DeepRepViz: Identifying potential confounders in deep learning model predictions

Roshan Prakash Rane[1,2][0000−0002−3996−2034], JiHoon Kim[3], Arjun Umesha[1], Didem Stark[1], Marc-André Schulz[1], and Kerstin Ritter[1][0000−0001−7115−0020]

[1] Charité - Universitätsmedizin Berlin, Department of Psychiatry and Neurosciences, Einstein Center for Neurosciences Berlin, Germany
[2] Humboldt-Universität zu Berlin, Department of Psychology, Germany
[3] Freie Universität Berlin, Department of Education and Psychology, Germany

Abstract. Deep Learning (DL) has emerged as a powerful tool in neuroimaging research. DL models predicting brain pathologies, psychological behaviors, and cognitive traits from neuroimaging data have the potential to discover the neurobiological basis of these phenotypes. However, these models can be biased from information related to age, sex, or spurious imaging artifacts encoded in the neuroimaging data. In this study, we introduce a lightweight and easy-to-use framework called 'DeepRepViz' designed to detect such potential confounders in DL model predictions and enhance the transparency of predictive DL models. DeepRepViz comprises two components - an online visualization tool (available at `https://deep-rep-viz.vercel.app/`) and a metric called the 'Con-score'. The tool enables researchers to visualize the final latent representation of their DL model and qualitatively inspect it for biases. The Con-score, or the 'concept encoding' score, quantifies the extent to which potential confounders like sex or age are encoded in the final latent representation and influences the model predictions. We illustrate the rationale of the Con-score formulation using a simulation experiment. Next, we demonstrate the utility of the DeepRepViz framework by applying it to three typical neuroimaging-based prediction tasks (n=12000). These include (a) distinguishing chronic alcohol users from controls, (b) classifying sex, and (c) predicting the speed of completing a cognitive task known as 'trail making'. In the DL model predicting chronic alcohol users, DeepRepViz uncovers a strong influence of sex on the predictions (Con-score=0.35). In the model predicting cognitive task performance, DeepRepViz reveals that age plays a major role (Con-score=0.3). Thus, the DeepRepViz framework enables neuroimaging researchers to systematically examine their model and identify potential biases, thereby improving the transparency of predictive DL models in neuroimaging studies.

Keywords: Deep Learning · Bias · Confound detection · Shortcut learning · Concept-based XAI · Neuroimaging .

Introduction

Deep Learning (DL) offers a promising avenue for neuroimaging research [28, 5] as it can be trained directly on high-dimensional ($p >> n$) neuroimaging modalities, such as structural Magnetic Resonance Imaging (MRI), functional Magnetic Resonance Imaging (fMRI), and Electroencephalogram (EEG). Particularly in population neuroscience research [17], DL models have been employed to understand the relationship between the brain and various psychological phenotypes [25], as well as brain pathologies [25, 5, 28]. For example, DL models are trained on brain MRI to predict the risk of developing Alzheimer's disease [12] or to identify subtypes in psychiatric disorders such as major depressive disorder or substance use disorders [4, 19]. However, instead of learning the brain-phenotype relationship relevant to the research, DL models can learn to use demographic factors or spurious biases encoded in the neuroimaging data [3]. These extraneous factors are referred to as 'confounders'. For example, consider the study of Thibeau-Sutre et al. [26], where they predict the risk of developing Alzheimer's disease from structural MRI data. In their study, most participants in the Alzheimer's group were scanned using a 3 Tesla scanner, while most controls were scanned using a 1.5 Tesla scanner [26]. Their DL model picked up on this spurious association and predicted all participants scanned with a 3 Tesla scanner as a high-risk group. Such a 'confounded' model that does not rely on any potential biomarkers of Alzheimer's disease to make its predictions would not be clinically useful. Several neuroimaging studies routinely report similar confounders influencing DL model predictions [10, 1, 22, 26]. Therefore, there is a pressing need in this field for a framework to detect potential confounders early in its life cycle of predictive DL models transparent.

In the field of neuroimaging, it is typical for researchers to select a set of 2 to 5 variables as potential confounders [10]. These commonly include age, sex, intracranial brain volume, educational level, and socio-economic status of the participants [10]. The variables are then controlled using confound control methods such as regressing out the confounder [23, 11], or counterbalancing the data [21, 20] to prevent them from influencing the model predictions. However, this practice may pose certain risks. Additional confounders may arise from the specific research question [18] or biases induced by the data sampling and image acquisition procedures [8]. Additionally, incorrectly controlling for certain variables that are not confounders can induce an artificial bias in the controlled data [18]. A more prudent practice would be to evaluate the variables associated with the study for their risk of confounding a model [8, 21]. This can be achieved by systematically predicting all the variables using the same modeling pipeline used for the phenotype prediction, as suggested by Görgen et al. [8]. Using our 'DeepRepViz' framework, researchers can perform such a systematic assessment for their DL models. The DeepRepViz framework enables this with two components:

1. A web-based visualization tool (`https://deep-rep-viz.vercel.app`) to visually inspect the latent representation learned by the DL model in its final

layer. The tool aids in qualitatively inspecting the model predictions against all the variables associated with the study.
2. A metric, called concept encoding score or 'Con-score', to quantify how strongly a variable is encoded in the latent representation of the DL model and how strongly this encoding influences the model prediction.

In the next section, we will outline the theoretical basis of DeepRepViz and the Con-score metric. In the Results section, we will demonstrate the efficacy of the DeepRepViz framework by applying it to three neuroimaging-based brain-phenotype prediction tasks.

Method

Consider a DL model used to predict a phenotype y from neuroimaging data X, i.e. $f_{model} : X \Rightarrow y$. Let's say that the DL model relies on a tertiary variable c^k (such as age or the scanner type), to make its predictions. The causal inference literature [18] states that c^k can influence the predictions $\hat{y}$ either as a confounder ($X \leftarrow c^k \rightarrow y$) or as a mediator ($X \rightarrow c^k \rightarrow y$)[18]. In this study, we focus on developing a framework that enables researchers to detect c^k when it influences the model predictions, irrespective of the direction of causality. Therefore, we propose (a) a visualization tool that assists in the easy detection of c^k and (b) a metric, Con-score, that quantifies how strongly c^k influences the predictions.

Related literature: Several metrics have been proposed in the past for detecting biased predictions in machine learning models, especially focused on preventing the unequal treatment of different demographic subgroups [15, 9]. They include observing the differential True Positive Rates (TPR), False Positive Rates (FPR), and prediction confidences across the different demographic subgroups. While these methods can identify bias across subgroups, they do not explain how different biases interact to produce the final predictions. The DeepRepViz framework offers a more comprehensive solution. The biases identified by the Con-score can be visualized as concept activation vectors [13] in the latent representation of the DL model. This allows researchers to visualize not only how different biases are encoded in the final latent representation, but also how they interact with each other and influence the final predictions $\hat{y}$. Thus, the DeepRepViz framework not only helps detect individual sources of bias but also enables researchers to interpret their DL model predictions as a function of different mediating features and confounders.

Con-score derivation: From the perspective of representation learning theory [2], the DL model f_{model} can be divided into two stages $X \Rightarrow^{nonlinear} H^{(l-1)} \Rightarrow^{linear} y$. In other words, the model first transforms X into a condensed latent representation H^{l-1} using a series of non-linear layers, and then the final layer generates the predictions by linearly mapping from $H^{(l-1)}$ to y, as demonstrated in Figure 1(a). Our goal is to estimate how strongly the variable c^k is influencing the model predictions $\hat{y}$. If c^k is influencing the predictions $\hat{y}$, then we would expect two things to be true: first, c^k would be linearly predictable from H^{l-1} (i.e., $X \Rightarrow H^{l-1} \Rightarrow c^k$ exists), and second, a linear model that predicts c^k

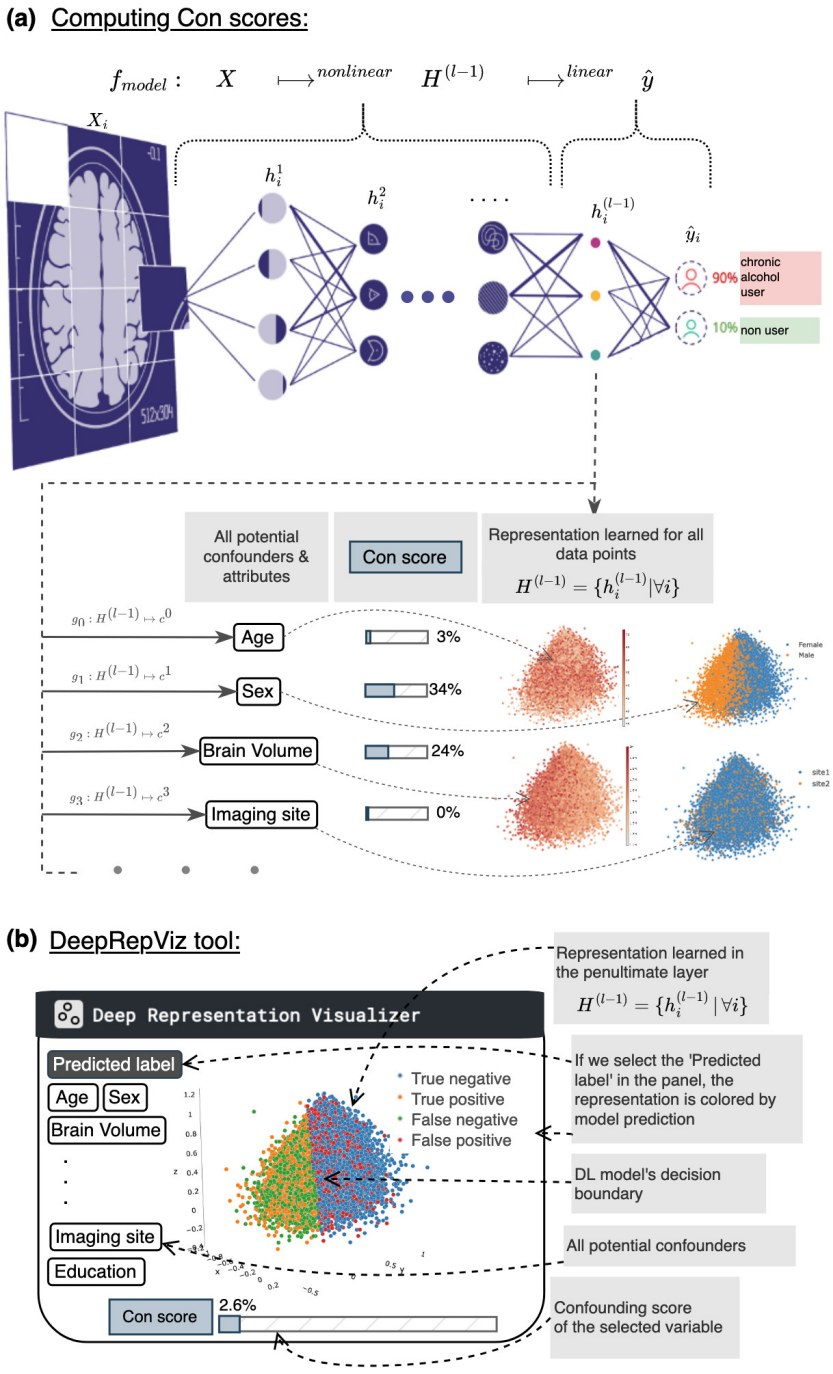

Fig. 1. A demonstration of DeepRepViz and the Con-score metric: Figure (a) shows a DL model that classifies chronic alcohol users from non-users using structural MRI of the brain. Using the final latent representation of the DL model ($H^{(l-1)}$), Con-score is computed for a set of variables associated with the dataset. It is highest for 'sex' when classifying chronic alcohol users. Figure(b) shows the DeepRepViz tool and how it can be used to inspect the learned representation $H^{(l-1)}$. When we select the predicted label $\hat{y}$ in the tool, we can see the linear decision boundary of the model in $H^{(l-1)}$. This decision boundary aligns with the representation of sex in $H^{(l-1)}$ shown in Figure(a). This is also reflected in the Con-score. Therefore, this implies that the model could be using the information about the participant's sex encoded in the MRI data as a proxy to predict chronic alcohol users.

($f_{c^k} : H^{l-1} \Rightarrow c^k$), would be similar to the final linear transformation of the DL model that predicts y ($f^{(l-1)}_{model} : H^{l-1} \Rightarrow y$). The Con-score combines these two criteria:

$$\text{Con}\,(c^k) = \; R^2_{c^k} \cdot \mid \cos\,(\theta_{cy}) \mid \tag{1}$$

Here, $R^2_{c^k}$ is the coefficient of multiple determination [16] of the linear model predicting c^k from H^{l-1} ($f_{c^k} : H^{l-1} \Rightarrow c^k$). If c^k is categorical, then we use McKelvey and Zavoina's *pseudo* R^2 [14] instead of R^2. The second term, $\cos(\theta_{cy})$, is the cosine similarity between the linear model predicting c^k and the final DL layer generating the predictions $\hat{y}$. Here, the θ_{cy} is obtained by taking the vector angle between the parameter of the linear model $f_{c^k} : H^{l-1} \Rightarrow c^k$ and the parameters of the DL layer performing $f^{(l-1)}_{model} : H^{l-1} \Rightarrow y$. The model predicting c^k, f_{c^k}, is constrained by the final DL layer $f^{(l-1)}_{model}$. That is, f_{c^k} always has the same number of parameters and the same activation function (or link function) as $f^{(l-1)}_{model}$. The final Con-score ranges between $[0, 1]$. The higher the Con-score, the higher the likelihood that the model is using the pathway $H^{l-1} \leftrightarrow c^k \rightarrow y$ for its prediction. A Con-score of 1 indicates that the DL model has learned all information about c_k in H^{l-1}, and that the linear prediction of y and the linear prediction of c_k are exactly the same.

DeepRepViz visualization tool: We offer an interactive web-based tool to inspect the latent representation of the DL model, H^{l-1}, with several variables associated with the prediction task. If the Con-score is high for c^k, then we can expect c^k to be clustered in the representation space H^{l-1} [7] and aligned with the prediction of y. Figure 1 demonstrates how the tool can be used to identify potential confounders using an example of a predictive DL model used to diagnose chronic alcohol use from MRI. Currently, the tool requires the representation space to be 3-dimensional (3D) $H^{(l-1)} \in \mathbb{R}^3$. However, the Con-score metric generalizes to representations of any arbitrary dimension $H^{(l-1)} \in \mathbb{R}^n$. Thus, latent representations higher than 3D can be reduced to 3D using dimensionality reduction methods such as PCA, UMAP, or t-SNE [27] before uploading it to the tool and compared using the Con-scores. Apart from the Con-score, the tool also computes other complimentary metrics such as the Silhouette Coefficient to quantify clusters of categorical variables, and correlation scores to quantify the strength of association with continuous variables (please click on 'other metrics' in the tool to see the list of available metrics[4]).

Experiment design: We begin by testing the Con-score metric on simulated data and assess its performance under controlled settings with different boundary conditions. The simulated dataset contains a binary label y, a binary variable c, and a 2-dimensional input data $H = \{h_0, h_1\}$. We generate eight instances of the dataset by systematically altering the correlation between y, c, and the input data H. In the figures, different colors represent the binary states of the label y, while different shapes denote the binary states of the variable c. The top row instances (numbered 1 to 4) are generated such that c can be easily predicted

[4] `https://deep-rep-viz.vercel.app/`

with a linear classification model from the input features H. In the bottom row instances (5 to 8), c classification becomes relatively difficult. Moving from instance number 1 to 4 or from 5 to 8, we incrementally change the correlation between c and y. For instance, in instance number 1, c and y are completely uncorrelated, while in instance 3, c and y become completely correlated.

Next, we evaluate the utility of the metric and the visualization tool on three brain-phenotype prediction tasks using neuroimaging data. On a subsample of $n = 12000$ from the UK Biobank dataset [24], we conduct three exemplary brain-phenotype prediction tasks using a state-of-the-art DL architecture, 3D ResNet-50 [6] available at `https://pytorch.org/hub/facebookresearch_pytorchvideo_resnet/`. We predict the participant's (1) alcohol use, (2) sex, and (3) performance at a cognitive task using the T1-weighted structural MRI data. In the first task, we classify chronic alcohol users from non-users of alcohol. In the second task, we predict the sex of participants from their structural MRI data. In the third task, we predict the time taken by the participants to complete the 'trail-making' cognitive test.

Results

The results of the experiments are presented in Figure 2. The simulated dataset experiment shows that as we systematically increase the similarity between a tertiary variable and the label, the Con-score of the variable also increases proportionately. Experiments on the neuroimaging dataset reveal potential confounders and features that influence the DL model predictions in three brain-phenotype prediction tasks.

Results on the simulated data: Figure 2 (a) shows the Con-score obtained for the eight instances of the simulated dataset generated with different settings of the variable c, a binary label y, and the input features $H = \{h_0, h_1\}$. The Con-scores are highest in the instance when the label classification boundary (black line) and the confound classification boundary (red line) align. In the top row, this occurs in instance 3 and in the bottom row, this occurs in instance 6. In these instances, c is highly correlated with the label y. The second term in the Con-score metric, $\cos(\theta)$, captures this correlation as seen in each instance's title in Figure 2(a). It is easier to linearly predict c in dataset instances on the top row (numbers 1 to 4) compared to the bottom row. The R^2 term of the Con-score captures this variation, as shown in the title of each instance in Figure 2(a). This is evident when we compare the R^2 of instance 3 with 6 or instance 1 with 8. In summary, the Con-scores are highest in the dataset instances where the correlation between c and y is high and c is encoded in the exposure H. For all other cases, the Con-score drops down proportionately.

Results on neuroimaging data: Figure 2 (b) shows the results obtained for the three brain-phenotype prediction tasks. For all the tasks, the Con-score is computed for eight variables from the UK Biobank dataset [24] as listed in the figure legend. When classifying high alcohol users, the DL model only achieves pseudo-$R^2 = 2.6\%$ and the Con-score is highest for the sex variable (0.35). When

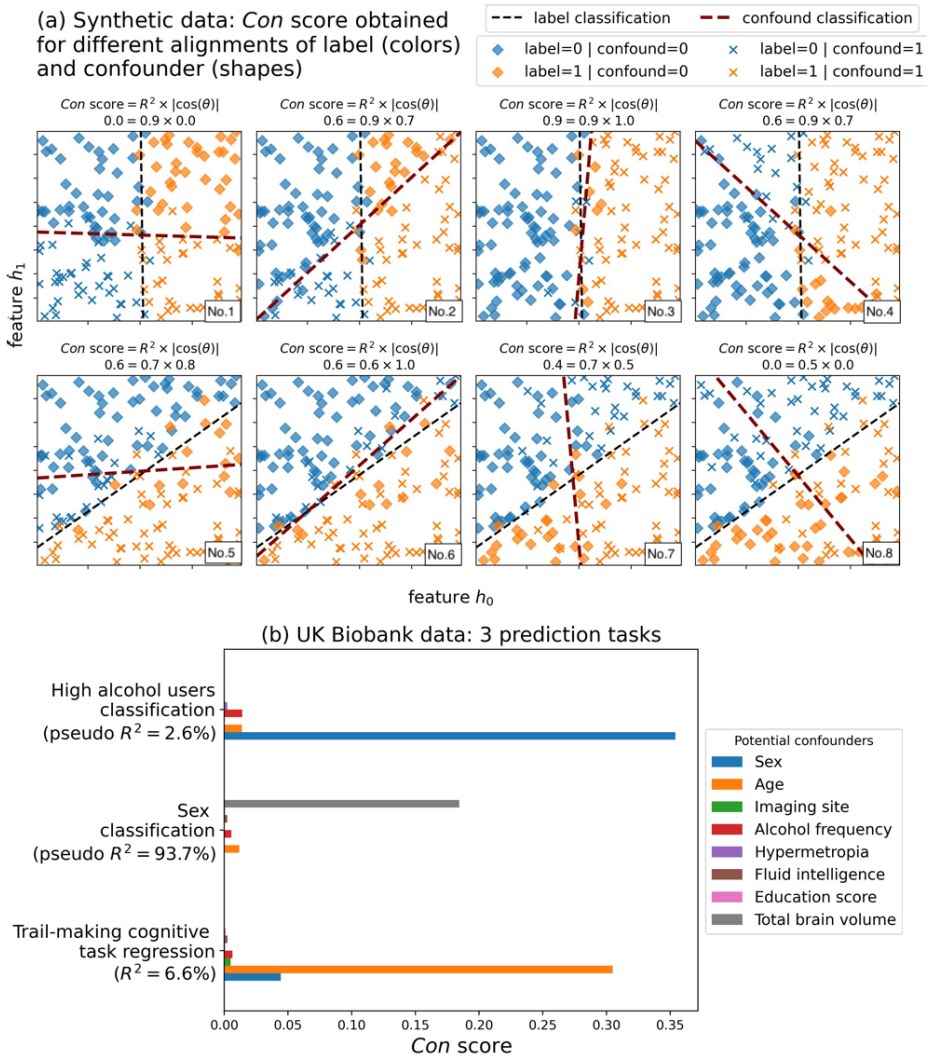

Fig. 2. Results of applying Con-score metric to (a) simulated dataset and (b) UK Biobank neuroimaging dataset: (a) shows Con-scores obtained on a simulated binary classification task for 8 different levels of correlations between a variable c, a binary label y, and the input features $H = \{h_0, h_1\}$. (b) shows the Con-scores obtained for eight potential confounder variables (see legend) in three brain-phenotype prediction tasks on the UK Biobank dataset.

we visualize the representation $H^{(l-1)}$ on DeepRepViz (refer to Figure 1), we find that the majority of the participants predicted as high alcohol users by the DL model are male. This reveals that the model is picking up on the sex bias present in the data. For the cognitive performance prediction task, the DL

model also achieves a low $R^2 = 6.6\%$ and the Con-score is highest for age (0.3). This suggests that the model tends to predict older participants as taking longer to complete the trail-making cognitive test. Visualization on the DeepRepViz tool also confirms that age is encoded in the latent representation and is aligned with the model's prediction of the label. Interestingly, hypermetropia or long-sightedness was not mediating or confounding this task, although one can expect that good eyesight is important for such visual cognitive tests. For sex prediction, the Con-scores reveal that information related to the 'total brain volume' of the participants is encoded in the final learned representation layer (Con-score = 0.18) but none of the other 7 variables are encoded. Whether total brain volume should be considered as a confounder or an explanation depends on the research question behind predicting sex from the brain MRI data and the causal graph [18].

Discussion and Conclusion

Confounders can pose a significant challenge when using predictive modeling techniques such as DL in population neuroscience research. As the size of observational neuroimaging datasets continues to grow, the issue of confounders is only going to worsen [10]. To address this challenge, we offer an easy-to-use framework, 'DeepRepViz', that can be used to examine the latent representation learned by the DL model and identify potential confounders.

Using the DeepRepViz framework in combination with predictive DL models offers several benefits. Firstly, it enables researchers to quickly and easily compare a battery of variables associated with the study and assess their impact on the model predictions [8]. Such a tool is especially useful for population neuroscience studies since psychological phenotypes often co-occur with various demographic, socioeconomic, and environmental factors [20]. Secondly, once confounders are identified, researchers can employ the DeepRepViz framework to validate a confound control method [21]. If a confound control method is successful, then it should reduce the Con-score of the controlled variable to zero. Lastly, the tool not only enables easy identification of potential confounders but also helps to develop an intuitive (qualitative) understanding of the model predictions. Researchers can comprehend their model decisions [20] in terms of human-understandable 'concepts' [13] such as sex, age, or socio-economic status. Additionally, DeepRepViz helps to detect training errors resulting from incorrect model configuration or optimization procedures. Please refer to the tool documentation for more information about these features [5].

Future work and limitations: We are developing the DeepRepViz framework with the goal of making it a generic model exploration tool that contains not only the Con-score but also a plethora of complimentary metrics. In the future, we aim to integrate other bias detection metrics such as the TPR, FPR, and Statistical Parity Difference, alongside the Con-score [15]. In this study,

[5] `https://deep-rep-viz.vercel.app/docs.html`

we validate the DeepRepViz framework on a DL model architecture with a 3D latent representation in its final layer, on two binary labels or one continuous label (regression). Future works should test DeepRepViz on more complex prediction schemes like multi-class classification for non-neuroimaging applications such as natural images. Future work should also test the framework on diverse DL architectures such as DenseNet, EfficientNet, and Vision Transformers. Finally, Con-score cannot differentiate between a mediator and a confounder [18]. A high Con-score is a necessary condition for a variable to be a confounder but it is not a sufficient condition. Nonetheless, the Con-score helps to detect the presence of a tertiary variable c^k influencing the model predictions. The researchers can combine this information with the knowledge of the structural causal graph of the application [18], to determine if c^k must be considered a confounder or interpreted as a crucial feature.

In conclusion, we present DeepRepViz as a versatile tool for validating DL models, opening a pathway for performing medical discovery from neuroimaging data using DL. The tool is publicly available at `https://deep-rep-viz.vercel.app`. Instructions on how to use the tool is available on the documentation page[5]. All the necessary code is made available at `https://github.com/ritterlab/DeepRepViz`. This includes a tutorial notebook demonstrating how DeepRepViz can be integrated into a predictive DL application.

Acknowledgments. This work was funded by the DeSBi Research Unit (DFG; KIFOR 5363; Project ID 459422098), the consortium SFB/TRR 265 Losing and Regaining Control over Drug Intake (DFG; Project ID 402170461), FONDA (DFG; SFB 1404; Project ID: 414984028) and FOR 5187 (DFG; Project ID: 442075332).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] Samaneh Abbasi-Sureshjani et al. "Risk of training diagnostic algorithms on data with demographic bias". In: *Interpretable and Annotation-Efficient Learning for Medical Image Computing*. Springer, 2020, pp. 183–192.

[2] Yoshua Bengio, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828.

[3] Richard J Chen et al. "Algorithmic fairness in artificial intelligence for medicine and healthcare". In: *Nature Biomedical Engineering* 7.6 (2023), pp. 719–742.

[4] Andrew T Drysdale et al. "Resting-state connectivity biomarkers define neurophysiological subtypes of depression". In: *Nature medicine* 23.1 (2017), pp. 28–38.

[5] Fabian Eitel et al. "Promises and pitfalls of deep neural networks in neuroimaging-based psychiatric research". In: *Experimental Neurology* 339 (2021), p. 113608.

[6] Christoph Feichtenhofer et al. "Slowfast networks for video recognition". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 6202–6211.

[7] Ben Glocker et al. "Algorithmic encoding of protected characteristics in image-based models for disease detection". In: *arXiv preprint arXiv:2110.14755* (2021).

[8] Kai Görgen et al. "The same analysis approach: Practical protection against the pitfalls of novel neuroimaging analysis methods". In: *Neuroimage* 180 (2018), pp. 19–30.

[9] Max Hort et al. "Bias mitigation for machine learning classifiers: A comprehensive survey". In: *ACM Journal on Responsible Computing* (2023).

[10] Courtland S Hyatt et al. "The quandary of covarying: A brief review and empirical examination of covariate use in structural neuroimaging studies on psychological variables". In: *NeuroImage* 205 (2020), p. 116225. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2019.116225. URL: https://www.sciencedirect.com/science/article/pii/S105381191930816X.

[11] W Evan Johnson, Cheng Li, and Ariel Rabinovic. "Adjusting batch effects in microarray expression data using empirical Bayes methods". In: *Biostatistics* 8.1 (2007), pp. 118–127.

[12] M Khojaste-Sarakhsi et al. "Deep learning for Alzheimer's disease diagnosis: A survey". In: *Artificial Intelligence in Medicine* 130 (2022), p. 102332.

[13] Been Kim et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)". In: *International conference on machine learning*. PMLR. 2018, pp. 2668–2677.

[14] Richard D McKelvey and William Zavoina. "A statistical model for the analysis of ordinal level dependent variables". In: *Journal of mathematical sociology* 4.1 (1975), pp. 103–120.

[15] Ninareh Mehrabi et al. "A survey on bias and fairness in machine learning". In: *ACM computing surveys (CSUR)* 54.6 (2021), pp. 1–35.

[16] Nico JD Nagelkerke et al. "A note on a general definition of the coefficient of determination". In: *Biometrika* 78.3 (1991), pp. 691–692.

[17] Tomáš Paus. "Population neuroscience: why and how". In: *Human brain mapping* 31.6 (2010), pp. 891–903.

[18] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[19] Roshan Prakash Rane, Andreas Heinz, and Kerstin Ritter. "AIM in Alcohol and Drug Dependence". In: *Artificial Intelligence in Medicine* (2022), pp. 1619–1628.

[20] Roshan Prakash Rane et al. "Eating-related variables partially explain the prospective prediction of binge drinking from structural brain features". In: (2023).

[21] Roshan Prakash Rane et al. "Structural differences in adolescent brains can predict alcohol misuse". In: *Elife* 11 (2022), e77545.

[22] Laleh Seyyed-Kalantari et al. "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations". In: *Nature medicine* 27.12 (2021), pp. 2176–2182.

[23] Lukas Snoek, Steven Miletić, and H Steven Scholte. "How to control for confounds in decoding analyses of neuroimaging data". In: *NeuroImage* 184 (2019), pp. 741–760.

[24] Cathie Sudlow et al. "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age". In: *PLoS medicine* 12.3 (2015), e1001779.

[25] Jing Sui et al. "Neuroimaging-based individualized prediction of cognition and behavior for mental disorders and health: methods and promises". In: *Biological psychiatry* 88.11 (2020), pp. 818–828.

[26] Elina Thibeau-Sutre et al. "MRI field strength predicts Alzheimer's disease: a case example of bias in the ADNI data set". In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2022, pp. 1–4.

[27] Laurens Van Der Maaten, Eric O Postma, H Jaap Van Den Herik, et al. "Dimensionality reduction: A comparative review". In: *Journal of Machine Learning Research* 10.66-71 (2009), p. 13.

[28] Sandra Vieira, Walter HL Pinaya, and Andrea Mechelli. "Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications". In: *Neuroscience & Biobehavioral Reviews* 74 (2017), pp. 58–75.