ETH zürich

Universität Zürich[UZH]

Translational Neuromodeling Unit

227-0973-00L: Translational Neuromodeling

Dynamic Causal Modelling of Resting-State fMRI to Predict Ketamine Response in Major Depressive Disorder

Keyshav Mor, Rico-Marcel Benning, Leon Schönleber

keymor@ethz.ch, rbenning@ethz.ch, scleon@ethz.ch

Supervising Professor: Prof. Klaas Stephan
Advisor: Herman Galioulline

June 1, 2025

***Abstract*—Rapid-acting antidepressants such as ketamine can provide symptom relief within hours, yet only a subset of patients benefit. Identifying responders from a single pre-treatment scan could spare non-responders an invasive, side-effect-prone intervention and personalised care. We therefore tested whether effective-connectivity patterns extracted with Dynamic Causal Modelling (DCM) from a baseline resting-state fMRI scan predict both response status and the magnitude of symptom change on the Montgomery–Åsberg Depression Rating Scale (MADRS) two days after ketamine infusion. We analysed the open-access NIMH "Ketamine Mechanism of Action" dataset comprising 33 patients with major depressive disorder and 22 healthy controls; after quality control, 26 patients with complete baseline and day-2 data remained, and response was defined as at least a 25 % MADRS reduction. Spectral DCM was inverted for three network granularities. A four-node default-mode network (DMN), an 11-node DMN plus attention network, and a 15-node DMN–attention–salience network. The resulting A-matrices served as generative embeddings for logistic, multinomial and random-forest classifiers as well as an elastic-net regressor evaluated under three-, four- and five-fold cross-validation. The four-node DMN embedding paired with logistic regression achieved the best binary classification (mean accuracy $= 73\,\%$, $F_1 = 0.71$, macro-recall $= 0.82$), whereas larger network models did not significantly improve performance ($\approx 65\,\%$ accuracy). Multilabel classifications were unreliable because of class imbalance, and prediction of continuous MADRS change showed high errors (MAE $\approx$ 8–10, RMSE $\approx$ 10–12). These results indicate that a parsimonious generative embedding derived from a simple DMN model contains sufficient information for modest prediction of who will benefit from a single ketamine infusion, while highlighting the need for larger, balanced cohorts to achieve clinically useful accuracy and enable finer-grained response stratification.**

I. Motivation

Major Depressive Disorder (MDD) is a severe and highly prevalent medical disease, with a significant negative impact on productivity and quality of life. [1]Effectively treating remains a challenge. Common antidepressants such as selective serotonin reuptake inhibitors (SSRI), serotonin and norepinephrine reuptake inhibitors (SNRI) and tricyclic antidepressants (TAC) show a response rate of 50% to 70% [2] [3]. It usually takes two to four weeks before they can show their effect [1] and non-response can be determined. Before administering an antidepressant, there is currently no reliable way to estimate whether treatment will have an effect or not. Finding the right medication for a patient can therefore be a long iterative process.

The Ketamine Mechanism of Action Study [4] attempts to study the effect of Ketamine infusion in depressed and healthy subjects in alleviating depression after a few days of the infusion. The study utilised the fMRI scans of the subjects to analyse the different brain regions and subject survey responses to ascertain the effect of Ketamine infusion on the depression levels in these subjects. Taking cue from this study, this project constructs DCMs for the fMRI data of the depressed subjects in the study's dataset and to use the effective connectivity matrix from the different DCMS constructed for each subject as generative embeddings to predict Ketamine treatment response on the subjects' depression levels as well as to identify treatment responders from nonresponders in a granular and coarse fashion. Traditional machine learning methods like logistic regression [5], multinomial regression [6], random forest classifier [7] and elastic net regression [8] are employed for analysing the predictive power of these generative embeddings.

A. Ketamine

Ketamine, typically known as a dissociative anesthetic, has recently gained attention for its potent, rapid-onset (within hours) and sustained (up to a week) antidepressant effects [9]. A recent meta-analysis that included 182 participants found a treatment response rate of 54% for ketamine one or two days after administration. With this relatively small body of evidence, Ketamine's antidepressant efficacy appears comparable to SSRIs, while its antidepressant effect occurs much faster [10]. Ketamine also is a drug of abuse with serious side effects such as sedation, induction of trance-like states, hallucinations and amnesia [11]. The combination of fast action and unwanted side effects makes it especially useful to estimate ketamine's treatment efficacy prior to use. Patients with a severe depressive episode can have their symptoms alleviated swiftly, if they can be identified as a responder. Predicting responsiveness to ketamine treatment is thus the goal of this project.

B. Heterogeneity of Depression

Depression is a multifaceted disorder characterized by significant heterogeneity. This heterogeneity is theorized to be the root cause for varying treatment responses and the need for individualized treatments. According to the fifth edition Diagnostic and Statistical Manual of Mental Disorders (DSM-5), five or more of nine symptoms must be present during a two-week period, for someone to be diagnosed with Major Depressive Disorder (MDD). This yields a total of 256 possible combinations of symptoms for the same diagnosis. This heterogeneity can be addressed using neuroimaging. Recent studies have proposed different depression biotypes based on functional connectivity patterns across brain circuits [12] [13]. These biotypes differ in functional connectivity patterns and treatment response rates. These findings indicated that treatment efficacies can be estimated from pre-treatment fMRI data. Building on this premise, this project uses Dynamic Causal Modeling (DCM) to estimate directed (effective) connectivity between brain regions prior to treatment. DCM is a Bayesian framework that models causal interactions among neuronal populations based on observed fMRI signals. We construct DCMs of brain circuits implicated in depression and ketamine's mode of action. The resulting connectivity estimates are used as generative embeddings or features for predicting individual treatment outcomes. The effective connectivity estimates obtained from the DCM models are used for classification and regression algorithms to predict treatment response.

II. The Dataset

The dataset comprises 58 participants (33 depressed (MDD), 3 Biploar (BP), and 22 healthy controls (HC)) enrolled in a double-blind, placebo-controlled, cross-over ketamine study at NIMH (IRP ID: 04-M-0222) [4]. All MDD subjects were medication-free for at least two weeks, with a baseline MADRS score ≥ 20. Participants were randomized to receive a single IV infusion of ketamine or saline. After two follow-up visits (Day 2 and Day 10), they crossed over to the alternate infusion and were re-assessed at the same intervals. Each subject was scheduled for up to five MRI sessions:

- Baseline ($\approx$ 2 days before first infusion)
- Post-infusion 1 (2 days after first infusion)
- Interim 1 (10 days after first infusion)
- Post-infusion 2 (2 days after second infusion/crossover)
- Interim 2 (10 days after second infusion/crossover)

Each session included a resting-state, task-free fMRI (8 min, eyes closed) and an anatomical T1-weighted scan, acquired on a 3 T MRI scanner. Diffusion weighted imaging was performed at the first assessment. Resting-state fMRI always included physiological recordings for most participants. Some subjects did not complete every planned scan (e.g., early dropout, technical issues, or MRI refusal), so only participants with complete resting-state runs and behavioral assessments are included in analyses. Behavioral/clinical measures (MADRS, HAMD-6, HAMD-17) were collected at each MRI visit. Although three BP subjects were enrolled, this report focuses on the 33 MDD and 22 HC participants. Only participants who completed all required scans and rating scales at each timepoint (Baseline, D2, D10, post-crossover D2, and post-crossover D10) were included in the final statistical analyses. The total is 26 subjects with MDD whose fMRI scans were used to obtain DCMs and their generative embeddings for the predictive tasks.

III. Dynamic Causal Models

When performing DCM analysis of task-free data, the classical DCM [14] formulation is augmented with a stochastic term to account for the fluctuation of endogenous activity. Spectral DCM [15], a variant of DCM that models cross-spectral densities, was chosen for its improved computational efficiency. Dynamic Causal Modeling requires prior assumptions about which brain regions interact to generate the observed signal. The brain regions were selected based on previous findings that implicate them in depression or indicate that they are altered by ketamine. In total, four fully connected DCMs were constructed.

- 4-node DCM including the default mode network (DMN)
- 11-node DCM including the DMN and the Attention Network (AN).
- 15-node DCM including DMN, AN and the Salience network(SN)
- 21-node DCM including DMN, AN, SN and regions implicated in Ketamine action

The 21-node DCM took too long to invert and was therefore not included in the final analysis.

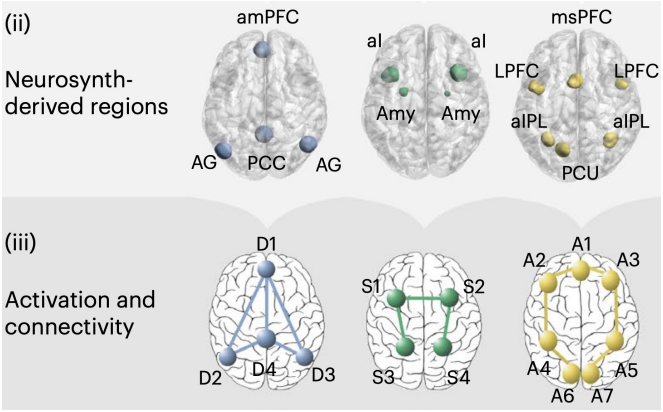

Fig. 1. Default Mode Network, Salience Network and Attention Network (left to right) by Tozzi et al. [12]

A. Regions of Interest

MDD is associated with altered static functional connectivity in the default mode network [16]. The frontostriatal salience network is expanded nearly twofold in the cortex of most individuals with depression [17]. Post-analysis literature research indicates that selective attention impairments are independent from overall depression severity. [18] This would suggest that connectivity in attention circuit is unlikely to predict improvements in depression severity. The triple network model of dysfunction hypothesis posits that depressive symptoms result from increased activity in the DMN, in contrast to reduced activity in the salience network and central executive network [19] [20]. For DCM generation, regions of interest by Tozzi et al. [12], were used. Tozzi et al. [12] used features from task based and task free fMRI data for their clustering analysis and the generation of depression biotypes. The features derived for task free data are derived from the default mode, salience and attention circuit. They used Neurosynth searches with the terms: ”default mode”, ”salience network” and ”frontoparietal attention network” to obtain their regions of interest. These are publically available as NifTi files on their GitHub [12]. For the 20-node DCM additional regions from the Brainnetome Atlas were used.

B. Ketamine informed DCM

For the fourth and biggest DCM, brain regions where ketamine exerts its effects were included. Our literature research indicates that ketamine’s antidepressant mechanism is diverse. Ketamines primary action as a NMDA receptor antagonist causes reduced inhibitory interneuron GABAergic transmisson, a glutmate surge, an AMPA-mediated increase in BDNF release and mTOR-dependent neuroplasticity [11] [10] [9]. Decreased levels of Brain-derived neurotrophic factor (BDNF) are linked to an increased occurence of depressive symptoms [21]. Rodent studies demonstrated that ketamine rapidly increases BDNF release and/or expression in the medial prefrontal cortex (mPFC) and hippocampus [9]. In response to ketamine, in order to retain NMDA receptor

activity Endocannabinoids are released. Augmentation of endocannabinoid signaling is proposed to have an antidepressant effect [22]. A higher frequency of Endocannabinoid Receptor one (CB1) is found throughout brain areas involved with attitude regulation, i.e. hippocampus, amygdala and prefrontal cortex. Ketamine primary action causes depression of certain cortical and thalamic functions and stimulation of parts of the limbic system. Which of these effects is most relevant for the antidepressant effect is a topic of debate [11]. A study on non-human primates found that ketamine administration significantly increases 5-HT1B receptor binding in the nucleus accumbens and ventral pallidum [23]. A study on humans found that actvity in the right caudate was lower in patients with treatment resistant depression compared to healthy controls. After ketamine treatment, greater connectivity of the right caudate was associated with improvement in depression severity [24]. The literature search implicates the following brain regions: mPFC, hippocampus, thalamus, amygdala, insula, nucleus accumbens, ventral palldium, caudate nucleus in ketamine's antidepressant mechanism of action. Caudate nucleus, nucleus accumbens and ventral palldium are all part of the basal ganglia. The amygdala and insula, which are part of the salience network, are already included in the regions in the 15-node DCM, as well as and parts of cortex. The 21-node DCM, in addition to the 15 nodes, includes the thalamus, hippocampus and the Basal Ganglia of each half of the brain.

C. DCM Pipeline

Pre-processing, Timeseries extraction, DCM generation, and inversion were performed in Matlab using the SPM toolbox [25]. After preprocessing, GLM analysis was performed to remove artifacts. FMRI time series were extracted for each region of intrest (ROI). These time series were used in the next step to generate the DCMs. Inversion was much faster for the 4-node DCM at less than a minute, compared to the larger DCMs. The 15-node DCM took 35 minutes, and the 21-node DCM took 5.5 hours for the inversion. The analysis for the 21-node DCM was not completed for time reasons. The inversion yields the estimated connectivity for the nodes of the DCM, the so-called A-matrix.

IV. Predictions using Generative Embeddings

The "A" matrix obtained from these parameter estimates is crucial for the subsequent study, as these A matrices are the generative embeddings that are used to analyze the predictive performance of the DCMs.

A. Objectives

The purpose of analyzing the predictive performance of these generative embeddings is two-fold: to analyze how powerful these embeddings are in distinguishing the different classes into which the depressed subjects of the ketamine study fall and also to analyze how well these generative embeddings perform in predicting the absolute Montgomery-Asberg Depression Rating Scale (MADRS) score among depressed subjects after two days of ketamine infusion, as well as the difference in the MADRS score between the baseline session (b0) and two days after the ketamine infusion (d2).

Keeping these objectives in focus, traditional machine learning methods, such as logistic regression [5], multinomial regression [6], and random forest classification [7], are applied to the generative embeddings for the classification tasks and elastic net regression [8] is applied to the generative embeddings for continuous value prediction tasks. These methods are expanded upon briefly in later sections.

B. Experiment Setup and Methodology

Of the 58 subjects who underwent the study, it was found that 22 subjects were healthy controls (HC), or, in other words, subjects not diagnosed with any degree or form of depression. Since this study focuses primarily on the performance of generative embeddings in classifying different classes of responses among subjects, as well as in predicting absolute MADRS scores and changes in MADRS scores between the baseline session (b0) and two days after ketamine infusion (d2), healthy control subjects are not included in the study. Furthermore, of the clinically diagnosed depressed subjects, only 26 subjects had the full spectrum of information in the data set i.e resting state fMRI scans and MADRS scores of sessions b0 and d2. Thus, the predictive performance of the generative embeddings is analyzed through the 3 DCM networks generated for these 26 subjects from their baseline (b0) session. For simplicity, these 3 DCM network models would be referred to as 4x4 DCM, 11x11 DCM and 15x15 DCM for the rest of this section.

To validate the classification performance of the generative embeddings using different classifiers, four different labeling modes were used. This was done to identify which on classification mode the generative embeddings perform the best in classification tasks. The labeling modes were as follows:

- **Binary:** Subjects' response to ketamine infusion treatment is divided into two categories namely- "Response" and "No Response". Response is the category for subjects who show a greater decrease in their MADRS scores than 25% between session b0 and session d2, whereas subjects with a decrease of less than 25% or an increase greater than 0% are considered nonresponders.
- **Tri-Label:** In this mode, subjects are divided into three categories. Subjects with less than 10% decrease or increase greater than 0% in MADRS scores between sessions b0 and d2 are considered nonresponders, subjects between 11% and 50% decrease in the MADRS score are considered partial responders and subjects with 50% or greater decrease in MADRS scores are considered in remission.
- **Quad Label:** In this mode, subjects are divided into four categories. Subjects with greater than 10% increase in MADRS scores between sessions b0 and d2 are considered deteriorating. Subjects between 10% increase and 10% decrease in MADRS scores between sessions

are considered stable. Subjects with greater than 10% but less than 50% decrease in MADRS scores are considered partial responders while subjects with greater than or equal to 50% decrease in MADRS scores are considered in remission.

- **Five Label:** In this mode, the subjects are divided into five categories. Subjects with greater than 10% increase in MADRS scores between sessions b0 and d2 are considered deteriorating. Subjects between 10% increase and 10% decrease in MADRS scores between sessions are considered stable. Subjects with greater than 10% but less than 50% decrease in MADRS scores are considered mild responders. Subjects with greater than 25% but less than 50% decrease in MADRS scores are considered strong responders while subjects with greater than or equal to 50% decrease in MADRS scores are considered in remission.

In addition to the classification performance, the predictive performance of the generative embeddings of the 3 DCM network models is also tested on continuous value prediction to predict the absolute MADRS scores for session d2 for the 26 subjects as well as for predicting the difference in MADRS scores between sessions b0 and d2. For this task, the ElasticNet regressor is used.

*1) **Cross Validation Strategy for Classification Performance**:* As mentioned earlier, we utilize the logistic, multinomial and random forest classifiers to test the predictive performance of the generative embeddings for all the above mentioned labeling modalities. Thus, for the 3 DCM network models the classification performance is validated for all four labeling modalities in a 3, 4 and 5-fold cross-validation format. Instead of testing just for 5-fold cross validation, 3 and 4-fold cross validation is carried out considering the small dataset of 26 and accounting for potential class imbalance that the classifier might come across.

*2) **Cross Validation Strategy for Regression**:* For continuous value prediction using the ElasticNet regressor applied to the generative embeddings of the 3 DCM network models, 5-fold cross validation is used and results are reported for the predictive performance of each fold.

C. Feature Engineering

The parameters in the A matrix obtained for each subject in these DCM networks capture the effective connectivity strengths among the different nodes of these DCM networks. These parameters function as our generative embeddings and also as features for the classification and regression models. The DMN DCM with 4x4 connectivity matrix results in 16 features with its 12 directed edges and 4 self-connections. Similarly, the 11x11 DCM results in 121 features with 110 directed edges and 11 self-connections. The 15x15 DCM results in 225 features with 210 directed edges and 15 self-connections.

The features of the respective models are fed into different classification models for cross-validation together with the relevant labeling modes, as explained earlier. For cross-validation on the ElasticNet regression performance of the generative embeddings, the feature values of the training and test set in each fold are normalized before being fed into the regression model. Naturally, the test set features are normalized with the mean and standard deviation values of the train set features to avoid data leakage issues. The normalization of features is not performed for classification tasks.

The b0 MADRS scores are not used in evaluating the performance of the generative embeddings. The ground truth target labels are obtained from the ketamine mechanism action study [4] with the baseline (b0) and post-infusion (d2) MADRS scores being used for hand-engineering classification labels mentioned in the four labeling modes. Additionally, the difference in MADRS scores between sessions b0 and d2 is obtained from the same dataset. These values and labels can found at the following GitLab link: Ketamine Study MADRS Profile on GitLab

D. Programming Language and Libraries

For the classification and regression tasks, Julia programming language [26] was used. Julia has strong library support for out-of-box machine learning tasks and given the time sensitivty of the project, this language was opted for. The MLJ library [27] has support for logistic, multinomial and random forest classifiers as well as for ElasticNet regression. Additionally, through the "evaluate" method available in MLJ library, automatic cross-validation for various models is made possible via automated splitting of training and test sets in each fold, fitting of the model and evaluation of the performance metrics. The MLJ library also provides pipeline support for normalization of the feature set with the "Standardizer" [27] method and this method was utilised for normalization of training and test features in each fold during the cross-validation of ElasticNet regression on different generative embeddings. MLJ inherits some models from the SciKitLearn Python library via the MLJSciKitLearnInterface library [28]. Some of the models that were implemented were inherited via this library. For cross-validation specifications, the "evaluate" method in the MLJ library splits the features and ground-truth labels in appropriate proportions depending on the number of folds specified. Thus, 3, 4 and 5-fold cross-validation splits the training and test sets during each fold in the ratio of 66.66%-33.33%, 75%-25% and 80%-20% respectively.

E. Machine Learning Models Overview

*1) **Logistic Regression**:* Logistic Regression or Logistic classification models is primarily used to model the probability of a binary outcome using the sigmoid function:

$$P(y = 1|\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b) = \frac{1}{1 + e^{-(\mathbf{w}^\top \mathbf{x} + b)}}$$

The model is trained by minimizing the binary cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

LogisticClassifier from MLJLinearModels in the MLJ library, supports multi-class classification via softmax internally. Thus, no manual adjustment is needed for handling multi-class problems, and it was used for this project without any modification or hyperparameter tuning.

*2) **Multinomial Logistic Regression**:* Multinomial logistic regression or classification extends logistic regression to multi-class problems:

$$P(y = k|\mathbf{x}) = \frac{e^{\mathbf{w}_k^\top \mathbf{x}+b_k}}{\sum_{j=1}^{K} e^{\mathbf{w}_j^\top \mathbf{x}+b_j}}$$

The model is trained by minimizing cateogorical cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\log(\hat{y}_{ik})$$

*3) **Random Forest Classifier**:* Random Forest is a method that uses an ensemble of decision trees to obtain majority voting for classification:

$$\hat{y} = \text{mode}(T_1(\mathbf{x}), T_2(\mathbf{x}), \ldots, T_M(\mathbf{x}))$$

Trees split nodes to minimize impurity which is commonly measured by Gini impurity:

$$G = 1 - \sum_{k=1}^{K} p_k^2$$

or entropy:

$$H = -\sum_{k=1}^{K} p_k \log(p_k)$$

*4) **Elastic Net Regression**:* Elastic Net is a linear regression model useful in continuous value prediction. It combines L1 (Lasso) and L2 (Ridge) regularization to avoid overfitting while maintaining learning stability:

$$\hat{y} = \mathbf{w}^\top \mathbf{x} + b$$

The objective function to minimize is:

$$\mathcal{L} = \frac{1}{2N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda_1\|\mathbf{w}\|_1 + \lambda_2\|\mathbf{w}\|_2^2$$

Here, $\|\mathbf{w}\|_1$ is the L1 norm (sparsity), and $\|\mathbf{w}\|_2^2$ is the squared L2 norm (shrinkage). Elastic Net is useful when features are highly correlated or when feature selection is desired. This is applicable to the case of validating predictive performance of generative embeddings since the parameters in generative embeddings are highly correlated.

F. Code Repository

The classification and regression notebook implementations and corresponding result plots and tables can be found at the following GitLab repository: Classification and Regression.

V. Results of Predictions using Generative Embeddings

Since rigorous validation of the predictive modeling performance of generative embeddings was carried out with the Logistic, Multinomial, Random Forest and ElasticNet models using multi-fold cross validation on the 4x4, 11x11 and 15x15 DCM network models, a large body of results was obtained. However, in the interest of brevity of the report and keeping it under 10 pages, only the most important results are presented in this section.

All the plots of the results can be viewed here: Result Plots. All the tables of the metrics can be viewed here: Performance Metric Tables

A. Model Evaluation Metrics

For classification tasks, the metrics used to compare the performance of different models, cross-validations, and labeling modes are: **Accuracy**, **F1 Score**, **Macro Precision**, and **Macro Recall**.

- **Accuracy**:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

- **F1 Score**:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Macro Precision** and **Macro Recall** are computed by averaging precision and recall across all classes, weighting them equally:

$$\text{Macro Precision} = \frac{1}{K}\sum_{k=1}^{K}\text{Precision}_k$$

$$\text{Macro Recall} = \frac{1}{K}\sum_{k=1}^{K}\text{Recall}_k$$

- **Precision**:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- **Recall**:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

For continuous value prediction tasks with the Elastic Net regressor, the metrics used are **Mean Absolute Error (MAE)** and **Root Mean Squared Error (RMSE)**:

- **Mean Absolute Error (MAE)**:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

- **Root Mean Squared Error (RMSE)**:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

Note: Sometimes the Macro Precision and Macro Recall values in our results are reported. These **NaN** values occur in

the final average precision and recall because of NaN values present in the folds which in turn occur due to class label imbalances that happen due to more labels per subject on a small dataset of 26 subjects. Thus, the results having NaN values in their precision and recall are not completely reliable even with high accuracy and F1 scores because even the F1 scores cannot be considered reliable. These unreliable scores are also reported here to ensure consistency across different results that are presented.

B. *Results for 4x4 DCM*

Fold	Accuracy	F1 Score	Macro Precision	Macro Recall
1	0.83	0.83	0.83	0.875
2	0.60	0.58	0.67	0.75
3	0.80	0.80	0.83	0.83
4	0.60	0.58	0.67	0.75
5	0.80	0.76	0.75	0.875
Average	**0.73**	**0.71**	**0.75**	**0.82**

TABLE I
4X4 DCM BINARY CLASSIFICATION METRICS ACROSS CROSS-VALIDATION FOLDS

*1) **Binary Classification**:* Upon verifying different results from the 3, 4 and 5-fold cross validation of the binary label case, it was found that the Logistic regression classifier performed the best for a 5-fold cross-validation case with an average accuracy of 73%, F1 score of 71%, average macro precision of 75% and average macro recall of 82%. The results are displayed in table I.

Fold	Accuracy	F1 Score	Macro Precision	Macro Recall
1	0.33	NaN	0.22	NaN
2	0.80	0.82	0.83	0.89
3	0.60	0.67	0.67	0.67
4	0.40	NaN	0.33	NaN
5	0.60	0.61	0.67	0.67
Average	**0.55**	**NaN**	**0.54**	**NaN**

TABLE II
4X4 DCM THREE LABEL CLASSIFICATION METRICS ACROSS FOLDS

*2) **Tri Label Classification**:* Upon verifying different results from the 3, 4 and 5-fold cross validation of the three label case, it was found that the Logistic and multinomial regression classifiers performed the best for a 5-fold cross-validation case with an average accuracy of 55%, F1 score as NaN average macro precision of 54% and average macro recall of **NaN**. The results are displayed in the table II. Both classifiers reported the same metrics.

Fold	Accuracy	F1 Score	Macro Precision	Macro Recall
1	0.285	NaN	NaN	NaN
2	0.285	NaN	NaN	NaN
3	0.670	NaN	0.625	NaN
4	0.330	NaN	0.375	NaN
Average	**0.39**	**NaN**	**NaN**	**NaN**

TABLE III
4X4 DCM FOUR LABEL CLASSIFICATION METRICS ACROSS FOLDS

*3) **Four Label Classification**:* Upon verifying different results from the 3, 4 and 5-fold cross validation of the four label case, it was found that the Logistic and multinomial regression classifiers performed the best for a 4-fold cross-validation case with an average accuracy of 39%, F1 score as NaN, average macro precision and average macro recall of **NaN**. Thus, these results are not entirely reliable. The results are displayed in table III. Both classifiers reported the same metrics.

Fold	Accuracy	F1 Score	Macro Precision	Macro Recall
1	0.33	NaN	0.23	NaN
2	0.44	NaN	NaN	0.50
3	0.625	NaN	0.60	NaN
Average	**0.47**	**NaN**	**NaN**	**NaN**

TABLE IV
4X4 DCM FIVE LABEL CLASSIFICATION METRICS ACROSS FOLDS

*4) **Five Label Classification**:* Upon verifying different results from the 3, 4 and 5-fold cross validation of the five label case, it was found that the random forest classifier performed the best for a 3-fold cross-validation case with an average accuracy of 47%, F1 score as NaN, average macro precision and average macro recall of **NaN**. Thus, these results are not entirely reliable. The results are displayed in table IV.

From the tables I, II, III, IV, it is evident that the logistic regression classifier for binary classification on the 4x4 DCM performs the best while other classifiers and other labeling modalities do not perform so well on the classification tasks but still perform better than random chance prediction. However, the classification modalities other than binary classification are also less reliable due to class imbalances and NaN values in key performance metrics.

*5) **ElasticNet Regression**:* 5-fold cross validation was carried out on the 4x4 DCM generative embeddings for absolute MADRS values and differences in the values between b0 and d2 sessions. The results are displayed in figures 2 and 3. From the plots, it is evident that the continuous value prediction errors for generative embeddings of 4x4 DCM using ElasticNet regression are very high and the performance is not as good as classification tasks.

Fig. 2. 4x4 DCM 5-fold Cross-Validation Metrics on Predicting Absolute MADRS Values

C. *Results for 11x11 DCM*

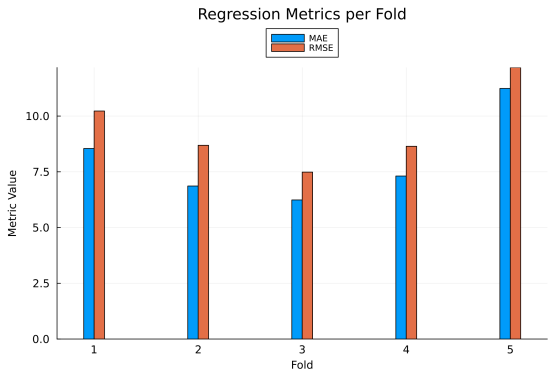

Fig. 3. 4x4 DCM 5-fold Cross-Validation Metrics on Predicting Delta MADRS Values

Fold	Accuracy	F1 Score	Macro Precision	Macro Recall
1	0.67	0.625	0.67	0.80
2	0.60	0.58	0.58	0.58
3	0.60	0.58	0.58	0.58
4	0.80	0.76	0.75	0.875
5	0.60	0.58	0.67	0.75
Average	**0.65**	**0.63**	**0.65**	**0.72**

TABLE V
11X11 DCM BINARY CLASSIFICATION METRICS ACROSS CROSS-VALIDATION FOLDS

*1) **Binary Classification**:* Upon verifying different results from the 3, 4 and 5-fold cross validation of the binary label case, it was found that the logistic regression classifier performed the best for a 5-fold cross-validation case with an average accuracy of 65%, F1 score of 63%, average macro precision of 65% and average macro recall of 72%. The results are displayed in the table V. The binary classification using generative embeddings from 11x11 DCM is not as good as it is for the 4x4 DCM; however, it is still better than random chance predictions.

Fold	Accuracy	F1 Score	Macro Precision	Macro Recall
1	0.67	0.72	0.77	0.833
2	0.20	NaN	0.17	NaN
3	0.20	NaN	0.17	NaN
4	0.20	NaN	0.17	NaN
5	0.40	NaN	0.50	NaN
Average	**0.33**	**NaN**	**0.355**	**NaN**

TABLE VI
11X11 DCM THREE LABEL CLASSIFICATION METRICS ACROSS FOLDS

*2) **Tri Label Classification**:* Upon verifying different results from the 3, 4 and 5-fold cross validation of the three label case, it was found that the random forest classifier performed the best for a 5-fold cross-validation case with an average accuracy of 33%, F1 score as NaN, average macro precision of 35.5% and average macro recall of **NaN**. Thus, these results are not entirely reliable. The results are displayed in table VI. This is a sub-optimal performance that indicates that the model did not perform better than a random chance predictor.

*3) **Four Label Classification**:* Upon verifying different results from the 3, 4 and 5-fold cross validation of the four label case, it was found that the random forest classifier performed the best for a 4-fold cross-validation case with an average accuracy of 53.5%, F1 score as NaN, average macro precision and average macro recall of **NaN**. Thus, these results are not entirely reliable. The results are displayed in the table VII. The accuracy of the four-label classification task indicates that it performs twice as well as a random chance predictor.

Fold	Accuracy	F1 Score	Macro Precision	Macro Recall
1	0.43	NaN	NaN	NaN
2	0.71	NaN	NaN	NaN
3	0.50	NaN	0.375	NaN
4	0.50	NaN	0.375	NaN
Average	**0.535**	**NaN**	**NaN**	**NaN**

TABLE VII
11X11 DCM FOUR LABEL CLASSIFICATION METRICS ACROSS FOLDS

Fold	Accuracy	F1 Score	Macro Precision	Macro Recall
1	0.43	NaN	NaN	NaN
2	0.86	NaN	NaN	NaN
3	0.17	NaN	0.20	NaN
4	0.33	NaN	0.20	NaN
Average	**0.45**	**NaN**	**NaN**	**NaN**

TABLE VIII
11X11 DCM FIVE LABEL CLASSIFICATION METRICS ACROSS FOLDS

*4) **Five Label Classification**:* Upon verifying different results from the 3, 4 and 5-fold cross validation of the case of five labels, it was found that logistic regression and multinomial classifiers performed the best for a case of 4-fold cross-validation with an average accuracy of 45%, a F1 score as NaN, an average macro precision and an average macro recall of **NaN**. Thus, these results are not entirely reliable. The results are displayed in table VIII. From the accuracy metrics, it can be inferred that the model performs slightly more than twice as well for the classification tasks as a random chance predictor.

*5) **ElasticNet Regression**:* 5-fold cross validation was carried out on the 11x11 DCM generative embeddings for absolute MADRS values and differences in the values between b0 and d2 sessions. The results are displayed in figures 4 and 5. From the plots, it is evident that the continuous value prediction errors for generative embeddings of 11x11 DCM using ElasticNet regression are very high and the performance is comparable to the performance seen for 4x4 DCM in figures 2 and 3.

D. Results for 15x15 DCM

Fold	Accuracy	F1 Score	Macro Precision	Macro Recall
1	0.78	0.78	0.78	0.78
2	0.56	0.55	0.55	0.55
3	0.63	0.56	0.63	0.79
Average	**0.65**	**0.63**	**0.65**	**0.70**

TABLE IX
15X15 DCM BINARY CLASSIFICATION METRICS ACROSS CROSS-VALIDATION FOLDS

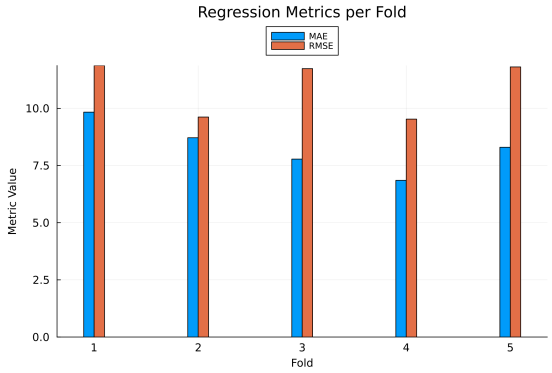

Fig. 4. 11x11 DCM 5-fold Cross-Validation Metrics on Predicting Absolute MADRS Values

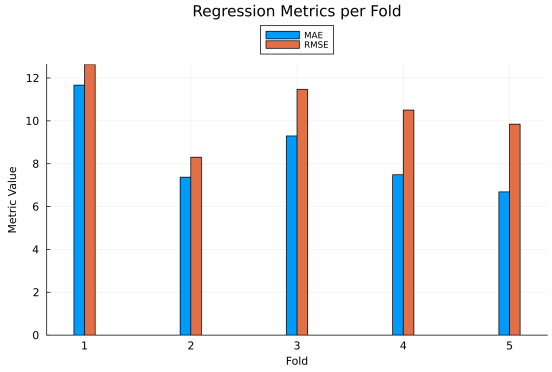

Fig. 5. 11x11 DCM 5-fold Cross-Validation Metrics on Predicting Delta MADRS Values

*1) **Binary Classification**:* Upon verifying different results from the 3, 4 and 5-fold cross validation of the binary label case, it was found that the multinomial classifier performed the best for a 3-fold cross-validation case with an average accuracy of 65%, F1 score of 63%, average macro precision of 65% and average macro recall of 70%. The results are displayed in the table IX. The binary classification using generative embeddings from 15x15 DCM for 3 fold cross validation is not as good as that for the 5-fold cross validation using logistic regression classifier on the 11x11 generative embeddings.

Fold	Accuracy	F1 Score	Macro Precision	Macro Recall
1	0.29	0.19	0.22	0.17
2	0.71	0.69	0.67	0.87
3	0.67	NaN	0.67	NaN
4	1.00	1.00	1.00	1.00
Average	**0.67**	**NaN**	**0.64**	**NaN**

TABLE X
15X15 DCM THREE LABEL CLASSIFICATION METRICS ACROSS FOLDS

*2) **Tri Label Classification**:* Upon verifying different results from the 3, 4 and 5-fold cross validation of the case of three labels, it was found that the logistic regression and multinomial classifiers performed the best for a case of 4-fold cross-validation with an average accuracy of 67%, a F1 score as NaN, average macro precision of 64% and an average macro recall of **NaN**. Thus, these results are not entirely reliable. The results are displayed in table X. This is a good performance when purely looked at from the accuracy metrics' perspective and it performs twice as well as a random choice predictor.

Fold	Accuracy	F1 Score	Macro Precision	Macro Recall
1	0.71	NaN	NaN	NaN
2	0.43	NaN	NaN	NaN
3	0.50	NaN	0.50	NaN
4	0.33	NaN	0.38	NaN
Average	**0.49**	**NaN**	**NaN**	**NaN**

TABLE XI
15X15 DCM FOUR LABEL CLASSIFICATION METRICS ACROSS FOLDS

*3) **Four Label Classification**:* Upon verifying different results from the 3, 4 and 5-fold cross validation of the case of four labels, it was found that logistic regression and multinomial classifiers performed the best for a case of 4-fold cross-validation with an average accuracy of 49%, a F1 score as NaN and an average macro precision and an average macro recall of **NaN**. Thus, these results are not entirely reliable. The results are displayed in the table XI.

Fold	Accuracy	F1 Score	Macro Precision	Macro Recall
1	0.57	NaN	NaN	NaN
2	0.29	NaN	NaN	NaN
3	0.50	NaN	0.40	NaN
4	0.33	NaN	0.30	NaN
Average	**0.42**	**NaN**	**NaN**	**NaN**

TABLE XII
15X15 DCM FIVE LABEL CLASSIFICATION METRICS ACROSS FOLDS

*4) **Five Label Classification**:* Upon verifying different results from the 3, 4 and 5-fold cross validation of the case of five labels, it was found that logistic regression and multinomial classifiers performed the best for a case of 4-fold cross-validation with an average accuracy of 42%, a F1 score as NaN, an average macro precision and an average macro recall of **NaN**. Thus, these results are not entirely reliable. The results are displayed in table XII. From the accuracy metrics, it can be inferred that the model performs slightly more than twice as well for the classification tasks as a random chance predictor. This performance is similar to the 11x11 DCM performance as seen in VIII.

*5) **ElasticNet Regression**:* 5-fold cross validation was carried out on the 15x15 DCM generative embeddings for absolute MADRS values and differences in the values between b0 and d2 sessions. The results are displayed in figures 6 and 7.

VI. CONCLUSION

From the performance results of various classification and regression models applied to the generative embeddings of the 4x4, 11x11 and 15x15 DCMs, it is evident that the generative embeddings can be really powerful for binary classification tasks (identifying responders of ketamine infusion from non-responders) as evidenced from the 3 and 5-fold cross validation metrics displayed in tables I, V and IX. This shows that the logistic regression and multinomial regression classifiers applied to generative embeddings from these DCMs can be

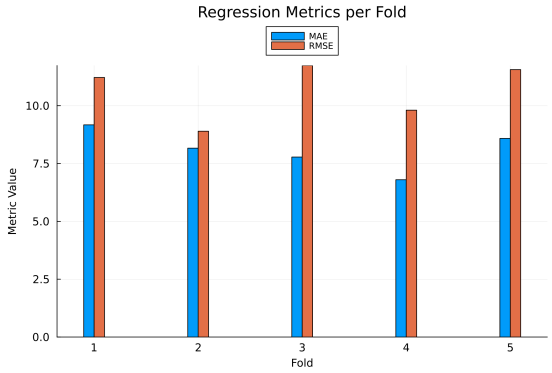

Fig. 6. 15x15 DCM 5-fold Cross-Validation Metrics on Predicting Absolute MADRS Values

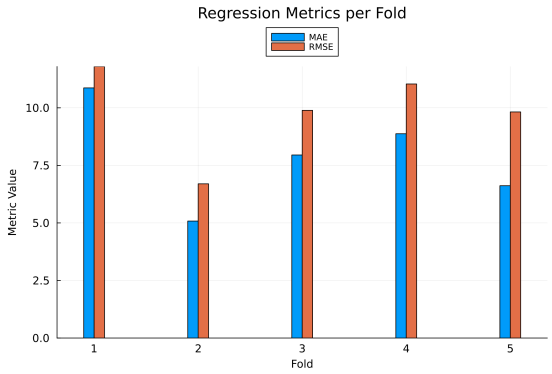

Fig. 7. 15x15 DCM 5-fold Cross-Validation Metrics on Predicting Delta MADRS Values

very powerful in identifying ketamine infusion responders from non-responders from a batch of depressed subjects. To this extent, testing smaller DCMs based on the Default Mode Network and the regions implicated for ketamine effects seems promising.

This can be an incredibly important step in identifying and selecting depressed patient subjects from a larger pool of subjects for effectively testing the effects of ketamine on the depression levels in these subjects. This kind of pre-selection can also help in selecting subjects for a longer study who are likely to respond well to the ketamine infusion treatment such that the study can also study how the dosage of ketamine over a longer period affects the depression levels in these patients. By pre-selecting the study for likely responders, such a study can be made more robust by studying the limitations of ketamine infusion treatment on each patient.

VII. Discussion & Limitations

26 subjects with depression is a small sample size. This means our results are likely not representative for a bigger set of data. From table VII and table VI, it is evident that generative embeddings for more granular and nuanced classification tasks can also be powerful but such a nuanced classification is hampered by the lack of fMRI scans and MADRS scores from baseline sessions for many subjects in the dataset [4]. If the dataset or similar datasets have a larger pool of depressed subjects with all the required data, these generative embeddings can be potentially quite powerful for multi-label, nuanced classification tasks as well or can be better analysed. The presence of NaN values in crucial performance metrics is a major limitation of our current work.

We can also conclude that while these generative embeddings are powerful for classification tasks and yield better than average results, from figures 2, 3 4, 5, 6 and 7 it is evident that these generative embeddings do not perform as well for continuous value prediction tasks when the elastic net regressor model is applied to them. The predictive performance of the elastic net regression on all generative embeddings is suboptimal.

VIII. Future Work

Potential future directions of work include building a bigger fMRI dataset of depressed subjects to study ketamine infusion to test more nuanced and granular classification methods like three, four, five or more category classification. Additionally, another direction of future work would be to fine-tune ElasticNet or other similar regression models to improve the performance of these generative embeddings in predicting continuous values. Another interesting direction for future work is to test these classification and regression models with the session b0 MADRS score as an additional feature vector along with the generative embeddings.

IX. Code and Results

All the code used for the project along with raw results can be found at the following GitLab Respository:Project 8: DCMs for Depression

Author Contributions

This project was a joint effort, with all team members contributing to design decisions, result interpretation, and writing.

Leon Schönleber focused on constructing the DCMs. He handled fMRI preprocessing in SPM, defined the ROIs, and inverted spectral DCMs for all network granularities. He also contributed to the DCM and neuroscience background sections of the report.

Keyshav Mor and Rico-Marcel Benning jointly developed the machine learning pipeline. They implemented classification and regression models in Julia, prepared the generative embeddings, defined label categories and evaluation strategies, and carried out cross-validation experiments. Keyshav also ran the DCM inversions locally to support scalability, while Rico performed additional permutation testing to verify the robustness of the classification results. Together, they wrote the methodology and results sections.

All authors collaborated closely on defining the research goals, interpreting results, and refining the manuscript.

REFERENCES

[1] W.-C. C. L. D. H. I. S. G. Z. C. Machado-Vieira R, Baumann J, "The timing of antidepressant effects: A comparison of diverse pharmacological and somatic treatments," *Pharmaceuticals (Basel)*, 2010.

[2] W. S.-N. A. W. D. R. L. N. G. H. R. L. B. M. P. S.-W. K. B. M. B. G. F. M. Trivedi MH, Rush AJ, "Evaluation of outcomes with citalopram for depression using measurement-based care in star*d: implications for clinical practice," *Am J Psychiatry*, vol. 18, 2006.

[3] J. C. J. J. D. Thomas R. Einarson, Steven R. Arikian, "Comparison of extended-release venlafaxine, selective serotonin reuptake inhibitors, and tricyclic antidepressants in the treatment of depression: A meta-analysis of randomized controlled trials," *Clinical Therapeutics*, vol. 21, 1999.

[4] C. A. Z. Jennifer W. Evans, Allison C. Nugent, "Nimh ketamine mechanism of action study," 2018,.

[5] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 20, no. 2, pp. 215–242, 1958.

[6] P. McCullagh, "Regression models for ordinal data," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 42, no. 2, pp. 109–142, 1980.

[7] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[8] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[9] R. S. D. Satoshi Deyama, "Neurotrophic mechanisms underlying the rapid and sustained antidepressant actions of ketamine," *Pharmacology Biochemistry and Behavior*, vol. 188, 2020.

[10] K. R. D. I. M. A. P. L. M. B. S. B. Nunez NA, Joseph B, "An update on the efficacy of single and serial intravenous ketamine infusions and esketamine for bipolar depression: A systematic review and meta-analysis," *Brain Sciences*, 2023.

[11] C. L. P. Z. Guo-liang Liu, Yun-feng Cui, "Ketamine a dissociative anesthetic: Neurobiology and biomolecular exploration in depression," *Chemico-Biological Interactions*, 2020.

[12] Z. X.-P. A. e. a. Tozzi, L., "Personalized brain circuit scores identify clinically distinct biotypes in depression and anxiety," *Nature Medicine*, vol. 30, 2024.

[13] D. J. D. K. M. F. M. Y. F. R. Z. B. O. D. E. A. S. A. S. K. K. J. M. H. G. F. A. G. F. M. P.-L. A. V. H. C. B. D. M. L. C. Drysdale AT, Grosenick L, "Resting-state connectivity biomarkers define neurophysiological subtypes of depression," *Nature Medicine*, vol. 23, 2017.

[14] W. P. K.J. Friston, L. Harrison, "Dynamic causal modelling," *NeuroImage*, vol. 19, 2003.

[15] B. B. A. R. Karl J. Friston, Joshua Kahan, "A dcm for resting state fmri," *NeuroImage*, vol. 94, 2014.

[16] M. L. P. A. e. a. Wise, T., "Instability of default mode network connectivity in major depression: a two-sample confirmation study," *Transl Psychiatry*, vol. 7, 2017.

[17] E. I. N. T. e. a. Lynch, C.J., "Frontostriatal salience network expansion in individuals in depression," *Nature*, vol. 633, 2024.

[18] W. L. Keller AS, Ball TM, "Deep phenotyping of attention impairments and the 'inattention biotype' in major depressive disorder," *Psychol Med*, vol. 50, 2020.

[19] V. Menon, "Large-scale brain networks and psychopathology: a unifying triple network model," *Trends in Cognitive Sciences*, vol. 15, 2011.

[20] J. W. e. a. Evans, "Default mode connectivity in major depressive disorder measured up to 10 days after ketamine administration," *Biological Psychiatry*, vol. 84, 2018.

[21] V. N. Correia AS, Cardoso A, "Bdnf unveiled: Exploring its role in major depression disorder serotonergic imbalance and associated stress conditions," *Pharmaceutics*, vol. 2081, 2023.

[22] H. C. Patel S, "Role of endocannabinoid signaling in anxiety and depression," *Curr Top Behav Neurosci*, 2009.

[23] Y. C. M. H. e. a. Yamanaka, H., "A possible mechanism of the nucleus accumbens and ventral pallidum 5-ht1b receptors underlying the antidepressant action of ketamine: a pet study with macaques," *Transl Psychiatry*, 2014.

[24] C. K. F. J. e. a. Murrough, J., "Regulation of neural responses to emotion perception by ketamine in individuals with treatment-resistant major depressive disorder," *Transl Psychiatry*, 2015.

[25] U. Wellcome Centre for Human Neuroimaging, "Statistical parametric mapping (spm)," 2024, accessed: June 1, 2025. [Online]. Available: https://www.fil.ion.ucl.ac.uk/spm/

[26] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, "Julia: A fresh approach to numerical computing," *SIAM Review*, vol. 59, no. 1, pp. 65–98, 2017.

[27] J. Blaom, T. L. Anthony *et al.*, "Mlj: A julia package for composable machine learning," *Journal of Open Source Software*, vol. 5, no. 55, p. 2704, 2020.

[28] M. S. Interface, "Mljscikitlearninterface.jl," https://github.com/alan-turing-institute/MLJScikitLearnInterface.jl, 2024, accessed: June 1, 2025.