resolution, and time-resolved decoding at each time point, which ignores temporal patterns of activity (Karimi-Rouzbahani et al., 2021b).

While this study provided insights about how neural codes might be detected from EEG activations, there remain two main limitations in understanding the nature of neural codes in EEG. First, physiological evidence is limited about how neurons produce, often such complicated codes, even in studies where the mathematical features of this study were first introduced. There are theories and mathematical justifications to explain why these complicated codes are helpful (Schaefer et al., 2006; Kayser et al., 2009; Schroeder and Lakatos, 2009, etc.) but not on how neurons produce them. Second, it seems unlikely that the distinctly-defined mathematical features necessarily extract distinct attributes/neural codes. In fact, many of the extracted features overlap: some of them are slightly different ways of quantifying similar characteristics of the neural activity (e.g., variance vs. power, which both quantify the strength of variability of the signal). Therefore, there are not necessarily distinct neural underpinnings for each feature.

There are several future directions for this research. First, as the encoding protocols for different cognitive processes might be different from object category processing (Panzeri et al., 2010), the generalization of our results to other domains of cognitive neuroscience needs to be evaluated. Second, previous results (Panzeri et al., 2010) suggest that different aspects of information (e.g., category processing, decision making and motor response) may be encoded using different encoding protocols. Our data did not allow us to tease those aspects apart, which is interesting area for future investigation. Third, following previous suggestions that even different aspects of *visual* information (e.g., color, variations, and task) might also be encoded using different encoding protocols (Gawne et al., 1996), the number of selected features might need to be varied from one dataset to another. Ideally, we would only keep the informative features above a certain threshold. Here, we chose an arbitrary threshold of 5 included, but it would be interesting to explore the impact of this parameter in the future.

The large-scale EEG analysis of this study aligns with the recent shift to cross-dataset meta-analyses for different human cognitive abilities such as working memory (Adam et al., 2020) and sustained attention (Langner and Eickhoff, 2013). Such studies lead to more generalizable conclusions and provide deeper insights into the human cognition. Here, across three very different datasets we showed that, the brain seems to implement a temporally and spatially flexible and multiscale encoding strategy rather than a combinatorial or switching encoding strategy, at least in object category processing.

DATA AVAILABILITY STATEMENT

Datasets 1 and 2 of this study are available online at https://osf.io/wbvpn/ and dataset 3 at https://exhibits.stanford.edu/data/catalog/tc919dd5388.

ETHICS STATEMENT

The datasets used in this study were obtained from experiments that were approved by Shahid Rajaee University Ethics Committee, Iran, and Institutional Review Board of Stanford University, United States. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

HK-R: conceptualization, methodology, formal analysis, writing – original draft, visualization, data curation, and funding acquisition. AW: writing – review and editing and funding acquisition. Both authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by the UK Royal Society's Newton International Fellowship NIF\R\192608 to HK-R and MRC intramural funding SUAG/052/G101400 to AW.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2022.825746/full#supplementary-material

REFERENCES

Adam, K. C., Vogel, E. K., and Awh, E. (2020). Multivariate analysis reveals a generalizable human electrophysiological signature of working memory load. *bioRxiv* [Preprint]. doi: 10.1111/psyp.13691

Ahmadi-Pajouh, M. A., Ala, T. S., Zamanian, F., Namazi, H., and Jafari, S. (2018). Fractal-based classification of human brain response to living and non-living visual stimuli. *Fractals* 26:1850069.

Alimardani, F., Cho, J. H., Boostani, R., and Hwang, H. J. (2018). Classification of bipolar disorder and schizophrenia using steady-state visual evoked potential based features. *IEEE Access* 6, 40379–40388.

Bastos, A. M., Vezoli, J., Bosman, C. A., Schoffelen, J. M., Oostenveld, R., Dowdall, J. R., et al. (2015). Visual areas exert feedforward and feedback influences through distinct frequency channels. *Neuron* 85, 390–401. doi: 10.1016/j.neuron.2014.12.018

Belitski, A., Gretton, A., Magri, C., Murayama, Y., Montemurro, M. A., Logothetis, N. K., et al. (2008). Low-frequency local field potentials and spikes in primary visual cortex convey independent visual information. *J. Neurosci.* 28, 5696–5709. doi: 10.1523/JNEUROSCI.0009-08.2008

Benucci, A., Ringach, D. L., and Carandini, M. (2009). Coding of stimulus sequences by population responses in visual cortex. *Nat. Neurosci.* 12, 1317–1324. doi: 10.1038/nn.2398

Carlson, T., Tovar, D. A., Alink, A., and Kriegeskorte, N. (2013). Representational dynamics of object vision: the first 1000 ms. *J. Vis.* 13:1. doi: 10.1167/13.10.1

Celebrini, S., Thorpe, S., Trotter, Y., and Imbert, M. (1993). Dynamics of orientation coding in area V1 of the awake primate. *Vis. Neurosci.* 10, 811–825. doi: 10.1017/s0952523800006052

Chan, A. M., Halgren, E., Marinkovic, K., and Cash, S. S. (2011). Decoding word and category-specific spatiotemporal representations from MEG and EEG. *Neuroimage* 54, 3028–3039. doi: 10.1016/j.neuroimage.2010.10.073

Cichy, R. M., Pantazis, D., and Oliva, A. (2014). Resolving human object recognition in space and time. *Nat. Neurosci.* 17:455. doi: 10.1038/nn.3635

Contini, E. W., Wardle, S. G., and Carlson, T. A. (2017). Decoding the time-course of object recognition in the human brain: from visual features to categorical decisions. *Neuropsychologia* 105, 165–176. doi: 10.1016/j.neuropsychologia.2017.02.013

da Silva, F. L. (2013). EEG and MEG: relevance to neuroscience. *Neuron* 80, 1112–1128. doi: 10.1016/j.neuron.2013.10.017

Dash, M., and Liu, H. (1997). Feature selection for classification. *Intell. Data Anal.* 1, 131–156.

DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron* 73, 415–434. doi: 10.1016/j.neuron.2012.01.010

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Front. Psychol.* 5:781. doi: 10.3389/fpsyg.2014.00781

Eckhorn, R., Bauer, R., Jordan, W., Brosch, M., Kruse, W., Munk, M., et al. (1988). Coherent oscillations: a mechanism of feature linking in the visual cortex? *Biol. Cybernet.* 60, 121–130. doi: 10.1007/BF00202899

Garrett, D. D., Epp, S. M., Kleemeyer, M., Lindenberger, U., and Polk, T. A. (2020). Higher performers upregulate brain signal variability in response to more feature-rich visual input. *Neuroimage* 217:116836. doi: 10.1016/j.neuroimage.2020.116836

Gawne, T. J., Kjaer, T. W., and Richmond, B. J. (1996). Latency: another potential code for feature binding in striate cortex. *J. Neurophysiol.* 76, 1356–1360. doi: 10.1152/jn.1996.76.2.1356

Gelman, A., and Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Comput. Stat.* 15, 373–390.

Gelman, A., Hill, J., and Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *J. Res. Educ. Effect.* 5, 189–211.

Gollisch, T., and Meister, M. (2008). Rapid neural coding in the retina with relative spike latencies. *Science* 319, 1108–1111. doi: 10.1126/science.1149639

Grootswagers, T., Wardle, S. G., and Carlson, T. A. (2017). Decoding dynamic brain patterns from evoked responses: a tutorial on multivariate pattern analysis applied to time series neuroimaging data. *J. Cogn. Neurosci.* 29, 677–697. doi: 10.1162/jocn_a_01068

Guo, L., Rivero, D., Seoane, J. A., and Pazos, A. (2009). "Classification of EEG signals using relative wavelet energy and artificial neural networks," in *Proceedings of the first ACM/SIGEVO Summit on Genetic and Evolutionary Computation*, Shanghai China, 177–184. doi: 10.3233/SHTI210538

Hebart, M. N., and Baker, C. I. (2018). Deconstructing multivariate decoding for the study of brain function. *Neuroimage* 180, 4–18. doi: 10.1016/j.neuroimage.2017.08.005

Hermundstad, A. M., Briguglio, J. J., Conte, M. M., Victor, J. D., Balasubramanian, V., and Tkačik, G. (2014). Variance predicts salience in central sensory processing. *Elife* 3:e03722. doi: 10.7554/eLife.03722

Houweling, A. R., and Brecht, M. (2008). Behavioural report of single neuron stimulation in somatosensory cortex. *Nature* 451, 65–68. doi: 10.1038/nature06447

Hung, C. P., Kreiman, G., Poggio, T., and DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science* 310, 863–866. doi: 10.1126/science.1117593

Iranmanesh, S., and Rodriguez-Villegas, E. (2017). An ultralow-power sleep spindle detection system on chip. *IEEE Trans. Biomed. Circuits Syst.* 11, 858–866. doi: 10.1109/TBCAS.2017.2690908

Isik, L., Meyers, E. M., Leibo, J. Z., and Poggio, T. (2014). The dynamics of invariant object recognition in the human visual system. *J. Neurophysiol.* 111, 91–102. doi: 10.1152/jn.00394.2013

Jacobs, A. L., Fridman, G., Douglas, R. M., Alam, N. M., Latham, P. E., Prusky, G. T., et al. (2009). Ruling out and ruling in neural codes. *Proc. Natl. Acad. Sci.* 106, 5936–5941. doi: 10.1073/pnas.0900573106

Jadidi, A. F., Zargar, B. S., and Moradi, M. H. (2016). "Categorizing visual objects; using ERP components," in *2016 23rd Iranian Conference on Biomedical Engineering and 2016 1st International Iranian Conference on Biomedical Engineering ICBME* (Piscataway, NJ: IEEE), 159–164.

Jeffreys, H. (1998). *The Theory Of Probability*. Oxford: Oxford University Press.

Joshi, D., Panigrahi, B. K., Anand, S., and Santhosh, J. (2018). Classification of targets and distractors present in visual hemifields using time-frequency domain EEG features. *J. Healthc. Eng.* 2018, 1–10. doi: 10.1155/2018/9213707

Kaneshiro, B., Guimaraes, M. P., Kim, H. S., Norcia, A. M., and Suppes, P. (2015). A representational similarity analysis of the dynamics of object processing using single-trial EEG classification. *PLoS One* 10:e0135697. doi: 10.1371/journal.pone.0135697

Karimi-Rouzbahani, H. (2018). Three-stage processing of category and variation information by entangled interactive mechanisms of peri-occipital and peri-frontal cortices. *Sci. Rep.* 8, 1–22. doi: 10.1038/s41598-018-30601-8

Karimi-Rouzbahani, H., Bagheri, N., and Ebrahimpour, R. (2017b). Hard-wired feed-forward visual mechanisms of the brain compensate for affine variations in object recognition. *Neuroscience* 349, 48–63. doi: 10.1016/j.neuroscience.2017.02.050

Karimi-Rouzbahani, H., Bagheri, N., and Ebrahimpour, R. (2017a). Average activity, but not variability, is the dominant factor in the representation of object categories in the brain. *Neuroscience* 346, 14–28. doi: 10.1016/j.neuroscience.2017.01.002

Karimi-Rouzbahani, H., Ramezani, F., Woolgar, A., Rich, A., and Ghodrati, M. (2021a). Perceptual difficulty modulates the direction of information flow in familiar face recognition. *Neuroimage* 233:117896. doi: 10.1016/j.neuroimage.2021.117896

Karimi-Rouzbahani, H., Shahmohammadi, M., Vahab, E., Setayeshi, S., and Carlson, T. (2021b). Temporal variabilities provide additional category-related information in object category decoding: a systematic comparison of informative EEG features. *Neural Comput.* 33, 3027–3072. doi: 10.1162/neco_a_01436

Karimi-Rouzbahani, H., Woolgar, A., and Rich, A. N. (2021c). Neural signatures of vigilance decrements predict behavioural errors before they occur. *ELife* 10:e60563. doi: 10.7554/eLife.60563

Karimi-Rouzbahani, H., Vahab, E., Ebrahimpour, R., and Menhaj, M. B. (2019). Spatiotemporal analysis of category and target-related information processing in the brain during object detection. *Behav. Brain Res.* 362, 224–239. doi: 10.1016/j.bbr.2019.01.025

Kayser, C., Montemurro, M. A., Logothetis, N. K., and Panzeri, S. (2009). Spike-phase coding boosts and stabilizes information carried by spatial and temporal spike patterns. *Neuron* 61, 597–608. doi: 10.1016/j.neuron.2009.01.008

Langner, R., and Eickhoff, S. B. (2013). Sustaining attention to simple tasks: a meta-analytic review of the neural mechanisms of vigilant attention. *Psychol. Bull.* 139:870. doi: 10.1037/a0030694

Lee, M. D., and Wagenmakers, E. J. (2005). Bayesian statistical inference in psychology: comment on trafimow (2003). *Psychol. Rev.* 112, 662–668. doi: 10.1037/0033-295X.112.3.662

Lempel, A., and Ziv, J. (1976). On the complexity of finite sequences. *IEEE Trans. Inform. Theor.* 22, 75–81.

Liu, H., Agam, Y., Madsen, J. R., and Kreiman, G. (2009). Timing, timing, timing: fast decoding of object information from intracranial field potentials in human visual cortex. *Neuron* 62, 281–290. doi: 10.1016/j.neuron.2009.02.025

Majima, K., Matsuo, T., Kawasaki, K., Kawai, K., Saito, N., Hasegawa, I., et al. (2014). Decoding visual object categories from temporal correlations of ECoG signals. *Neuroimage* 90, 74–83. doi: 10.1016/j.neuroimage.2013.12.020

Mazaheri, A., and Jensen, O. (2008). Asymmetric amplitude modulations of brain oscillations generate slow evoked responses. *J. Neurosci.* 28, 7781–7787. doi: 10.1523/JNEUROSCI.1631-08.2008

Miyakawa, N., Majima, K., Sawahata, H., Kawasaki, K., Matsuo, T., Kotake, N., et al. (2018). Heterogeneous redistribution of facial subcategory information within and outside the face-selective domain in primate inferior temporal cortex. *Cereb. Cortex* 28, 1416–1431. doi: 10.1093/cercor/bhx342

Montemurro, M. A., Rasch, M. J., Murayama, Y., Logothetis, N. K., and Panzeri, S. (2008). Phase-of-firing coding of natural visual stimuli in primary visual cortex. *Curr. Biol.* 18, 375–380. doi: 10.1016/j.cub.2008.02.023

Murphy, B., Poesio, M., Bovolo, F., Bruzzone, L., Dalponte, M., and Lakany, H. (2011). EEG decoding of semantic category reveals distributed representations for single concepts. *Brain Lang.* 117, 12–22. doi: 10.1016/j.bandl.2010.09.013

Muthukumaraswamy, S. (2013). High-frequency brain activity and muscle artifacts in MEG/EEG: a review and recommendations. *Front. Hum. Neurosci.* 7:138. doi: 10.3389/fnhum.2013.00138

Namazi, H., Ala, T. S., and Bakardjian, H. (2018). Decoding of steady-state visual evoked potentials by fractal analysis of the electroencephalographic (EEG) signal. *Fractals* 26:1850092.

Ng, B. S. W., Logothetis, N. K., and Kayser, C. (2013). EEG phase patterns reflect the selectivity of neural firing. *Cereb. Cortex* 23, 389–398. doi: 10.1093/cercor/bhs031

Orbán, G., Berkes, P., Fiser, J., and Lengyel, M. (2016). Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron* 92, 530–543. doi: 10.1016/j.neuron.2016.09.038

Panzeri, S., Brunel, N., Logothetis, N. K., and Kayser, C. (2010). Sensory neural codes using multiplexed temporal scales. *Trends Neurosci.* 33, 111–120. doi: 10.1016/j.tins.2009.12.001

Pouryazdian, S., and Erfanian, A. (2009). "Detection of steady-state visual evoked potentials for brain-computer interfaces using PCA and high-order statistics," in *Proceddings of the World Congress on Medical Physics and Biomedical Engineering, September 7-12, 2009* (Munich: Springer), 480–483.

Pulini, A. A., Kerr, W. T., Loo, S. K., and Lenartowicz, A. (2019). Classification accuracy of neuroimaging biomarkers in attention-deficit/hyperactivity disorder: effects of sample size and circular analysis. *Biol. Psychiatr. Cogn. Neurosci. Neuroimag.* 4, 108–120. doi: 10.1016/j.bpsc.2018.06.003

Qin, Y., Zhan, Y., Wang, C., Zhang, J., Yao, L., Guo, X., et al. (2016). Classifying four-category visual objects using multiple ERP components in single-trial ERP. *Cogn. Neurodynam.* 10, 275–285. doi: 10.1007/s11571-016-9378-0

Racine, R. (2011). *Estimating the Hurst Exponent*. Zurich: Mosaic Group.

Richman, J. S., and Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *Am. J. Physiol. Heart Circ. Physiol.* 278, H2039–H2049. doi: 10.1152/ajpheart.2000.278.6.H2039

Ritchie, J. B., Tovar, D. A., and Carlson, T. A. (2015). Emerging object representations in the visual system predict reaction times for categorization. *PLoS Comput. Biol.* 11:e1004316. doi: 10.1371/journal.pcbi.1004316

Roffo, G. (2016). Feature selection library (MATLAB toolbox). *arXiv* [Preprint]. 1607.01327

Roth, Z. N., Ryoo, M., and Merriam, E. P. (2020). Task-related activity in human visual cortex. *PLoS Biol.* 18:e3000921. doi: 10.1371/journal.pbio.3000921

Rouder, J. N., Morey, R. D., Speckman, P. L., and Province, J. M. (2012). Default Bayes factors for ANOVA designs. *J. Math. Psychol.* 56, 356–374.

Rupp, K., Roos, M., Milsap, G., Caceres, C., Ratto, C., Chevillet, M., et al. (2017). Semantic attributes are encoded in human electrocorticographic signals during visual object recognition. *Neuroimage* 148, 318–329. doi: 10.1016/j.neuroimage.2016.12.074

Schaefer, A. T., Angelo, K., Spors, H., and Margrie, T. W. (2006). Neuronal oscillations enhance stimulus discrimination by ensuring action potential precision. *PLoS Biol.* 4:e163. doi: 10.1371/journal.pbio.0040163

Schroeder, C. E., and Lakatos, P. (2009). Low-frequency neuronal oscillations as instruments of sensory selection. *Trends Neurosci.* 32, 9–18. doi: 10.1016/j.tins.2008.09.012

Seeliger, K., Fritsche, M., Güçlü, U., Schoenmakers, S., Schoffelen, J. M., Bosch, S. E., et al. (2018). Convolutional neural network-based encoding and decoding of visual object recognition in space and time. *Neuroimage* 180, 253–266. doi: 10.1016/j.neuroimage.2017.07.018

Shourie, N., Firoozabadi, M., and Badie, K. (2014). Analysis of EEG signals related to artists and nonartists during visual perception, mental imagery, and rest using approximate entropy. *BioMed Res. Int.* 2014:764382. doi: 10.1155/2014/764382

Simanova, I., Van Gerven, M., Oostenveld, R., and Hagoort, P. (2010). Identifying object categories from event-related EEG: toward decoding of conceptual representations. *PLoS One* 5:e14465. doi: 10.1371/journal.pone.0014465

Stewart, A. X., Nuthmann, A., and Sanguinetti, G. (2014). Single-trial classification of EEG in a visual object task using ICA and machine learning. *J. Neurosci. methods* 228, 1–14. doi: 10.1016/j.jneumeth.2014.02.014

Storey, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc.* 64, 479–498.

Szczepański, J., Amigó, J. M., Wajnryb, E., and Sanchez-Vives, M. V. (2003). Application of Lempel–Ziv complexity to the analysis of neural discharges. *Network* 14:335.

Taghizadeh-Sarabi, M., Daliri, M. R., and Niksirat, K. S. (2015). Decoding objects of basic categories from electroencephalographic signals using wavelet transform and support vector machines. *Brain Topogr.* 28, 33–46. doi: 10.1007/s10548-014-0371-9

Torabi, A., Jahromy, F. Z., and Daliri, M. R. (2017). Semantic category-based classification using nonlinear features and wavelet coefficients of brain signals. *Cogn. Comput.* 9, 702–711.

Tzovara, A., Murray, M. M., Plomp, G., Herzog, M. H., Michel, C. M., and De Lucia, M. (2012). Decoding stimulus-related information from single-trial EEG responses based on voltage topographies. *Pattern Recognit.* 45, 2109–2122.

Victor, J. D. (2000). How the brain uses time to represent and process visual information. *Brain Res.* 886, 33–46. doi: 10.1016/s0006-8993(00)02751-7

Vidaurre, D., Myers, N. E., Stokes, M., Nobre, A. C., and Woolrich, M. W. (2019). Temporally unconstrained decoding reveals consistent but time-varying stages of stimulus processing. *Cereb. Cortex* 29, 863–874. doi: 10.1093/cercor/bhy290

Voloh, B., Oemisch, M., and Womelsdorf, T. (2020). Phase of firing coding of learning variables across the fronto-striatal network during feature-based learning. *Nat. Commun.* 11, 1–16. doi: 10.1038/s41467-020-18435-3

Wairagkar, M., Zoulias, I., Oguntosin, V., Hayashi, Y., and Nasuto, S. (2016). "Movement intention based Brain Computer Interface for Virtual Reality and Soft Robotics rehabilitation using novel autocorrelation analysis of EEG," in *Proceedings of the 2016 6th IEEE International Conference on Biomedical Robotics and Biomechatronics (BioRob)* (Piscataway, NJ: IEEE), 685–685.

Wang, C., Xiong, S., Hu, X., Yao, L., and Zhang, J. (2012). Combining features from ERP components in single-trial EEG for discriminating four-category visual objects. *J. Neural Eng.* 9:056013. doi: 10.1088/1741-2560/9/5/056013

Wang, Y., Wang, P., and Yu, Y. (2018). Decoding english alphabet letters using EEG phase information. *Front. Neurosci.* 12:62. doi: 10.3389/fnins.2018.00062

Wark, B., Fairhall, A., and Rieke, F. (2009). Timescales of inference in visual adaptation. *Neuron* 61, 750–761. doi: 10.1016/j.neuron.2009.01.019

Waschke, L., Kloosterman, N. A., Obleser, J., and Garrett, D. D. (2021). Behavior needs neural variability. *Neuron* 109, 751–766. doi: 10.1016/j.neuron.2021.01.023

Watrous, A. J., Deuker, L., Fell, J., and Axmacher, N. (2015). Phase-amplitude coupling supports phase coding in human ECoG. *Elife* 4:e07886.

Williams, M. A., Dang, S., and Kanwisher, N. G. (2007). Only some spatial patterns of fMRI response are read out in task performance. *Nat. Neurosci.* 10, 685–686. doi: 10.1038/nn1900

Wong, K. F. K., Galka, A., Yamashita, O., and Ozaki, T. (2006). Modelling non-stationary variance in EEG time series by state space GARCH model. *Comput. Biol. Med.* 36, 1327–1335. doi: 10.1016/j.compbiomed.2005.10.001

Woolgar, A., Dermody, N., Afshar, S., Williams, M. A., and Rich, A. N. (2019). Meaningful patterns of information in the brain revealed through analysis of errors. *bioRxiv* [Preprint].

Zellner, A., and Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. *Trabajos De Estadística y De Investigación Operativa* 31, 585–603.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 10 March 2022
doi: 10.3389/fnins.2022.755988

Caveats and Nuances of Model-Based and Model-Free Representational Connectivity Analysis

Hamid Karimi-Rouzbahani[1,2]*, *Alexandra Woolgar*[1], *Richard Henson*[1,3] *and Hamed Nili*[4,5]*

[1] MRC Cognition and Brain Sciences Unit, University of Cambridge, Cambridge, United Kingdom, [2] Department of Computing, Macquarie University, Sydney, NSW, Australia, [3] Department of Psychiatry, University of Cambridge, Cambridge, United Kingdom, [4] Department of Excellence for Neural Information Processing, Center for Molecular Neurobiology (ZMNH), University Medical Center Hamburg-Eppendorf (UKE), Hamburg, Germany, [5] Wellcome Centre for Integrative Neuroimaging, University of Oxford, Oxford, United Kingdom

OPEN ACCESS

Edited by:
Reza Lashgari,
Shahid Beheshti University, Iran

Reviewed by:
Alex Clarke,
University of Cambridge,
United Kingdom
Alessandro Principe,
Hospital del Mar, Spain

***Correspondence:**
Hamid Karimi-Rouzbahani
hamid.karimi-rouzbahani@mrc-cbu.cam.ac.uk
Hamed Nili
h.nili@uke.de

Specialty section:
This article was submitted to
Brain Imaging Methods,
a section of the journal
Frontiers in Neuroscience

Received: 09 August 2021
Accepted: 02 February 2022
Published: 10 March 2022

Citation:
Karimi-Rouzbahani H, Woolgar A, Henson R and Nili H (2022) Caveats and Nuances of Model-Based and Model-Free Representational Connectivity Analysis. Front. Neurosci. 16:755988. doi: 10.3389/fnins.2022.755988

Brain connectivity analyses have conventionally relied on statistical relationship between one-dimensional summaries of activation in different brain areas. However, summarizing activation patterns within each area to a single dimension ignores the potential statistical dependencies between their multi-dimensional activity patterns. Representational Connectivity Analyses (RCA) is a method that quantifies the relationship between multi-dimensional patterns of activity without reducing the dimensionality of the data. We consider two variants of RCA. In model-free RCA, the goal is to quantify the shared information for two brain regions. In model-based RCA, one tests whether two regions have shared information about a specific aspect of the stimuli/task, as defined by a model. However, this is a new approach and the potential caveats of model-free and model-based RCA are still understudied. We first explain how model-based RCA detects connectivity through the lens of models, and then present three scenarios where model-based and model-free RCA give discrepant results. These conflicting results complicate the interpretation of functional connectivity. We highlight the challenges in three scenarios: complex intermediate models, common patterns across regions, and transformation of representational structure across brain regions. The article is accompanied by scripts (https://osf.io/3nxfa/) that reproduce the results. In each case, we suggest potential ways to mitigate the difficulties caused by inconsistent results. The results of this study shed light on some understudied aspects of RCA, and allow researchers to use the method more effectively.

Keywords: representational connectivity analysis, multi-dimensional connectivity, functional connectivity, multivariate pattern analysis, representational similarity analysis

INTRODUCTION

To study the neural underpinnings of cognitive processes, we need not only to characterize the response of individual brain regions but understand the functional connectivity between them. This is critical to understand how brain regions interact in giving rise to cognition (Bressler and Menon, 2010). Functional connectivity across the brain has been conventionally evaluated using univariate/one-dimensional analyses (Bastos and Schoffelen, 2016). In these analyses, responses in each brain region is initially summarized by a one-dimensional metric (Biswal et al., 1995). If the 1D metrics for different regions are statistically related, we then infer functional connectivity between them (Bastos and Schoffelen, 2016). For example, methods such as gamma-band synchronization (Gregoriou et al., 2009), phase covariance across regions (Bar et al., 2006), frequency coupling (Karimi-Rouzbahani et al., 2021b), and differential equations (Friston et al., 2013) have been used to evaluate connectivity after summarizing the activation patterns across vertices (sensors or voxels) within each region. However, univariate connectivity analysis can miss connectivity if the pairs of regions are statistically related through multi-dimensional patterns of activation rather than the summarized (e.g., averaged) activation within each region (Coutanche, 2013; Basti et al., 2019, 2020). For example, for heterogeneous ROIs, where multiple response modes co-exist, projecting multivariate response patterns on a line (one dimension) could lead to strong distortions. This has led to a recent shift from univariate to multi-dimensional (multivariate) connectivity analyses (Coutanche and Thompson-Schill, 2013; Goddard et al., 2016; Anzellotti and Coutanche, 2018; Basti et al., 2019, 2020; Karimi-Rouzbahani et al., 2021a,c; Shahbazi et al., 2021). One approach to multi-dimensional connectivity is Representational Connectivity Analysis (RCA; Kriegeskorte et al., 2008), which utilizes the versatility of Representational Similarity Analysis (RSA) to move from the direct comparison of representations to the comparison of representational geometries (Kriegeskorte et al., 2008). Recent implementations of RCA can be divided into model-free [e.g., Information Flow Analysis (Goddard et al., 2016), RSA-Granger Analysis (Kietzmann et al., 2019), static RSA (Karimi-Rouzbahani et al., 2021c), and jackknife-resampling RCA (Coutanche et al., 2020)] and model-based (Clarke et al., 2018; Karimi-Rouzbahani et al., 2021a) methods, each having specific characteristics. Here, we describe model-free and model-based RCA and point out their differences. Specifically, we present three simple scenarios where model-free and model-based RCA provide inconsistent connectivity results, flagging the situations where they should be used with caution and adding nuance to how the results of each should be interpreted.

One key feature of RCA is that, rather than activations (Anzellotti et al., 2017a,b; Basti et al., 2019), it evaluates the statistical dependency between the geometry/structure of neural representations across areas. Accordingly, RCA relies on the distinctiveness (i.e., dissimilarity) of patterns across conditions, which is conceived in terms of "information encoding/representation," rather than the activity patterns themselves. Therefore, one prerequisite for performing RCA is to have enough distinct experimental conditions to obtain the geometry of representations in the neural data [see, however, how we performed RCA on a single condition across time (Karimi-Rouzbahani et al., 2021c)]. This usually precludes RCA from being used to test functional connectivity in resting-state data (single, continuous fMRI, or M/EEG time series), which dominates univariate functional connectivity analyses. On the flip side, however, the representational nature of RCA provides several advantages over activity-based connectivity analyses. First, RCA allows the evaluation of connectivity across any two regions with different number of response channels (i.e., vertices, voxels, sensors, or sources; Kriegeskorte et al., 2008). Second, it allows one to ask how information (e.g., sensory, cognitive, etc.), rather than activation, is potentially transferred across areas. Third, model-based RCA allows one to target specific aspects of information, based on hypotheses about how a specific aspect of information is transferred, avoiding the influence from undesired confounders on connectivity (Clarke et al., 2018; Karimi-Rouzbahani et al., 2021a). Despite these advantages, under some circumstances representational connectivity analysis can miss true connectivity or erroneously detect non-existing (false) connectivity. This necessitates further investigation of RCA methods before they are more widely used as measures of multi-dimensional connectivity.

From a broad perspective, model-free RCA (Basti et al., 2020; Coutanche et al., 2020; Karimi-Rouzbahani et al., 2021c; Shahbazi et al., 2021) evaluates whether there is any commonality in the distributed patterns of activity for two brain regions. The commonality might reflect shared information due to similar encoding in the two regions. But it might also be due to the encoding of nuisance factors that are shared across regions. On the contrary, model-based RCA asks whether the two regions have shared information with regards to a specific hypothesis defined by a model. Having the model(s) can give a more specific picture of the multi-dimensional regional interactions. In this article, we compare model-free and model-based RCA and explain their pros and cons. In particular, we raise some cautions for using each method by showing simulated cases where one method fails to capture functional connectivity between two regions with shared information.

METHODS AND RESULTS

General Simulation Details

We generated multidimensional patterns of activity using scripts from the Matlab RSA toolbox (Nili et al., 2014; mean = 0; variance = 0.5; same statistics for every simulated subject). The Matlab script for reproducing the results can be downloaded from https://osf.io/3nxfa/. We simulated activity patterns for 16 stimuli in two brain regions. The number of vertices/voxels were set to 120 and 150 for regions of interest (ROIs) 1 and 2, respectively. The 16 conditions can be thought of as corresponding to four peripheral positions of the visual field (e.g., top left, top right, bottom left, and bottom right) of four semantically distinct visually presented object categories (e.g., animals, faces, fruits, and objects). For simpler explanation and interpretation of the results one can think of region of interest (ROI) 1 as visual area 2 (V2) and ROI 2 as inferior temporal cortex (ITC). Accordingly, ROI 1 dominantly represents position (i.e., regardless of the category of the objects) and ROI 2 dominantly represents semantic categories (i.e., regardless of the position of the stimuli). **Figure 1A** depicts the arrangements of the conditions in the Representational Dissimilarity Matrix (RDM).

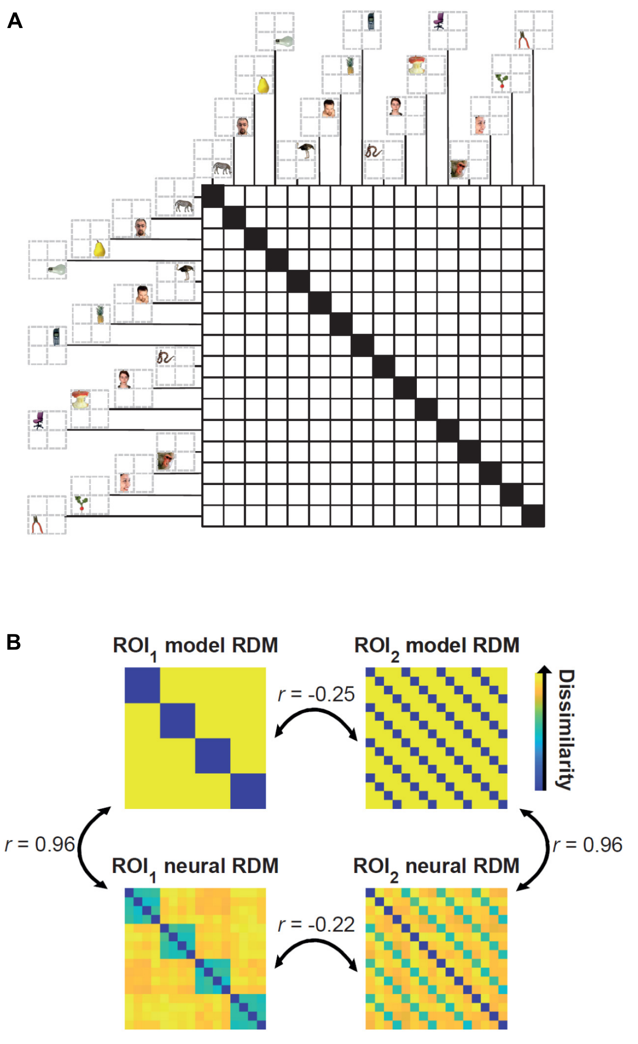


FIGURE 1 | Arrangement of the simulated conditions in the representational dissimilarity matrices (RDMs), model and neural RDMs and their pairwise correlations. **(A)** There are 16 conditions in the RDM consisting of four semantic categories of objects in four distinct positions. Example stimuli are from Kriegeskorte et al. (2008) available at http://www.cns.nyu.edu/kianilab/Datasets.html. **(B)** Top matrices show the ground truth for the two ROIs that represent "positions" (left) and "semantic categories" (right). The neural RDMs shown in the bottom were generated by adding noise to the activity patterns used to generate the two top model RDMs. Therefore, the simulated neural RDMs are highly correlated to the ground truth (r = 0.96). The correlations between the (model and/or neural) RDMs of the two different ROIs are negative ($r < -0.2$) implying no connectivity between them.

RDMs are generated by calculating the dissimilarity (here 1-correlation coefficient) of activity patterns across all experimental conditions and characterize the geometry of representations in the representational space (Kriegeskorte and Kievit, 2013). **Figure 1B** shows the ground-truth of the RDMs in the two ROIs and neural RDMs for a simulated subject which are different from the ground truths due to the added noise.

We used Pearson's (linear) correlation for comparing RDMs. Accordingly, we only considered significantly positive correlations as indicating representational connectivity. We performed significance testing using a one-sided Wilcoxon's signed rank test (Wilcoxon, 1992) across subjects and applied a threshold of 0.001 for statistical significance. Note that as RDMs are symmetric matrices, we only analyzed the elements in the upper triangle excluding the diagonal. We simulated data for N = 20 subjects to match it to the conventional number of subjects in real-life neuroimaging experiments and performed the statistical tests at group level.

Simulation 1: Model-Based Representational Connectivity Analysis Tests Connectivity Through the Lens of Model(s)

Problem Statement

Model-based RCA is designed to test whether two ROIs are related with regards to a specific model[1]. A model privileges a specific direction in the dissimilarity space so that all comparisons would be made with respect to the direction specified by the model. This allows us to test whether two regions share particular information. Although there have been (two) different implementations of model-based RCA (Clarke et al., 2018; Karimi-Rouzbahani et al., 2021a), here we use a minimalistic implementation to raise concerns about caveats of model-based RCA as clearly as possible. A model-free approach to test for representational connectivity would be to directly compare the RDMs in the two ROIs. Here, we use linear correlation to perform model-free RCA.

Despite the potential benefits of model-based RCA, it has some limitations that should be considered with caution. For example, consider the scenarios depicted in **Figure 2A** (note that although RDMs can generally reside in a high dimensional dissimilarity space, we illustrate the main point with 2D figures). **Figure 2A** shows a case where RDMs from two ROIs have a positive correlation to a model RDM and in fact identical similarities to it (e.g., Pearson correlation of 0.7). Conceptually, model-based RCA asks whether correlations of two brain RDMs to a model RDM are similar. It would be tempting to conclude that two ROIs share the information captured by a model if they are equally close (similar) to it. However, in this example, the RDMs in the two ROIs are in fact orthogonal, so this conclusion would be erroneous. Unlike model-based RCA, model-free RCA (e.g., the correlation of the two RDMs) would correctly conclude no functional connectivity.

Conversely, consider the case depicted in **Figure 2B**. The neural RDMs in the two ROIs have a positive correlation (e.g., a Pearson correlation of 0.7). This means that model-free RCA would indicate representational connectivity. However, one ROI has no relationship to the model (correlation of 0) and another ROI has a positive correlation to it (r = 0.7), and therefore from the perspective of the model, the two ROIs would not be connected. While the two regions share some information, this is not the same information that is captured by the model.

[1] We use the term "model" in a general sense: it can be a conceptual model, a computational model or a third brain region, etc.

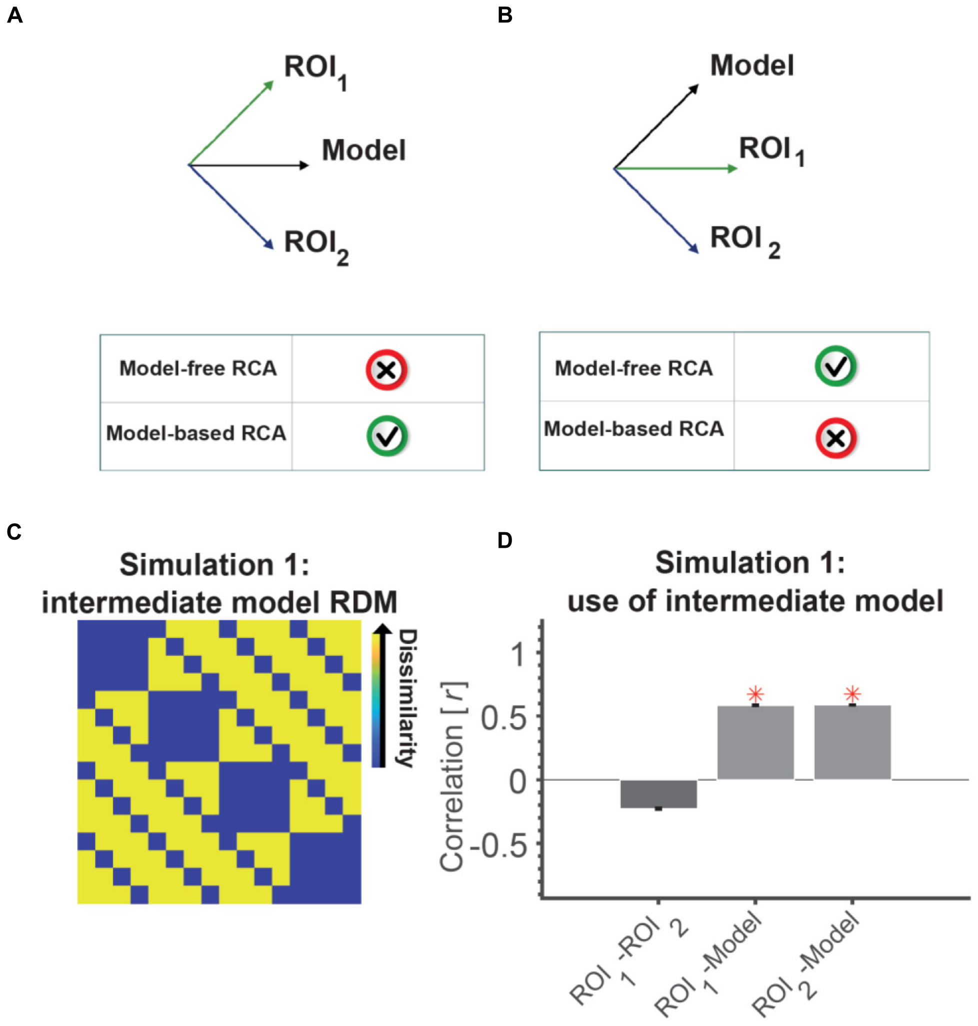


FIGURE 2 | An intermediate model RDM can lead to the wrong conclusion that two uncorrelated ROIs (RDMs) are connected. RDMs can be represented as a vector emanating from the origin in the N-dimensional space [N = number of elements in the RDM, i.e., n(n–1)/2 with n being the number of conditions]. **(A)** Shows a situation where a model RDM has equal angles to two orthogonal (unrelated) neural RDMs. Looking from the lens of this model can leave the impression that the two neural RDMs are similar as they project similarly on the model RDM. **(B)** Shows a situation where two neural RDMs are positively correlated, but might look unrelated (disconnected) when looking at them through the lens of the model RDM. This is because only one has a component along the model RDM's vector and the other is orthogonal. **(C)** The model RDM used in Simulation 1, which has components of the representations in ROI 1 (position) and ROI 2 (semantic category). This model RDM has equal correlations (= 0.6) to the models used to generate neural RDMs of the two ROIs (shown in **Figure 1**, top); almost similar correlations to the neural RDMs (0.6). **(D)** The correlation between the neural RDM of the two ROIs (model-free RCA) and between the neural and model RDMs of each ROI (model-based RCA). Red asterisks show significant above-chance correlation values (connectivity) as evaluated by a one-sided Wilcoxon's signed rank test against zero.

These examples show that model-based and model-free RCA can easy lead to different results. It follows from the fact that model-free RCA is based on direct comparison of neural RDMs in the original dissimilarity space and that model-based RCA is based on comparison of projected RDMs on a line (projection), defined by the model. The degree of inconsistency depends on the neural RDMs and the direction defined by the model.

In Simulation 1 we present a scenario (similar to **Figure 2A**) where applying model-based RCA to two ROIs which represent independent aspects of visual stimuli could result in the wrong conclusion that the two are functionally connected.

Simulation Details

The general simulation details are provided in section "General Simulation Details." For model-based RCA, we used a model RDM that incorporates both aspects of the stimuli (i.e., position and object category). This "intermediate" model hypothesized a larger pattern dissimilarity for two conditions that are different in both position and object category. The model had equal level of correlation/similarity to the neural RDM in each of the two ROIs (r = 0.6 between the model in **Figure 2C** and the RDMs shown in top and/or bottom panels of **Figure 1**). To implement model-based RCA, for each ROI we calculated the correlation between the neural RDM and the model for each subject. Then, we statistically compared the correlation results across subjects for the two ROIs. Specifically, two ROIs are considered to reflect model-based representational connectivity if they show significant positive correlations to a model and their correlations to the model are not significantly different. To implement the model-free RCA, we calculated the correlation between the neural RDMs of the two ROIs directly. Therefore, statistically positive correlation between neural RDMs shows model-free representational connectivity.

Simulation Results

Two ROIs that represent statistically unrelated information are not connected and do not have any shared information. However, looking at the two ROIs through the lens of an intermediate model in model-based RCA can leave the impression that the ROIs are connected. There was significant positive correlation for the two ROIs with the intermediate model, and the correlations between the model and the two ROIs were not statistically different (Wilcoxon's signed rank test; p = 0.94, **Figure 2D**). This (incorrectly) suggests that the two ROIs are connected, by virtue of sharing the information captured in the intermediate model. However, model-free RCA (i.e., direct correlation between the two ROIs) correctly showed no positive correlation between the ROIs suggesting no connectivity. It might be worth adding that had we used a simple model instead of the intermediate model, for example, one of the two models illustrated in the top panel of **Figure 1B**, we would have correctly observed no model-based RCA. Therefore, the issue relates to the representational structures of the two ROIs as well as the model used for examining their connectivity.

Potential Solutions

To avoid false conclusion about connectivity across ROIs, it is important to evaluate it using both model-based and model-free RCA, see that if the results agree, and interpret accordingly. Where possible, for model-based RCA, it may also help to use minimal models where only one, rather than several, aspects of information is captured. In our simulation, the fact that our intermediate model had components from both aspects of stimuli (i.e., position and category) made it possible to capture variances explained by different processes, i.e., independent encoding of each aspect. Simpler models, for example models that correspond to simple hypotheses, might help to untangle representational connectivity along different dimensions of information transfer. However, it might be difficult to know these models in advance, unless the tasks are simple, and the underlying representations are already well characterized.

It is of note that, while we implemented a simplified version of RCA here, implementations in the literature have incorporated other parameters, such as time and delay, and other techniques such as multi-linear regression and partial correlation (Goddard et al., 2016, 2021; Karimi-Rouzbahani, 2018; Karimi-Rouzbahani et al., 2019, 2021a) each of which may affect the results. However, both the previous published implementations of model-based RCA (Clarke et al., 2018; Karimi-Rouzbahani et al., 2021a) ultimately rely on assessing the similarity in model fits between regions, so are subject to the concern we have demonstrated.

Simulation 2: Spurious Connectivity From Common Input to Regions With Distinct Representations Can Be Avoided Using Model-Based Representational Connectivity Analysis With Appropriate Models

Problem Statement

There can be situations where common uninformative patterns are present along with the informative representations in the pair of ROIs considered for connectivity analysis. The common patterns can be as simple as measurement or neural noise which might be statistically dependent across areas and/or the leakage or feeding of activations from a third ROI to both ROIs as a result of proximity and/or poor spatial resolution (e.g., in EEG and MEG). On the other hand, it can also be the case that the two ROIs encode/represent some shared aspects, which are either task-irrelevant or not the target of study. For example, both position-selective early visual area (V2) and the semantically selective area (ITC) can be sensitive to low-level image statistics such as the spatial frequencies of the stimulus due to connections from V1. This shared information may lead to apparent connectivity if their RDMs are directly compared (as in model-free RCA), but may not reflect shared information of interest to the researcher. In general, we are not interested in capturing commonality in noise, and may not be interested in capturing this low-level information (i.e., spatial frequency) which are represented in both ROIs, but rather by the particular information for which we have hypotheses. In this simulation we ask whether model-based RCA is robust to this type of shared information and allows us to draw a specific conclusion about the shared information of interest to the researcher.

Below we simulate the impact of adding common patterns of activation to a pair of ROIs which otherwise represent distinct information, and show how model-free RCA, and some implementations of model-based RCA, can be affected. We show that using appropriate models that match the dominant representations of the two ROIs can mitigate the false connectivity.

Simulation Details

The neural patterns generated here are the same as Simulation 1 (with no connectivity between the two ROIs) except that now we also include the time course of representations to be able to implement more realistic model-free and model-based RCAs (rather than the simplified ones implemented in Simulation 1). We added the temporal dimension so that correlations could be computed over time. Please note that, however, ROI representations at different time points were consistent with the same structures depicted in **Figure 1B**. In other words, the information did not change over time but experienced some additive Gaussian noise (zero-mean; variance = 0.5). We simulated the activity patterns of the two ROIs over 200 time samples. The two ROIs were simulated to encode the two above-mentioned distinct aspects of information (i.e., position and semantic categories).

We performed model-free RCA by calculating the direct correlation between RDMs of the two ROIs at every time point and then averaging the resultant correlations over the simulated time window.

In this Simulation (and also the next simulation), we consider two versions of model-based RCA that have different motivations. In either case, we first obtained the correlation between the neural and the corresponding model RDM of each ROI at every time point and then calculated the correlation between the time courses of neural-model correlations for the two ROIs.

In the first version, we considered a common model for the two ROIs (similar to Simulation 1) and in the second version we used ROI-specific models (i.e., one model per ROI). The motivation for the first approach (1-model RCA) was that the experimenter might simply want to evaluate the information exchange reflecting a single known aspect of information (e.g., familiarity information across occipital vs. frontal areas: Karimi-Rouzbahani et al., 2021a). On the other hand, the experimenter might hypothesize that the dominant aspect of information represented in each of the two ROIs is different (e.g., visual information in lower visual areas vs. semantic information in ITC: Clarke et al., 2018). In this case it might be more suitable to compare each ROI to a specific model of itself (ROI-specific models) and use a 2-model RCA. In this case the interpretation of 2-model RCA results would be different and will be explained below (Simulation 3). For our implementation of 1-model RCA in the simulated example, we used the position model for both ROIs. For the 2-model RCA, we used the position model for ROI 1 and a semantic-category model for ROI 2; therefore, the models perfectly matched what was dominantly represented in each ROI.

We added a non-structured (noise) pattern to both ROIs and evaluated its impact on connectivity (**Figure 3A**). To generate the common pattern, we used Gaussian noise (zero-mean; variance = 7) and a random transformation matrix (containing random numbers from a zero-mean unit-variance Gaussian distribution) to impose correlated noise across areas. Similar to Basti et al. (2020), we first simulated the added noise for one ROI and then transformed it via a multivariate linear mixing matrix to obtain the noise in the other ROI (**Figure 3B**).

Simulation Results

The results are shown in **Figure 4A**. As expected, before adding the common patterns to the ROIs, the three connectivity measures were either negative (model-free RCA) or around-zero (1-model and 2-models RCA), suggesting no connectivity between the ROIs (**Figure 4A**). However, the addition of common noise patterns to the two ROIs led to spurious connectivity for model-free and 1-model RCA, with both showing significantly above-chance connectivity. This was expected for the model-free RCA because it relies on shared information across ROIs, which become correlated by the added correlated noise (similar to the example depicted in **Figure 3B**). The researcher would conclude that the ROIs were connected, when in fact the positive result reflected shared information of no interest.

This result was especially interesting for 1-model RCA because the method required the two ROIs to be temporally correlated to show connectivity. This confirms that the common pattern has not only correlated the patterns of the two ROIs on every time point, but it has also added temporal correlations to the patterns of the two ROIs making them fluctuate similarly over time (which is key for our model-based connectivity). We also observed that this spurious connectivity for 1-model RCA was not specific to the particular model we used and remained when using any arbitrarily defined random models (results not shown). Specifically, we observed that even models unrelated to the representational structure of one of the two ROIs (e.g., position and/or semantic categories) could lead to false connectivity in 1-model RCA. This can be explained by the fact that the time-locked common input will make the RDMs of the two ROIs be more similar to each other and also to any random RDM.

Finally, despite the correlations imposed on the contents of representations and the temporal patterns across ROIs, the 2-model RCA (correctly) showed negative correlations between the ROIs after the common input suggesting no connectivity (**Figure 4A**). This negative correlation can be explained by the fact that the two correlated representations will have a negative correlation when evaluated against two negatively correlated models. 2-model RCA is determined by the relationship between the neural RDMs of the two ROIs after projecting them on their relevant model RDMs. More specifically, it suggests that if the model RDMs for two ROIs do not correlate, or correlate negatively (suggesting distinct codes represented in each of them), they can remain immune to the added common noise. To test this hypothesis, we generated random models for the two ROIs of each simulated subject and calculated the level of increase in correlation/connectivity from before to after adding the common noise. Results showed a direct relationship between the similarities of the two models (and the corresponding ROIs) and the change in (2-model RCA) connectivity under the influence of the added noise (**Figure 4B**). Specifically, the more correlated the two models were, the larger the influence of adding correlated patterns. For negatively correlated models, like those in our simulation, adding correlated noise reduced the connectivity, so would not lead to spurious results. Therefore, 2-model RCA can be robust to shared noise or other common signals of no interest, but only if the 2-models are orthogonal or negatively correlated (and negative correlations are not interpreted).

Potential Solution

Both model-free and model-based RCA are affected by common-inputs to the two ROIs. This is particularly important for model-free RCA and 1-model RCA where it will always be the case, and should be taken into consideration when interpreting results. However, in 2-model RCA, where the two ROIs originally represented two distinct aspects of the task, the results were robust to the added common noise. This result was dependent on the chosen model RDMs not being positively correlated. Therefore, for cases where two regions are hypothesized to represent distinct information, the use of 2-model RCA with orthogonal or negatively correlated models can avoid spurious

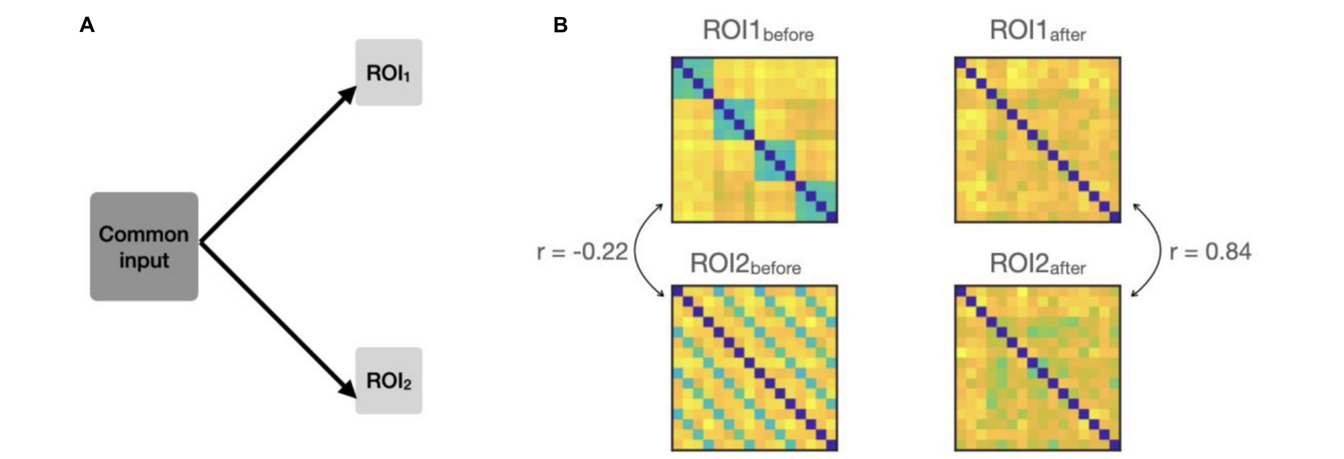


FIGURE 3 | (A) Simulation settings, a common input is added to the responses in both ROIs, so that the patterns in the two ROIs covary at each time-point. **(B)** Example RDMs from the two ROIs before and after adding the common input. RDMs originally had a negative correlation (–0.22). The correlations went up from –0.22 to 0.84 due to the common input.

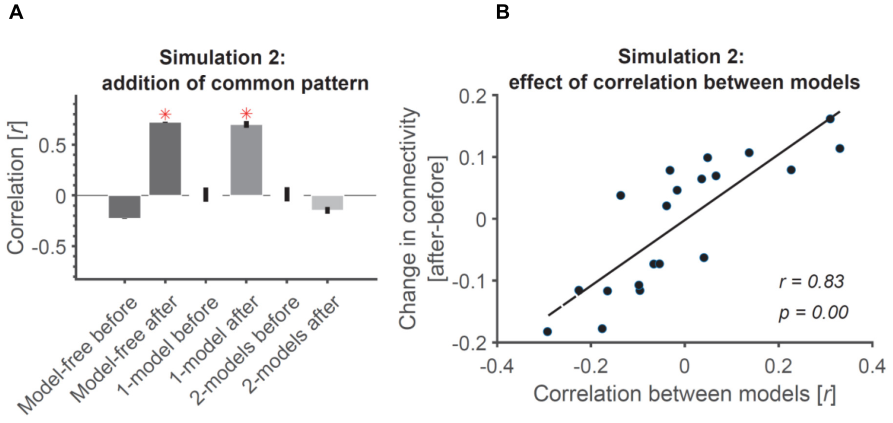


FIGURE 4 | Addition of common pattern to pair of ROIs can make them look connected in model-free and 1-model RCA but not 2-model RCA if the ROIs are distinct enough. **(A)** Addition of common non-structured (noise) patterns to both ROIs leads to significant connectivity when evaluated using model-free and 1-model RCA, but not 2-model RCA because the two ROIs dominantly represent information that is negatively correlated (see **Figure 1B**). Red asterisks show significant above-chance correlation values (connectivity) as evaluated by one-sided Wilcoxon's signed rank test against zero. **(B)** The more distinct the dominant information represented across the ROIs, the less the effect of added common noise on their connectivity. Dots show the amount of change in connectivity as a function of correlation between the information represented in the two model RDMs used in 2-model RCA. Each dot represents data from a single simulated subject. The line shows the best linear fit to the data. The correlation and the significance of correlations are also shown as calculated using Pearson's linear correlation.

connectivity caused by common patterns of activation such as correlated noise.

Simulation 3: Model-Based Representational Connectivity Analysis With Region of Interest-Specific Models Can Detect Transformation of Information Across Region of Interests

Problem Statement

There can be situations where the structure of the information is transformed from one ROI to the next. In fact, it seems unlikely that information remains intact ("copied") between any two ROIs in the brain. Therefore, direct comparison of neural representations, as implemented in model-free RCA, can miss such potential connectivity simply because the statistical relationship may be lost in transformation. However, model-based RCA may allow us to detect the connectivity between two areas, which encode distinct information, based on their temporal statistical congruency. Below we simulate two ROIs that represent two distinct aspects of information, with dynamics that are either temporally congruent or incongruent between ROIs. Specifically, the information about the stimulus position initially appears in the source ROI (V2) and is followed by the semantic-category information which appears in the destination ROI (ITC)[2]. This

[2] We consider the case where the information is reliably transferred from one ROI to another as temporally congruent and cases where there is no transfer

scenario resembles a study which found evidence in support of causal information transfer and transformation from early visual to ITC areas (Clarke et al., 2018). As the detection of connectivity using model-based RCA with ROI-specific models also needs the adoption of correct model for each ROI, because the information is transformed, we also examine the effect of choosing the correct model for each ROI.

Simulation Details

We used simulations to investigate the transformation of information using model-based RCA. The details of information representation in the two ROIs in this simulation are identical to Simulation 2, with the exception that the information does not appear throughout the simulation window but rather for a fixed period of time in each ROI (samples 30–60 in ROI 1; solid black curve in **Figure 5A**). There was a delay of ±20 samples between ROIs 1 and 2 (positive for congruent and negative for incongruent case) which was jittered between 0 and 10 samples (uniform random distribution) across the simulated subjects (N = 20). This led to information appearing in ROI 1 before ROI 2 in congruent cases, and in ROI 2 before ROI 1 in incongruent cases (**Figure 5A**). Specifically, patterns could appear between samples of 40 and 80 in ROI 2 in the congruent case and between samples of 0 and 40 in ROI 2 in the incongruent case. The activity patterns of the two ROIs did not contain any information in the samples outside the mentioned windows. Note that similar to the previous simulations, the information which was dominantly represented in the two ROIs was different [position encoding in V2 (source) and semantic category encoding in ITC (destination)]. This scenario simulated information flow from the source to the destination area that has been evaluated in previous studies using both model-free and model-based RCA (Goddard et al., 2016, 2021; Clarke et al., 2018; Karimi-Rouzbahani, 2018; Karimi-Rouzbahani et al., 2019, 2021a). In this analysis, the onset of information in each ROI predicts the direction of information transfer (e.g., potential information flow from V2 to ITC). Here we only evaluate the feed-forward information flow/connectivity from ROI 1 to ROI 2 (e.g., as in the ventral visual stream) and not vice versa. In testing the connectivity for both model-free and model-based RCA methods, we set the analysis delay-time (i.e., lag) between ROIs to be 20 (no jitter) for all our subjects. This parameter is usually set by the researcher and fixed across subjects (Goddard et al., 2016, 2021; Karimi-Rouzbahani, 2018; Karimi-Rouzbahani et al., 2019, 2021a). We performed model-free RCA by calculating the direct correlation between the RDM of the source ROI at time t and the RDM of the destination ROI at time $t + \tau$ where τ refers to the delay (= 20 samples) and then averaged the time course of correlations within each subject. For model-based RCA, we calculated the correlation between the neural and model RDMs for each ROI on every time point as in Simulation 2 (note that we considered the two cases of having one model RDM or two different model RDMs), shifted the model-correlation time course of ROI 2 by 20 (jittered between 0 and 10 samples) relative to ROI 1, and computed their correlation coefficient. In both model-free and model-based analyses, the incorporation of the delay compensated for the inter-ROI delay in the data.

Our assumption here is that two ROIs that encode/represent statistically unrelated information can be considered connected if their temporal information-encoding profiles are statistically related/congruent (representations appear in the destination after the source ROI at around the hypothesized delay). We ask whether such a relationship would be detected using model-free, 1-model and 2-model RCA.

Simulation Results

Figure 5A shows the time courses of correlations between the RDM of each ROI with its corresponding specific model RDM. In congruent trials, correlations between RDMs from ROI1 and model1 (black solid curve) peak reliably earlier than correlations between ROI2 and model2 RDMs (gray solid curve).

Simulation results show that model-free RCA did not detect any connectivity between the two ROIs (**Figure 5B**). The 1-model RCA also failed to detect the connectivity whether the representations in ROIs appeared congruently or incongruently. The reason is that the representations that were transformed from the source to the destination ROI no longer matched the common model in the destination ROI (we used the model RDM from ROI 1 for both ROIs). The 2-model RCA also failed to detect the connectivity when the representations appeared incongruently across the ROIs (first in the destination followed by the source) because information time courses in one did not reliably follow the other according to the hypothesized lag. However, the 2-model RCA could detect the connectivity when the representations appeared congruently across the ROIs. Therefore, for the transformed information to be detected, one needs to have both accurate models of information representations as well as correct prior knowledge about temporal dynamics and direction of information flow across ROIs.

Note that in these simulations, we incorporated the delay in our analysis and the two ROIs followed the temporal profiles of representations shown in **Figure 5A**. Therefore, the absence of connectivity in the model-free and 1-model RCA cannot be explained by the fact that we used lagged correlations in the 2 model case. Specifically, we incorporated the delay in ***all*** our RCA measures here to avoid any systematic difference in RCA across methods.

Potential Solutions

Model-free RCA is only sensitive to direct statistical relationship between neural RDMs, and fails to detect the connectivity if the two ROIs do not statistically relate. However, 2-model RCA allows detection of congruent inter-ROI statistical dependencies by having models that capture the representational structure of each ROI. Importantly, as 2-model RCA relies on hypotheses about the representations in source and destination areas, it will be less affected by confounders such as noise which are generally represented similarly across the two ROIs. Similar to the observation made in Simulation 2, it might be that

of information or the transformation is not reliable/consistent as temporally incongruent.

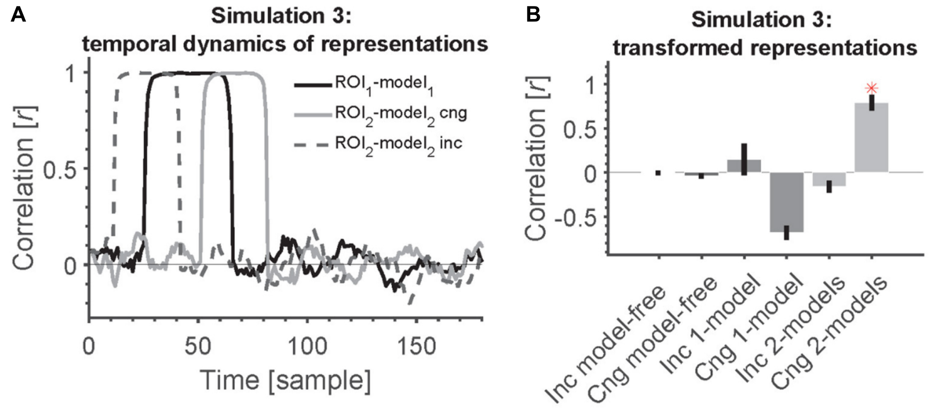


FIGURE 5 | 2-model RCA allows us to detect transformation of information if the temporal dynamics across ROIs are statistically related/congruent. **(A)** Time course of information encoding in the two ROIs at a delay of 20 samples [congruent (cng), solid gray line] and a delay of –20 from ROI 1 to ROI 2 [incongruent (inc), dashed gray line]. The delay was variable across simulated subjects. The time courses show the correlation between each ROI and its corresponding model (position model for ROI 1 and semantic-category model for ROI 2). **(B)** Transformation of information across ROIs causes all model-free and model-based RCAs to miss the connectivity except when the information appears congruently across ROIs (first in ROI 1 followed by ROI 2) and using 2-model RCA. Red asterisk shows significant above-chance correlation value (connectivity) as evaluated by one-sided Wilcoxon's signed rank test against zero.

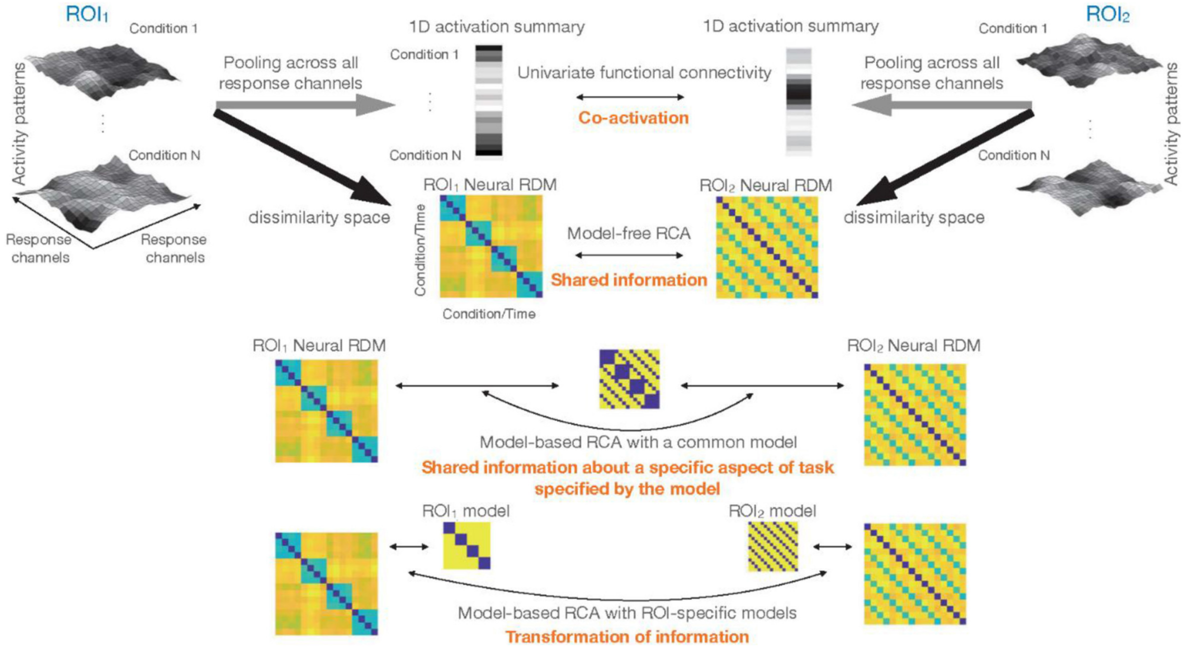


FIGURE 6 | Different types of inference about functional connectivity: top left and top right show the response patterns for N experimental conditions in two ROIs. Larger activations in a voxel are shown by lighter colors. One classical approach would be to reduce the dimensionality of data in each ROI to 1, and summarize the rich patterns of activity by a single vector containing one number for each experimental condition (or time-point for the case of resting-state data). Significant correlation between these vectors implies co-activation, i.e., that activations in ROI1 and ROI2 co-vary. Multi-dimensional connectivity methods that we consider in this article characterize the response patterns for different conditions by a representational dissimilarity matrix (RDM). Direct comparison of the RDMs (model-free RCA) tests for shared information (i.e., whether the two sets of response patterns in the two ROIs have any shared information with regards to the experimental conditions). Incorporation of models, i.e., model-based RCA, when a common model is used for both ROIs (1-model RCA) tests for shared information about a specific aspect of task/stimuli. This hypothesis in RCA is specified in the ROI-common model. Finally, model-based RCA with ROI-specific models (2-model RCA) detects potential transformation of information.

common task-irrelevant patterns in both ROIs obscure the shared information as captured by 1-model RCA or the transformation of information as captured by 2-model RCA. A solution to this would be to remove their contribution by regressing out the RDM of the common pattern from the RDM of each ROI at each time-point. However, for this one needs the knowledge about the structure of the common patterns, which is not often known *a priori*. Researchers should be aware of these limitations so that they can choose their analysis method and interpretation accordingly.

Another solution to the failure of model-free RCA to detect connectivity under transformed representations might be to use non-linear mapping functions. Such functions allow more flexible relationships to be detected between areas despite drastic transformation of the representational structure. Such non-linear mapping functions include distance correlation (Geerligs and Henson, 2016), projection to a Riemannian manifold (Shahbazi et al., 2021) or more general functions estimated by artificial neural networks (Anzellotti et al., 2017b). These potential solutions are not investigated here.

DISCUSSION AND CONCLUSION

Multi-dimensional connectivity is a rapidly developing area of brain connectivity analysis. One of the approaches to multi-dimensional connectivity is representational connectivity analysis (RCA). RCA quantifies the similarity of inter-relationship between the neural representations across experimental conditions for distributed patterns of activity of two brain regions (Kriegeskorte et al., 2008). This allows us to track "information" (by the representational geometry in the multi-dimensional response space) rather than the mere similarity of average response levels across two regions. Despite its versatility, a better understanding of the situations that can challenge and/or mislead RCA is needed. In this manuscript, we explain two main approaches of RCA. One is model-free RCA that directly compares the representational geometries of two brain ROIs. Model-free RCA can tell us whether two brain ROIs have any shared information in their multi-dimensional response patterns. The other is model-based RCA. In this article, we make a further distinction between two approaches to model-based RCA: using a single or multiple models. We think this distinction is important, since besides the difference in technical details and implementation, they entail different interpretations about regional interactions. The first variant of model-based RCA, which uses a common model (1-model RCA), tests whether the representational geometries of the two ROIs are similarly concordant to a hypothesized geometry (i.e., the model). This can tell us whether two brain regions have shared information with regard to a specific aspect of the stimuli/task. The other variant of model-based RCA uses ROI-specific models, which, with time-resolved data, tells us whether information in one region is transformed into different information in another region. Therefore, while this also pertains to functional connectivity, it does not explicitly get at shared information. **Figure 6** provides an overview of the distinctions explained in the article.

Model-free and model-based RCA can potentially provide inconsistent results in certain circumstances. These inconsistencies depend on many factors, some of which are the spatiotemporal structure of neural representations and the choice of the model(s) used in the analysis, and inform interpretation. Here, we focused on three simulations where model-free and model-based RCA provided opposing connectivity results.

First, we simulated a situation where the neural representations across a pair of regions showed unrelated information. As expected, model-free RCA showed no connectivity between the pair of regions. Interestingly, however, we observed that using model-based RCA with an intermediate model, which contains information about the representations in both regions, can leave the false impression that the two regions are connected. Specifically, the two regions showed almost equal, positive and significant correlation to the intermediate model suggesting that from the "lens" of the selected model, the two regions appear to be connected.

There are a few considerations. First, although for simplicity we did not directly implement either of the two published methods of model-based RCA (Clarke et al., 2018; Karimi-Rouzbahani et al., 2021a), the problem we pointed out here can affect both those methods. This is because they compare the correlation between the models either explicitly (Clarke et al., 2018), or implicitly within the formulation of partial correlation (Karimi-Rouzbahani et al., 2021a). Second, although we used a two-component model for this simulation to simplify the interpretation, this situation is not limited to two-component models. In fact, any other models that share roughly equal amounts of variance with different components encoded in two areas would lead to a similar situation. Third, the false connectivity observed in this scenario is not driven by the specific similarity metric we used (i.e., Pearson's linear correlation). Although different similarity metrics show different characteristics (Walther et al., 2016; Shahbazi et al., 2021), as long as the selected metric provides similar values for the similarity between two different neural and a given RDM model, the same effect will be observed. The reason is that all similarity metrics summarize a high-dimensional representational space into a single-dimensional space, which inevitably leads to loss of information. Finally, at the other end of the spectrum, there can be cases where two regions represent one or several very similar aspects of information, but they still look unrelated/disconnected through the lens of a particular model. However, this case seems less problematic since the main reason behind using model-based rather than model-free RCA is to limit the representations to the desired information (Karimi-Rouzbahani et al., 2021a). Nonetheless, it would be good practice to do perform both types of RCA (together with RSA information mapping) and to compare the results while being aware of the limitations and caveats of each.

In the second simulation, we modeled a situation where the addition of statistically related patterns of activity to a pair of statistically unrelated regions imposed a statistical relationship between them. This led to apparent connectivity in model-free RCA and when using 1-model RCA. However, the common pattern did not affect apparent connectivity when using 2-model RCA, as long as the two models were orthogonal and the two ROIs represented distinct information. Please note that the added common pattern can be non-structured or structured. Although we have seen that both common noise (non-structured) and structured patterns (data not shown) led to similar results, the structure of the common pattern can affect the connectivity as a result of interaction with the representations in the regions and models. It is also of note that the addition of common patterns does not always inflate the connectivity (e.g., in model-free or 1-model RCA); it can also decrease it leading to missing the connectivity. For example, if two regions are perfectly correlated, the addition of common noise (if not perfectly identical but only statistically related across regions) could lead to a decline in model-free RCA as a result of distorting the patterns. Generally, both model-free and model-based RCA can be affected by the noise as a result of the complex interaction between the representation in each region, the structure of the added pattern, the models, and the temporal dynamics of representations. Therefore, despite the situation shown in Simulation 2, these methods we are still far from remaining immune to common noise. We can, however, understand where we are most susceptible to it. One simple remedy for the effect of common patterns would be to regress out its contribution from the RDMs of the two ROIs prior to computing connectivity measures. This is in spirit similar to our recent implementation of model-based RCA (using partial correlation), where we partialled out the effect of additional low-level image statistics from the two regions under study (Karimi-Rouzbahani et al., 2021a). However, as the structure of the added pattern (noise or common structure of no interest) is usually unknown, this will not always be an option.

In the third simulation, we showed a situation where two regions encoded different types of information that were either temporally congruent or incongruent. In other words, the information initially appeared in one region and after some delay in the other region (temporally congruent). Model-based RCA with proper choices of models can capture this relationship. This may be useful as transformation of information seems an integral part of brain connectivity as it seems unlikely that information would remain intact from one brain region to another (Lahaye et al., 2003; Hlinka et al., 2011). Transformations of information have already been reported in visual system of human and monkey brain (DiCarlo et al., 2012; Kietzmann et al., 2019) and are implemented by other sensory hierarchies as well (Winkowski and Kanold, 2013). For example, it has been suggested that visual information is moved from low- to a high-dimensional space along the ventral visual stream and brought back to the low-dimensional space in later stages of the stream to compensate for variations of visual objects and form semantically categorized object clusters (DiCarlo et al., 2012; Karimi-Rouzbahani et al., 2017a,b). Using model-based RCA, previous work has found that information transforms from visual to semantic brain areas (Clarke et al., 2018). In our simulation, the drastic transformation of information simulated in Simulation 3 meant that the connectivity was missed by model-free RCA and 1-model RCA. However, 2-model model-based RCA detected the connectivity as a result of its simultaneous sensitivity to targeted region-specific information representation and the temporally congruent patterns of information representation. Therefore, a hypothesis-driven method of RCA allows us to detect information that is transformed as it passes between brain regions.

This simulation also demonstrated the importance of the delay in connectivity analysis matching the data. The delay in the analysis potentially captures the neural lag in information transfer in the brain (Cichy et al., 2014). The delay is generally set *a priori*, meaning that choice of improper delays (negative vs. positive; which also determines the direction of information) can lead to missing the connectivity. A more principled way of estimating the delay would be to partition the data and estimate the optimal delay from one half and apply it to the other half. However, this requires independent measurements of the same task in each subject. A more extended version of the RCA could be to perform Granger causality to examine Granger-causal relationships between areas as in previous studies (Goddard et al., 2016, 2021; Clarke et al., 2018; Karimi-Rouzbahani, 2018; Karimi-Rouzbahani et al., 2019; Kietzmann et al., 2019). That would also be subject to similar considerations. However, comparing the different approaches at a conceptual and mathematical level is beyond the scope of the current study.

It is generally desired that a connectivity method determines the transferred *content*, *direction,* and *temporal dynamics* of information flow. To that end, previous studies implemented techniques including partial correlation (Goddard et al., 2016, 2021; Karimi-Rouzbahani, 2018; Karimi-Rouzbahani et al., 2019, 2021a) and regression (Kietzmann et al., 2019), or tested for Granger causal relationship between areas (Goddard et al., 2016, 2021; Clarke et al., 2018; Karimi-Rouzbahani, 2018; Karimi-Rouzbahani et al., 2019), or used models to measure the contribution of one area to another in the direction of the task (Karimi-Rouzbahani et al., 2021a) or incorporated autoregressive approaches to estimate proper delay between areas (Clarke et al., 2018). In our most recent effort, to bring together the advantages of the mentioned methods, we proposed a variant of model-based RCA which provided information about the content of the transferred information, its direction and temporal dynamics simultaneously (Karimi-Rouzbahani et al., 2021a). This method showed distinct dynamics and direction of face familiarity-information flow across peri-frontal and peri-occipital cortices for different levels of perceptual uncertainty. Despite our minimalist approach in the current study, the insights and cautions provided by this work can be generalized to more complex implementations of RCA as well.

Additionally, one could also consider other extensions to model-free RCA. Similar to "information connectivity" (Coutanche, 2013) where multi-dimensional connectivity is established by correlating time series of classification-accuracies across regions, one can compare time courses of the exemplar discriminability index (EDI, Nili et al., 2020) across regions. EDI is a model-free RSA statistic in each region and quantifies the extent to which different experimental conditions elicit distinct patterns of activation. Similar to the implementation of model-free RCA, however, this definition of model-free RCA also does not shed light into the content of shared information.

One limitation of the current study is that we only evaluated connectivity using linear, rather than non-linear, relationships. While this simplification allowed us to make more intuitive predictions about the relationship between brain responses and the models, a more general approach would be to incorporate non-linear connectivity between areas as well. While we believe that the cases evaluated in Simulations 1 and 2 will not be affected by using a non-linear connectivity metric, non-linear mapping functions in Simulation 3 (Geerligs and Henson, 2016; Anzellotti et al., 2017b; Basti et al., 2020; Shahbazi et al., 2021) may allow for detecting non-linear relationships between areas. Therefore, future studies will need to evaluate the impact of non-linear mapping functions in RCA.

This work takes initial steps toward better characterization of the model-free and model-based RCA approaches that have been increasingly used in recent years. We tried to make the simulations as general and ideal as possible (no nuisance factors, e.g., measurement noise, leakage incorporated), so that the insights can be generalized to different implementations of the two general classes of model-free and model-based RCA. Therefore, the points made here can provide insight when studying brain connectivity using variety of neural recording modalities such as EEG, MEG, multi-electrode electrophysiology, and fMRI. Specifically, apart from Simulation 1, which presents a conceptual point applicable to all multivariate imaging/recording modalities, the methods implemented in Simulations 2 and 3 can directly be applied to EEG and MEG data.

DATA AVAILABILITY STATEMENT

The dataset used in this study is auto-generated using the Matlab script available at https://osf.io/3nxfa/ which generates the simulation figures as well.

AUTHOR CONTRIBUTIONS

HK-R: conceptualization, methodology, software, formal analysis, funding acquisition, writing – original draft, and writing – review and editing. AW and RH: conceptualization, methodology, and writing – review and editing. HN: conceptualization, methodology, software, formal analysis, writing – original draft, and writing – review and editing. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by the UK Royal Society's Newton International Fellowship NIF\R1\192608 to HK-R and MRC intramural funding SUAG/052/G101400 and SUAG/046/G101400 to AW and RH, respectively.

REFERENCES

Anzellotti, S., and Coutanche, M. N. (2018). Beyond functional connectivity: investigating networks of multivariate representations. *Trends Cogn. Sci.* 22, 258–269. doi: 10.1016/j.tics.2017.12.002

Anzellotti, S., Caramazza, A., and Saxe, R. (2017a). Multivariate pattern dependence. *PLoS Comput. Biol.* 13:e1005799. doi: 10.1371/journal.pcbi.1005799

Anzellotti, S., Fedorenko, E., Kell, A. J., Caramazza, A., and Saxe, R. (2017b). Measuring and modeling nonlinear interactions between brain regions with fMRI. *bioRxiv* [Preprint]. bioRxiv, 074856,

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., et al. (2006). Top-down facilitation of visual recognition. *Proc. Natl. Acad. Sci. U.S.A.* 103, 449–454.

Basti, A., Mur, M., Kriegeskorte, N., Pizzella, V., Marzetti, L., and Hauk, O. (2019). Analysing linear multivariate pattern transformations in neuroimaging data. *PLoS one* 14:e0223660. doi: 10.1371/journal.pone.0223660

Basti, A., Nili, H., Hauk, O., Marzetti, L., and Henson, R. N. (2020). Multi-dimensional connectivity: a conceptual and mathematical review. *Neuroimage* 221:117179. doi: 10.1016/j.neuroimage.2020.117179

Bastos, A. M., and Schoffelen, J. M. (2016). A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. *Front. Syst. Neurosci.* 9:175. doi: 10.3389/fnsys.2015.00175

Biswal, B., Zerrin Yetkin, F., Haughton, V. M., and Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magn. Reson. Med.* 34, 537–541. doi: 10.1002/mrm.1910340409

Bressler, S. L., and Menon, V. (2010). Large-scale brain networks in cognition: emerging methods and principles. *Trends Cogn. Sci.* 14, 277–290. doi: 10.1016/j.tics.2010.04.004

Cichy, R. M., Pantazis, D., and Oliva, A. (2014). Resolving human object recognition in space and time. *Nat. Neurosci.* 17, 455–462. doi: 10.1038/nn.3635

Clarke, A., Devereux, B. J., and Tyler, L. K. (2018). Oscillatory dynamics of perceptual to conceptual transformations in the ventral visual pathway. *J. Cogn. Neurosci.* 30, 1590–1605. doi: 10.1162/jocn_a_01325

Coutanche, M. N. (2013). Distinguishing multi-voxel patterns and mean activation: why, how, and what does it tell us? *Cogn. Affect. Behav. Neurosci.* 13, 667–673. doi: 10.3758/s13415-013-0186-2

Coutanche, M. N., Akpan, E., and Buckser, R. R. (2020). Representational connectivity analysis: identifying networks of shared changes in representational strength through jackknife resampling. *bioRxiv*[Preprint]. doi: 10.1101/2020.05.28.103077

Coutanche, M. N., and Thompson-Schill, S. L. (2013). Informational connectivity: identifying synchronized discriminability of multi-voxel patterns across the brain. *Front. Hum. Neurosci.* 7:15. doi: 10.3389/fnhum.2013.00015

DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron* 73, 415–434. doi: 10.1016/j.neuron.2012.01.010

Friston, K., Moran, R., and Seth, A. K. (2013). Analysing connectivity with Granger causality and dynamic causal modelling. *Curr. Opin. Neurobiol.* 23, 172–178. doi: 10.1016/j.conb.2012.11.010

Geerligs, L., and Henson, R. N. (2016). Functional connectivity and structural covariance between regions of interest can be measured more accurately using multivariate distance correlation. *Neuroimage* 135, 16–31. doi: 10.1016/j.neuroimage.2016.04.047

Goddard, E., Carlson, T. A., and Woolgar, A. (2021). Spatial and feature-selective attention have distinct effects on population-level tuning. *J. Cogn. Neurosci.* 34, 1–23. doi: 10.1162/jocn_a_01796

Goddard, E., Carlson, T. A., Dermody, N., and Woolgar, A. (2016). Representational dynamics of object recognition: feedforward and feedback information flows. *Neuroimage* 128, 385–397. doi: 10.1016/j.neuroimage.2016.01.006

Gregoriou, G. G., Gotts, S. J., Zhou, H., and Desimone, R. (2009). High-frequency, long-range coupling between prefrontal and visual cortex during attention. *Science* 324, 1207–1210. doi: 10.1126/science.1171402

Hlinka, J., Paluš, M., Vejmelka, M., Mantini, D., and Corbetta, M. (2011). Functional connectivity in resting-state fMRI: is linear correlation sufficient? *Neuroimage* 54, 2218–2225. doi: 10.1016/j.neuroimage.2010.08.042

Karimi-Rouzbahani, H. (2018). Three-stage processing of category and variation information by entangled interactive mechanisms of peri-occipital and peri-frontal cortices. *Sci. Rep.* 8, 1–22. doi: 10.1038/s41598-018-30601-8

Karimi-Rouzbahani, H., Bagheri, N., and Ebrahimpour, R. (2017a). Hard-wired feed-forward visual mechanisms of the brain compensate for affine variations in object recognition. *Neuroscience* 349, 48–63. doi: 10.1016/j.neuroscience.2017.02.050

Karimi-Rouzbahani, H., Bagheri, N., and Ebrahimpour, R. (2017b). Invariant object recognition is a personalized selection of invariant features in humans, not simply explained by hierarchical feed-forward vision models. *Sci. Rep.* 7, 1–24. doi: 10.1038/s41598-017-13756-8

Karimi-Rouzbahani, H., Shahmohammadi, M., Vahab, E., Setayeshi, S., and Carlson, T. (2021b). Temporal variabilities provide additional category-related information in object category decoding: a systematic comparison of informative EEG features. *Neural Comput.* 33, 3027–3072. doi: 10.1162/neco_a_01436

Karimi-Rouzbahani, H., Ramezani, F., Woolgar, A., Rich, A., and Ghodrati, M. (2021a). Perceptual difficulty modulates the direction of information flow in familiar face recognition. *Neuroimage* 233:117896. doi: 10.1016/j.neuroimage.2021.117896

Karimi-Rouzbahani, H., Vahab, E., Ebrahimpour, R., and Menhaj, M. B. (2019). Spatiotemporal analysis of category and target-related information processing in the brain during object detection. *Behav. Brain Res.* 362, 224–239. doi: 10.1016/j.bbr.2019.01.025

Karimi-Rouzbahani, H., Woolgar, A., and Rich, A. N. (2021c). Neural signatures of vigilance decrements predict behavioural errors before they occur. *eLife* 10:e60563. doi: 10.7554/eLife.60563

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K., Cichy, R. M., Hauk, O., and Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proc. Natl. Acad. Sci. U.S.A.* 116, 21854–21863. doi: 10.1073/pnas.1905544116

Kriegeskorte, N., and Kievit, R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends Cogn. Sci.* 17, 401–412. doi: 10.1016/j.tics.2013.06.007

Kriegeskorte, N., Mur, M., and Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* 2:4. doi: 10.3389/neuro.06.004.2008

Lahaye, P. J., Poline, J. B., Flandin, G., Dodel, S., and Garnero, L. (2003). Functional connectivity: studying nonlinear, delayed interactions between BOLD signals. *Neuroimage* 20, 962–974. doi: 10.1016/S1053-8119(03)00340-9

Nili, H., Walther, A., Alink, A., and Kriegeskorte, N. (2020). Inferring exemplar discriminability in brain representations. *PLoS One* 15:e0232551. doi: 10.1371/journal.pone.0232551

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS Comput. Biol.* 10:e1003553. doi: 10.1371/journal.pcbi.1003553

Shahbazi, M., Shirali, A., Aghajan, H., and Nili, H. (2021). Using distance on the Riemannian manifold to compare representations in brain and in models. *Neuroimage* 239:118271. doi: 10.1016/j.neuroimage.2021.118271

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., and Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *Neuroimage* 137, 188–200. doi: 10.1016/j.neuroimage.2015.12.012

Wilcoxon, F. (1992). "Individual comparisons by ranking methods," in *Breakthroughs in Statistics*, eds S. Kotz and N. L. Johnson (New York, NY: Springer), 196–202. doi: 10.1007/978-1-4612-4380-9_16

Winkowski, D. E., and Kanold, P. O. (2013). Laminar transformation of frequency organization in auditory cortex. *J. Neurosci.* 33, 1498–1508. doi: 10.1523/JNEUROSCI.3101-12.2013

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer AC declared a shared affiliation, with several of the authors, HK-R, AW, and RH, to the handling editor at the time of the review.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 21 March 2022
doi: 10.3389/fnins.2022.701632

Identification and Classification of Parkinsonian and Essential Tremors for Diagnosis Using Machine Learning Algorithms

Xupo Xing[1]†, Ningdi Luo[2]†, Shun Li[1], Liche Zhou[2], Chengli Song[1] and Jun Liu[2]**

[1] *Shanghai Institute for Minimally Invasive Therapy, School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai, China,* [2] *Department of Neurology and Institute of Neurology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China*

Edited by:
Reza Lashgari,
Shahid Beheshti University, Iran

Reviewed by:
Farheen Syeda,
Consultant, Baton Rouge, LA,
United States
Arun Singh,
University of South Dakota,
United States

****Correspondence:***
Chengli Song
csong@usst.edu.cn
Jun Liu
jly0520@hotmail.com

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Neural Technology, a section of the journal Frontiers in Neuroscience

Received: *28 April 2021*
Accepted: *02 March 2022*
Published: *21 March 2022*

Citation:
Xing X, Luo N, Li S, Zhou L, Song C and Liu J (2022) Identification and Classification of Parkinsonian and Essential Tremors for Diagnosis Using Machine Learning Algorithms. Front. Neurosci. 16:701632. doi: 10.3389/fnins.2022.701632

Due to overlapping tremor features, the medical diagnosis of Parkinson's disease (PD) and essential tremor (ET) mainly relies on the clinical experience of doctors, which often leads to misdiagnosis. Seven predictive models using machine learning algorithms including random forest (RF), eXtreme Gradient Boosting (XGBoost), support vector machine (SVM), logistic regression (LR), ridge classification (Ridge), backpropagation neural network (BP), and convolutional neural network (CNN) were evaluated and compared aiming to better differentiate between PD and ET by using accessible demographics and tremor information of the upper limbs. The tremor information including tremor acceleration and surface electromyogram (sEMG) signals were collected from 398 patients (PD = 257, ET = 141) and then were used to train the established models to separate PD and ET. The performance of the models was evaluated by indices of accuracy and area under the curve (AUC), which indicated the ensemble learning models including RF and XGBoost showed the best overall predictive ability with accuracy above 0.84 and AUC above 0.90. Furthermore, the relative importance of sex, age, four postures, and five tremor features was analyzed and ranked showing that the dominant frequency of sEMG of flexors, the average amplitude of sEMG of flexors, resting posture, and winging posture had a greater impact on the diagnosis of PD, whereas sex and age were less important. These results provide a reference for the intelligent diagnosis of PD and show promise for use in wearable tremor suppression devices.

Keywords: Parkinsonian tremor, essential tremor, tremor differentiation, machine learning algorithms, upper limb posture

INTRODUCTION

Parkinson's disease (PD) and essential tremor (ET) are two common diseases usually accompanied by tremors of the upper limbs, which may severely impair motor function and have a negative influence on patients, especially in the aging population (Helmich et al., 2013). The symptoms of PD are complex and severe in the later stages; therefore, early diagnosis and effective treatment are crucial (Mark, 2007).

Owing to overlapping tremor features, it remains difficult to distinguish between PD and ET (Algarni and Fasano, 2018). Given that there is currently no gold standard to differentiate between PD and ET, the diagnosis of the two diseases mainly relies on the clinical experience of doctors (Thenganatt and Jankovic, 2016). Individuals diagnosed with PD typically have gradual development of non-motor symptoms for years before movement symptoms begin, but often they will not mention these symptoms unless specifically queried (Armstrong and Okun, 2020). Dopamine replacement therapy works better to diagnose PD. However, it could be difficult in the early stage of the disease and thus approximately a quarter of PD are misdiagnosed as ET, which usually causes the optimal medical treatments of the two diseases to be overlooked (Rizzo et al., 2016; Reich and Savitt, 2019; Armstrong and Okun, 2020).

Some efficient and accessible non-invasive biomarkers such as tremor signals including tremor acceleration and surface electromyogram (sEMG) have been investigated for the differentiation between PD and ET (Meigal et al., 2013; Barrantes et al., 2017). And a series of statistical characteristics of tremor signals including the dominant frequency and peak value were extracted and studied for distinguishing PD and ET (Hossen et al., 2010; Thanawattano et al., 2015; De Oliveira Andrade et al., 2020).

Artificial intelligence technology is widely used to solve problems in the medical field, including differentiating between PD and ET (Xiao et al., 2019; Duque et al., 2020). Based on various extracted statistical characteristics of tremor signals and methodologies of machine learning, a series of machine learning algorithms, such as linear models (logistic regression, ridge classification, etc.), ensemble learning models (random forest, XGBoost, etc.), the kernel-based model (support vector machine, etc.), and neural network models (backpropagation neural network, convolutional neural network, etc.) have been introduced for the diagnosis and progression prediction of PD and ET (Ai et al., 2011; Hossen, 2013; Ahmadi Rastegar et al., 2019; Hssayeni et al., 2019; Qin et al., 2019).

Tremors of the upper limbs in PD patients are mainly manifested as a resting tremor which can be used as an important symptom to distinguish between PD and ET, however, only 20% of ET patients suffer from that (Oren Cohen et al., 2003; Jankovic, 2008; Helmich et al., 2013). In addition to resting posture, stretching posture and some novel postures were introduced and investigated to evaluate their ability to discriminate PD from ET showing that tremors information collected from various postures behaves more effectively in differentiating between PD and ET compared to a single posture (Zhang et al., 2018).

Although research has been carried out by using tremor information of the upper limbs to differentiate PD and ET, the influence of various upper limb postures, tremor features, and demographics on the diagnosis has been rarely studied. To help clinicians better distinguish between PD and ET, we evaluated and compared seven prediction models using machine learning algorithms. Based on the results, we analyzed and compared the relative importance of various upper limb postures, tremor features, and demographics in the diagnosis of the two diseases.

MATERIALS AND METHODS

Subjects and Data Collection

A total of 398 patients confirmed PD or ET with upper limb tremors were recruited for the experiment from June 2020 to November 2020 by the Department of Neurology of Rui Jin Hospital (Shanghai, China). With the help of a medical device

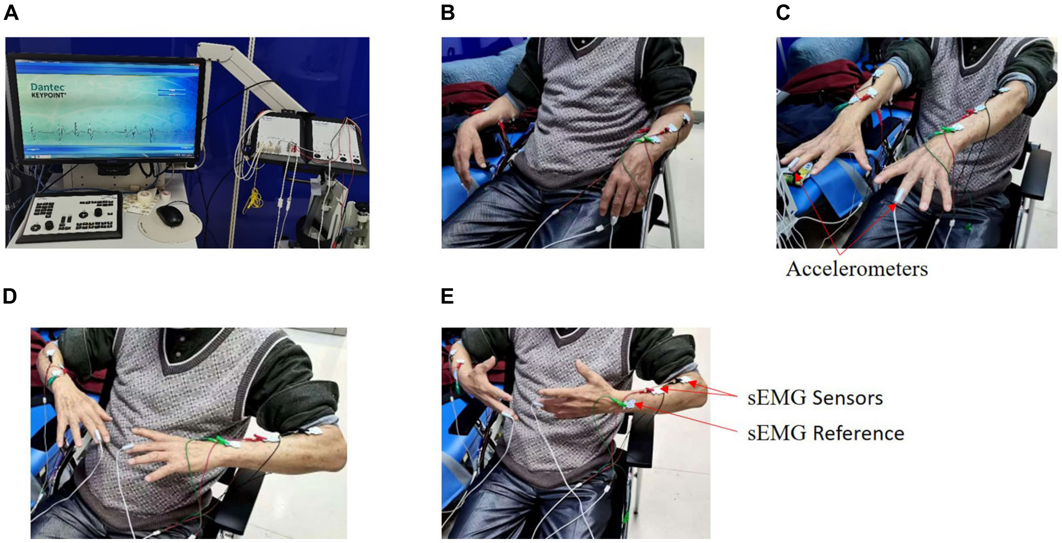


FIGURE 1 | Experimental setup. Tremor information was collected from four postures by a medical device system called Dantec® Keypoint® G4 for each patient. **(A)** Dantec® Keypoint® G4. **(B)** Resting posture. **(C)** Stretching posture. **(D)** Winging posture. **(E)** Vertically winging posture.

system (Dantec® Keypoint® G4, Natus Medical Inc.), the tremor information, including acceleration and sEMG, was collected from four postures for each subject. And most of the subjects were tested on medication.

Two accelerometers were fixed onto the distal finger of both hands, respectively, and six sEMG sensors were fixed onto the extensor and flexor muscles on both sides. In this experiment, each patient performed four respective postures (**Figure 1**): resting, stretching, winging, and vertically winging, meanwhile acceleration measurements and sEMG measurements were acquired.

For each patient, the sensor signals were measured for 30 s in each posture and sampled at a rate of 12,000 Hz. Patients were asked to avoid unrelated behaviors, and irrelevant personnel were removed from the room throughout the experiment. The demographics of age and sex for each patient were also recorded.

For each posture, five tremor features (each tremor feature with two tremor variables), including the dominant frequency of the acceleration signals, the dominant frequency of sEMG (extensor), the dominant frequency of sEMG (flexor), the average amplitude of sEMG (extensor), and the average amplitude of sEMG (flexor), were acquired by the Dantec® Keypoint® G4 medical device system. Finally, a total of 40 tremor variables (**Table 1**) were obtained from the four postures. Our study was approved by the local ethics committee of Shanghai Jiao Tong University.

Establishment of Models

Based on several predictive models widely adopted in many clinical applications, seven predictive models, including random forest (RF), eXtreme gradient boosting (XGBoost), support vector machine (SVM), backpropagation neural network (BP), ridge classification (Ridge), logistic regression (LR), and convolution neural network (CNN), were established and compared to differentiate PD and ET using tremor information collected from upper limbs.

For the linear models, LR and Ridge were selected. For the ensemble learning models, such as RF and XGBoost, multiple evaluators were established using the sample, and an output response was obtained after considering and aggregating the results of multiple evaluators. And a traditional machine learning algorithm, SVM, was built. Finally, the neural network models, including BP and CNN, were selected due to their powerful non-linear learning ability and extensive application to diagnose and predict the progression of PD (Hossen, 2013).

Because of different principles and usage between the CNN model and the other six models, the raw sensor signals, including the acceleration measurement and sEMG measurement of upper limbs, were used to train the CNN model to differentiate PD and ET. Due to the large volume of the time-series data which needs to be further processed for CNN, we did not combine demographic data to train the model. For the other six models, 40 tremor variables acquired from the Dantec® Keypoint® G4 medical device system, as well as two demographics (sex and age), were used to train these models. Therefore, for CNN and the other six models, the data preprocessing and training of the models were different.

TABLE 1 | Demographic data of 398 patients.

Cases (*n* = 398, Male 196, Female 22)			**Mean**	**SD**
Age			66.23	40.85
Resting posture	Dominant frequency	Acc (L)	3.41	3.17
		Flexor (L)	8.83	4.21
		Extensor (L)	8.77	4.26
		Acc (R)	3.48	2.97
		Flexor (R)	8.31	4.53
		Extensor (R)	8.95	4.11
	Average amplitude	Flexor (L)	212.38	173.45
		Extensor (L)	201.94	119.85
		Flexor (R)	202.28	120.14
		Extensor (R)	171.96	88.05
Stretching posture	Dominant frequency	Acc (L)	3.99	2.90
		Flexor (L)	9.21	4.30
		Extensor (L)	10.33	4.55
		Acc (R)	3.38	2.78
		Flexor (R)	9.38	3.93
		Extensor (R)	10.23	4.53
	Average amplitude	Flexor (L)	167.54	76.38
		Extensor (L)	203.56	73.17
		Flexor (R)	173.75	97.12
		Extensor (R)	210.59	76.67
Winging posture	Dominant frequency	Acc (L)	6.25	2.33
		Flexor (L)	7.59	4.09
		Extensor (L)	9.18	4.43
		Acc (R)	3.53	2.53
		Flexor (R)	8.68	3.91
		Extensor (R)	10.35	5.56
	Average amplitude	Flexor (L)	196.71	103.60
		Extensor (L)	213.69	79.56
		Flexor (R)	188.90	119.53
		Extensor (R)	217.25	81.16
Vertically winging posture	Dominant frequency	Acc (L)	4.17	2.33
		Flexor (L)	8.77	3.95
		Extensor (L)	9.70	4.33
		Acc (R)	3.66	7.88
		Flexor (R)	8.55	3.96
		Extensor (R)	9.68	4.39
	Average amplitude	Flexor (L)	196.01	98.92
		Extensor (L)	190.21	84.42
		Flexor (R)	182.90	83.32
		Extensor (R)	188.41	70.05

SD, standard deviation; L, left; R, right.

Acc (L) affiliated to "Dominant frequency" attached to "Resting posture": the dominant frequency of the acceleration signal collected from the left hand; Flexor (L) affiliated to "Dominant frequency" attached to "Resting posture": the dominant frequency of the surface EMG signal collected from the flexor on the left hand; Extensor (L) affiliated to "Dominant frequency" attached to "Resting posture": the dominant frequency of the surface EMG signal collected from the extensor on the left hand. And the others have similar meanings.

For these six models (RF, XGBoost, SVM, BP, Ridge, and LR), data preprocessing was performed as follows. For each patient, 40 tremor variables and two demographics (sex and age) were used as the variables with the diagnosis of either PD or ET as the labels, resulting in a total of 398 samples. **Table 1** indicates the two demographics (sex and age) and a total of 40 tremor variables affiliated to four postures, with each posture having ten tremor variables. First, we filled in the null values with the mean value of each variable (Zheng and Casari, 2018; Géron, 2019). Then, we scaled the data using Z-score normalization (Eq. 1) to enhance the predictive ability of the model and prevent overfitting (Géron, 2019).

$$z = \frac{x - u}{\sigma} \quad (1)$$

Where$u$is the mean of the variable and$\sigma$is the standard deviation.

For the CNN model, data preprocessing was performed as follows. Raw acceleration and sEMG measurements were used to train the CNN model. The middle 25 s of each signal was selected to avert potential noise in the experimental procedure, and then the extracted data were down-sampled to 120 Hz for ease of calculation, following which these down-sampled signals were converted to the frequency domain using a fast Fourier transform (FFT). Because the frequency band of pathological tremors is mainly in the 2–20 Hz range, the FFT signals at 2–20 Hz were finally chosen. The 24 converted signals from the acceleration measurement and sEMG measurement were stacked along the vertical axis to form a two-dimensional array for CNN input (**Figure 2**), and they were scaled using Eq. 1 (Kim et al., 2018).

Training of Models

Some parameters were selected and adjusted using the grid search method to acquire the best parameter combination for each model. **Table 2** lists the technical parameters of the models. First, the data were preprocessed as described above and then randomly divided into a training set (80%) and a validation set (20%). The proportion of PD and ET in the training set was consistent with that in the validation set.

TABLE 2 | Tuning parameters of the seven models.

Models	Tuning
RF	n_estimators (subtrees)
XGBoost	max_depth(maximum depth of number)
SVM	γ(Gaussian kernel), C(Cost)
BP	Size (hidden layer units); α(Regulation parameter)
Ridge	α(Regulation parameter)
LR	C (reciprocal of Regulation parameter)
CNN	The number of convolutional layers, the number of kernels

RF, random forest; XGBoost, eXtreme Gradient Boosting; SVM, support vector machine; BP, backpropagation neural network; LR, logistic regression; Ridge, ridge classification; CNN, convolutional neural network.

Ten-fold cross-validation was applied to the training set to obtain the optimal model parameters. The training set was divided into ten parts, nine of which were used to train the model in turn; the remaining one was used to test the model. The average value of AU-ROC, which was calculated ten times, was used as an indicator to evaluate the model for determining the different parameter combinations for each model. A forecast flow chart is shown in **Figure 3**. Because of the high sampling frequency and lack of good connectivity between muscles and sensors in some aged patients, some acceleration measurements or sEMG measurements were corrupted and became distorted, which led to only 188 samples could finally being used to train the CNN model.

For the CNN model, a specially formulated structure (**Figure 4**) containing several layers of neural networks was established to distinguish between PD and ET. The first layer

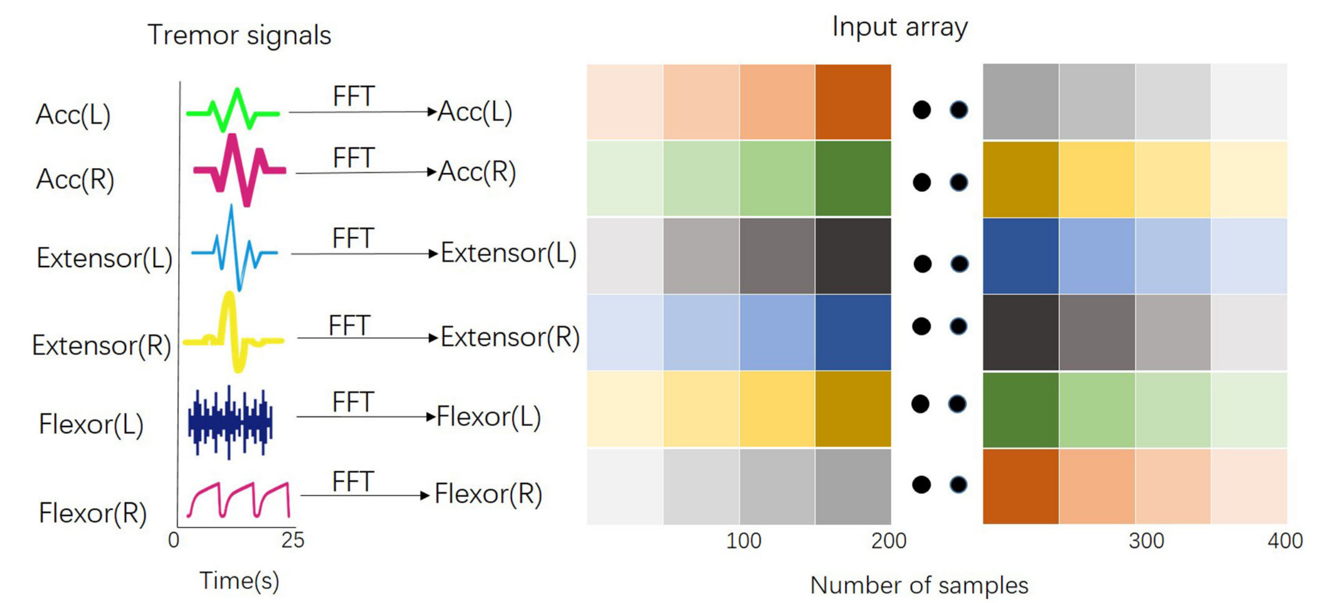


FIGURE 2 | Input array for the training of the CNN model. All signals from the acceleration measurement and sEMG measurement have been converted into the frequency domain by the Fast Fourier Transform and stacked along the vertical axis to form a two-dimensional array for CNN input.

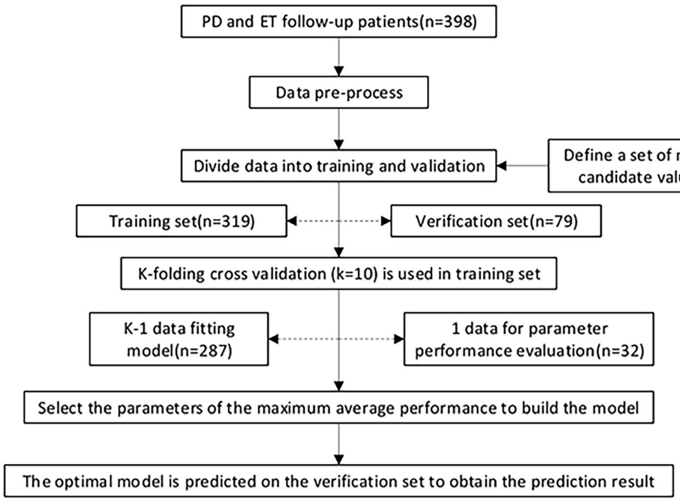

PD and ET follow-up patients(n=398)
Data pre-process
Divide data into training and validation
Define a set of model parameter candidate values to evaluate
Training set(n=319)
Verification set(n=79)
K-folding cross validation (k=10) is used in training set
K-1 data fitting model(n=287)
1 data for parameter performance evaluation(n=32)
Select the parameters of the maximum average performance to build the model
The optimal model is predicted on the verification set to obtain the prediction result

FIGURE 3 | Model training, parameter adjustment, and performance evaluation. 398 patients were recruited in the current study. The data were pre-processed and randomly divided into a training set (80%) and a validation set (20%), and the proportion of the two class proportions in each set is the same. In the training set, k-fold cross-validation (k = 10) is used, and various parameter combinations are exhausted by grid search. Performance evaluation index of AUC was adopted to judge the average predictive performance of the model. The average performance maximum is used as the best performance tuning parameter, and the prediction is finally performed on the test set.

of the convolutional neural network received a normalized two-dimensional input array, and 4 × 20 convolution kernels with 4 × 5 strides were used to fuse the local signal information from a signal sensor with the output size of 6 × 73. The second convolutional layer with 2 × 10 convolution filters and 2 × 2 strides was used to extract the sensor information. After each convolutional layer, a batch normalization layer and a dropout layer with a 30% dropout rate were used to avoid overfitting. Finally, a fully connected layer and a softmax classifier were used to distinguish between PD and ET.

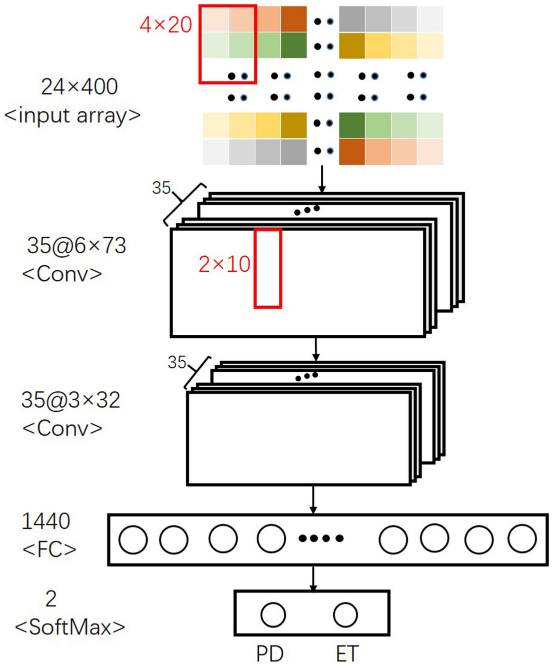


FIGURE 4 | Final CNN architecture for separating PD from ET.

Evaluation of Models

Evaluation indicators, including the confusion matrix, accuracy, area under the curve (AUC), recall (TPR, sensitivity), specificity, F1, false positive rate (FPR,1- specificity), and precision calculated by true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), were used to evaluate the performance of each model (Eqs 2–7). And higher AUC value indicates a better overall performance of the current feature, ς.

$$TPR = \frac{TP}{TP + FN} \tag{2}$$

$$FPR = \frac{FP}{FP + TN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{5}$$

$$AUC = \int_{-\infty}^{\infty} TPR(\varsigma) - FPR(\varsigma) d\varsigma \tag{6}$$

$$F1 = 2 \times \frac{Recall \times Precision}{Recall + Precision} \tag{7}$$

where AUC denotes the area under the curve value of the variable ς .

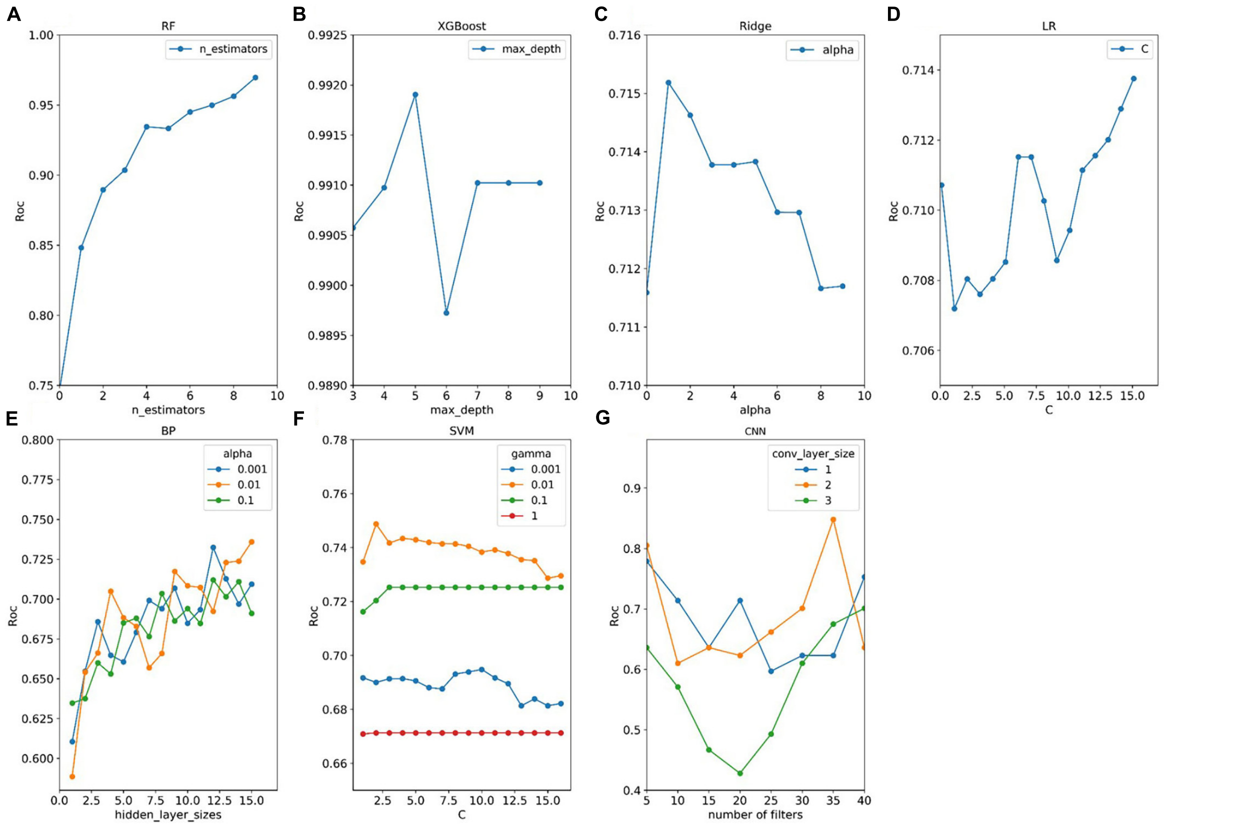


FIGURE 5 | Tuning results of model parameters. **(A–G)** Four models (RF, XGBoost, Ridge, and LR) have one adjustment parameter, and three models (BP, SVM, and CNN) have two adjustment parameters. For each set of parameters, the model parameters were evaluated for fit using the procedure described in panel **Figure 2**. The optimal parameters for each model are selected by obtaining the parameters that the model evaluates to the maximum.

Furthermore, we analyzed the relative importance of the variables in each model, except for CNN. The models XGBoost and RF allowed the importance of variables to be derived during model training; the coefficients of the Ridge model were used as the importance factor.

For models, such as LR, BP, and SVM, wherein the importance of variables was difficult or impossible to extract, the mean decrease accuracy was obtained by directly measuring the effect of each feature on the accuracy of the model. Briefly, the model was fitted, and parameter adjustment was performed to predict the validation set to obtain the model performances. Then, the feature values were disturbed to establish a new disturbance prediction set. Obviously, for the unimportant variables, the

TABLE 3 | Confusion matrices of seven models.

Confusion matrix	Actual	Prediction	
		PD	**ET**
RF	PD	44	7
	ET	6	22
XGBoost	PD	49	2
	ET	10	18
SVM	PD	50	1
	ET	27	1
BP	PD	41	10
	ET	12	16
Ridge	PD	38	13
	ET	16	12
LR	PD	40	11
	ET	9	19
CNN	PD	19	3
	ET	5	10

AUC, area under the curve; PD, Parkinson's disease; ET, essential tremor.

TABLE 4 | Evaluation summary based on AUC, recall, specificity, accuracy, FPR and precision.

Models	AUC	Recall	Specificity	Accuracy	FPR	Precision	F1
RF	0.90	0.86	0.79	0.84	0.21	0.88	0.87
XGBoost	0.95	0.96	0.64	0.85	0.36	0.83	0.89
SVM	0.81	0.98	0.04	0.65	0.96	0.65	0.78
BP	0.75	0.80	0.57	0.72	0.43	0.77	0.78
Ridge	0.71	0.75	0.43	0.63	0.57	0.70	0.72
LR	0.73	0.78	0.68	0.75	0.32	0.82	0.80
CNN	0.77	0.86	0.67	0.78	0.33	0.79	0.83

FPR, false positive rate.

scrambling order has little effect on the accuracy of the model, but for the important variables, the scrambled order will reduce the accuracy of the model. Finally, the relative importance ratio of all the eigenvalues was given a weight between 0 and 1 according to the overall proportion.

We added the relative importance of the ten tremor variables affiliated to each posture as the relative importance of the four postures, respectively. In addition, we added the relative importance of the two tremor variables affiliated to each tremor feature attached to the four postures as the relative importance of the five tremor features, respectively, thereby obtaining the effect sizes.

RESULTS

Tuning of Parameters

The average AU-ROC for different models and their parameters are listed (**Figure 5**). In these models, XGBoost obtained the best overall performance, and the parameter max_depth of five was optimal. RF achieved optimal performance as the parameter n_estimators reached nine. A two-layered CNN architecture with 35 convolution kernels was developed (**Figure 4**). The other four models had a similar performance, with a maximum performance index of approximately 0.7. The cost (C) of SVM was two, and the parameter gamma of 0.01 produced the best performance. For LR, parameter C (reciprocal of the regulation parameter) of 15 performed the best. For BP, parameter hidden layer sizes of 15 and an alpha of 0.01 produced the best performance. The alpha of the Ridge was one, which enabled the optimal performance.

Validation of the Training Set

The confusion matrices of the seven models are displayed in **Table 3**. The number of actual subjects of PD and ET in the confusion matrix is 51 and 28, respectively.

For RF and XGBoost, the sum of false negatives (FNs) and false positives (FPs) could be controlled within 13, while the others had a sum of FNs and FPs above 20 (79 validation samples). For CNN, the sum of FNs and FPs was eight (37 validation samples). The evaluation indices, including recall (TPR, sensitivity), specificity, accuracy, FPR (1-specificity), and F1 for each model, are displayed in **Table 4**. For the ensemble learning models, RF and XGBoost show a better performance, with an accuracy rate equal to and above 0.84. XGBoost has a higher accuracy rate than RF. However, the specificity of RF is higher, which means that it has a higher accuracy rate in identifying ET patients. For the neural networks, the accuracy of BP and CNN reaches 0.72 and 0.78, respectively. Compared with BP, the CNN model has a stronger non-linear predictive ability. In this study, the accuracy of CNN was also higher than that of BP. However, the neural network did not perform

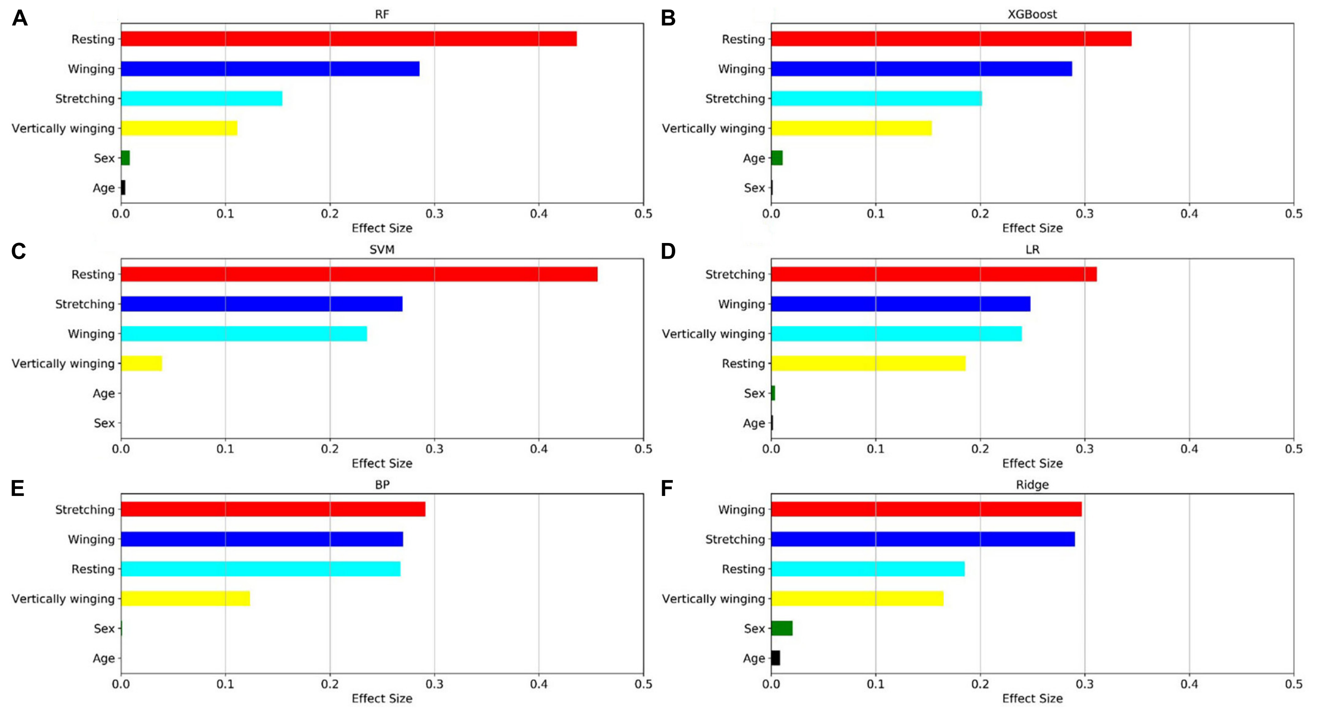


FIGURE 6 | Factors effect size. The **(A–F)** histogram displays the proportion of the factoric importance of sex, age, and four postures calculated by the models. For each model, the relative importance is quantified by assigning a weight between 0 and 1 for each variable and then the relative importance of the four postures is calculated by the sum of the factoric importance of the corresponding variables affiliated to that posture. The models XGBoost and RF allow the importance of variables to be derived during model training; the coefficients of the Ridge model are used as the basis for factor importance; the LR, BP, and SVM models are obtained by the Mean decrease accuracy method.

well owing to the limited number of samples. The Ridge linear model obtained the lowest accuracy rate of 0.63 and the lowest AU-ROC value of 0.71.

Important Features

The relative importance of sex, age, and the four postures (resting, stretching, winging, and vertically winging), were calculated using the models displayed in **Figure 6**. The relative importance of sex, age, and the five tremor features, including the dominant frequency of acceleration of distal fingers (Dom_fre_acc), the dominant frequency of sEMG of extensors (Dom_fre_ext), the dominant frequency of sEMG of flexors (Dom_fre_fle), the average amplitude of sEMG of extensors (Ave_amp_ext), and the average amplitude of sEMG of flexors (Ave_amp_fle), were calculated by the models as displayed in **Figure 7**.

Among the seven established models, the ensemble learning models, including RF and XGBoost showed the best prediction capabilities. Thus, the relative importance obtained from these two models was adopted. In the two models, the relative levels of importance of sex, age, the four postures, and the five tremor features were ranked showing that resting posture, winging posture, Dom_fre_fle, and Ave_amp_fle had a significant influence on the predictability of the models, whereas sex and age had a slight impact on the prediction.

DISCUSSION

Most PD and ET patients suffer from tremors of the upper limbs (Zhang et al., 2018; Duque et al., 2020). Owing to the overlapping tremor features, misdiagnosis between PD and ET is common. As a non-invasive biomarker, the tremor information of upper limbs, including acceleration and sEMG, has been investigated to distinguish PD from ET. Although some tremor features (tremor amplitude, dominant frequency, etc.) from various upper limb postures are extracted for the differentiation of PD and ET, the relative importance of the tremor features and various upper limb postures have been less frequently investigated.

In this study, we applied the tremor signals, including the acceleration measurements and sEMG measurements, which were collected from the four upper limb postures and two

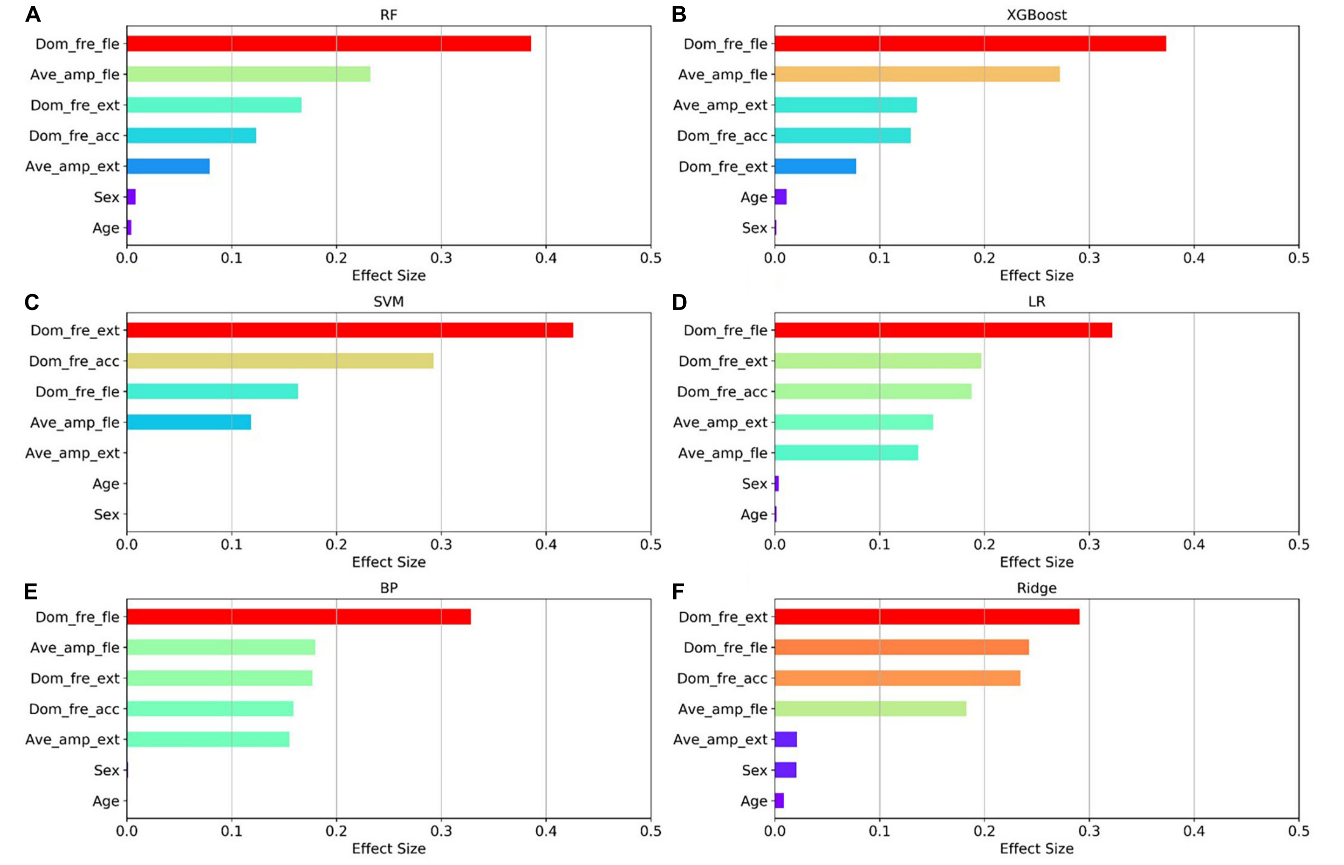


FIGURE 7 | Factors effect size. The **(A–F)** histogram displays the proportion of the factoric importance of sex, age, and five tremor features calculated by the models. For each model, the relative importance is quantified by assigning a weight between 0 and 1 for each variable and then the relative importance of the five tremor features is calculated by the sum of the factoric importance of the corresponding variables affiliated to that tremor feature. The models XGBoost and RF allow the importance of variables to be derived during model training; the coefficients of the Ridge model are used as the basis for factor importance; the LR, BP, and SVM models are obtained by the Mean decrease accuracy method.

demographics (sex and age) to distinguish PD from ET using seven machine learning algorithms. The ensemble learning models RF and XGBoost provided a rapid classification of outpatients. Various complex models could be established, and accurate decisions could be made using machine learning algorithms when given certain data. In this study, we used a dataset with a size of 398 and 42 dimensions. It was proved that the ensemble learning models performed better than the other models and fulfilled the clinical needs.

It may be considered that the current sample was not sufficient to support the result owing to the limited sample size. In the case of small data size and high data dimensions, the ensemble learning classifier XGBoost and RF could separate samples more effectively, whereas the other models of SVM, LR, BP, Ridge, and CNN exhibited a lower accuracy. Owing to the high data dimensions, SVM had a low predictive ability, resulting in most samples being predicted as PD, and Ridge had the lowest accuracy rate. The more complex neural network model with a powerful non-linear learning ability also did not perform well. In this study, among the seven established models, the ensemble learning models RF and XGBoost performed ideally, while the other five models lacked a significant predictive ability.

Although some assistive engineering approaches using tremor information of the upper limbers collected by wearable sensors have been proposed to differentiate between PD and ET, the results are less convincing limited by a few subjects. In this paper, we evaluated seven classification models using machine learning algorithms to differentiate PD and ET by using accessible demographics and tremor information of the upper limbs collected from various postures. The results with AUC above 0.90 and accuracy above 0.84 for RF and XGBoost models are convincing because more subjects (398 cases) were collected and the data was adequate compared with previous studies. Furthermore, we firstly analyzed and ranked the relative importance of sex, age, the four postures, and the five tremor features for differentiating PD and ET, which could help the diagnosis of PD in the early stage.

Recent progress in artificial intelligence and wearable technology has made wearable tremor suppression devices for PD a potentially viable alternative for tremor management. The relative importance of sex, age, the four postures, and the five tremor features, provides a reference for the intelligent diagnosis of PD and shows promise for use in wearable tremor suppression devices. To further enhance the performance of the established models, more ET subjects will be recruited in the subsequent study.

CONCLUSION

In this study, seven models were evaluated and compared for separation of PD from ET by using the tremor information of the upper limbs in various postures. It was determined that the ensemble learning models, including RF and XGBoost, had the greatest overall predictive ability and could effectively distinguish PD and ET. We also found that the dominant frequency of flexor sEMG, the average amplitude of flexor sEMG, the resting posture, and the winging posture had a greater impact on the predictability of the models, whereas the other predictors, specifically sex and age, were less important. These results provide a reference for the intelligent diagnosis of PD and are promising for use in wearable tremor suppression devices. This study investigating the differentiation between PD and ET using machine learning algorithms was preliminary. With the further acquisition of data of ET subjects in future work, the performance of models will be further improved and more valuable results will be obtained.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Local Ethics Committee of Shanghai Jiao Tong University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

XX, NL, LZ, CS, and JL designed the experiments. XX and NL analyzed the dataset and drafted the manuscript. SL performed the experiments. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (51901137 and 51735003) and the Medicine-Engineering Cross-disciplinary Fund of the University of Shanghai for Science and Technology.

REFERENCES

Ahmadi Rastegar, D., Ho, N., Halliday, G. M., and Dzamko, N. (2019). Parkinson's progression prediction using machine learning and serum cytokines. *NPJ Parkinsons Dis.* 5:14. doi: 10.1038/s41531-019-0086-4

Ai, L., Wang, J., and Yao, R. (2011). Classification of Parkinsonian and essential tremor using empirical mode decomposition and support vector machine. *Digit. Signal Process.* 21, 543–550. doi: 10.1016/j.dsp.2011.01.010

Algarni, M., and Fasano, A. (2018). The overlap between essential tremor and Parkinson disease. *Parkinsonism Relat. Disord.* 46, S101–S104. doi: 10.1016/j.parkreldis.2017.07.006

Armstrong, M. J., and Okun, M. S. (2020). Diagnosis and treatment of Parkinson disease: a review. *JAMA* 323, 548–560. doi: 10.1001/jama.2019.22360

Barrantes, S., Sánchez Egea, A. J., González Rojas, H. A., Martí, M. J., Compta, Y., Valldeoriola, F., et al. (2017). Differential diagnosis between Parkinson's disease and essential tremor using the smartphone's accelerometer. *PLoS One* 12:e0183843. doi: 10.1371/journal.pone.0183843

De Oliveira Andrade, A., Paixão, A. P. S., Cabral, A. M., Rabelo, A. G., Luiz, L. M. D., Dionísio, V. C., et al. (2020). Task-specific tremor quantification in a clinical setting for Parkinson's disease. *J. Med. Biol. Eng.* 40, 821–850. doi: 10.1007/s40846-020-00576-x

Duque, J. D. L., Egea, A. J. S., Reeb, T., Rojas, H. A. G., and Gonzalez-Vargas, A. M. (2020). Angular velocity analysis boosted by machine learning for helping in the differential diagnosis of Parkinson's disease and essential tremor. *IEEE Access* 8, 88866–88875. doi: 10.1109/ACCESS.2020.2993647

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*. California, CA: O'Reilly Media.

Helmich, R. C., Toni, I., Deuschl, G., and Bloem, B. R. (2013). The pathophysiology of essential tremor and Parkinson's tremor. *Curr. Neurol. Neurosci. Rep.* 13:378. doi: 10.1007/s11910-013-0378-8

Hossen, A. (2013). A neural network approach for feature extraction and discrimination between Parkinsonian tremor and essential tremor. *Technol. Health Care* 21, 345–356. doi: 10.3233/THC-130735

Hossen, A., Muthuraman, M., Raethjen, J., Deuschl, G., and Heute, U. (2010). Discrimination of Parkinsonian tremor from essential tremor by implementation of a wavelet-based soft-decision technique on EMG and accelerometer signals. *Biomed. Signal Process. Control* 5, 181–188. doi: 10.1016/j.bspc.2010.02.005

Hssayeni, M. D., Jimenez-Shahed, J., Burack, M. A., and Ghoraani, B. (2019). Wearable sensors for estimation of Parkinsonian tremor severity during free body movements. *Sensors (Basel)* 19:4215. doi: 10.3390/s19194215

Jankovic, J. (2008). Parkinson's disease: clinical features and diagnosis. *J. Neurol. Neurosurg. Psychiatry* 79, 368–376. doi: 10.1136/jnnp.2007.131045

Kim, H. B., Lee, W. W., Kim, A., Lee, H. J., Park, H. Y., Jeon, H. S., et al. (2018). Wrist sensor-based tremor severity quantification in Parkinson's disease using convolutional neural network. *Comput. Biol. Med.* 95, 140–146. doi: 10.1016/j.compbiomed.2018.02.007

Mark, L. (2007). Overview of Parkinson‘s disease. *Pharmacotherapy* 27(12 Pt 2), 155S–160S.

Meigal, A. Y., Rissanen, S. M., Tarvainen, M. P., Airaksinen, O., Kankaanpää, M., and Karjalainen, P. A. (2013). Non-linear EMG parameters for differential and early diagnostics of Parkinson's disease. *Front. Neurol.* 4:135. doi: 10.3389/fneur.2013.00135

Oren Cohen, M. D., Seth Pullman, M. D., EvaJurewicz, B. A., Dryden Watner, M. A., and Elan D.Louis, M. D. M. S. (2003). Rest tremor in patients with essential tremor prevalence, clinical correlates, and electrophysiologic characteristics. *Arch. Neurol.* 60, 405–410.

Qin, Z., Jiang, Z., Chen, J., Hu, C., and Ma, Y. (2019). sEMG-based tremor severity evaluation for Parkinson's disease using a light-weight CNN. *IEEE Signal Process. Lett.* 26, 637–641. doi: 10.1109/LSP.2019.2903334

Reich, S. G., and Savitt, J. M. (2019). Parkinson's disease. *Med. Clin. North Am.* 103, 337–350. doi: 10.1016/j.mcna.2018.10.014

Rizzo, G., Copetti, M., Arcuti, S., Martino, D., Fontana, A., and Logroscino, G. (2016). Accuracy of clinical diagnosis of Parkinson disease: a systematic review and meta-analysis. *Neurology* 86, 566–576. doi: 10.1212/WNL.0000000000002350

Thanawattano, C., Pongthornseri, R., Anan, C., Dumnin, S., and Bhidayasiri, R. (2015). Temporal fluctuations of tremor signals from inertial sensor: a preliminary study in differentiating Parkinson's disease from essential tremor. *Biomed. Eng. Online* 14, 101–114. doi: 10.1186/s12938-015-0098-1

Thenganatt, M. A., and Jankovic, J. (2016). The relationship between essential tremor and Parkinson's disease. *Parkinsonism Relat. Disord.* 22(Suppl 1), S162–S165. doi: 10.1016/j.parkreldis.2015.09.032

Xiao, J., Ding, R., Xu, X., Guan, H., Feng, X., Sun, T., et al. (2019). Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. *J. Transl. Med.* 17:119. doi: 10.1186/s12967-019-1860-0

Zhang, B., Huang, F., Liu, J., and Zhang, D. (2018). A novel posture for better differentiation between Parkinson's tremor and essential tremor. *Front. Neurosci.* 12:317. doi: 10.3389/fnins.2018.00317

Zheng, A., and Casari, A. (2018). *Feature Engineering for Machine Learning*. California, CA: O'Reilly Media.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 01 August 2022
DOI 10.3389/fnins.2022.744737

OPEN ACCESS

EDITED BY
Hamid R. Rabiee,
Sharif University of Technology, Iran

REVIEWED BY
Bertille Somon,
Université de Toulouse, France
Kamran Avanaki,
University of Illinois at Chicago,
United States

*CORRESPONDENCE
S. Hamid Amiri
s.hamidamiri@sru.ac.ir

†These authors have contributed equally to this work and share the first authorship

SPECIALTY SECTION
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

RECEIVED 20 July 2021
ACCEPTED 14 July 2022
PUBLISHED 01 August 2022

CITATION
Sarailoo R, Latifzadeh K, Amiri SH, Bosaghzadeh A and Ebrahimpour R (2022) Assessment of instantaneous cognitive load imposed by educational multimedia using electroencephalography signals. *Front. Neurosci.* 16:744737. doi: 10.3389/fnins.2022.744737

Assessment of instantaneous cognitive load imposed by educational multimedia using electroencephalography signals

Reza Sarailoo[1†], Kayhan Latifzadeh[1†], S. Hamid Amiri[1*†], Alireza Bosaghzadeh[1] and Reza Ebrahimpour[1,2]

[1]Artificial Intelligence Group, Faculty of Computer Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran, [2]School of Cognitive Sciences, Institute for Research in Fundamental Sciences, Tehran, Iran

The use of multimedia learning is increasing in modern education. On the other hand, it is crucial to design multimedia contents that impose an optimal amount of cognitive load, which leads to efficient learning. Objective assessment of instantaneous cognitive load plays a critical role in educational design quality evaluation. Electroencephalography (EEG) has been considered a potential candidate for cognitive load assessment among neurophysiological methods. In this study, we experiment to collect EEG signals during a multimedia learning task and then build a model for instantaneous cognitive load measurement. In the experiment, we designed four educational multimedia in two categories to impose different levels of cognitive load by intentionally applying/violating Mayer's multimedia design principles. Thirty university students with homogenous English language proficiency participated in our experiment. We divided them randomly into two groups, and each watched a version of the multimedia followed by a recall test task and filling out a NASA-TLX questionnaire. EEG signals are collected during these tasks. To construct the load assessment model, at first, power spectral density (PSD) based features are extracted from EEG signals. Using the minimum redundancy - maximum relevance (MRMR) feature selection approach, the best features are selected. In this way, the selected features consist of only about 12% of the total number of features. In the next step, we propose a scoring model using a support vector machine (SVM) for instantaneous cognitive load assessment in 3s segments of multimedia. Our experiments indicate that the selected feature set can classify the instantaneous cognitive load with an accuracy of $84.5 \pm 2.1\%$. The findings of this study indicate that EEG signals can be used as an appropriate tool for measuring the cognitive load introduced by educational videos. This can be help instructional designers to develop more effective content.

KEYWORDS

instantaneous cognitive load, EEG, classification, educational video, multimedia design

Introduction

Cognitive load is defined as the load being imposed on working memory while performing a cognitive task (Paas et al., 2004). There are three types of cognitive load: intrinsic, which is dependent on the nature of the task and cannot be modified by the designer; extraneous, which is related to the design of the task and can be altered by formatting the materials being presented; germane load which is associated with the amount of mental effort for building the schema in working memory (Sweller et al., 2019). Cognitive load assessment has a critical role in different areas such as education (Sweller, 2018) and human-computer interaction (HCI) designing (Zagermann et al., 2016). Multimedia plays an essential role in modern education. Keeping the amount of cognitive load at an optimum level is crucial in instructional design (Mutlu-Bayraktar et al., 2019). Mayer (2002), in his book Multimedia Learning, introduced twelve principles that help multimedia designers to minimize the amount of cognitive load on learners. Among these principles, five of them are devoted to extraneous processing, a type of cognitive processing in instructional multimedia learning, originating from the extra material in multimedia without any relevance to the instructional goal. The five principles for reducing extraneous processing are (1) *Coherence Principle*: extraneous words, images, and sounds should be excluded (e.g., attractive but non-related images); (2) *Signaling Principle*: essential materials should be highlighted with a cue (e.g., color highlight); (3) *Redundancy Principle*: in the presence of graphics and narration, the on-screen text should be excluded; (4) *Spatial Contiguity Principle*: corresponding words and images should be presented near to each other; (5) *Temporal Contiguity Principle*: corresponding words and images should be presented simultaneously, not successively. The effect of the introduced rules on cognitive load has been investigated based on behavioral, self-reported, and performance test data (Mayer and Mayer, 2005).

Cognitive load can be measured in five levels, within or between distinct tasks: overall, accumulated, average, peak, and instantaneous load (Antonenko et al., 2010). Instantaneous load reflects the amount of imposed cognitive load in each moment of a cognitive task (Antonenko et al., 2010). In general, there are two methods for cognitive load assessment: subjective [e.g., NASA-TLX questionnaire (Hart and Staveland, 1988)], and objective [e.g., electroencephalography (EEG) (Antonenko et al., 2010), eye-tracking (Pomplun and Sunkara, 2003; Barrios et al., 2004; Chen et al., 2011; Kruger and Doherty, 2016; Dalmaso et al., 2017; Latifzadeh et al., 2020), and fMRI (Tomasi et al., 2006)]. Subjective methods which are based on self-reporting have limitations for instantaneous or online assessment of cognitive load, and they are mainly being used for overall and average assessment of mental workload (Anmarkrud et al., 2019). In contrast, physiological measurements as objective methods have the advantage of measuring the cognitive load continuously and online during a cognitive task (Antonenko et al., 2010), such as video-based learning.

Electroencephalography as a neurophysiological measure with a high temporal resolution (approximately 1 ms) is a well-suited candidate for the assessment of cognitive load in educational environments because this method is objective, non-invasive, and less restricted in comparison to other neuroimaging methods (Antonenko et al., 2010). Nowadays, many portable EEG devices can be easily used in classrooms for cognitive load assessment (Xu and Zhong, 2018). Moreover, it has a high temporal resolution which is a good property for the assessment of instantaneous cognitive load. This ability may provide the opportunity to monitor the dynamics of cognitive load on working memory during a cognitive task such as multimedia learning. During the past decades, cognitive load has mainly been measured using subjective methods and behavioral data such as reaction times and error rates to perform specific tasks. According to the literature, EEG band power spectra (i.e., delta, theta, alpha, and beta) at different brain regions have been introduced to assess cognitive workload. Specially, theta and alpha have been linked to cognitive workload studies (Mazher et al., 2017; Puma et al., 2018; Castro-Meneses et al., 2020).

Several recent studies have empirically examined the relationship between cognitive demands and EEG activity at different frequency bands and brain regions. These studies have used EEG, alone or along with other subjective and objective measures, to assess participants' cognitive workload in different environments, including performing the arithmetic task (Borys et al., 2017; Plechawska-Wójcik et al., 2019), engaging in a virtual reality space (Dan and Reiner, 2017; Tremmel et al., 2019; Baceviciute et al., 2020), and being in a multitasking situation (Puma et al., 2018). Moreover, most studies utilized statistical analysis to assess cognitive states/conditions based on subjective, behavioral, and physiological measure (Baceviciute et al., 2020; Castro-Meneses et al., 2020; Scharinger et al., 2020). However, recent studies have been focused on the usage of machine learning methods to improve the performance of cognitive load measurements (Plechawska-Wójcik et al., 2019; Appriou et al., 2020; Rojas et al., 2020).

Borys et al. (2017) applied several classification methods on different combinations of EEG and eye-tracking features to classify cognitive workload states on arithmetic task. They calculated power spectra of three frequency bands (theta, alpha, and beta) acquired from five scalp locations (Cz, F3, F4, P3, and P4) as EEG features. Their results showed that none of the EEG features were used in the best classification model. One limitation of this research was concentration on the specific brain regions with low effect in reducing workload. In a study carried out by Dan and Reiner (2017), they focused on EEG-based measures for cognitive load assessment related to event processing in 2D displays against 3D virtual reality environments. They calculated the ratio of the average power of the middle frontal theta (Fz) and the central parietal alpha (Pz) as cognitive load indicator. They found that the cognitive load of processing 3D information is lower than 2D. In a subsequent study, Tremmel et al. (2019) evaluated the feasibility of passive monitoring of cognitive workload*via*EEG while performing a classical n-back task in an interactive VR environment. They extracted EEG spectral powers of four frequency bands (theta, alpha, beta, and gamma) from eight electrode positions (Fz, F3, F4, C3, C4, P3, P4, and Pz). The Results revealed the positive correlation of alpha activity in the parietal area with workload levels. In another experimental paradigm, Puma et al. (2018) used theta and alpha band power to assess cognitive workload in a multitasking environment. In this task, the participants completed a task commonly used in airline pilot recruitment, with an increasing number of concurrent sub-tasks from one phase to the next phase of the task. They conducted their EEG analysis only on five electrodes centered in the frontal area (Fz, F3, F4, F7, and F8) for the theta rhythm and five electrodes centered in the parietal area (Pz, P3, P4, P7, and P8) for the alpha rhythm. Besides these EEG features, the researchers collected performance, subjective (NASA-TLX) and pupillometry measurements as overall cognitive workload indicators. According to the results, the power of both theta and alpha bands increased with task difficulty, indicating the direct effect of these bands in cognitive load. Although different indicators have been proposed in the literature, it is essential to explore the most optimal indices for assessing cognitive load in a specific research area such as multimedia learning environments.

In addition, there are a few studies on using EEG for cognitive load assessment in multimedia and video-based learning. Wang et al. (2013) used EEG frequency bands to classify two videos labeled confusing and non-confusing based on the participants' self-reported feelings. They obtained an accuracy of 0.67 using a Gaussian Naïve Bayes classifier. In another study, Mazher et al. (2017) displayed identical video-based multimedia to their participants in three different sessions followed by a performance test. They assumed that by repeating the same content, cognitive load decreases. They also divided EEG signals into 10 s sections as the samples of their study. Using partial directed coherence (PDC) and support vector machine (SVM) classifiers, they inferred that the alpha band in the frontal and parietal lobes of the brain cortex could be a good indicator of cognitive load in multimedia learning. Lin and Kao (2018) showed that using Power Spectral Density (PSD) of all channels in EEG signal can discriminate different levels of mental effort in online educational videos. They examine three other models, including ANN, SVM, and decision tree. In a recent study, Castro-Meneses et al. (2020) assigned different levels of cognitive load based on the linguistic complexity of the presented content. They showed that theta oscillations are potentially an objective indicator of cognitive load.

In comparison to the previous related works, we follow an approach to reach the most informative brain regions and frequency bands associated with cognitive load. We assume that multimedia learning is a complex task in which different parts of the brain and may be different frequency bands are involved. Thus, it is hard to claim that only one or two regions of the brain in specific bands are important for measuring cognitive workload. Furthermore, we try to simulate the different conditions of instantaneous cognitive load in instructional videos by applying/violating the principles of multimedia which has rarely been attempted in the previous related works. We also investigate different time windows to find the optimal time frame for cognitive load assessment.

In this study, we aim to quantitatively measure the instantaneous cognitive load in multimedia learning using EEG signals. To this end, we design an experiment by applying/violating multimedia design principles to have two levels of cognitive load. Then, we build a classification model on the most informative spectral features. Using this model, we reach the goal of this manuscript, instantaneous cognitive load assessment. The rest of the manuscript is organized as follows. In the next section, we describe the materials of our study, including the educational videos, and the procedure of the experiment, and the methods that have been applied in our analyses. In section "Results," we report the results of the current study, and finally, in section "Discussion," a discussion on the results will be provided.

Materials and methods

Participants

Thirty-six university students between the ages of 18 and 25 participated in our experiment. Except for two, all other participants were male. The data acquired from six of them were discarded due to failure in recordings. The final set of our subjects includes thirty participants. We only excluded participants whose data were entirely corrupted. Thus, we tried to preserve as much data as possible for analysis. They are divided into two groups randomly to perform the task in two separate sessions. According to **Figure 1A**, 16 of them are in group 1 (LV1HV2) and 14 in group 2 (LV2HV1). Unfortunately, some participants participated only in one session and refused to continue the experiment due to their preferences. Thus, nine participants from group 1 and five participants from group 2 only watch one multimedia (see **Supplementary material** for detailed information about data management approach). The native language of all participants is Farsi (Persian), all of them are in the range of 23–32 in terms of listening skills of English which is evaluated by simulating the listening part of the International English

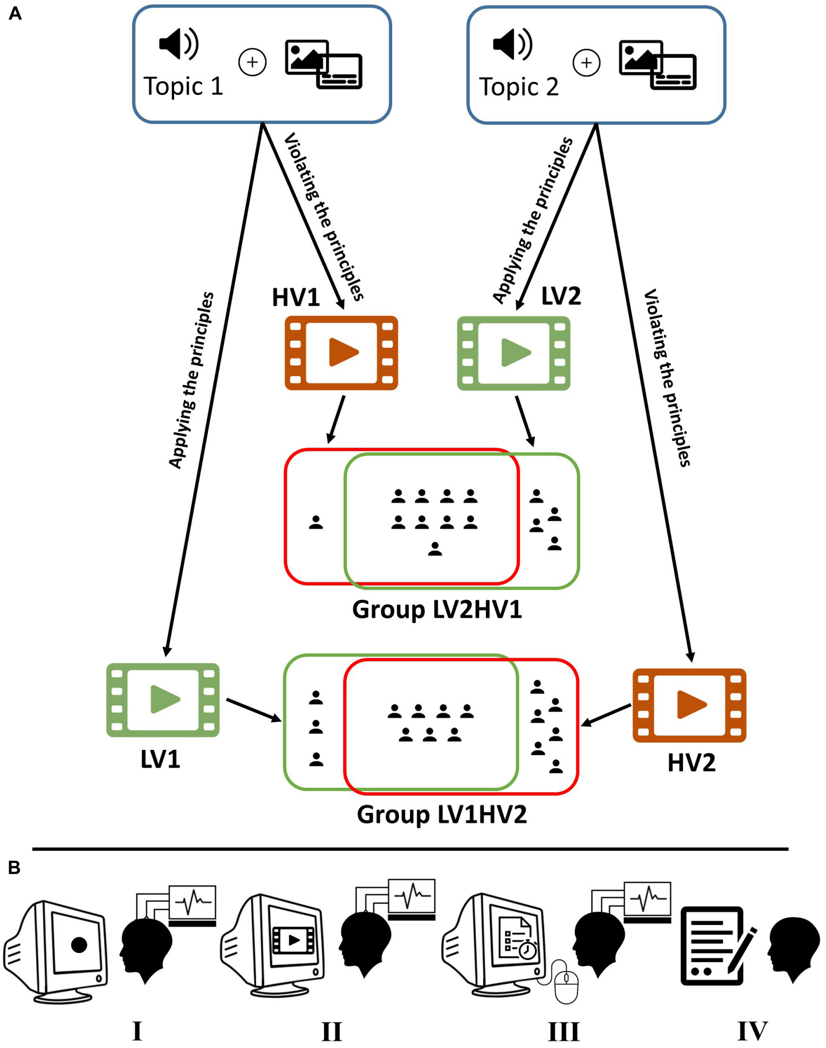


FIGURE 1
Experiment design. **(A)** Based on two audio narrations, four versions of videos (two for each narration) were created: LV1 (HV1) and LV2 (HV2) by applying (violating) the principles of multimedia design. Low load and high load video identified by green and red, respectively. Participants are randomly divided into two groups: LV1HV2 and LV2HV1. In our experiment, the LV1HV2 (LV2HV1) watched LV1 (LV2) and HV2 (HV1) videos in two separate sessions. As illustrated in the figure, in each group, some subjects participated only in one session. **(B)** The procedure of the experiment (left to right): first, looking at a black-filled circle for recording baseline data; second, watching the multimedia (no interaction); third, taking part in the recall test (via mouse interface); and finally completing the NASA-TLX questionnaire (paper-based version). In the first and second steps, electroencephalography (EEG) signals are collected.

Language Testing System (IELTS) exam. All participants were right-handed and had normal or corrected-to-normal eye vision. All participants signed informed written consent before attending the study. The experimental protocols were approved by the ethics committee of the Iran University of Medical Sciences.

Educational multimedia

We created four multimedia. In two of them, we apply the multimedia design principles to impose a minimum amount of extraneous cognitive load on our participants. In contrast, the other two multimedia are created by violating these principles to impose a higher amount of cognitive load on the subjects in our study. We selected two chapters of Open Forum 3 (Duncan and Parker, 2007) which are listening comprehension tasks; lesson 6 and lesson 11 (for online access to the resources, see https://elt.oup.com/student/openforum/3?cc=ir&selLanguage=en hosted on Oxford University Press). Using the audio of each lesson, we created two versions of motion-graphic-animation (low-load and high-load) as two multimedia for that lesson (see **Figure 2**). The videos corresponding to lessons 6 and 11 have the length of 290 and 342 s, respectively. Two

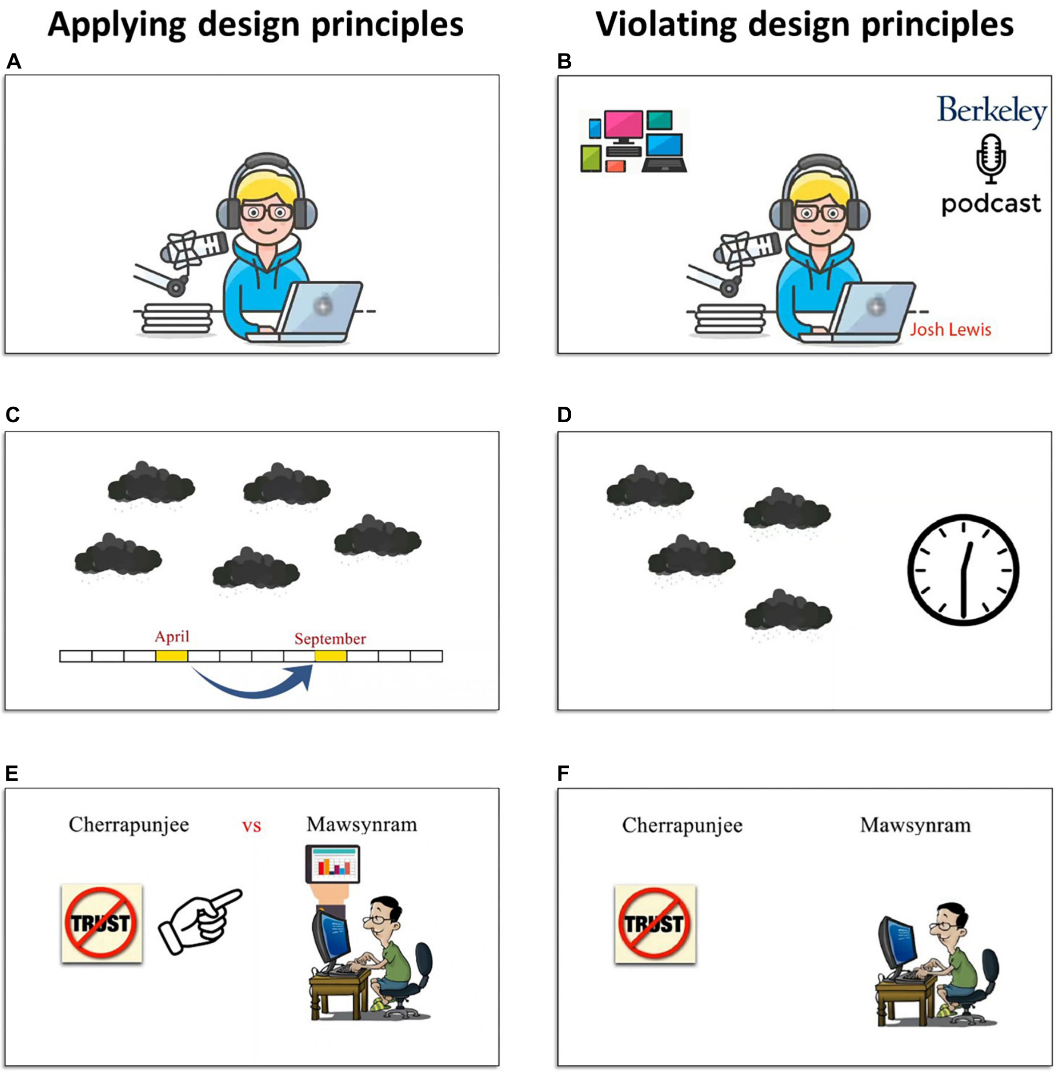


FIGURE 2
Educational multimedia. Examples of application (i.e., left panels: **A,C,E**)/violation (i.e., right panels: **B,D,F**) of multimedia design principles.

linguists in English language teaching devised the scenario for making the instructional videos and arranging the materials (texts and images). Then, all four videos have been created by a motion graphic specialist in Adobe After Effects CC 2017 v14.2.1.34 environment. We name the low-load versions of lesson 6 and lesson 11 as LV1 and LV2, respectively. Also, the high-load versions of lesson 6 and lesson 11 are named HV1 and HV2, respectively.

Recall test and subjective questionnaire

We designed a multiple-options-question (MCQ) as a computer-based recall test with twelve identical questions for LV1 and HV1 and twelve identical questions for LV2 and HV2. The recall test has been designed by two linguists in the field of English language teaching. In addition to the recall test, we use the classic paper-based version of NASA-TLX (Hart and Staveland, 1988) as a subjective measure to compare the overall cognitive load between two conditions (i.e., low-load and high-load) in our study. NASA-TLX is a self-report index of cognitive load in the range of 0–100. Although the NASA-TLX is often used to measure general workload, a study (Mutlu-Bayraktar et al., 2019) that systematically reviews the cognitive load research literature in multimedia learning environments introduces NASA-TLX as a subjective indicator and performance outcomes as an indirect objective indicator for assessing cognitive load (the detailed information about NASA-TLX subscale values is provided as **Supplementary material**).

Baseline

Before performing the main experiment, all subjects are requested to look at a black-filled circle ($r = 5$ mm) at the center of a gray screen for approximately 20 s. They are asked to keep relax and not think about anything special. We record EEG signals during this task and use the middle 10-s of the signals as our baseline in the analysis.

Experiment design

After setting up the EEG cap on the participant's head by a technician, the recording was started. The participant was alone in the semi-dark room, sitting 57 cm away from a 17-inch monitor with a refresh rate of 60 Hz. After a few seconds, when a timer in the center of the screen ends, the multimedia was played automatically. We asked participants to pay attention to the concepts presented in the video. There was no interaction between the person and the computer during the playing video. A few seconds after the multimedia is over; the recall test was started automatically. The participants could answer the questions in 420 *svia*a mouse interface. Participants had this option to leave any question unanswered. Moreover, there was the feasibility of moving between questions at any time, but only one question with all its options was displayed on the screen at a time. Also, the subject could terminate the recall test before the end of the timer. But by stopping the timer, the test phase was being finished automatically. The software platform for presenting the multimedia and recall test has been written in Java (for more details, see https://github.com/K-Hun/multimedia-learning-hci hosted on GitHub). After these steps, the EEG was stopped, and then the paper-based NASA-TLX was given to them. To make sure participants are familiar with the procedure and software environment of the experiment, we designed a trial phase before the experiment. In the trial phase, EEG signals are not recorded and also the multimedia is a 1-min video that is quite different in content and topic from the main multimedia of the experiment.

We assigned all thirty participants into two groups randomly, called LV1HV2 and LV2HV1 groups. Each subject participated in two distinct sessions of the experiment. The conditions in each group were counterbalanced across participants. Subjects in the LV1HV2 (LV2HV1) group performed the experiment in a session with LV1 (LV2) multimedia and in another session with HV2 (HV1) (some starting with the low load condition, and others with the high load one). Using this arrangement, each participant will not observe two multimedia with the same topic and audio and thus the concept of each multimedia is new to her/him. We summarized the experiment procedure in **Figure 1**.

Electroencephalography recording and preprocessing

To collect EEG data, we use a portable 32-channels eWave amplifier (Karimi-Rouzbahani et al., 2017a,b; Shooshtari et al., 2019) paired with eProbe v6.7.3.0 software. In this study, we recorded EEG data from 29 passive wet electrodes (FP1, FP2, FPz, F3, F4, F7, F8, Fz, FC1, FC2, FC5, FC6, C3, C4, Cz, T7, T8, CP1, CP2, CP5, CP6, P3, P4, P7, P8, Pz, O1, O2, and Oz) according to the 10–20 system of electrode placement, plus two bilateral mastoids (M1: left and M2: right) as the online reference for EEG signal potentials (see **Figure 3A**). The system has 24-bits data resolution with capturing 1K samples per second. Electrode impedances were kept below 5 KΩ in all recordings and electrode sites.

Analysis of EEG data and preprocessing are performed using the EEGLAB Toolbox v2020.0 and scripting in the MATLAB (R2019b) environment as shown in **Figure 3B**. As the first step, the basic FIR band-pass filter in the range of 1–30 Hz is applied to remove DC and high-frequency noise. Mastoid referencing makes EEG signals prone to external experimental artifacts. These artifacts come from the unstable connection of the EEG sensor to the mastoids, generating large spikes that are several orders of magnitude more prominent than the neural response produced by EEG. Therefore, in the next step, to reduce the effect of these artifacts, we apply the re-referencing part of the PREP pipeline algorithm (Bigdely-Shamlo et al., 2015) to estimate the true reference. Next, we utilize the Artifact Subspace Reconstruction (ASR) algorithm (Mullen et al., 2013) to correct corrupted parts of EEG data. ASR is being used to detect and remove high-amplitude components such as eye blinks, muscle movements, and sensor motion (Mullen et al., 2015). We perform ASR using *Clean_Rawdata* plug-in with default settings. A visual examination of the signals indicates that there are still some artifacts related to eye movements in the data. Thus, in the last step of preprocessing, independent component analysis (ICA) is applied using fastICA algorithm and the remaining artifacts (i.e., eye movements) are removed from the data using IC Label with threshold of 90% (Pion-Tonachini et al., 2019).

Segment length analysis

One challenge in the assessment of instantaneous cognitive load is selecting the most appropriate segment length. This issue has not been clearly answered in the previous related studies, so different time interval has been adopted as segment length. Here, we are faced with a content-oriented task (i.e., multimedia learning). To this end, we are seeking to achieve the smallest meaningful and informative interval in the multimedia learning task by analyzing the optimal time window selection. Hence, we consider the average time spent to convey a meaningful phrase to learners as a metric to determine the segment length. For this purpose, we use the silent moments in the audio narrations of the multimedia as an appropriate situation for learners to understand the contents presented before these moments. We use WavePad Sound Editor v12.4 to find silence points with minimum duration of 300 ms and below 25 dB level. The segments with audio narration that conveys some words without silent interruptions in two multimedia are shown in **Figure 4**. The figure illustrates the number of time frames with audio narration for each segment length. As shown in this figure, it is desirable to choose a segment length in the range of 2.5–4 s. Thus, in the following, we assess the instantaneous cognitive load for segments with a length of 3 s.

Feature extraction

We adopt a time-frequency-based analysis approach for feature extraction. For each participant's EEG data, the PSD in each channel is estimated by calculating the squared magnitude of the fast-Fourier transform (FFT) (Semmlow, 2011) from 50% overlapping windows, which is tapered by the Hanning window to reduce the spectral leakage. A window size contains 1,000 sample points (1 s) and an overlap of 500 sample points (500 ms) (see **Figure 3C**). Next, relative band power (*rBP*) of 3 s segments are extracted in each frequency band: delta ($\delta : 2\text{–}3$ Hz), theta ($\theta : 4\text{–}7$ Hz), alpha ($\alpha : 8\text{–}12$ Hz) and beta ($\beta : 13\text{–}30$ Hz). In order to extract these frequency bands, for each segment, we performed the decibel (dB) conversion (Cohen, 2014). The dB conversion is a baseline normalization method that quantifies the ratio of the median PSD in each band and the median PSD of the baseline on a logarithmic scale. In this way, we overcame the positively skewed distribution of EEG power data. By applying this method, power values are often normally distributed and thus parametric statistical analysis is an appropriate approach for feature extraction (Cohen, 2014).

To calculate the *rBP*, we use Eq. (1) where $rBP^{i}_{ch,b}$ is the median power of *i*-th segment *seg* in the channel *ch* ($ch \in \{1,\ 2,\ \ldots,\ 29\}$) and the band *b* ($b \in \{\delta,\ \theta,\ \alpha,\ \beta\}$) relative to the median power of the baseline *base* in same channel and band. Moreover, seg_i indicates EEG data of the *i*-th segment.

$$rBP^{i}_{ch,b} \ = \ 10log10\left(\frac{median\ PSD^{seg_i}_{ch,b}}{median\ PSD^{base}_{ch,b}}\right) \tag{1}$$

By concatenating the extracted features for the *i*-th segment, a feature vector (FV_i) is constructed for that segment. This feature vector consists of 116 elements (4 *rBP*s in 29 channels), as follows:

$$FV_i \ = \ [rBP^{i}_{1,\delta},\ rBP^{i}_{1,\theta},\ rBP^{i}_{1,\alpha},\ rBP^{i}_{1,\beta},\ \ldots,\ rBP^{i}_{29,\delta},\ rBP^{i}_{29,\theta},\ rBP^{i}_{29,\alpha},\ rBP^{i}_{29,\beta}]_{1\ \times\ 116} \tag{2}$$

Extracted features of each participant in all segments are illustrated in **Figure 3D**.

Feature selection

In the next step, we select the best discriminative feature set with the highest prediction accuracy. Also, it is essential to determine the regions of the brain and frequency bands that are highly informative for predicting cognitive load. To address this goal, we use the minimum redundancy-maximum relevance (MRMR) algorithm (Peng et al., 2005), which is a mutual information-based feature selection method. The algorithm follows an incremental search method iteratively. At each iteration, the candidate feature will be examined whether it has: (1) maximum relevance with respect to the class label, and (2) minimum redundancy with respect to the features selected at previous iterations. To evaluate the importance of features, a score is calculated for each feature according to these two criteria. Next, the MRMR algorithm will rank the features based on the scores in descending order. This process returns the ranking of 116 features which indicates the importance of each frequency band and channel. However, the limitation of this process is that the best feature set is not determined, and the optimal feature set must be selected by evaluating the ranked list with respect to the classification performance. To this end, we evaluated the ranked features by applying Linear Discriminant Analysis (LDA) (McLachlan, 2004) to samples in the following manner, to achieve the best set that improves the performance of classification. At first, the samples of all segments are split into 10 folds such that one fold is considered as the test set and the remaining folds are used to train the LDA model. Then, by increasing the number of features for every sample from 1 to 116 according to the ranking obtained by the MRMR algorithm, the LDA model is trained using selected features and prediction accuracy is computed on the test set. This process is repeated 10 times by considering each fold as a test set. Finally, by averaging over prediction accuracy of different folds, the final accuracy is computed for a subset of features (from 1 to 116) (see **Figure 3E**).

Classification of cognitive load

In this phase, in order to assess the instantaneous cognitive load, we follow an approach that classifies segments into two conditions (i.e., low-load and high-load). Our goal is to assign a score of cognitive load to each 3 s segment based on the distances between the samples and decision boundary (see **Figure 3F**).

To perform classification and assign scores to segments, we use the SVM algorithm. The algorithm has been widely used for non-linear binary classification problems in machine learning. It has achieved desirable results in cognitive and mental task applications (Amin et al., 2017). SVM transforms input data into higher dimensional space by applying the kernel trick, after which it finds the hyperplane with the best generalization capabilities by maximizing the margins (Wang et al., 2009). SVM with the kernel is extremely sensitive to hyperparameters, so it must be tuned to achieve a good level of performance. Hence, we apply the radial basis function (RBF) kernel, which only needs to optimize two hyperparameters (i.e., C as the penalty parameter and γ as the kernel width parameter) (Hsu et al., 2003). We examine various pairs of (C, γ) values using the Bayesian optimization algorithm, and the one set with the lowest cross-validation loss is selected. In the next step, in order to measure the performance of the optimized classifier and extract classification scores, we randomly select 70% of samples as a training set, and the rest of the samples are considered as a test set. Then, the classification scores are computed as mental workload scores. Indeed, these scores indicate the signed distance between a sample and the decision boundary. The score (s_i) for i-th segment is computed as follows:

$$s_i = \sum_{j=1}^{n} p_j y_j G\left(sv_j, seg_i\right) + q \tag{3}$$

where $G\left(sv_j, seg_i\right)$ is a non-linear transformation with radial basis function (RBF) which is defined in Eq. (4).

$$G\left(sv_j, seg_i\right) = \exp\left(-\left|\left|sv_j - seg_i\right|\right|^2\right) \tag{4}$$

where n is the number of support vectors, sv_j is j-th support vector, $y_j \in \{-1,1\}$ (i.e., low-load: −1 and high-load: +1) is the label of j-th support vector, p_j is the estimated SVM parameter for j-th support vector and q is the bias term. For more details on the estimation of $(p_1, \ldots, p_j, q)$ see Cristianini and Shawe-Taylor (2000).

Three values of the score (s) would be possible based on the position of each sample: (1) zero value ($s = 0$) when the sample is located on the decision boundary (hyperplane); (2) positive value ($s > 0$) when the sample has been correctly classified; (3) negative value ($s < 0$) otherwise. Once the scores are determined, we will normalize them to the range of 0–1 using the min-max normalization method as follows:

$$SC_i = \frac{s_i - \min(S)}{\max(S) - \min(S)} \tag{5}$$

where s_i and S are the scores of the i-th segment and the set of all segments' scores obtained after SVM classification, respectively.

Results

In this section, first, we validate the experimental conditions. Second, we examine the appropriate time interval for assessing the cognitive load imposed by the educational videos. Third, we evaluate the selected features and identify the most important frequency bands and brain regions for distinguishing two mental workload conditions. Finally, we present the results of the scoring model for instantaneous cognitive load assessment and investigation of its generalizability.

Validation of experimental conditions

To validate two experimental conditions (i.e., low-load and high-load), we performed statistical analysis on NASA-TLX scores and recall test. The assumption is that applying/violating multimedia design principles imposes different levels of cognitive load on learners. As a result, a two-sided independent samples t-test was used to investigate statistical differences for the two experimental conditions. The average and standard deviation of NASA-TLX scores and recall test scores in each group are presented in **Figure 5**. This analysis on NASA-TLX scores indicates a significant difference between cognitive load imposed by the different instructional design in multimedia, $t(18) = -4.87$, $p < 0.0002$ and $t(24) = -6.07$, $p < 0.0001$ for multimedia 1 (i.e., LV1 and HV1) and multimedia 2 (i.e., LV2 and HV2), respectively. Also, the same analysis on recall test shows that $t(18) = 6.41$, $p < 0.0001$ and $t(24) = 6.22$, $p < 0.0001$ for multimedia 1 and multimedia 2, respectively. Thus, two groups in both multimedia have significantly different performances. These results validate the assumption that the different mental demands are elicited due to the experimental conditions.

Evaluation of selected features and activated cerebral regions

The goal of feature selection is to extract the optimal feature set by reducing redundancy while keeping the information of gathered data. After performing the method described in section "Feature selection," we select the top 14 features of the MRMR algorithm as the best subset. This feature set can achieve the highest classification accuracy of 78.34 ± 1.3% using the LDA method for two load conditions. The best feature set is ordered in Eq. (6), where each element represents the selected channel with the band in the subscription.

$$\begin{aligned} Best\ Features = \{&O1_\alpha, C3_\alpha, P3_\theta, P7_\theta, CP1_\delta, P7_\beta, O2_\delta, FC5_\alpha, \\ &CP1_\beta, FPz_\alpha, FC6_\alpha, C4_\theta, F7_\alpha, F7_\delta\} \end{aligned} \tag{6}$$

For evaluating the selected features, we investigated the overall brain topographic difference between two experimental conditions in each frequency band. For this purpose, first, we

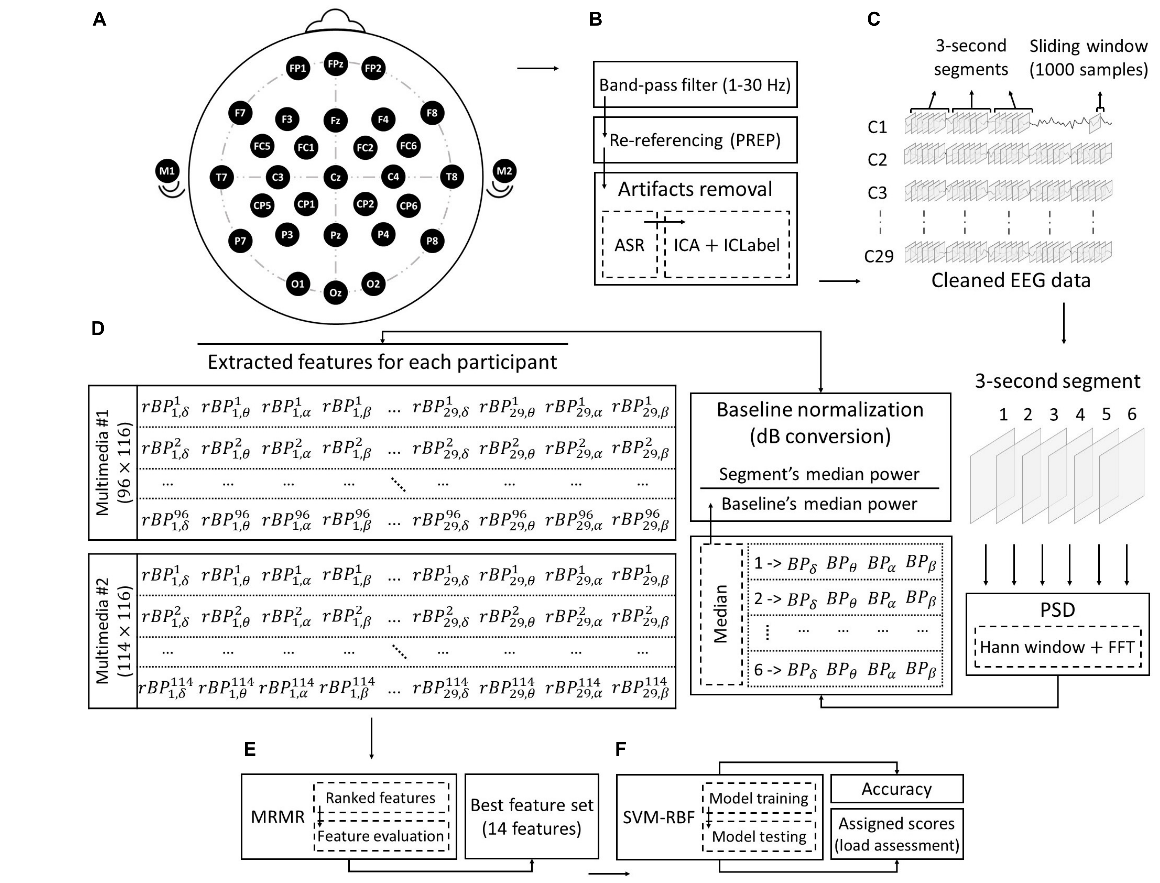

A
FP1
FPz
FP2
F7
F3
Fz
F4
F8
FC5
FC1
FC2
FC6
M1
T7
C3
Cz
C4
T8
M2
CP5
CP1
CP2
CP6
P7
P3
Pz
P4
P8
O1
Oz
O2
B
Band-pass filter (1-30 Hz)
Re-referencing (PREP)
Artifacts removal
ASR
ICA + ICLabel
C
3-second segments
Sliding window (1000 samples)
C1
C2
C3
C29
Cleaned EEG data
D
Extracted features for each participant
Multimedia #1 (96 × 116)
$rBP^1_{1,\delta}$ $rBP^1_{1,\theta}$ $rBP^1_{1,\alpha}$ $rBP^1_{1,\beta}$ … $rBP^1_{29,\delta}$ $rBP^1_{29,\theta}$ $rBP^1_{29,\alpha}$ $rBP^1_{29,\beta}$
$rBP^2_{1,\delta}$ $rBP^2_{1,\theta}$ $rBP^2_{1,\alpha}$ $rBP^2_{1,\beta}$ … $rBP^2_{29,\delta}$ $rBP^2_{29,\theta}$ $rBP^2_{29,\alpha}$ $rBP^2_{29,\beta}$
$rBP^{96}_{1,\delta}$ $rBP^{96}_{1,\theta}$ $rBP^{96}_{1,\alpha}$ $rBP^{96}_{1,\beta}$ … $rBP^{96}_{29,\delta}$ $rBP^{96}_{29,\theta}$ $rBP^{96}_{29,\alpha}$ $rBP^{96}_{29,\beta}$
Multimedia #2 (114 × 116)
$rBP^1_{1,\delta}$ $rBP^1_{1,\theta}$ $rBP^1_{1,\alpha}$ $rBP^1_{1,\beta}$ … $rBP^1_{29,\delta}$ $rBP^1_{29,\theta}$ $rBP^1_{29,\alpha}$ $rBP^1_{29,\beta}$
$rBP^2_{1,\delta}$ $rBP^2_{1,\theta}$ $rBP^2_{1,\alpha}$ $rBP^2_{1,\beta}$ … $rBP^2_{29,\delta}$ $rBP^2_{29,\theta}$ $rBP^2_{29,\alpha}$ $rBP^2_{29,\beta}$
$rBP^{114}_{1,\delta}$ $rBP^{114}_{1,\theta}$ $rBP^{114}_{1,\alpha}$ $rBP^{114}_{1,\beta}$ … $rBP^{114}_{29,\delta}$ $rBP^{114}_{29,\theta}$ $rBP^{114}_{29,\alpha}$ $rBP^{114}_{29,\beta}$
Baseline normalization (dB conversion)
Segment's median power
Baseline's median power
Median
1 -> BP_δ BP_θ BP_α BP_β
2 -> BP_δ BP_θ BP_α BP_β
6 -> BP_δ BP_θ BP_α BP_β
3-second segment
1 2 3 4 5 6
PSD
Hann window + FFT
E
MRMR
Ranked features
Feature evaluation
Best feature set (14 features)
F
SVM-RBF
Model training
Model testing
Accuracy
Assigned scores (load assessment)

FIGURE 3
Electroencephalography (EEG) analysis workflow. **(A)** EEG acquisition: data collected from 29 channels for each participant during displaying the multimedia. **(B)** Pre-processing: includes band-pass filter, re-referencing, and artifacts removal processes. **(C)** Data segmentation: a sliding window (size = 1,000 *ms*; 50% overlappings) moves on the signal of each channel. Each of the six adjacent windows forms a 3s segment. **(D)** Feature extraction: by performing Hann window and by using the FFT method, PSD of all frequency sub-bands is calculated for each window. Next, the ratio of the median power of each 3s segment to the median power of the baseline is considered as the relative band power (rBP) of that segment. For each rBP, the superscript shows the segment number and the subscripts show the channel number and the band, respectively. Finally, the extracted features of each participant for each of the multimedia will be formed in 96 × 116 and 114 × 116 dimensions for multimedia 1 (i.e., LV1 and HV1) and multimedia 2 (i.e., LV2 and HV2), respectively. **(E)** Feature selection: the best set of features will be selected by evaluating the importance of the features which is ranked by the MRMR algorithm. **(F)** Classification: an SVM (kernel: RBF) is built to assign a score to each segment (assessment of instantaneous cognitive load).

calculated the average *rBP* [see Eq. (1)] of all 3 s segments of each condition (i.e., low-load and high-load) in each band and then subtracted the average of low-load average from the average of high-load. Figure 6 illustrates the difference between the *rBP* averages of two conditions in each band. The powers in each band are scaled to the range of −1 to +1. According to this figure, active cortical areas are different in each band, and we can determine active cerebral regions for each band as below where the superscription (i.e., L: low-load and H: high-load) indicates the corresponding condition.

$$\delta_{ACTIVE} = \{F7^H, CP1^H, FC5^H, FC2^L, P3^L, FPz^L, CP2^L, Oz^L, CP6^L, Fz^L, O2^L\} \tag{7}$$

$$\theta_{ACTIVE} = \{P7^H, F3^H, FC5^H, T7^H, T8^L, Oz^L, Fz^L, FP2^L, O1^L, C4^L, P3^L\} \tag{8}$$

$$\alpha_{ACTIVE} = \{C3^H, FC5^H, F7^H, P4^H, FC1^H, P3^L, O2^L, Oz^L, FC6^L, FPz^L, O1^L\} \tag{9}$$

$$\beta_{ACTIVE} = \{P7^H, FC1^L, P4^L, T8^L, FPz^L, CP1^L, P8^L\} \tag{10}$$

The results show that the selected features are consistent with the active cerebral regions in different locations and bands. It is inferred from the comparison of the best feature set [as mentioned in Eq. (6)] and the active cerebral regions [as stated

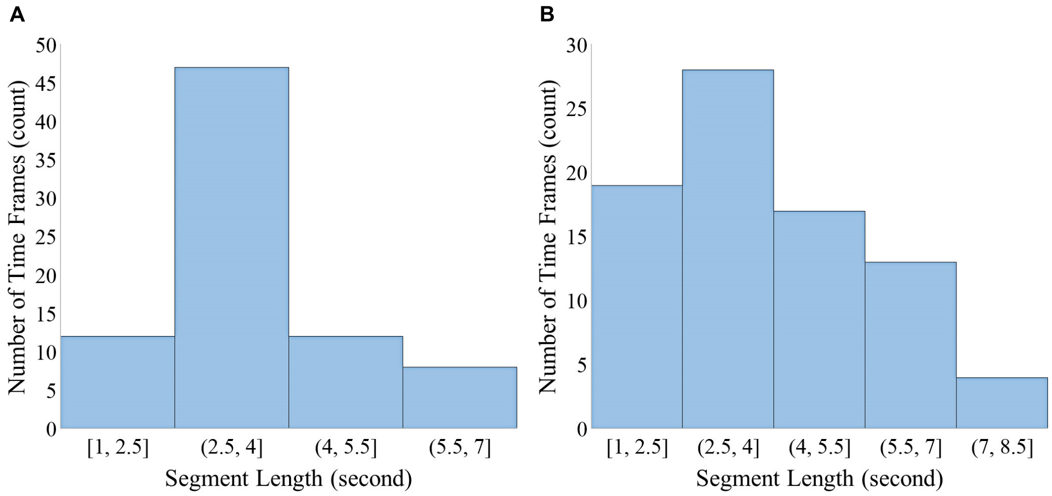


FIGURE 4
Segment length analysis in **(A)** multimedia 1 and **(B)** multimedia 2. Histograms show the frequency of number of meaningful timeframes regarding segment length for multimedia 1 **(A)** and multimedia 2 **(B)**.

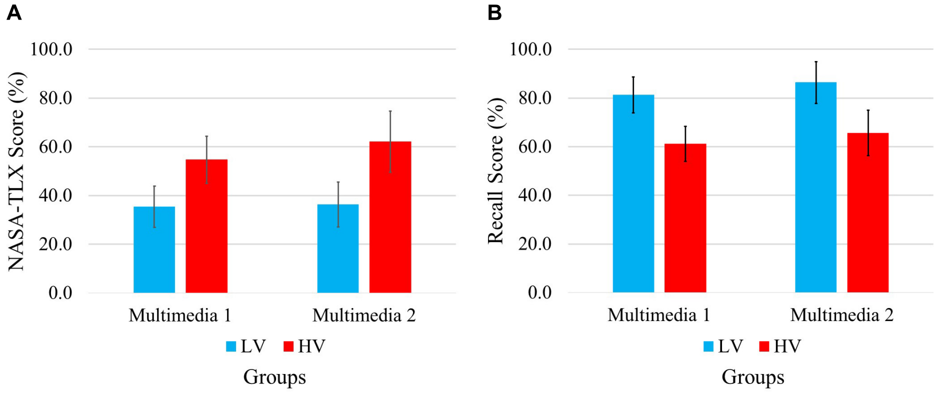


FIGURE 5
Comparison of **(A)** NASA-TLX scores and **(B)** recall test scores in two experimental conditions. In each graph the scores [NASA-TLX scores in panel **(A)** and Recall scores in panel **(B)**] are compared between two conditions (LV and HV) for each multimedia (Multimedia 1 and Multimedia 2). The scores have been scaled in the range [0, 100].

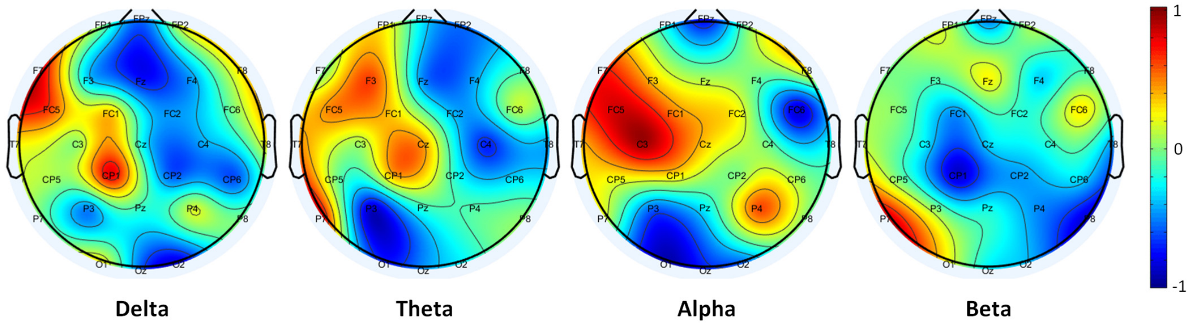


FIGURE 6
Differences between average relative band powers of electroencephalography (EEG) features (bands and locations) in two conditions.

TABLE 1 Classification accuracy and standard deviation (std) of electroencephalography (EEG) band powers using linear discriminant analysis (LDA).

	Frequency bands			
	Delta (δ)	**Theta (θ)**	**Alpha (α)**	**Beta (β)**
Accuracy	72.97	68.33	73.85	68.22
Std (±)	2.29	2.23	2.73	2.23

Results are presented in percentage.

above in Eqs (7–10)]. So that, all the selected features were selected from the active cortical areas. This indicates that the feature selection method effectively selects a combination of informative and relevant features to cognitive load with respect to the brain activity map.

In order to identify which frequency band can distinguish two cognitive load conditions more effectively, we perform the classification task in each frequency band separately by selecting the feature subset associated with that band. Again, we apply 10-fold cross-validation using the LDA method on data. As presented in Table 1, the alpha is the best frequency band for predicting mental workload. The predictive power of the alpha feature set is 73.85 ± 2.73%. Figure 7 illustrates brain topographies of relative alpha power distribution in two conditions compared to the baseline. According to this figure, the diagonal activity of alpha power in each condition attracts attention. In low-load condition, most alpha activation is concentrated in the left lateral posterior to the right lateral anterior cortices. Conversely, in high-load condition, this pattern is localized in the right lateral posterior to the left lateral anterior cortical areas. By comparing Eqs (6) and (9), it is found that alpha power suppression in prefrontal midline (*FPz*), right lateral frontal (*FC6*), and left lateral occipital (*O1*) cortices have a more significant impact on increasing cognitive load. Also, activation of alpha power in the left lateral frontal (*FC5*, *F7*) and left central (*C3*) cortical areas synchronize by increasing cognitive load.

Instantaneous cognitive load scoring model results

After evaluating the best feature set, we evaluate the performance of the classification method presented in section "Classification of cognitive load" for assigning scores to segments. Thus, we compute the average and standard deviation of classifier accuracy to assess the SVM model performance. The performance of the model is achieved 84.5 ± 2.1%. As described previously, assigned scores are converted into normalizing scores (*SCs*) using Eq. (5).

Then, the cognitive load imposed at each moment of each multimedia is calculated by averaging over the normalized scores obtained by the SVM in the corresponding segments at that moment. Figure 8 displays predicted workload scores in two multimedia over time. As depicted in this figure, the average of predicted scores corresponding to two load conditions is significantly different across multimedia timeline. These scores for LV1, HV1, LV2, and HV2 are 29, 43, 46, and 60, respectively.

Discussion

In this study, based on the most informative feature set, we construct an SVM model for assessing instantaneous cognitive load. To impose low or high levels of cognitive load on the participants, we designed an experiment with two versions of multimedia by applying or violating the principles of multimedia design. The conditions of our

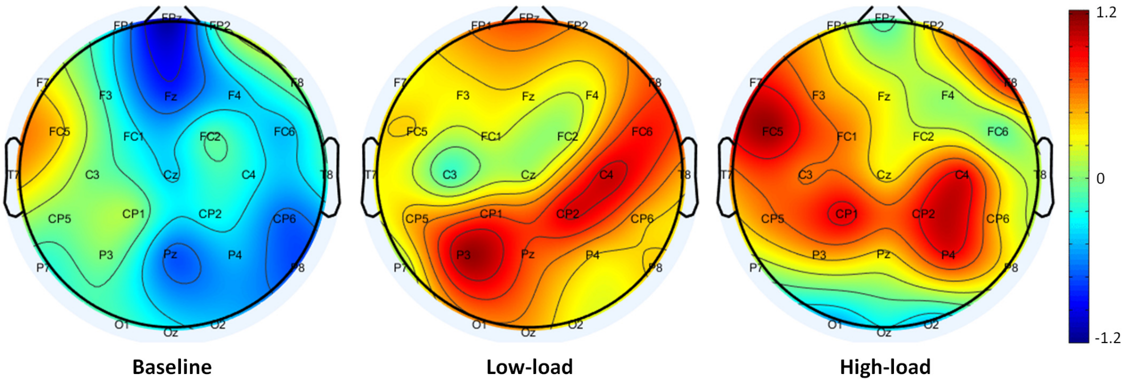


FIGURE 7
Brain topographies of alpha power distribution in two conditions compared to baseline. From left to right, the topographies represent the average relative alpha power for the eye-opened baseline, low-load, and high-load conditions.

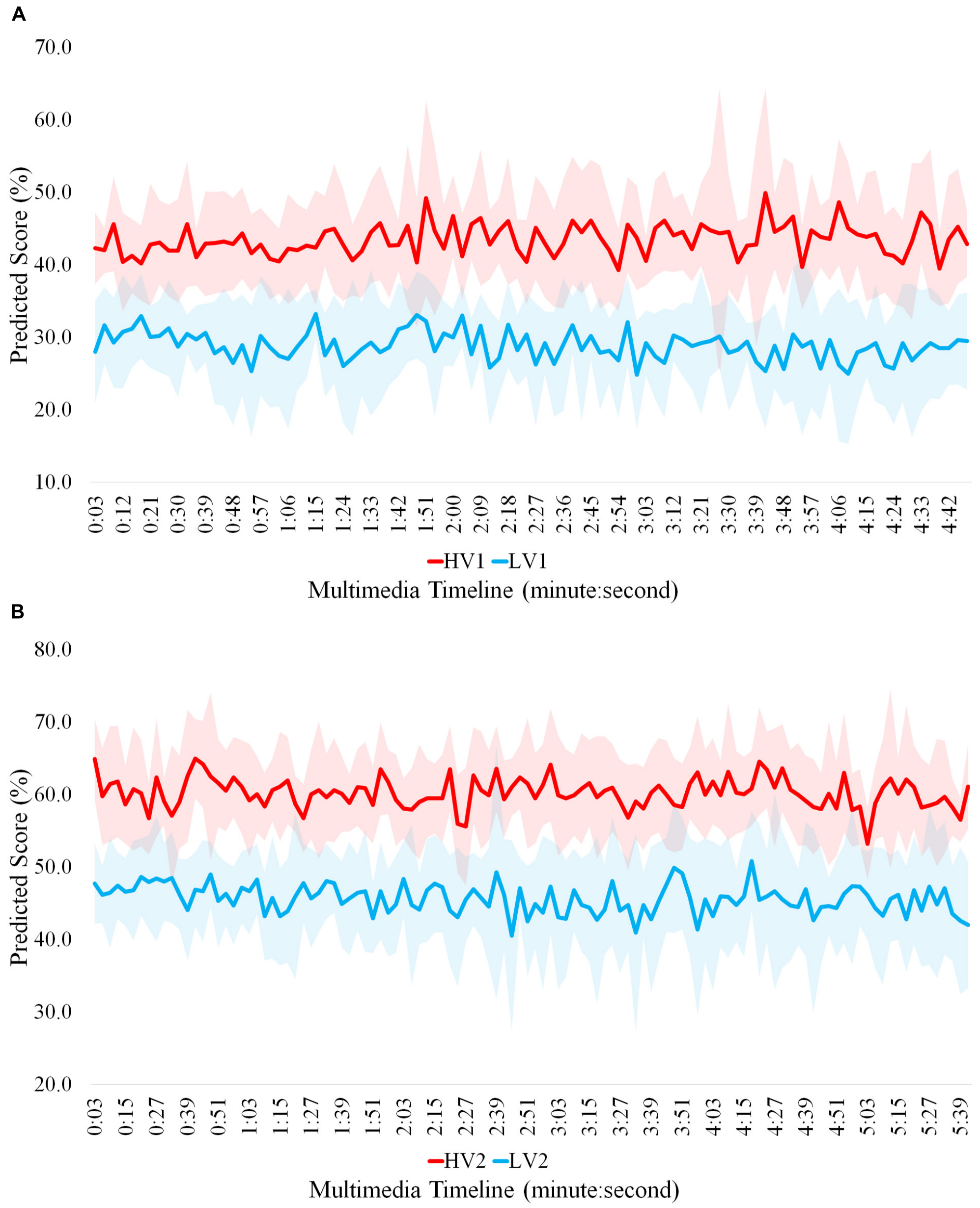


FIGURE 8
Predicted cognitive load scores in **(A)** multimedia 1 and **(B)** multimedia 2 over time.

experiment are evaluated by a recall test and a NASA-TLX as a subjective measurement of cognitive load. As a result, applying the principles leads to lower NASA-TLX scores and improvement of performance tests, indicating that this experimental condition will induce a lower cognitive load in comparison to the condition of violating design principles.

In order to extract the informative and relevant EEG features as an objective measurement of cognitive load, first, we calculated the PSD of the common frequency bands. Then, we extracted the optimal feature set by using the MRMR algorithm, which is a ranking method based on mutual information. The main advantage of this feature selection method is the effective reduction of redundant features while preserving relevant features. In addition, compared to other dimensionality reduction techniques such as PCA, the readability and interpretability of the features are held, and no changes are made to the data.

The selected feature set includes less than 12% of the total features. These 14-top features confirmed the different conditions of the cognitive load imposed on the subjects very good. The selected features show a remarkable combination of activated frequency bands in different brain regions associated with executive functions of brain which are referred to as supervisory cognitive processes (e.g., attention, cognitive inhibition, or learning) because they involve higher level organization and execution of complex thoughts and behavior (Alvarez and Emory, 2006; Whelan, 2007). Especially, in multimedia learning, verbal information (e.g., words or sentences) and visual part (e.g., illustrations, photographs, or diagrams) are merged (Gyselinck et al., 2008). These audio/visual signals may arise conflicting effects and overloads on the overall brain, and thus it is expected to have a simultaneous activation of different areas of the cerebral cortex. Given the specified locations/frequencies, it can be possible to find cognitive load differences at these locations/frequencies using simple statistical analysis.

Most of the selected features are from the frontal region (FPz_α, $FC5_\alpha$, $FC6_\alpha$, $F7_\alpha$, and $F7_\delta$). Except for one of them ($F7_\delta$), the other mentioned features belong to the alpha band. In addition, two features have been selected from the centro-parietal ($C3_\alpha$) and the occipital ($O1_\alpha$) regions. This result is in line with previous studies that link cognitive processes to the frontal and parieto-occipital regions (e.g., Puma et al., 2018 for review), and alpha band activity (e.g., Foxe and Snyder, 2011 for review). According to the literature, activation of alpha indicate two opposite behaviors related to cognitive processing: active processing associated with memory maintenance and inhibition of irrelevant information (Jensen and Mazaheri, 2010). In fact, the increase in cognitive workload may be due to either of these two reasons or both of them. In this study, we observed that the power of alpha band in the low-load condition (i.e., applying design principles) is higher than the high-load condition (i.e., violating design principles), prominently in the prefrontal and the occipital regions. The increases of alpha spectral power seems to reflect the top-down control of the parieto-frontal attention network. As reported in recent studies, this mechanism inhibits irrelevant information flow from the visual perception system and internal cognitive processing (Pi et al., 2021). In this way, the information is transferred from task-irrelevant regions to task-relevant ones (Jensen and Mazaheri, 2010). Therefore, the decrease in alpha power near Broca's area, which plays a significant role in language comprehension (Novick et al., 2005), suggests the effective engagement of cognitive resources related to the task.

After feature analysis, we propose a scoring model to measure instantaneous cognitive load in 3s segments of multimedia. The model can predict the mental workload scores in multimedia across time at appropriate accuracy. In other words, applying (violating) principles at each moment has caused that the predicted cognitive load score for LV1 (HV1) and LV2 (HV2) is lower (higher) than HV1 (LV1) and HV2 (LV2) at that moment. This allows us to monitor and manage learners' cognitive status while watching multimedia at each moment. In this way, we can evaluate the quality of presented instructional materials and design principles in multimedia across time. Also, it can be possible to measure the effect size and impact of applying each principle. Therefore, by detecting the segments of multimedia that impose a great cognitive load on learners, we can provide the optimal load and improve learning performance by applying appropriate instructional materials and effective design principles. Moreover, a comparison of several multimedia that convey the same content can be feasible. This ability facilitates the production or selection of appropriate educational multimedia based on cognitive neurophysiological indicators.

Several limitations in the current research should be noticed. The first limitation of this study is the use of gel-based EEG equipment to collect data. The sensitivity of this device to get good contact of electrodes to scalp sites makes data prone to noise, resulting into extra time for preprocessing and increase in data loss rate. Moreover, for future studies, it might be useful to evaluate some cognitive-related abilities of subjects such as short-term memory capacity, visual attention, auditory and visual processing, etc. These abilities can be evaluated by common psychometric tests. In addition, it is a good idea to consider the cognitive and learning styles of participants in future studies. Another limitation to be mentioned here is the restriction of the analytical method. We assessed cognitive load by analyzing features extracted from the electrodes individually. Therefore, the interconnected functionality of the brain during a cognitive task is not considered. It is essential to consider the brain connectivity analysis approach in future researches to investigate information flows that are important in cognitive processes.

Conclusion

In this study, we investigated the possibility of instantaneous assessment of cognitive load in educational multimedia using EEG data as an objective measure. Our experimental conditions, which impose two distinct levels of cognitive load by applying/violating multimedia design principles to learners, were validated by using the result of the NASA-TLX and recall test. We extracted the relative band powers for common frequency bands in each cerebral area. The most informative and relevant feature set for measuring cognitive load was selected using the MRMR method. We constructed an SVM classification model to predict cognitive load scores at 3s moments. The proposed model was validated for generalization from one multimedia to another. This capability can significantly help educational multimedia designers to construct multimedia by imposing an optimal amount of cognitive load on learners. In short, our main contributions in this study can be considered as (1) investigation of active cortical areas and major frequency bands associated with cognitive load in learning task, (2) instantaneous assessment of cognitive load in educational multimedia using objective indicators, and (3) generalizability of the workload scoring model from one multimedia to another.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by Ethics Committee of the Iran University of Medical Sciences. The patients/participants provided their written informed consent to participate in this study.

Author contributions

RS and KL performed the experiment. KL implemented the HCI software platform. RS analyzed the data and wrote the manuscript under the supervision of SA and in consultation with AB. SA and AB determined the methodology, including signal processing methods and machine learning approaches. RE conceptualized and provided domain knowledge in this study and conducted the research direction. All authors contributed to the manuscript and approved the submitted version.

Acknowledgments

We thank our colleagues from Shahid Rajaee Teacher Training University, Akbar Keshavarz-nasab and Araz Farkish for assistance with EEG data collection and pre-processing, and Amirhossein Asadi for comments that greatly improved the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2022.744737/full#supplementary-material

References

Alvarez, J. A., and Emory, E. (2006). Executive function and the frontal lobes: a meta-analytic review. *Neuropsychol. Rev.* 16, 17–42. doi: 10.1007/s11065-006-9002-x

Amin, H. U., Mumtaz, W., Subhani, A. R., Saad, M. N. M., and Malik, A. S. (2017). Classification of EEG signals based on pattern recognition approach. *Front. Comput. Neurosci.* 11:103. doi: 10.3389/fncom.2017.00103

Anmarkrud, Ø, Andresen, A., and Bråten, I. (2019). Cognitive load and working memory in multimedia learning: conceptual and measurement issues. *Educ. Psychol.* 54, 61–83. doi: 10.1080/00461520.2018.1554484

Antonenko, P., Paas, F., Grabner, R., and Van Gog, T. (2010). Using electroencephalography to measure cognitive load. *Educ. Psychol. Rev.* 22, 425–438. doi: 10.1007/s10648-010-9130-y

Appriou, A., Cichocki, A., and Lotte, F. (2020). Modern machine-learning algorithms: for classifying cognitive and affective states from electroencephalography signals. *IEEE Syst. Man Cybern. Mag.* 6, 29–38. doi: 10.1109/MSMC.2020.2968638

Baceviciute, S., Mottelson, A., Terkildsen, T., and Makransky, G. (2020). "Investigating representation of text and audio in educational VR using learning outcomes and EEG," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (New York, NY: Association for Computing Machinery), 1–13. doi: 10.1145/3313831.3376872

Barrios, V. M. G., Gütl, C., Preis, A. M., Andrews, K., Pivec, M., Mödritscher, F., et al. (2004). "Adele: a framework for adaptive e-learning through eye tracking," in *Proceedings of the IKNOW*, Graz, 609–616.

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K.-M., and Robbins, K. A. (2015). The PREP pipeline: standardized preprocessing for large-scale EEG analysis. *Front. Neuroinformat.* 9:16. doi: 10.3389/fninf.2015.00016

Borys, M., Plechawska-Wójcik, M., Wawrzyk, M., and Wesołowska, K. (2017). "Classifying cognitive workload using eye activity and EEG features in arithmetic tasks," in *Proceedings of the International conference on information and software technologies*, (Berlin: Springer), 90–105. doi: 10.1007/978-3-319-67642-5_8

Castro-Meneses, L. J., Kruger, J.-L., and Doherty, S. (2020). Validating theta power as an objective measure of cognitive load in educational video. *Educ. Technol. Res. Dev.* 68, 181–202. doi: 10.1007/s11423-019-09681-4

Chen, S., Epps, J., Ruiz, N., and Chen, F. (2011). "Eye activity as a measure of human mental effort in HCI," in *Proceedings of the 16th International Conference on Intelligent user Interfaces*, Palo Alto, CA, 315–318. doi: 10.1145/1943403.1943454

Cohen, M. X. (2014). *Analyzing Neural Time Series Data: Theory and Practice*. Cambridge, MA: MIT press.

Cristianini, N., and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and other Kernel-Based Learning Methods*. Cambridge: Cambridge university press.

Dalmaso, M., Castelli, L., Scatturin, P., and Galfano, G. (2017). Working memory load modulates microsaccadic rate. *J. Vis.* 17:6. doi: 10.1167/17.3.6

Dan, A., and Reiner, M. (2017). EEG-based cognitive load of processing events in 3D virtual worlds is lower than processing events in 2D displays. *Int. J. Psychophysiol.* 122, 75–84. doi: 10.1016/j.ijpsycho.2016.08.013

Duncan, J., and Parker, A. (2007). *Open Forum 3: Academic Listening and Speaking*. Oxford: Oxford University Press.

Foxe, J. J., and Snyder, A. C. (2011). The role of alpha-band brain oscillations as a sensory suppression mechanism during selective attention. *Front. Psychol.* 2:154. doi: 10.3389/fpsyg.2011.00154

Gyselinck, V., Jamet, E., and Dubois, V. (2008). The role of working memory components in multimedia comprehension. *Appl. Cognit. Psychol.* 22, 353–374. doi: 10.1002/acp.1411

Hart, S. G., and Staveland, L. E. (1988). "Development of NASA-TLX (Task Load Index): results of empirical and theoretical research," in *Advances in Psychology*, eds P. A. Hancock and N. Meshkati (Amsterdam: Elsevier), 139–183. doi: 10.1016/S0166-4115(08)62386-9

Hsu, C.-W., Chang, C.-C., and Lin, C.-J. (2003). *A Practical Guide to Support Vector Classification*. Taipei: National Taiwan University

Jensen, O., and Mazaheri, A. (2010). Shaping functional architecture by oscillatory alpha activity: gating by inhibition. *Front. Hum. Neurosci.* 4:186. doi: 10.3389/fnhum.2010.00186

Karimi-Rouzbahani, H., Bagheri, N., and Ebrahimpour, R. (2017b). Hard-wired feed-forward visual mechanisms of the brain compensate for affine variations in object recognition. *Neuroscience* 349, 48–63. doi: 10.1016/j.neuroscience.2017.02.050

Karimi-Rouzbahani, H., Bagheri, N., and Ebrahimpour, R. (2017a). Average activity, but not variability, is the dominant factor in the representation of object categories in the brain. *Neuroscience* 346, 14–28. doi: 10.1016/j.neuroscience.2017.01.002

Kruger, J.-L., & Doherty, S. (2016). Measuring cognitive load in the presence of educational video: Towards a multimodal methodology. *Australasian Journal of Educational Technology*, 32, 19–31. doi: 10.14742/ajet.3084

Latifzadeh, K., Amiri, S. H., Bosaghzadeh, A., Rahimi, M., and Ebrahimpour, R. (2020). Evaluating cognitive load of multimedia learning by eye-tracking data analysis. *Technol. Educ. J.* 15, 33–50.

Lin, F.-R., and Kao, C.-M. (2018). Mental effort detection using EEG data in E-learning contexts. *Comput. Educ.* 122, 63–79.

Mayer, R., and Mayer, R. E. (2005). *The Cambridge Handbook of Multimedia Learning*. Cambridge: Cambridge university press.

Mayer, R. E. (2002). "Multimedia learning," in *Psychology of Learning and Motivation*, ed. R. E. Mayer (Amsterdam: Elsevier), 85–139. doi: 10.1016/S0079-7421(02)80005-6

Mazher, M., Abd Aziz, A., Malik, A. S., and Amin, H. U. (2017). An EEG-based cognitive load assessment in multimedia learning using feature extraction and partial directed coherence. *IEEE Access* 5, 14819–14829. doi: 10.1109/ACCESS.2017.2731784

McLachlan, G.J., *Discriminant analysis and statistical pattern recognition*. Vol. 544. 2004: John Wiley & Sons.

Mullen, T., Kothe, C., Chi, Y. M., Ojeda, A., Kerth, T., Makeig, S., et al. (2013). "Real-time modeling and 3D visualization of source dynamics and connectivity using wearable EEG," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Osaka: IEEE), 2184–2187. doi: 10.1109/EMBC.2013.6609968

Mullen, T. R., Kothe, C. A., Chi, Y. M., Ojeda, A., Kerth, T., Makeig, S., et al. (2015). Real-time neuroimaging and cognitive monitoring using wearable dry EEG. *IEEE Transac. Biomed. Eng.* 62, 2553–2567. doi: 10.1109/TBME.2015.2481482

Mutlu-Bayraktar, D., Cosgun, V., and Altan, T. (2019). Cognitive load in multimedia learning environments: a systematic review. *Comput. Educ.* 141:103618. doi: 10.1016/j.compedu.2019.103618

Novick, J. M., Trueswell, J. C., and Thompson-Schill, S. L. (2005). Cognitive control and parsing: reexamining the role of Broca's area in sentence comprehension. *Cognit. Affect. Behav. Neurosci.* 5, 263–281. doi: 10.3758/cabn.5.3.263

Paas, F., Renkl, A., and Sweller, J. (2004). Cognitive load theory: instructional implications of the interaction between information structures and cognitive architecture. *Instruct. Sci.* 32, 1–8. doi: 10.1023/B:TRUC.0000021806.17516.d0

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 1226–1238. doi: 10.1109/TPAMI.2005.159

Pi, Z., Zhang, Y., Zhou, W., Xu, K., Chen, Y., Yang, J., et al. (2021). Learning by explaining to oneself and a peer enhances learners' theta and alpha oscillations while watching video lectures. *Br. J. Educ. Technol.* 52, 659–679. doi: 10.1111/bjet.13048

Pion-Tonachini, L., Kreutz-Delgado, K., and Makeig, S. (2019). ICLabel: an automated electroencephalographic independent component classifier, dataset, and website. *Neuroimage* 198, 181–197. doi: 10.1016/j.neuroimage.2019.05.026

Plechawska-Wójcik, M., Tokovarov, M., Kaczorowska, M., and Zapała, D. (2019). A three-class classification of cognitive workload based on EEG spectral data. *Appl. Sci.* 9:5340. doi: 10.1109/TNSRE.2019.2913400

Pomplun, M., and Sunkara, S. (2003). "Pupil dilation as an indicator of cognitive workload in human-computer interaction," in *Proceedings of the International Conference on HCI*, Toronto.

Puma, S., Matton, N., Paubel, P.-V., Raufaste, É., and El-Yagoubi, R. (2018). Using theta and alpha band power to assess cognitive workload in multitasking environments. *Int. J. Psychophysiol.* 123, 111–120. doi: 10.1016/j.ijpsycho.2017.10.004

Rojas, R. F., Debie, E., Fidock, J., Barlow, M., Kasmarik, K., Anavatti, S., et al. (2020). Electroencephalographic workload indicators during teleoperation of an unmanned aerial vehicle shepherding a swarm of unmanned ground vehicles in contested environments. *Front. Neurosci.* 14:40. doi: 10.3389/fnins.2020.00040

Scharinger, C., Schüler, A., and Gerjets, P. (2020). Using eye-tracking and EEG to study the mental processing demands during learning of text-picture combinations. *Int. J. Psychophysiol.* 158, 201–214. doi: 10.1016/j.ijpsycho.2020.09.014

Semmlow, J. (2011). *Signals and Systems for Bioengineers: a MATLAB-based Introduction*. Cambridge, MA: Academic Press.

Shooshtari, S. V., Sadrabadi, J. E., Azizi, Z., and Ebrahimpour, R. (2019). Confidence representation of perceptual decision by EEG and eye data in a random dot motion task. *Neuroscience* 406, 510–527. doi: 10.1016/j.neuroscience.2019.03.031

Sweller, J. (2018). Measuring cognitive load. *Perspect. Med. Educ.* 7, 1–2. doi: 10.1007/s40037-017-0395-4

Sweller, J., Van Merriënboer, J. J., and Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educ. Psychol. Rev.* 31, 261–292. doi: 10.1007/s10648-019-09465-5

Tomasi, D., Ernst, T., Caparelli, E. C., and Chang, L. (2006). Common deactivation patterns during working memory and visual attention tasks: an intra-subject fMRI study at 4 Tesla. *Hum. Brain Mapp.* 27, 694–705. doi: 10.1002/hbm.20211

Tremmel, C., Herff, C., Sato, T., Rechowicz, K., Yamani, Y., and Krusienski, D. J. (2019). Estimating cognitive workload in an interactive virtual reality environment using EEG. *Front. Hum. Neurosci.* 13:401. doi: 10.3389/fnhum.2019.00401

Wang, B., Wong, C. M., Wan, F., Mak, P. U., Mak, P. I., and Vai, M. I. (2009). "Comparison of different classification methods for EEG-based brain computer interfaces: a case study," in *Proceedings of the 2009 International Conference on Information and Automation* (Zhuhai: IEEE), 1416–1421. doi: 10.1109/ICINFA.2009.5205138

Wang, H., Li, Y., Hu, X., Yang, Y., Meng, Z., and Chang, K.-M. (2013). "Using EEG to improve massive open online courses feedback interaction," in *Poster at the AIED Workshops* (Pittsburgh, PA: Carnegie Mellon University)

Whelan, R. R. (2007). Neuroimaging of cognitive load in instructional multimedia. *Educ. Res. Rev.* 2, 1–12. doi: 10.1016/j.edurev.2006.11.001

Xu, J., and Zhong, B. (2018). Review on portable EEG technology in educational research. *Comput. Hum. Behav.* 81, 340–349. doi: 10.1016/j.chb.2017.12.037

Zagermann, J., Pfeil, U., and Reiterer, H. (2016). "Measuring cognitive load using eye tracking technology in visual computing," in *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, Baltimore, MD, 78–85. doi: 10.1145/2993901.2993908

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership