Computer Methods and Programs in Biomedicine 242 (2023) 107771

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Identifying predictive biomarkers for repetitive transcranial magnetic stimulation response in depression patients with explainability

Matthew Squires [a,*], Xiaohui Tao [a,*], Soman Elangovan [b], Raj Gururajan [c], Xujuan Zhou [c], Yuefeng Li [d], U. Rajendra Acharya [e]

[a] School of Mathematics, Physics and Computing, University of Southern Queensland, Toowoomba, Australia
[b] Belmont Private Hospital, Brisbane, Australia
[c] School of Business, University of Southern Queensland, Springfield, Australia
[d] School of Computer Science, Queensland University of Technology, Brisbane, Australia
[e] School of Mathematics, Physics and Computing, University of Southern Queensland, Springfield, Australia

ARTICLE INFO

Keywords:
Repetitive transcranial magnetic stimulation
Deep learning
Explainable AI
Depression

ABSTRACT

Repetitive Transcranial Magnetic Stimulation (rTMS) is an evidence-based treatment for depression. However, the patterns of response to this treatment modality are inconsistent. Whilst many people see a significant reduction in the severity of their depression following rTMS treatment, some patients do not. To support and improve patient outcomes, recent work is exploring the possibility of using Machine Learning to predict rTMS treatment outcomes. Our proposed model is the first to combine functional magnetic resonance imaging (fMRI) connectivity with deep learning techniques to predict treatment outcomes before treatment starts. Furthermore, with the use of Explainable AI (XAI) techniques, we identify potential biomarkers that may discriminate between rTMS responders and non-responders. Our experiments utilize 200 runs of repeated bootstrap sampling on two rTMS datasets. We compare performances between our proposed feedforward deep neural network against existing methods, and compare the average accuracy, balanced accuracy and F1-score on a held-out test set. The results of these experiments show that our model outperforms existing methods with an average accuracy of 0.9423, balanced accuracy of 0.9423, and F1-score of 0.9461 in a sample of 61 patients. We found that functional connectivity measures between the Subgenual Anterior Cingulate Cortex and Centeral Opercular Cortex are a key determinant of rTMS treatment response. This knowledge provides psychiatrists with further information to explore the potential mechanisms of responses to rTMS treatment. Our developed prototype is ready to be deployed across large datasets in multiple centres and different countries.

1. Introduction

Depression is a highly prevalent and debilitating mental illness [1]. As such, finding effective and efficient treatments for depression is a high priority. Repetitive Transcranial Magnetic Stimulation (rTMS) is an evidence-based treatment for depression [2–4]. rTMS involves electromagnetic stimulation of the brain that aims to alter its underlying structures to improve a patient's symptoms [7]. However, the patterns of response to this treatment are inconsistent [5]. Evidence [6,9,5] suggests that the distribution of response to rTMS is bimodal. For some patients, rTMS treatment will lead to a significant reduction in depression severity. However, others see minimal improvement in their depression rating scale scores post-treatment. Given this disparity, current work is investigating the potential of using artificial intelligence (AI) to predict treatment outcomes and personalize mental healthcare.

To date, existing research has sought to predict response to rTMS treatment using machine learning (ML) algorithms. These systems aim to delineate between responders and non-responders in rTMS treatment. Thus, the problem can be defined as a supervised binary classification task. Existing works [21,22,11–13,15–17,10,9,18] have applied a variety of algorithms to predict the response to rTMS treatment. Methods include linear support vector machines [22,11,12], linear regression [13]

* Corresponding authors.
E-mail addresses: matthew.squires@usq.edu.au (M. Squires), xiaohui.tao@usq.edu.au (X. Tao), soman.elangovan@healthecare.com.au (S. Elangovan), Raj.Gururajan@usq.edu.au (R. Gururajan), xujuan.zhou@usq.edu.au (X. Zhou), y2.li@qut.edu.au (Y. Li), Rajendra.Acharya@usq.edu.au (U.R. Acharya).

https://doi.org/10.1016/j.cmpb.2023.107771
Received 8 June 2023; Received in revised form 12 August 2023; Accepted 19 August 2023
Available online 25 August 2023

and k-nearest neighbours [15]. These surveyed methods vary from those that rely on features collected after treatment has begun, to emerging methods that utilize pre-treatment measures only.

Predicting the treatment outcome before it begins is the goal of personalized mental healthcare. However, such examples are in the minority. For example, only Hopman et al. [11] and Hasanzadeh et al. [15] predicted treatment outcomes before starting treatment. In their work, Hasanzadeh et al. [15] utilized pre-treatment EEG features to accurately predict rTMS treatment outcomes in roughly 90% of cases. Hopman et al. [11] instead used pre-treatment functional magnetic resonance imaging (fMRI) to predict treatment outcomes. These results suggest that there is scope for the use of more advanced techniques, such as deep neural networks (DNN), to predict the outcome of rTMS in patients [7]. This observation is echoed by [12], who assert that future work could explore the efficacy of deep learning (DL) algorithms for predicting the treatment response. For example, the linear support vector machine (SVM) used by Hopman et al. [11] performed excellently during cross-validation, however, Hopman et al. [11] reported sharp declines in predictive performance on a held-out test set. Therefore, an opportunity exists to explore more sophisticated algorithms, such as DNNs, which are known - under the right settings - to generalize well on unseen data. As such, our work seeks to address the following research questions:

1. Can a deep neural network improve upon existing methods for predicting treatment response on a held-out test set?
2. Which features most predict treatment response?
3. In what circumstances is the proposed network vulnerable to misclassification?
4. Are there any commonalities in misclassification errors that can be communicated to the end user to improve clinical utility?

To address these research questions, we compare empirically existing shallow ML methods against our proposed DNN. Furthermore, with the aim of increasing the value of work for end users and in collaboration with domain experts, we utilize explainable artificial intelligence (XAI) techniques to identify the features that are most predictive of treatment response. In addressing this research question, we aim to identify candidate biomarkers indicative of treatment response. Additionally, to support the potential implementation of our model, we present model knowledge, which is an extension of the 'model facts labels' presented by [20]. Model knowledge is our process of rigorously evaluating model performance, including potential limitations. By gathering model knowledge, we can enhance the clinical utility by explicitly declaring circumstances such as when the model performs well or is vulnerable to prediction errors, which in turn promotes trust in end users. Lack of trust in AI models is seen as a key barrier to its implementation in healthcare [19]. Through addressing these research questions, our work makes the following contributions:

- A robustly validated regularised deep feedforward neural network that predicts the treatment outcome of rTMS before treatment commences.
- A robust analysis of rTMS treatment response patterns through the use of two datasets, namely, the differences in the predictive power of self-reported psychometric data against fMRI connectivity measures.
- The use of XAI techniques, including SHAPLEY values, to identify candidate biomarkers indicative of response to rTMS treatment.

These findings help to provide confidence to clinicians in our model while also uncovering new knowledge for depression researchers.

The current work proposes a DNN model for predicting treatment response to rTMS. We use multi-modal data to predict treatment outcomes and explore XAI techniques to add support and robustness to our model. As such, our work aims to produce the first DNN model to model rTMS treatment outcomes that includes explainability. The paper is structured as follows. The following section reviews existing strategies for treatment response prediction used to evaluate rTMS. Given the dearth of literature exploring DL architectures in rTMS, we include an exploration of DL systems applied in other medical contexts. Additionally, we survey some methods used to produce interpretable AI systems. In Section 2, we present the research problem, a summary of the dataset, formally define the research problem and introduce the notation. Section 2 also introduces our proposed model, the baseline models, and model hyperparameters. Details on the experiment design, performance measuring schemes and experiment results are included in Section 3. Finally, comments about our findings and proposed future directions from this research topic are included in Section 4.

1.1. Related work

Previous studies [21,22,11–13,15–17,10,9,18] have applied a variety of techniques to predict treatment outcomes to rTMS. To date, ML algorithms have performed well on this binary classification task. A summary of the current literature is shown in Table 1.

This Table shows that several feature modalities, including electroencephalogram (EEG), fMRI, psychological, and demographic features, have been used to predict treatment response to rTMS. Existing work has relied on shallow machine learning methods like linear SVM (LSVM) and k-nearest neighbours algorithms (KNN). Recently, Shadabi et al. [22] became the first to apply DL methods to rTMS response prediction when they explored the ability of a convolutional neural network (CNN) to predict treatment response to rTMS using EEG features. CNNs are well suited to the temporal data collected by EEG, with the authors reporting a 97.1% average accuracy after 10 fold cross validation.

Previously, existing research has relied on shallow ML methods. For example, Hopman et al. [11] deployed a LSVM using features collected via fMRI. They used connectivity features between the subgenual anterior cingulate cortex and lateral occipital cortex, superior parietal lobule, frontal pole and central opercular cortex. During five-fold cross validation, the authors presented a training accuracy of $\approx 97\%$ however, on a unseen test set, model performance dropped to an average of $\approx 87\%$, with a 95% confidence interval from 100% to roughly 70% accuracy. Further implementations of a LSVM include Bailey et al. [12], who built a LSVM classifier composed of 54 features. These features consisted of a combination of mood and EEG measurements collected at baseline and after one week of treatment. In addition to measurements collected at these two-time points, features were extracted for the change between week 1 and baseline. Each feature was standardized, which is a common technique in ML. Testing of the final classifier was validated against 5000 runs of five-fold validation. For this LSVM, Bailey et al. [12] reported a mean balanced accuracy of 86.60%. As part of their conclusions, they felt that the efficacy of existing algorithms for the prediction of treatment response could be improved [12].

DL algorithms are capable of modelling complex relationships and yield high classification performances. Moving from ML to DL to predict rTMS treatment outcomes is the potential next step in mental healthcare. The strength of DL architectures is the ability to model complex multi-variable relationships with improved accuracy [26]. Hence, the opportunity exists for DL methods to be applied to rTMS modelling using fMRI connectivity features. However, the challenge in applying DL methods to critical domains such as mental health care is the distrust toward DL methods due to their lack of interpretability [19]. XAI is a field of AI research that focuses on the inner workings of complex DL models. DL models are more powerful for identifying relationships than more interpretable shallow methods. However, there is a trade-off between performance and interpretability. XAI techniques aim to eliminate that trade-off by increasing the interpretability of DL models. Furthermore, Hopman et al. [11] observed that their LSVM failed to generalise well to unseen data. By contrast, DNNs are known to gener-

Table 1
rTMS depression treatment response prediction.

Author	Modality	Features	Algorithm	Performance	Validation
Ebrahimzadeh et al. [21]	EEG	EEG beta power, Correlation Dimension (CD), Permutation entropy (PE), Fractal dimension (FD), Lempel-Ziv Complexity (LZC), Power spectral density, Frontal and prefrontal cordance	SVM*	94.31% average accuracy after cross validation	10 fold cross validation
Shahabi et al. [22]	EEG	Continuous Wavelet Transform	CNN	97.1% average accuracy after cross validation	ten-fold cross validation
Hopman et al. [11]	fMRI	Connectivity features: subgenual anterior cingulate cortex, lateral occipital cortex, superior parietal lobule, frontal pole and central opercular cortex	LSVM	87% accuracy on a held out test set	five-fold cross validation
Bailey et al. [12]	EEG and Mood	Alpha power, theta power, alpha connectivity, theta connectivity, theta cordance, individualised alpha peak frequency (iAPF) and MADRS	LSVM	Mean balanced accuracy of 86.60%	5000 runs of five-fold cross validation
Fan et al. [13]	fMRI	Network Segregation of the Salience Network*	Regression	Coefficient of determination of 0.27	NA
Hasanzadeh et al. [15]	EEG	Power of beta*	K-NN	Accuracy of 91.3%	Leave-one-out cross validation
Bailey et al. [17]	Mood, Behaviour and EEG	Alpha power, theta power, gamma power, alpha connectivity, theta connectivity, gamma connectivity, theta gamma coupling, MADRS, working memory and reaction time	LSVM	F1 score = 0.93	200000 runs of five fold cross validation
Drysdale et al. [9]	fMRI	Connectivity features	Hierarchical Clustering and SVM	Balanced accuracy of 90.39%	Leave-one-out cross validation

* Best performing model.

alise well to unseen data, which potentially addresses the performance decline found in studies such as [11]. Additionally, the inclusion of explainability can improve trust in end users and take advantage of DNN's improved performance over existing methods.

At present, there is a dearth of literature exploring XAI and rTMS treatment. A recent review by [19] argued for the importance of including XAI through methodologies like SHAP values to enhance trust in DL methods. Thus, our work is motivated to enhance trust in DL methods from psychiatrists through both improving performance in rTMS response prediction, and including XAI in our approach.

2. Materials and methods

This section outlines the problem statement, the datasets used and defines the research problem to be explored. Here, we provide some background information on DNNs and their development, before we present our model, which uses quantitative data to evaluate the treatment effects of rTMS. Additionally, this section includes details about our strategies for reducing overfitting and internally validating our model.

2.1. Problem statement

The effectiveness of rTMS for the treatment of depression is now well-established [27]. Significant evidence shows rTMS to be a safe and effective intervention for treatment-resistant depression [2–4]. Despite this effectiveness, some patients will see no significant improvement in their depression severity following rTMS treatment [5]. To address these inconsistent response patterns, we are exploring ways to better target rTMS treatment toward patients who are likely to see the most benefit. In order to personalize care, AI can be deployed to support psychiatrists [28]. The aim of our work is to explore the potential of a DNN architecture to predict response to rTMS treatment and identify any potential biomarkers indicative of treatment response.

2.2. Research design

The current work aims to test the efficacy of a DNN to predict the treatment outcome of rTMS. Utilising empirical experiments, we seek to investigate whether DL offers any improvement over existing methods. As part of this work, we identify the features that provide the most information for treatment response in the hope of identifying the key biomarkers. Additionally, our experiments compare self reported measures or fMRI connectivity features for predicting treatment response. By utilising XAI techniques, we present new knowledge that can aid clinicians in prescribing treatments.

2.3. Datasets

To address whether a DNN can provide robust predictions compared to existing methods, we utilise two datasets. The first dataset was used in published work by Hopman et al. [11] and made publicly available in [29]. The data includes several fMRI features, along with a patient's treatment outcomes. A summary of the mean connectivity measures across response type can be seen in Table 2. Further detailed summary statistics of this dataset, including associated ethics approval, can be found in [29].

The second dataset is new data collected from a large private hospital in Australia that specializes in the delivery of rTMS care. A summary of the relevant psychological variables collected in this data is shown in Table 3. This Table shows the mean survey scores between groups. The psychological health information in this Table was collected using

Table 2
Mean connectivity measurements by group in Dataset 1: Hopman [29].

	Responders	Non-responders
N (n = 61)	33	28
Frontal Pole Connectivity	0.0252	−0.1026
Occipital Cortex Connectivity	0.0667	−0.0467
Superior Parietal Lobule Connectivity	0.0802	−0.0450
Centeral Opercular Cortex Connectivity	0.1142	0.0266
Left Lateral Occipital Cortex Connectivity	0.0607	−0.0440
Right Lateral Occipital Cortex Connectivity	0.0386	−0.0476

Table 3
Mean DASS measurements by groups in Dataset 2: data collected from Belmont hospital.

	Responders	Non-responders
N (n = 133)	83	50
Depression Baseline	29.4819	27.4000
Anxiety Baseline	17.0843	16.3600
Stress Baseline	25.4578	23.4800
Depression after 10 sessions	17.1566	25.9800
Anxiety after 10 sessions	11.4698	13.7200
Stress after 10 sessions	15.3253	20.4400

the Depression, Anxiety and Stress Subscale [DASS 30]. DASS is a self-report survey measuring three dimensions of mental health: depression, anxiety and stress. Patients are required to complete a baseline survey prior to treatment, then in the rTMS program, they complete the DASS survey 3 times during treatment. An additional survey is completed after 10 sessions of rTMS, and a final measurement about patients is collected following treatment.

In the current work, the DASS-21 was used, a short form of the 42 item DASS. Each dimension of mental health in the DASS-21 has a maximum score of 42. For the depression dimension, a score of greater than 21 is deemed severe depression [23]. Participants in this study consented to DASS data being used for the study of rTMS treatment. Ethics approval was obtained from the University's Human Research Ethics Committee to use and analyse collected data.

2.4. Problem definition

The current work seeks a function that optimizes the classification of patients as responders or non-responders to rTMS treatment. In addition, this function provides a model of the treatment effect of rTMS dependent on either psychological or neuroimaging based variables.

Formally, let X be a dataset containing Patients P and class label Y, where:

$$X = \{(p_1, y_1), (p_2, y_2), (p_3, y_3) \dots (p_n, y_n)\} \tag{1}$$

Each patient p has a set of connectivity measures C such that

$$C = \{FP, OC, SPL, COC, lOCL, lOCR\}$$

$$y_i = \begin{cases} 1 & \text{if } \frac{\Delta d}{d_0} \leq -0.5 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Patients are assigned a class label according to the function in Equation (2). Any patient who experiences a greater than 50% reduction in depression severity, $\Delta d \leq -0.5$, is classified as a responder and assigned the class label $y_i = 0$. Conversely, patients who see a less than 50% reduction, $\Delta d > -0.5$, are classed as non responders and receive the label $y_i = 0$. The target variable for this binary classification task is y. As such, we seek a classifier

$$h[p] = y$$

which minimizes the prediction error between classes. See Table 4.

Table 4
Symbol descriptions.

Symbol	Description
X	Dataset
Y,*y*	Set of class labels and a patient's class label
p	A patient
d_0, Δd	Depression severity at baseline, change in depression severity following treatment
FP	Functional connectivity measure between Subgenual Anterior Cingulate Cortex and Frontal Pole
OC	Functional connectivity measure between Subgenual Anterior Cingulate Cortex and Occipital Cortex
SPL	Functional connectivity measure between Subgenual Anterior Cingulate Cortex and Superior Parietal Lobule
COC	Functional connectivity measure between Subgenual Anterior Cingulate Cortex and Centeral Opercular Cortex
lOCL	Functional connectivity measure between Subgenual Anterior Cingulate Cortex and Left Lateral Occipital Cortex
lOCR	Functional connectivity measure between Subgenual Anterior Cingulate Cortex and Right Lateral Occipital Cortex

2.5. Deep neural networks

DL is a subfield of ML that builds upon existing neural network architectures by increasing the number of hidden layers of a network [24]. This increased depth allows for modelling of increasingly complex nonlinear functions [25]. The complexity makes it possible for models to learn complex representations of existing data, which may not be observable using traditional inferential statistics or standard ML techniques.

The basis for the artificial neural network (ANN) is found in the seminal work of Rosenblatt et al. [31]. Where initially a single perceptron defined a linear decision boundary between a binary set of classes, the multilayer perceptron (MLP) adds the concept of a hidden layer. The hidden layer involves multiple perceptrons, with each perceptron sharing an edge with each node in the hidden layer. This increase in model complexity increases the model's predictive power beyond linear functions to learning complex nonlinear decision boundaries between classes [33]. The MLP is a feedforward neural network applied to classification and regression [32]. An MLP with several hidden layers is referred to as a DNN [8]

A crucial aspect of the performance of the MLP is the training of the network. Training refers to the model weights being tuned so that predicted outputs match the expected outputs or ground truth values of the data [33]. The process of tuning these weights or parameters can be referred to as learning. Rarely can a model match all examples with their ground truth labels, therefore, we need a function to monitor the performance of the model during training. Training aims to minimize a loss function to obtain the weights so that the difference between the expected and predicted outcomes is minimized [32].

2.6. Regularisation

Modern solutions have enabled the fitting of increasingly complex functions to data. However, the added complexity of networks with several hidden layers increases the risk of overfitting. That is, where functions simply memorise datasets. Regularisation encompasses a class of tools used to reduce the risk of model overfitting. Common strategies for reducing the risk of overfitting include: early stopping, weight regularisation and dropout [24].

Early stopping involves monitoring a metric during training and ending training when the selected value stops improving [34]. In addition to an unseen test dataset, we use a validation set during model training in our project. Validation loss is monitored throughout training, and for each trained model as part of our bootstrap resampling, patience was set to 100. We set a minimum improvement in validation loss of 0.05 as being required to continue model training.

Table 5
Model hyperparameters.

Hidden Layers	4
Layer Width	10
Activation Function	reLU
Loss Function	Binary Crossentopy
Regularisation Layers	4
Test set size	20%
Epochs	2000 or until early stopping criteria met

Dropout involves turning a proportion of parameter weights down to zero. Conceptually, we can perceive this as 'dropping' edges between nodes. Srivastava et al. [35] first proposed dropout as a regularisation strategy to add noise to a neural network. The introduction of noise through 'dropping' connections between neurons forces the network being trained on the data to identify the true nature of the signal within the data. In turn, this reduces an overparametrised network's ability to memorize the dataset. The benefit in identifying the true signal from the data means a greater potential for identifying meaningful patterns within the data. As such, each layer of our trained model includes a dropout probability of 0.3.

The final hyperparameters of our model are shown in Table 5. These final hyperparameters were selected after an iterative model building process. Through several cycles of experiments, we monitored how changes in model hyperparameters impacted model performance. Through continual refinement and the aim of creating a model robust to overfitting, we settled on the final model hyperparameters. These selected values achieve the goal of a model that generalizes well to unseen data when compared against existing methods.

2.7. Experimental design

This section provides an overview of the empirical experiments used to explore the research questions outlined in Section 1. The current work presents two experiment arms. In the first arm, we test our proposed DNN on data collected by [11]. This data includes fMRI connectivity features from 61 patients suffering from depression treated by rTMS. In this experiment arm, we assess the ability of our model to discriminate rTMS responders from non-responders neuroimaging features.

Our second set of experiments utilises a privately collected dataset from Belmont Private Hospital, Brisbane, Australia. This data includes the records of 133 patients who undertook rTMS treatment. However, in contrast to the first experiment, the features of the second experiment arm include only features collected through a self-reported questionnaire.

Rigorous validation of ML and DL algorithms is essential to ensure the robustness of reported results. The validation of AI systems for healthcare is an important step in the transition to clinical practice [36]. Harrel [37] asserts that the strongest form of internal validation is repeated bootstrap resamples, with analysis of the target variable repeated for each resample. This process ensures that a relationship between input variables and target variables exists, thus increasing the robustness of the results.

These experiments are designed to compare the performance of self reported psychometric measures of depression severity against quantitative fMRI measures, in predicting treatment outcomes to rTMS. Through these experiments, we aim to identify candidate biomarkers that explain the patterns of response to rTMS treatment.

2.8. Baseline models

Our baseline models include a LSVM, as proposed by [21,11,12], and a KNN classifier [21]. Additionally, we include XGBoost and random forests as baseline models, which are widely used in healthcare with explainability [19]. The hyperparameters of all baseline models were optimized using grid search. In our final experiments, we used the best hyperparameter set found during grid search to compare against our proposed DNN. Full baseline model hyperparameters are listed in Appendix A, Table A.11.

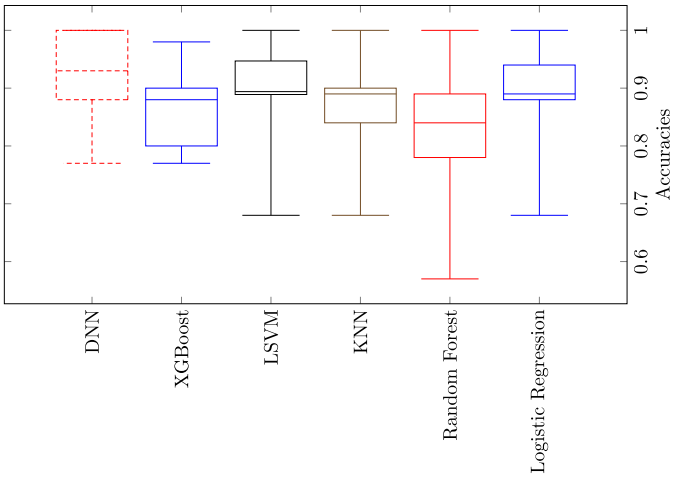

Fig. 1. Box plot showing the accuracies obtained using various algorithms.

For our second dataset, we explore the potential of DNNs to model changes in depression severity from psychometric questionnaires. For this experiment, we utilize the only model that includes the same features. This baseline model is provided by Feffer et al. [6].

As outlined in Section 1.1, Feffer et al. used early symptom improvement to predict treatment response to rTMS. In their study, they proposed that patients with $< 20\%$ reduction in depression severity after 2 weeks of treatment (10 sessions) are unlikely to respond to treatment. As such, inline with the model proposed by Feffer et al., patients with $\leq 20\%$ improvement in symptoms after 10 sessions are classed as non-responders, and the remaining cases are defined as responders. The Feffer et al. [6] model reported high sensitivity but low specificity.

2.9. Performance measuring schemes

Performance metrics are required to evaluate models and make comparisons between them. These metrics differ slightly depending on the nature of the outcome variable. Common metrics used for the evaluation of classification models in psychiatry include F1 score and accuracy, as used in Chang et al. [38]. Additionally, in line with Bailey et al. [12], we have included balanced accuracy to assess performance in both the positive and negative cases.

3. Results

We present the results of our two experiment arms (Experiments 1 and 2) in the two sub-sections below.

3.1. Experiment 1: fMRI connectivity measures to predict rTMS treatment outcomes

Recent work by Hopman et al. [11] proposed a LSVM for the early prediction of treatment response to rTMS. Their works combined fMRI features with a Linear SVM to predict treatment outcomes. Our experiments compare the performance of several baseline models against our proposed DNN architecture over 200 repeated bootstrap samples. The distribution of test set accuracy for each algorithm is shown in Fig. 1.

From this diagram we see while most algorithms have an upper limit of correctly predicting all cases in the test set. The DNN finds this optimal solution more frequently across all trials. Followed by the logistic regression, and the LSVM. With the observed LSVM performance closely mirroring the performance that [11] obtained on the same dataset.

Table 6
Summary of average model performance matrices obtained from 200 bootstrap resamples.

Model	Accuracy	F1 score	Balanced accuracy
DNN	0.9423 (0.0605)	0.9461 (0.0561)	0.9423 (0.0618)
XGBoost	0.7813 (0.0823)	0.7916 (0.0794)	0.7804 (0.0830)
LSVM	0.9107 (0.0588)	0.9127 (0.0584)	0.9115 (0.0587)
KNN	0.8913 (0.06842)	0.8942 (0.06753)	0.8918 (0.0686)
Random Forests	0.8416 (0.0822)	0.8506 (0.0790)	0.8404 (0.0830)
Logistic Regression	0.9047 (0.0636)	0.9071 (0.0647)	0.9052 (0.0635)

Table 7
Summary of average model performance in Experiment 2.

Model type	Feature Set	F1-score	Accuracy	Balanced accuracy
Feffer et al. [6]	Domain Knowledge	0.857	0.815	0.791
DNN	DASS Scores	0.772	0.630	0.500

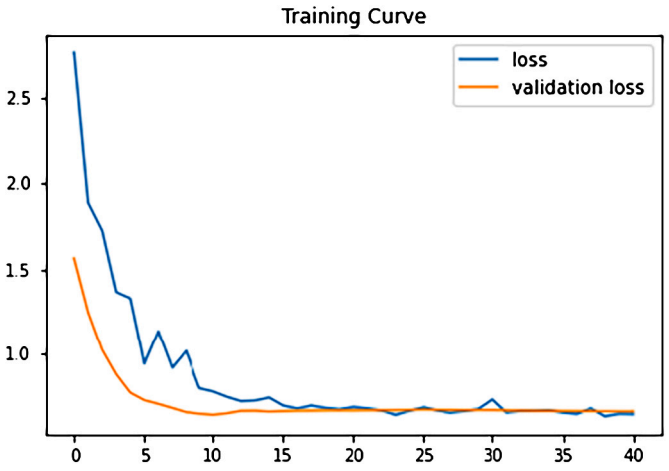

Fig. 2. A graph of training loss against validation loss.

Table 6 shows the summary of results obtained using our model compared against our baseline models. Reported results are the average of 200 bootstrap resamples with standard deviations included in parentheses. Again, our proposed DNN outperformed all baseline models across the reported metrics.

In addition to the described summary metrics obtained using 200 samples, the LSVM identified the optimal solution for correctly classifying the unseen test set 21 times compared to 64 times using the DNN. To demonstrate the robustness of the DNN model we have also included training curves. One way to ensure the robustness of DNN performance is to monitor training loss compared to validation loss. The significant divergence between training and validation loss, when validation loss deteriorates significantly compared to training loss, indicates that the model is overfitting. The training curve from 1 of the 200 trained networks is shown in Fig. 2. The figure shows no significant divergence between losses. During our testing when regularisation was removed, the model was prone to overfitting. This was demonstrated by a significant divergence between validation loss and training loss.

It may be noted from the performance of both the LSVM and DNN that a signal exists between variables and patterns of response. This motivated us to further explore which variables are most significant for correctly predicting treatment response. Based on the results of our experiments, a DNN with the hyperparameters described in Table 5 outperforms the existing baseline models across all metrics.

3.2. Experiment 2: self-reported DASS scores to predict final rTMS treatment outcome

Extending our current work, we investigate the potential for self-reported measures to predict rTMS treatment outcomes. Existing work has shown early changes in symptom severity to be a reasonable predictor of rTMS treatment outcome. Extending upon this work, we explore whether a DNN can identify the relationship between self-reported depression severity and treatment response.

The results shown in Table 7 highlight that when using DASS scores, the preferred method to predict treatment response is domain knowledge as described in [6]. These results highlight that the fMRI connectivity features are superior to self reported DASS scores. Surprisingly, the DNN was unable to pick up on the relationship between early symptom improvement and final treatment outcome.

3.3. Explainable AI (XAI) approaches to identify potential biomarkers indicative of response to treatment

This paper has emphasized the importance of understanding model performance. This position is echoed by Tjoa and Guan [41], who asserted that when DNNs and AI models are applied to non-trivial tasks, improving model understanding is imperative. Methods for assessing feature importance vary from global to local explanations. Global methods explore feature importance from a global scope [42]. In contrast, local methods provide an explanation as to which variables are contributing to the prediction of an individual case within the dataset.

We consider two methods for ranking feature importance: a global and a local method. A global post-hoc method that is commonly used to interpret AI methods is permutation feature importance (PFI) [25]. A PFI score involves the shuffling of one variable within the testing set before the data containing the shuffled feature is input into the trained model [25]. Similar to ablation, the more significant the decline in the model's performance metrics, the greater the relative importance to the model. This process is then repeated throughout all variables in the dataset. One limitation of this approach is that any correlation between a shuffled feature and an unshuffled feature may lead to underestimating the importance of a feature [25]. This issue is similar to the issue of colinearity in simple linear regression.

Utilizing the PFI score, Table 8 shows the relative performance declines in performances associated with each variable. This Table highlights that shuffling of COC leads to the most significant performance decline in model performance. This is measured by the change in test set accuracy by iteratively shuffling each variable. For clarity, the relative performance change attributed to each feature is shown visually in Fig. 3.

Table 8
Summary of feature importance scores.

Feature	Test set accuracy	Percentage drop in accuracy
FP	92.3077	−7.6923
OC	76.9230	−23.0769
SPL	61.5384	−38.4616
COC	53.8462	−46.1538
LOCL	92.3077	−7.6923
LOCR	92.3077	−7.6923

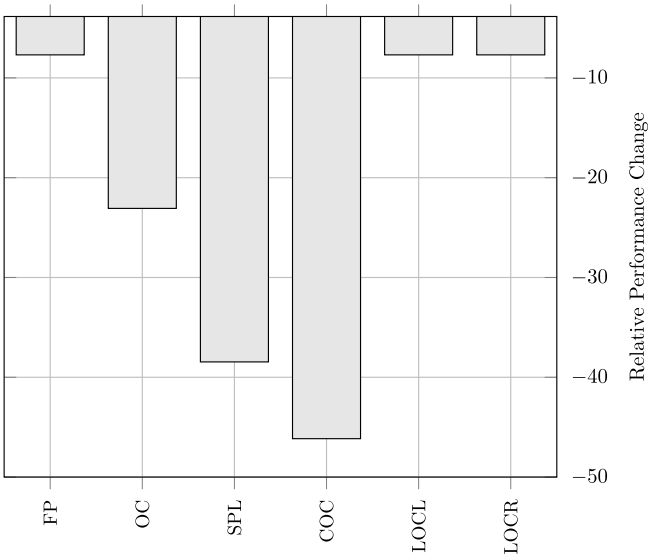

Fig. 3. Relative change in the performance due to various features.

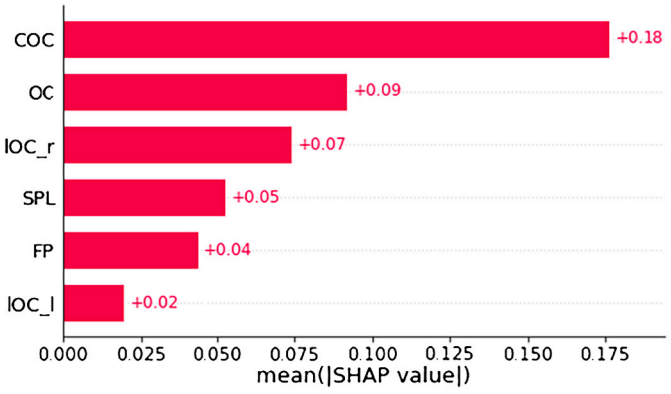

Fig. 4. Average SHAP values on Training Set.

3.3.1. Shapley (SHAP) values

One local method used for assessing feature importance is the SHAP value [43]. Inspired by the seminal work of Shapley [44], Lundberg and Lee [43] introduced the SHAP value. The Shapley value is a game theoretic approach to measure a player's contribution to an end goal in an n-player cooperative game. SHAP values then provide a local explanation for the contribution of each feature to a final output.

Using the SHAP values calculated in Python's SHAP package offers support for computing PFI score results. Fig. 4 shows that COC contributes significantly to model predictions, followed by OC. These findings mirror the results of the PFI score except for SPL, which is ranked much lower by SHAP value when compared to the results in Fig. 3.

One strength of local approaches to XAI is the ability to investigate SHAP values for individual cases. We can use this to instill greater trust from clinicians in our model to support its use.

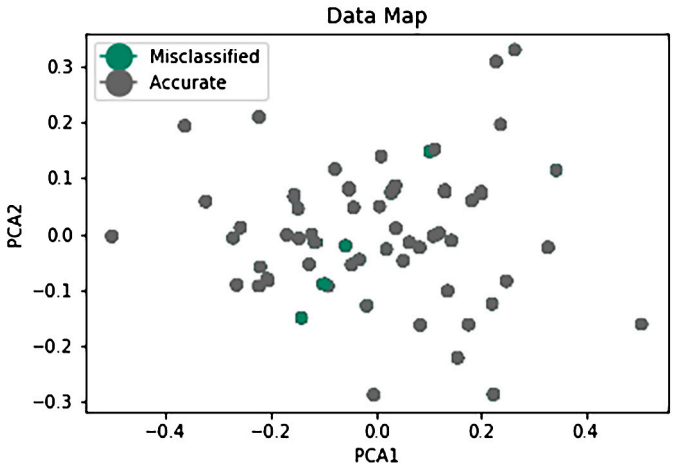

Fig. 5. Plot of Principal Component 2 (PCA2) versus Principal Component 1 (PCA1).

3.4. Model limitations

To the best of our knowledge, we are the first group to use a DNN with fMRI connectivity features for the classification of rTMS response. By using connectivity measures collected before treatment, our network reliably predicts the final treatment outcome. Motivated by Sendak et al. [20], we have included a detailed overview of our model, including potential limitations and the relative contribution of each feature to the model's overall performance. These contributions are an essential step to support the transition from research to clinical use.

Given the high accuracy reported in Section 2.7, it is useful to give some attention to any misclassified examples. By aiming to understand their occurrences, this investigation provides end users - in our case, psychiatrists - with a complete understanding of the model's behaviour.

Amershi et al. [39] present several guidelines for human-AI interaction. These guidelines emphasise the importance of setting clear expectations for the quality and capability of AI systems. Additionally, [39] highlight the importance of making the user aware of situations when an AI system may make mistakes. Formalizing this process, Sendak et al. [20] present model facts, a systematic approach to documenting a ML model designed for clinicians, including advice on interpreting model outputs and warnings. As noted by Sendak et al. [20], warnings regarding the use of an AI model are rarely discussed in the literature. Model limitations must be acknowledged if proposed models are to have an impact in clinical practice.

Exploring the performance of our model, we investigated similarities between commonly misclassified cases. Swayamdipta et al. [40] present a novel methodology for recognizing areas of uncertainty within a large corpus of text. Existing works focus on identifying mislabelled examples within the training data. In contrast, our work focuses on identifying portions of the data that are mislabelled in the test set. Using this novel adaptation, we identified a portion in the lower left-hand quadrant of Fig. 5 that is vulnerable to misclassification. The clustering of these misclassified examples motivated us to explore the hypothesis that these points may share commonalities.

3.4.1. Extreme values are vulnerable to being misclassified

Given the clustering of misclassified examples in Fig. 5, we hypothesise that these values may share some similarities. Identifying these commonalities is an important step in communicating the potential limitations of our model to clinicians. Analysis of the proposed model indicated that after 200 runs, our model accurately predicts treatment response on a held-out test set $\approx 92\%$ of the time. This leaves roughly 8% of cases being misclassified. A variable-wise comparison of distributions by t-test is shown in Table 9. It can be noted from the Table that,

Table 9
Comparison of variable distributions in correctly labelled and mislabelled cases.

Variable	t	p-value
FP	−0.8274	0.4113
OC	1.9119	0.0607
SPL	−0.1517	0.8800
COC	0.5420	0.5899
lOCL	1.2261	0.2250
lOCR	1.2907	0.2018

Table 10
Exploring the differences in OC between examples correctly labelled and mislabelled examples.

		95% Confidence interval	
Group	Mean	Lower bound	Upper bound
Correct	0.0297	−0.0044	0.06372
Misclassified	−0.0313	−0.0749	0.01220

although marginally short of the level of significance, there is some difference between groups in the OC variable.

Through further analysis of group differences as shown in Table 10, we can see the differences between values that were accurately classified and misclassified. Misclassified values had, on average, lower connectivity measures in the OC variable compared to the correctly classified values.

Rerunning an analysis of our model with the removal of the OC variable shows a drop in performance over 200 bootstrap resamples. Highlighting OC is valuable in discerning between classes, however, it does have some observed failure cases. These are important considerations, given each rejected positive case is a patient who may be denied access to treatment when they may actually benefit from it, or conversely, a patient who commits time to receive treatment and sees no benefit. As such, we include the limitation or warning of misclassifications in our model between the 95% confidence interval of −0.0749 to 0.0122. These model limitations can be communicated to end users.

4. Discussions

The current work demonstrates that a feedforward DNN model can accurately predict the treatment outcome of rTMS before treatment. With rigorous internal validation, our work shows a DNN using fMRI connectivity features outperforms existing baseline methods. In our experiments, the performance of prominent ML algorithms like XGBoost and random forests was disappointing. It may be noted that tree-based algorithms like XGBoost have under-performed when the number of samples is less than 500 [14]. Furthermore, the baseline LSVM reproduces the findings of Hopman et al. [11], offering additional support for the use of fMRI features and their ability to predict rTMS treatment outcomes. These findings further emphasize the potential of fMRI connectivity measures as biomarkers for response to rTMS treatment. Furthermore, the current work reiterates that demographic and psychometric variables alone are insufficient to identify patterns of response to rTMS treatment. Even when using sophisticated algorithms, psychometric variables could not improve on the existing rule-based methods proposed by Feffer et al. [6]. Using these psychometric variables, a DL model was unable to identify the association between early change in depression severity and treatment outcome, similar to Feffer et al. [6].

Our work utilizes high levels of internal validation to ensure robust results in an important setting: the psychiatric care of those suffering from depression. Along with this validation, we demonstrated the significant impact of regularisation on model performance to reduce the risks of overfitting. These initial findings will become increasingly significant as larger rTMS datasets become available to further explore the potential of verifying these results against independent datasets. Our results highlight the benefits of using DNNs with several hidden layers compared against shallow ML methods in modelling complex relationships. The proposed architecture outperforms other shallow methods in terms of F1 score, balanced accuracy, and accuracy. This superior predictive performance may be due to the ability of DL algorithms to model complex multi-variable relationships. Shallow methods, such as traditional linear algorithms, are unable to recognize these complicated relationships. In practice, the interplay between treatment, psychiatrists and patient variables is more complex than can be modelled using linear models. The proposed model here consists of 1191 parameters, highlighting the complexity of the model compared to shallow methods.

One thing to note is that the current work has been developed using a limited number of complete records. In the future, we plan to impute the missing records to increase the size of the data. Furthermore, participants for whom data is incomplete may have left the study due to a lack of improvement in their psychological health, leaving only patients who benefited. The risk, then, is that the remaining sample is not truly representative of the true population of patients receiving rTMS treatment. Also, there is a possibility that the model may be overfitting to the current distribution of patients. Further work involving data collected from multiple centres could help to improve the robustness of this model. While the current work is completed using data where classes are relatively balanced, it is not known how the current method would perform if training data was imbalanced. Future work could attempt to incorporate methods that are robust to uneven class distributions of the target variable.

5. Conclusions

In this work, we have proposed a novel DL architecture to predict the outcome of rTMS treatment using fMRI connectivity features. To the best of our knowledge, we are the first to apply both a DNN and combine a DNN with XAI to rTMS response prediction using fMRI connectivity features. Through empirical experiments, we showed that a DNN using fMRI connectivity measures outperforms existing state-of-the-art algorithms. In our repeated bootstrap simulations, we demonstrate our model finds the optimal solution in an unseen test set more frequently than other methods. The demonstrated robustness of this model moves the field closer to clinical implementation over existing shallow methods. Furthermore, our work demonstrates neuroimaging variables are superior to psychometric variables in predicting treatment response to rTMS. Additionally, using XAI techniques, our work shows functional connectivity measures between the Subgenual Anterior Cingulate Cortex and Centeral Opercular Cortex to be a key determinant for rTMS treatment response. These findings are validated using both SHAP values and relative feature importance. The current work improves upon existing methods by including XAI and predicting treatment outcomes before the start of treatment. However, the main limitation of this work is that only a small dataset has been used to develop and test the model. In the future, we plan to use larger datasets from various centres and ethnicities to improve the accuracy of our work.

Statement of ethical approval

Ethical approval for this project was granted by the Universities Human Research Ethics Committee (H21REA026).

Funding

This work is partially funded by The Cannan Institute, Belmont Private Hospital, Brisbane. The authors declare no competing interests.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We gratefully acknowledge the support from Belmont Private Hospital team members, especially, Ms Mary Williams (CEO), Rachel Stark (Area Manager), Dr Mark Spelman (Psychiatrist), Dr Sean Gills (Psychiatrist), and Dr Tom Moore (Psychiatrist). Without their kind support, this work wouldn't be possible.

Appendix A

A.1. Baseline model hyperparameters

Table A.11
Baseline models and their hyperparameters.

Model	Hyperparameter	Value
Linear SVM	C	100
	Gamma	1
	Kernel	Linear
KNN	Distance Metric	Manhattan
	k	4
	Weight	Uniform
XGBoost	colsample_bytree	0.7
	learning_rate	0.01
	max_depth	3
	n_estimators	200
	subsample	1
Random Forest	bootstrap	True
	max_depth	10
	max_features	auto
	min_samples_leaf	1
	min_samples_split	2
	n_estimators	50
Logistic Regression	C	5.4287
	Penalty	L1

References

[1] D. Schofield, M. Cunich, R. Shrestha, R. Tanton, L. Veerman, S. Kelly, M. Passey, Indirect costs of depression and other mental and behavioural disorders for Australia from 2015 to 2030, BJPsych Open 5 (3) (May 2019).

[2] P.B. Fitzgerald, K.E. Hoy, J. Reynolds, A. Singh, R. Gunewardene, C. Slack, S. Ibrahim, Z.J. Daskalakis, A pragmatic randomized controlled trial exploring the relationship between pulse number and response to repetitive transcranial magnetic stimulation treatment in depression, Brain Stimul. 13 (1) (Jan. 2020) 145–152.

[3] C.A. Conelea, N.S. Philip, A.G. Yip, J.L. Barnes, M.J. Niedzwiecki, B.D. Greenberg, A.R. Tyrka, L.L. Carpenter, Transcranial magnetic stimulation for treatment-resistant depression: naturalistic treatment outcomes for younger versus older patients, J. Affect. Disord. 217 (Aug. 2017) 42–47.

[4] C.L. Hovington, A. McGirr, M. Lepage, M.T. Berlim, Repetitive transcranial magnetic stimulation (rTMS) for treating major depression and schizophrenia: a systematic review of recent meta-analyses, Ann. Med. 45 (4) (May 2013) 308–321.

[5] P.B. Fitzgerald, K.E. Hoy, R.J. Anderson, Z.J. Daskalakis, A study of the pattern of response to rTMS treatment in depression, Depress. Anxiety 33 (8) (Apr. 2016) 746–753.

[6] K. Feffer, H.H. Lee, F. Mansouri, P. Giacobbe, F. Vila-Rodriguez, S.H. Kennedy, Z.J. Daskalakis, D.M. Blumberger, J. Downar, Early symptom improvement at 10 sessions as a predictor of rTMS treatment outcome in major depression, Brain Stimul. 11 (1) (Jan. 2018) 181–189.

[7] M. Squires, X. Tao, S. Elangovan, R. Gururajan, X. Zhou, U.R. Acharya, A Novel Genetic Algorithm Based System for the Scheduling of Medical Treatments, Expert Systems with Applications, vol. 195, Elsevier BV, Jun. 2022, p. 116464.

[8] O.A. Montesinos López, A. Montesinos López, J. Crossa, Fundamentals of artificial neural networks and deep learning, in: Multivariate Statistical Machine Learning Methods for Genomic Prediction, Springer International Publishing, 2022, pp. 379–425.

[9] A.T. Drysdale, L. Grosenick, J. Downar, K. Dunlop, F. Mansouri, Y. Meng, R.N. Fetcho, B. Zebley, D.J. Oathes, A. Etkin, A.F. Schatzberg, K. Sudheimer, J. Keller, H.S. Mayberg, F.M. Gunning, G.S. Alexopoulos, M.D. Fox, A. Pascual-Leone, H.U. Voss, B. Casey, M.J. Dubin, C. Liston, Erratum: resting-state connectivity biomarkers define neurophysiological subtypes of depression, Nat. Med. 23 (2) (Feb. 2017) 264.

[10] N. Koutsouleris, T. Wobrock, B. Guse, B. Langguth, M. Landgrebe, P. Eichhammer, E. Frank, J. Cordes, W. Wölwer, F. Musso, G. Winterer, W. Gaebel, G. Hajak, C. Ohmann, P.E. Verde, M. Rietschel, R. Ahmed, W.G. Honer, D. Dwyer, F. Ghaseminejad, P. Dechent, B. Malchow, P.M. Kreuzer, T.B. Poeppl, T. Schneider-Axmann, P. Falkai, A. Hasan, Predicting response to repetitive transcranial magnetic stimulation in patients with schizophrenia using structural magnetic resonance imaging: a multisite machine learning analysis, Schizophr. Bull. 44 (5) (Aug. 2017) 1021–1034.

[11] H. Hopman, S. Chan, W. Chu, H. Lu, C.-Y. Tse, S. Chau, L. Lam, A. Mak, S. Neggers, Personalized prediction of transcranial magnetic stimulation clinical response in patients with treatment-refractory depression using neuroimaging biomarkers and machine learning, J. Affect. Disord. 290 (Jul. 2021) 261–271.

[12] N. Bailey, K. Hoy, N. Rogasch, R. Thomson, S. McQueen, D. Elliot, C. Sullivan, B. Fulcher, Z. Daskalakis, P. Fitzgerald, Differentiating responders and non-responders to rTMS treatment for depression after one week using resting EEG connectivity measures, J. Affect. Disord. 242 (Jan. 2019) 68–79.

[13] J. Fan, I.F. Tso, D.F. Maixner, T. Abagis, L. Hernandez-Garcia, S.F. Taylor, Segregation of salience network predicts treatment response of depression to repetitive transcranial magnetic stimulation, NeuroImage Clin. 22 (2019) 101719.

[14] M. Zou, W-G. Jiang, Q-H. Qin, Y-C. Liu, M-L. Li, Optimized XGBoost model with small dataset for predicting relative density of Ti-6Al-4V parts manufactured by selective laser melting, Materials, MDPI AG 15 (15) (Aug. 2022) 5298, https://doi.org/10.3390/ma15155298.

[15] F. Hasanzadeh, M. Mohebbi, R. Rostami, Prediction of rTMS treatment response in major depressive disorder using machine learning techniques and nonlinear features of EEG signal, J. Affect. Disord. 256 (Sep. 2019) 132–142.

[16] A. Zandvakili, N.S. Philip, S.R. Jones, A.R. Tyrka, B.D. Greenberg, L.L. Carpenter, Use of machine learning in predicting clinical response to transcranial magnetic stimulation in comorbid posttraumatic stress disorder and major depression: a resting state electroencephalography study, J. Affect. Disord. 252 (Jun. 2019) 47–54.

[17] N. Bailey, K. Hoy, N. Rogasch, R. Thomson, S. McQueen, D. Elliot, C. Sullivan, B. Fulcher, Z. Daskalakis, P. Fitzgerald, Responders to rTMS for depression show increased fronto-midline theta and theta connectivity compared to non-responders, Brain Stimul. 11 (1) (Jan. 2018) 190–203.

[18] T.T. Erguzel, S. Ozekes, S. Gultekin, N. Tarhan, G.H. Sayar, A. Bayram, Neural network based response prediction of rTMS in major depressive disorder using QEEG cordance, Psychiatry Investig. 12 (1) (2015) 61.

[19] H.W. Loh, C.P. Ooi, S. Seoni, P.D. Barua, F. Molinari, U.R. Acharya, Application of Explainable Artificial Intelligence for Healthcare: A Systematic Review of the Last Decade (2011–2022), Computer Methods and Programs in Biomedicine, vol. 226, Elsevier BV, Nov. 2022, p. 107161.

[20] M.P. Sendak, M. Gao, N. Brajer, S. Balu, Presenting machine learning model information to clinical end users with model facts labels, npj Digit. Med. 3 (1) (Mar. 2020).

[21] E. Ebrahimzadeh, F. Fayaz, L. Rajabion, M. Seraji, F. Aflaki, A. Hammoud, Z. Taghizadeh, M. Asgarinejad, H. Soltanian-Zadeh, Machine Learning Approaches and Non-linear Processing of Extracted Components in Frontal Region to Predict rTMS Treatment Response in Major Depressive Disorder, Frontiers in Systems Neuroscience, vol. 17, Frontiers Media SA, Mar. 2023.

[22] M.S. Shahabi, A. Shalbaf, R. Rostami, R. Kazemi, A convolutional recurrent neural network with attention for response prediction to repetitive transcranial magnetic stimulation in major depressive disorder, Sci. Rep. 13 (1) (Jun. 2023), https://doi.org/10.1038/s41598-023-35545-2, Springer Science and Business Media LLC.

[23] I.N. Beaufort, G.H. De Weert-Van Oene, V.A.J. Buwalda, J.R.J. De Leeuw, A.E. Goudriaan, The depression, anxiety and stress scale (DASS-21) as a screener for depression in substance use disorder inpatients: a pilot study, Eur. Addict. Res. 23 (5) (2017) 260–268, https://doi.org/10.1159/000485182, S. Karger AG.

[24] F. Chollet, Deep Learning with Python, second edition, Manning Publ., Dec. 2021, [Online]. Available: https://www.ebook.de/de/product/40499536/francois_chollet_deep_learning_with_python_second_edition.html.

[25] Y. han Sheu, Illuminating the black box: interpreting deep neural network models for psychiatric research, Front. Psychiatry 11 (Oct. 2020).

[26] S. Itani, M. Rossignol, At the crossroads between psychiatry and machine learning: insights into paradigms and challenges for clinical applicability, Front. Psychiatry 11 (Sep. 2020).

[27] P.B. Fitzgerald, M.S. George, S. Pridmore, The evidence is in: repetitive transcranial magnetic stimulation is an effective, safe and well-tolerated treatment for patients with major depressive disorder, Aust. N.Z. J. Psychiatry (Aug. 2021) 000486742110430.

[28] P.M. Doraiswamy, C. Blease, K. Bodner, Artificial intelligence and the future of psychiatry: insights from a global physician survey, Artif. Intell. Med. 102 (Jan. 2020) 101753.

[29] H. Hopman, S. Chan, W. Chu, H. Lu, C-Y. Tse, S. Chau, L. Lam, A. Mak, S. Neggers, Personalized prediction of repetitive transcranial magnetic stimulation clinical response in medication-refractory depression data, Data Brief 37 (Aug. 2021) 107264, https://doi.org/10.1016/j.dib.2021.107264, Elsevier BVc.

[30] S. Lovibond, P.F. Lovibond, Manual for the Depression Anxiety Stress Scales, 2nd ed., Psychology Foundation, Sydney, 1995.

[31] F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain, Psychol. Rev. 65 (6) (1958) 386–408.

[32] H. Ramchoun, M. Amine, J. Idrissi, Y. Ghanou, M. Ettaouil, Multilayer perceptron: architecture optimization and training, Int. J. Interact. Multimed. Artif. Intell. 4 (1) (2016) 26.
[33] E. Aldana-Bobadila, A. Kuri-Morales, I. Lopez-Arevalo, A.B. Rios-Alvarado, An unsupervised learning approach for multilayer perceptron networks, Soft Comput. 23 (21) (Nov. 2018) 001.
[34] X. Ying, An overview of overfitting and its solutions, J. Phys. Conf. Ser. 1168 (Feb. 2019) 022022.
[35] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (Jan. 2014) 1929–1958.
[36] C. Birkenbihl, M.A. Emon, H. Vrooman, S. Westwood, S. Lovestone, M. Hofmann-Apitius, H. F., Differences in cohort study data affect external validation of artificial intelligence models for predictive diagnostics of dementia - lessons for translation into clinical practice, EPMA J. 11 (3) (Jun. 2020) 367–376.
[37] F.E. Harrell Jr, Regression Modeling Strategies: with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, Springer, 2015.
[38] B. Chang, Y. Choi, M. Jeon, J. Lee, K.-M. Han, A. Kim, B.-J. Ham, J. Kang, ARPNet: antidepressant response prediction network for major depressive disorder, Genes 10 (11) (Nov. 2019) 907.
[39] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P.N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, E. Horvitz, Guidelines for human-AI interaction, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, ACM, May 2019.
[40] S. Swayamdipta, R. Schwartz, N. Lourie, Y. Wang, H. Hajishirzi, N.A. Smith, Y. Choi, Dataset cartography: mapping and diagnosing datasets with training dynamics, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2020, pp. 9275–9293, Online. https://doi.org/10.18653/v1/2020.emnlp-main.746.
[41] E. Tjoa, C. Guan, A survey on explainable artificial intelligence (XAI): toward medical XAI, IEEE Trans. Neural Netw. Learn. Syst. 32 (11) (Nov. 2021) 4793–4813.
[42] L. Gianfagna, A.D. Cecco, Explainable AI with Python, Springer International Publishing, 2021.
[43] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 4768–4777.
[44] L.S. Shapley, A value for n-person games, in: Contributions to the Theory of Games (AM-28), Volume II, Princeton University Press, Dec. 1953, pp. 307–318.