# Applying Model-Agnostic Meta Learning for Deep Learning Schizophrenia Diagnosis

**Ashwin Ahuja**
aa2001
Gonville & Caius College

## Abstract

Throughout psychiatry, small sample sizes pose a significant challenge, especially to novel deep learning approaches. In particular, it means that tasks better solved separately are combined to have enough data to solve the overarching problem. This reduces the model's efficacy and reliability in solving the individual task. In this proposal, we present the idea of using Model-Agnostic Machine Learning (MAML), using the clinical problem of diagnosing schizophrenia using fMRI image connectivity matrices as an example. The meta-learning approach works by first training a base model on all data. Prior to testing on an fMRI image of a particular individual, we first adapt this base model using data of only individuals of a similar age. In a preliminary experiment, we demonstrate that the meta-learning approach surpasses the performance of baseline Convolutional Neural Networks (CNNs) and Support Vector Machines (SVMs) by 24.8% and 19.4% respectively in diagnosis accuracy on the COBRE dataset. This is used as a springboard to define research aims that the proposed work using the larger PSYSCAN dataset will complete.

## 1 Introduction

In recent years, there has been a shift towards using Deep Learning models to help with psychiatric diagnosis using medical imaging. Qiu et al. [1] shows one example of schizophrenia diagnosis using a Convolutional Neural Network (CNN). The potential of Deep Learning has been replicated around a number of problems in psychiatry, be that for diagnosis or treatment [2].

However, in general, as demonstrated by Schultz et al. [3], Deep Learning tends to under-perform (performing worse than SVMs) with a small sample size. Small sample sizes are the norm in psychiatry, where data is difficult and expensive to collect, especially for medical imaging. This problem often leads to the creation of models to solve more than one specific task to ensure the data is sufficient. This impacts the accuracy and reliability of the model. Hence, we propose using Model Agnostic Meta-Learning (MAML) to maximally utilise all data, whilst producing better models for a specific task. For example, the problem of diagnosing schizophrenia for all humans can be divided by age and sex. Meta-learning allows us to produce a base model for all humans, before adapting this model to a particular task (i.e. diagnosing schizophrenia in men aged between 30 and 40).

The potential of MAML is considered in this proposal with a primary experiment using a small dataset from the Center for Biomedical Research Excellence (COBRE) [4][5][6][7]. We demonstrate that MAML performs statistically significantly better than a non meta-learning approach for diagnosing schizophrenia using fMRI images. It also performs better than an SVM approach, which is surprising given the small dataset size. We then discuss future research aims building on this preliminary experiment. These range from validating the results presented here, to testing with a range of clinical problems, task definitions and pre-processing methods. The aims are largely enabled by the larger and more comprehensive data provided by the PSYSCAN project.

## 2 Background

Meta-learning is best described as learning how to learn. It allows us to use deep learning to solve a range of tasks, where each problem may have minimal data but the method to solve each task is very similar. In this proposal, we consider the overarching problem as schizophrenia diagnosis using fMRI connectivity matrices. The individual tasks involve splitting this problem into tasks for differing demographic groups. Schizophrenia diagnosis is a reasonably well researched area, with papers showing promising results using both SVMs [8] and deep learning approaches [9]. We aimed to implement both of these as baseline method with which we compare our novel MAML method. Hence, in this section, we look at how each of these approaches works. We also look at how the novel MAML approach works, building on existing CNN schizophrenia diagnosis methods. Finally, we look at the challenge of limited sample sizes and other existing solutions.

### 2.1 Support Vector Machines (SVM)

SVMs attempt to find a hyperplane that separates different classes (in this case, having schizophrenia and not having schizophrenia) in feature space, where the features are the fMRI connectivity matrices. Unlike logistic regression, SVMs focus only on boundary points (support vectors) and use these to find the largest margin. Since in many cases there is no perfect separation plane, we attempt to find the hyperplane that almost separates the data using a soft margin. In training, we obtain parameters for both the hyperplane and for slack variables (normally controlled by a tuning parameter). This optimisation problem can be efficiently solved as a convex quadratic program using Lagrange Multipliers. The feature space can be defined by either the features or combinations of the features. The kernel functions which define the combinations result in non-linear decision boundaries in the original space.

#### 2.1.1 SVMs with matrices as training data

SVM's take vectors as inputs. Our use, passing in fMRI connectivity matrices alongside their corresponding diagnosis as the expected class, requires the conversion of the matrix to a vector. Historically, methods simply flatten the matrix [10]. This is true in psychiatry as well, with this idea being described for three-dimensional brain voxels by LaConte et al. [11]. These flattened vectors can correspond to all possible brain voxels [12] or use a subset [13], deciding on the subset using a feature selection algorithm.

However, these all ignore the two-dimensional structure of the initial data. Structure can be inferred using wavelets that represent spatial correlations [14]. We can also use k-means clustering to convert fMRI connectivity matrices to feature vectors. In soft (fuzzy) clustering, we represent each point by the contribution it makes to each cluster. Hence, each point is represented by a vector of length k. This clustering can be completed in a number of different methods, including using weighted k-means, as used by Kerdprasop et al. [15] or Gaussian Mixture Models (Biernacki et al. [16]). Prior to clustering, Mwangi et al. [17] also proposed using t-distributed stochastic neighbour embedding (t-SNE). This offers dimensionality reduction whilst preserving and exacerbating clusters.

#### 2.1.2 Schizophrenia Diagnosis

There has been lots of existing work using SVMs diagnose schizophrenia using models trained using fMRI connectivity data. Steardo Jr et al. [18] completed a review of 22 of the most promising works. The accuracy of these models varied from between 69% (Ji et al. [19]) to over 99.3% (Qureshi et al. [20]). The best performing models selected specific regions of the brain. The only module using the connectivity of the whole brain, as we conducted in our preliminary experiment was by Orban et al. [21], which had an accuracy of 84%, but used a much larger sample size than our experiment. These papers help to corroborate that any improvement on a poorly performing baseline SVM model in our preliminary experiment could be powerful. With more pre-processing, for example more careful feature selection, any improvement made by MAML could be exacerbated.

### 2.2 Convolutional Neural Networks (CNN)

Hierarchical deep learning is a branch of machine learning inspired by the human brain. These use computational models that show similar characteristics as the the neo-cortex, in order to aim to emulate the human brain's learning mechanism. The development of CNNs by Fukushima et al [22]

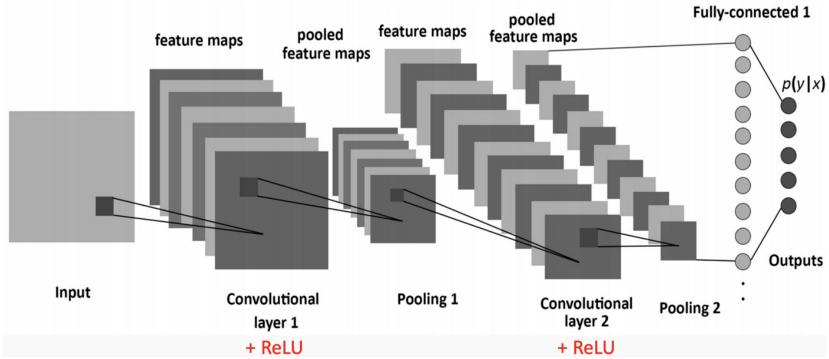

Figure 1: Diagram of Convolutional Neural Network structure for classification. Image taken from [23]

was driven by the human vision system. As opposed to other neural networks, they are designed to take in two (or three) dimensional images, using spatial relationships to reduce required training parameters. Where a multilayer perceptron would use one perceptron per pixel, CNNs instead use filters which act on the entire image. This CNN architecture offers invariance to shift, scale and rotation, instead looking at elementary features including oriented edges and corners.

CNNs are made from (often multiple) convolutional units. Each convolutional units includes a convolutional layer, pooling layer, normalisation layer and linear layer. The parameters of the convolutional layer are a set of filters. In a forward pass, we convolve each filter with the input, producing an activation map for each. During training, using back-propagation of losses, the network learns of useful filters that activate for a particular feature. The pooling layer prevents over-fitting by choosing the largest values on feature maps to pass to future layers. Figure 1 demonstrates the general CNN architecture, using two convolutional units.

### 2.2.1 Schizophrenia Diagnosis

There have been a few papers to use CNNs to complete diagnosis of schizophrenia using fMRI data. Some of these use three dimensional imaging features, whilst others use two dimensional connectivity matrices (similarly to our preliminary experiment). Irrelevant of the input data, most papers used a generic CNN, with differing numbers of convolutional units. Qureshi et al [9] use a five layer system, whilst Qiu et al [1] use two layers. Both however, include significant pre-processing to select regions of interest before training. They result in accuracies of 98.1% and 72.7% respectively. For our baseline for the CNN in the preliminary experiment, we choose to emulate Qiu et al.'s work since it uses similar input data, but we exclude the complex feature selection due to time constraints.

## 2.3 Model Agnostic Meta-Learning

In meta-learning [24], we have two specific stages. In the first, we use a training dataset consisting of all data across all sub-problems (tasks) to train a base model. In the second stage, we pass in a few pieces of data (known as shots) using these to adapt the model for the specific sub-problem that is being encountered for the test. MAML works for any deep neural network that can use gradient descent and uses only a few gradient descent steps to adapt the model.

In base-model training, the model learns how to adapt to a new task with only a few shots. For each task $T_i$, we samples a support set $S_{T_i}$ of length k (where k is the number of shots) and use the remainder for the query set $Q_{T_i}$. The support set is used to train the task specific parameters whilst this adapted model is then evaluated on the query set. This uses a loss function. For a classification problem, this is normally the cross-entropy loss.

$$\mathcal{L}_{\mathcal{T}_i}\left(f_\theta\right) = \sum_{(\boldsymbol{x}_j, \boldsymbol{y}_j) \in S_{\mathcal{T}_i}} \boldsymbol{y}_j \log f_\theta\left(\boldsymbol{x}_j\right) + \left(1 - \boldsymbol{y}_j\right) \log f_\theta\left(1 - \boldsymbol{x}_j\right) \tag{1}$$

This is used to update the task-specific parameters of the base model using a few ($\approx 5$) gradient descent steps. The task learning rate is set reasonably high ($\approx 0.1$) to enforce fast adaptation of the model to very little data.

$$\theta'_{\mathcal{T}_i} = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}\left(f_\theta\right) \tag{2}$$

To use these task-specific parameters to update the parameters of the base model, we define a meta-objective. We choose to minimise the sum of task losses of the query sets, that is:

$$\arg\min_{\theta} \sum_{\mathcal{T}_i} \mathcal{L}_{\mathcal{T}_i}\left(f_{\theta'_{\mathcal{T}_i}}\right) \quad \text{where } \theta'_{\mathcal{T}_i} = \theta - \alpha\nabla_\theta \mathcal{L}_{\mathcal{T}_i}\left(f_\theta\right) \tag{3}$$

We update the parameter of the base model $\theta$ by minimising the meta objective with stochastic gradient descent.

$$\theta \leftarrow \theta - \beta\nabla_\theta \sum_{\mathcal{T}_i} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_{\mathcal{T}'_i}}) \tag{4}$$

In the second (testing) stage, for each test item, we adapt the base model using a couple of pieces of data. This data is randomly taken from the original query and support set of the class of the test item. This is used to adapt the base model using the same adaptation procedure described above. We then test the adaptation on the withheld test item.

### 2.3.1 Use in Psychiatry

MAML has been used across research topics, from mobile sensing [24] to NLP [25]. However, the vast majority of uses have been in computer vision [26]. This lends itself therefore to working well for fMRI analysis, since there has been widespread success in using computer vision approaches in psychiatry. For example, wavelet approaches are now widely used for preprocessing of medical imaging. However, the uses of MAML in psychiatry are very limited. We could find only a single paper using a variant of MAML. This was a paper by Bontonou et al. [27] where MAML is used to identify what type of action is being performed by a subject during an fMRI scan (e.g. audio sentence, speech sound).

Our usage of MAML is very different to this, instead using MAML to define separate models for people of different demographics, attempting to improve accuracy and reliability for a problem with a very small sample size per demographic group. The idea of improving the performance of diagnostic models by separating subjects by demographics is not new. Most papers (e.g. Cooper et al. [28]) produce models for men and women separately. From a theoretical perspective, Bluhm et al. [29] demonstrate that brain connectivity is significantly impacted by the person's age and sex. Thus, models produced for a specific demographic will likely outperform models trained on all demographics. Our results to the preliminary experiment in Section 3 demonstrate that this is true for all modelling techniques. This comes however, with a reliability reduction for SVMs. MAML offers this higher accuracy with the reliability that comes from using all available data.

## 2.4 Sample Size Issues

Psychiatry has an ongoing problem with small sample sizes. This comes from the cost and time associated with data collection. This is partially fixed by the PSYSCAN project which is offering a larger dataset than previously available, however, this is still reasonably small (500 images). The use of a small sample size, as discussed by Faber et al. [30], 'undermines the internal and external validity'. It results in differences which might be small turn into statistically significant ones, whilst other differences which may exist are not sufficiently shown. It also combines with a publication bias to exacerbate a low reliability in papers' results [31]. The issue of a small sample size could lead to papers combining individual tasks to produce an overarching task for which more data exists. This is the motivation for considering the impact of MAML for psychiatry, since it could more effectively utilise all data, whilst still solving each task individually. Other solutions split largely into two types. The first is data augmentation, whilst the second is careful modelling techniques.

In data augmentation, we can use a number of methods. For CNN diagnosis of Alcohol use Disorder (AUD), Wang et al. [32] use random rotation, translation, gamma correction and noise injection to augment a 235-image dataset to 13,100 images for training. Koskinen et al. [33] propose a Bayesian approach using the posterior distribution of existing data, whilst Wright et al. [34] wrote a methodology for Bootstrap sampling (using the plug-in principle) for psychiatry. However, it is important to note that these methods could be complimentary to MAML rather than alternatives.

The second category of solutions for dealing with insignificant data involves carefully choosing the modelling method. For example, Mativo et al. [35] show that SVMs perform better than Multiple Linear Regression, whilst Schultz et al. [3] consider the differing impact of sample size on SVMs and CNNs. They demonstrate that SVMs perform better with a smaller sample size than CNNs, with the gap closing as the sample size increases.

## 3 Preliminary Experiment

In our preliminary experiment, we took fMRI connectivity data as well as schizophrenia diagnoses and used it to develop a model for the function:

$$f : \text{fMRI Connectivity Matrix} \longrightarrow \text{Diagnosis} \in \{0, 1\}$$

We first looked at the distribution of data across age ranges and genders. Figure 2 demonstrates that the majority of test subjects from the COBRE dataset are male. In fact, there appears to be very few females within each of the age ranges, defined as 20s, 30s, etc. Therefore, we decided for this experiment to concentrate on only males, and look at the impact of creating models for (1) the entire male dataset, (2) specifically for age ranges and (3) using a meta-learning approach to target adaptation for each age range.

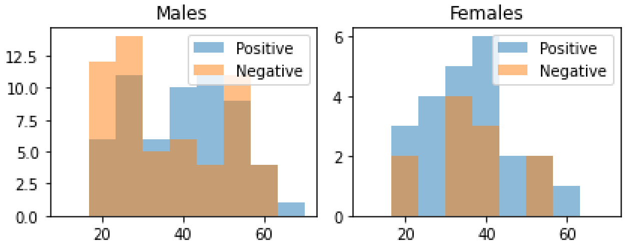

Figure 2: Histogram showing distribution of ages and sex of the data-set

Figure 2 also demonstrates the lack of balance of data. This would have a significant impact on the CNNs. Therefore, data was probabilistically included a number of times to ensure that there was a similar number of data-points for each class (schizophrenia positive and negative), for each age range.

### 3.1 Methodology

Three different forms of models are trained. For SVM models, 10-fold leave one out cross-validation is used. The mean and variance diagnosis accuracy, precision and recall of the withheld test-set is reported. For the CNN and MAML models, only one test is completed due to time constraints. The mean diagnosis accuracy, precision and recall of the model on a random withheld test-set is reported. Statistical significance testing is done with a t test with an $\alpha$ value of 0.05.[1]. All code for this experiment is available online [36].

1. **Support Vector Machine**: The scikit-learn library is used to create and test an SVM. A polynomial kernel of order three is used, based on empirical testing. A number of SVMs are fitted, using a range of input data. One is created that uses all possible fMRI data and one for every age range (20s, 30s, 40s, 50s). Patients outside of these ranges are not considered.

   Since this input data used is a matrix, we considered three methods of converting this into a vector (as per section 2.1.1). We (1) flattened the matrix, (2) used k-means soft clustering (with an empirically chosen k value of 5) using Gaussian Mixtures and (3) used t-SNE before clustering.

2. **Convolutional Neural Network**: A basic two layer CNN was created using PyTorch. The diagram of the network is shown in Figure 3. We trained the model using the same subsets of data as for SVMs. Training was completed for 100 epochs, using Cross Entropy as the loss function. Figure 4 shows the change in loss during training for an example model. This demonstrates that whilst the loss function does continue to drop over time, we can reasonably use the results at 100 epochs to represent the model's accuracy. fMRI connectivity matrices were scaled down from (293, 293) to (128, 128) which empirically proved to be a good dimension reduction to conserve fMRI information, processing time and power. Training and evaluation took around fifteen minutes per model.

[1] The null hypothesis is that the systems being compared are the same

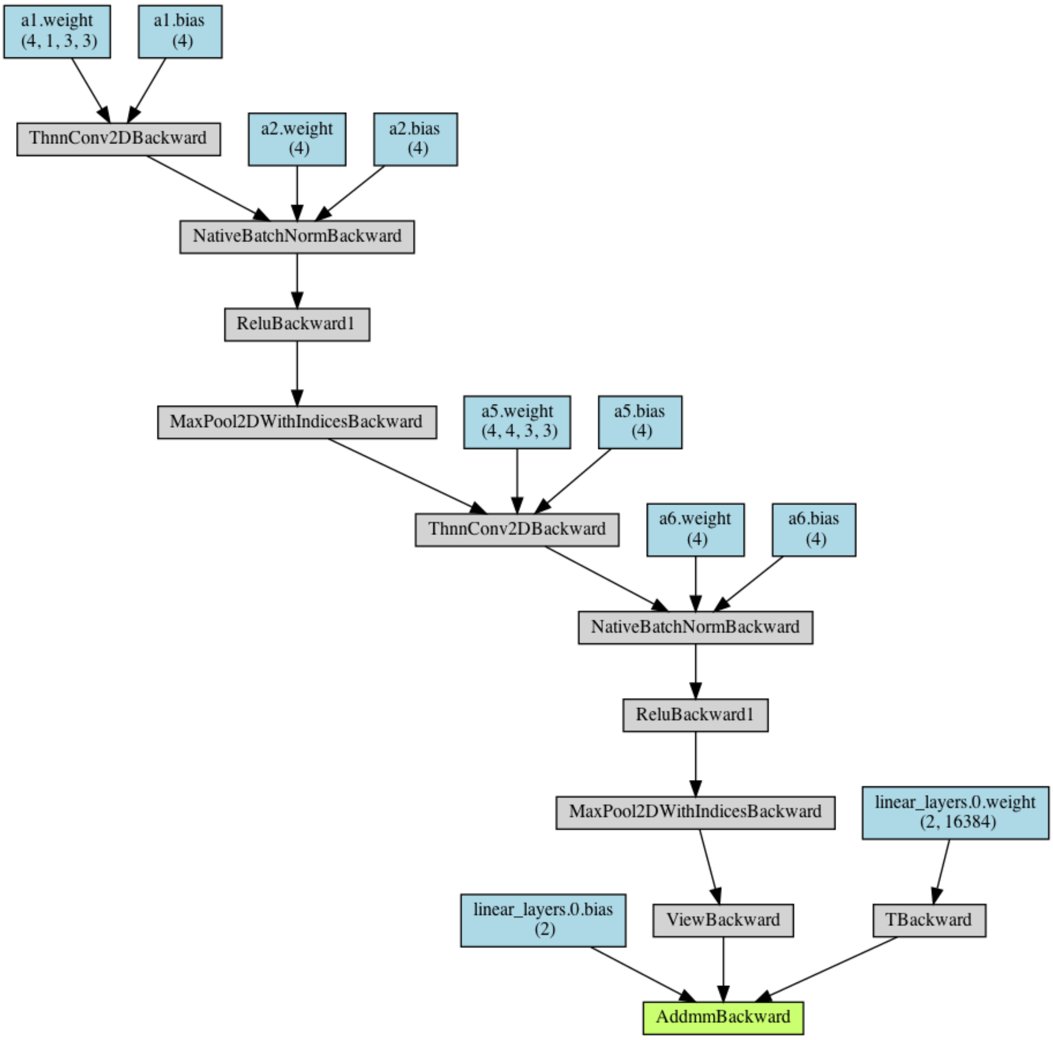

Figure 3: CNN Model Diagram

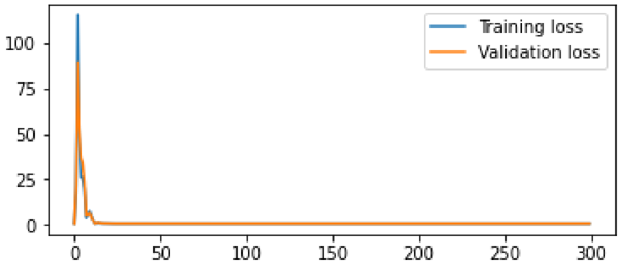

Figure 4: CNN Losses

3. **Model-Agnostic Meta Learning**: The PyMeta and PyTorch-MAML libraries were used and modified to produce an MAML implementation. The same model architecture described above was used for this approach using the same input data and loss function. The input dataset was split into three parts: a support set, query set and withheld test set. The former two are used for training whilst the latter two are used for testing, with the old query set being used for demographic specific adaptation during tests. We used five data-points for adaptation, taking five gradient descent steps.

### 3.2 Results

| | | All Data | | | 20s | | | 30s | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Mean Diagnosis Accuracy / % | Precision | Recall | Mean Diagnosis Accuracy / % | Precision | Recall | Mean Diagnosis Accuracy / % | Precision | Recall |
| SVM | Flattened | 60.7 (15.5) | 62.4 (20.4) | 55.6 (15.8) | 66.7 (15.8) | 91.7 (18.6) | 66.7 (25.4) | 66.7 (28.6) | 85.7 (22.6) | 88.1 (19.3) |
| | k-means clustering | 65.4 (15.9) | 58.8 (22.9) | 86.1 (12.5) | 67.5 (21.7) | 90.0 (20.0) | 60.0 (22.6) | 60.0 (37.4) | 83.3 (23.6) | N/A |
| | t-SNE and k-means clustering | 66.8 (8.71) | 65.7 (16.5) | 73.1 (15.6) | 61.7 (27.7) | 90.0 (20.0) | 60.0 (22.6) | 65.0 (32.0) | 81.3 (24.2) | 93.8 (16.5) |
| Convolutional Neural Network | | 61.5 | 62.5 | 71.4 | 76.9 | 57.1 | 100 | 88.4 | 66.7 | 100 |
| Model Agnostic Meta-Learning (CNN) | | 76.8 | 70.0 | 83.3 | N/A | | | | | |

| | | 40s | | | 50s | | |
|---|---|---|---|---|---|---|---|
| | | Mean Diagnosis Accuracy / % | Precision | Recall | Mean Diagnosis Accuracy / % | Precision | Recall |
| SVM | Flattened | 53.3 (25.1) | 85.7 (22.6) | 76.2 (28.0) | 63.3 (30.7) | N/A | N/A |
| | k-means clustering | 75.0 (25.0) | 93.8 (16.5) | 87.5 (21.7) | 63.3 (26.7) | N/A | N/A |
| | t-SNE and k-means clustering | 60.0 (37.4) | 81.3 (24.2) | 93.8 (16.5) | 60.0 (31.8) | N/A | N/A |
| CNN | | 66.7 | 66.7 | 66.7 | 66.7 | 50 | 100 |
| MAML (CNN) | | N/A | | | | | |

Table 1: Preliminary Results Table. Standard Deviations, when completing cross-validation are in brackets. Some precisions and recalls are not available when the denominator is zero.

### 3.3 Discussion

We can first evaluate the general results of baseline SVMs shown by 1 and 5. These demonstrate that there is a potential gain by creating a model for specific age groups. A number of age-group models statistically significantly outperform the all data models. However, a number of models also perform significantly worse than the all data baseline. Additionally, the standard deviations are higher than the all data baselines, showing the reduced reliability. Compared to existing literature, the all data models perform worse than other diagnosis systems described by Steardo Jr et al. [18]. However, this is with significantly less pre-processing and smaller sample size.

k-means clustering adds an improvement (with improvements in all but one age category) but this difference is not statistically significant (p=0.21). This would require more data to verify if the performance difference is significant. This could be done using the new PSYSCAN data. There are also interesting effects made by the SVM models on the precisions and recalls. The age-specific models have a significantly higher precision (average p=0.02) than the all models, with no statistically significant difference in the recalls (average p=0.4). The increase in precision is promising, but this is less important for a diagnosis method than increasing the recall.

The CNN initially trained on all data appears to perform as well as the SVM. It does under-perform compared to Qiu et al.'s [1] work. This is expected, since our sample size is lower and we use less pre-processing. It is interesting to note that with no exceptions, models trained for each of the age groups performed better than the SVMs. The results also show a statistically significant difference using the one sample t test with the recall being higher (p=0.01). This is a useful property for a diagnosis method, since the cost of a false negative is far higher than that of a false positive.

The results for MAML are incredibly promising. It has a higher average accuracy than that the SVMs or CNNs without meta-learning (average p=0.02 for SVMs). It even out-performs most results for each age category, under-performing compared to only CNNs in the 20s and 30s. This demonstrates that the method could potentially get the best of both worlds, gaining the advantages of using all data, whilst adapting to the most relevant data for each test case. Finally in comparison to other methods from literature, the method outperforms the CNN by Qiu et al. [1] and ranks highly in the SVM methods evaluated by Steardo Jr et al. [18], both of which used far more pre-processing.

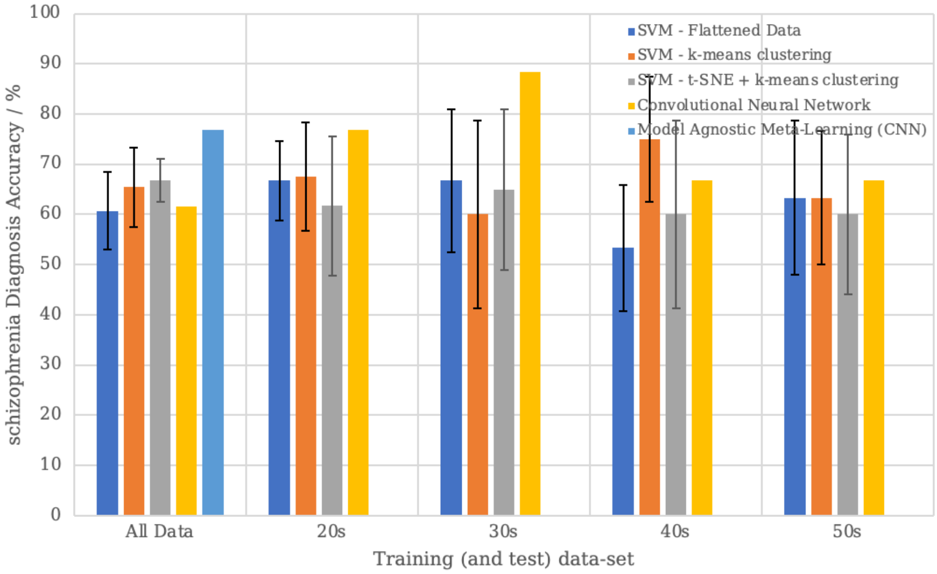

Figure 5: Graph of results for mean diagnostic accuracy

## 4 Research Proposal Aims

The preliminary experiment demonstrated the potential impact of a meta-learning approach to maximise data usage when producing a model using the COBRE dataset. Whilst the effects were shown only for subjects of different ages, we expect that sex could also show similar results. The experiment also allows us to utilise knowledge that our brains, and the way in which disease affects our brain varies significantly with age. A meta-learning approach to diagnosis allows us to use both the commonalities and differences in order to produce the best possible model.

There is lots of work that can be completed using the larger PSYSCAN dataset to further test this hypothesis. Thus, our research would have the following aims:

1. **Validate preliminary experiment results**: By adding more data and following the same method, we can validate whether the improvements shown by MAML over both CNNs and SVMs continues. PSYSCAN will have 250 individuals, with two scans each, whilst COBRE had only 148.
2. **Testing different task definitions**: In the preliminary experiment, we demonstrated that there was value in separating subjects by age. PSYSCAN will have information with regards to age, sex, IQ and years in education. These should be systematically tested to investigate whether separating by these will offer any value to increase diagnosis accuracy. We can also define a task per individual since PSYSCAN will have two fMRI scans per person.
3. **Testing different models**: The preliminary experiment only looked at one possible model type, a CNN. The major advantage of MAML over existing meta-learning and transfer learning is the flexibility to use any model types. These should be tested.
4. **Testing prepossessing methods**: The best performing methods evaluaded by Steardo Jr et al. [18] such as the one by Qureshi et al. [20] use significant pre-processing, including specific feature selection. These should be incorporated to evaluate whether these further improve the diagnosis accuracy of the MAML model.
5. **Considering other clinical problems**: MAML could be considered with regards for other clinical problems. For example, another task which could benefit from a meta-learning approach is predicting the level of psychosis in a patient using fMRI images. The initial idea here would be to denote each individual as a different task. Hence, you would train on all possible images, passing in an fMRI image and a PANSS score. Then, during adaptation, you could pass in the baseline fMRI and PANSS score and use the adapted model to predicted the PANSS score at one year. Other possible problems involve predicting the progression of psychosis and predicting the difference between resting state and T1w fMRI scans.

## 5 Conclusion

The results of Section 3 demonstrates the potential of MAML for diagnosing schizophrenia, based on the hypothesis that different age patients will exhibit different brain responses to psychosis. Our preliminary experiment demonstrates this hypothesis is true, with age-specific CNNs performing 21.4% better than models which use all data. A small (3.0%), but not statistically significant difference also exists for SVMs.

We then use meta-learning to produce a base model using all data before adapting it to each age group. This outperforms the CNN by 24.8% and the average SVM by 19.4% in terms of diagnosis accuracy. It also offers the desirable property of a high (16.6% higher than CNN and 16.3% higher than average SVM) recall. It even surpasses the performance of the average model trained only on data from a particular age range by an average of 20.9% for SVMs and 2.8% for CNNs.

The results lead to a number of future research questions that can be answered with the addition of the new PSYSCAN data. These include looking at how the results change with differing data volumes, preprocessing, task definitions, neural network models. Finally, we propose extending the results to alternative clinical questions.

## References

[1] Y. Qiu, Q.-H. Lin, L.-D. Kuang, W.-D. Zhao, X.-F. Gong, F. Cong, and V. D. Calhoun, "Classification of schizophrenia patients and healthy controls using ica of complex-valued fmri data and convolutional neural networks," in *International Symposium on Neural Networks*, pp. 540–547, Springer, 2019.

[2] D. Durstewitz, G. Koppe, and A. Meyer-Lindenberg, "Deep neural networks in psychiatry," *Molecular psychiatry*, vol. 24, no. 11, pp. 1583–1598, 2019.

[3] M.-A. Schulz, B. T. Yeo, J. T. Vogelstein, J. Mourao-Miranada, J. N. Kather, K. Kording, B. Richards, and D. Bzdok, "Different scaling of linear models and deep learning in ukbiobank brain images versus machine-learning datasets," *Nature communications*, vol. 11, no. 1, pp. 1–15, 2020.

[4] J. M. Stephen, B. A. Coffman, R. E. Jung, J. R. Bustillo, C. Aine, and V. D. Calhoun, "Using joint ica to link function and structure using meg and dti in schizophrenia," *Neuroimage*, vol. 83, pp. 418–430, 2013.

[5] A. R. Mayer, D. Ruhl, F. Merideth, J. Ling, F. M. Hanlon, J. Bustillo, and J. Canive, "Functional imaging of the hemodynamic sensory gating response in schizophrenia," *Human brain mapping*, vol. 34, no. 9, pp. 2302–2312, 2013.

[6] F. M. Hanlon, J. M. Houck, C. J. Pyeatt, S. L. Lundy, M. J. Euler, M. P. Weisend, R. J. Thoma, J. R. Bustillo, G. A. Miller, and C. D. Tesche, "Bilateral hippocampal dysfunction in schizophrenia," *Neuroimage*, vol. 58, no. 4, pp. 1158–1168, 2011.

[7] V. D. Calhoun, J. Sui, K. Kiehl, J. A. Turner, E. A. Allen, and G. Pearlson, "Exploring the psychosis functional connectome: aberrant intrinsic networks in schizophrenia and bipolar disorder," *Frontiers in psychiatry*, vol. 2, p. 75, 2012.

[8] L. Su, L. Wang, H. Shen, G. Feng, and D. Hu, "Discriminative analysis of non-linear brain connectivity in schizophrenia: an fmri study," *Frontiers in human neuroscience*, vol. 7, p. 702, 2013.

[9] M. N. I. Qureshi, J. Oh, and B. Lee, "3d-cnn based discrimination of schizophrenia using resting-state fmri," *Artificial intelligence in medicine*, vol. 98, pp. 10–17, 2019.

[10] X. Li, L. Wang, and E. Sung, "Multilabel svm active learning for image classification," in *2004 International Conference on Image Processing, 2004. ICIP'04.*, vol. 4, pp. 2207–2210, IEEE, 2004.

[11] S. LaConte, S. Strother, V. Cherkassky, J. Anderson, and X. Hu, "Support vector machines for temporal classification of block design fmri data," *NeuroImage*, vol. 26, no. 2, pp. 317–329, 2005.

[12] J. Mourao-Miranda, A. L. Bokde, C. Born, H. Hampel, and M. Stetter, "Classifying brain states and determining the discriminating activation patterns: support vector machine on functional mri data," *NeuroImage*, vol. 28, no. 4, pp. 980–995, 2005.

[13] X. Wang, R. Hutchinson, and T. M. Mitchell, “Training fmri classifiers to detect cognitive states across multiple human subjects,” *NIPS03*, vol. 16, 2003.

[14] D. Van De Ville, T. Blu, and M. Unser, “Surfing the brain,” *IEEE engineering in medicine and biology magazine*, vol. 25, no. 2, pp. 65–78, 2006.

[15] K. Kerdprasop, N. Kerdprasop, and P. Sattayatham, “Weighted k-means for density-biased clustering,” in *International Conference on Data Warehousing and Knowledge Discovery*, pp. 488–497, Springer, 2005.

[16] C. Biernacki, G. Celeux, and G. Govaert, “Assessing a mixture model for clustering with the integrated completed likelihood,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 22, no. 7, pp. 719–725, 2000.

[17] B. Mwangi, J. C. Soares, and K. M. Hasan, “Visualization and unsupervised predictive clustering of high-dimensional multimodal neuroimaging data,” *Journal of neuroscience methods*, vol. 236, pp. 19–25, 2014.

[18] L. Steardo Jr, E. A. Carbone, R. de Filippis, C. Pisanu, C. Segura-Garcia, A. Squassina, P. De Fazio, and L. Steardo, “Application of support vector machine on fmri data as biomarkers in schizophrenia diagnosis: A systematic review,” *Frontiers in Psychiatry*, vol. 11, p. 588, 2020.

[19] L. Ji, S. A. Meda, C. A. Tamminga, B. A. Clementz, M. S. Keshavan, J. A. Sweeney, E. S. Gershon, and G. D. Pearlson, “Characterizing functional regional homogeneity (reho) as a b-snip psychosis biomarker using traditional and machine learning approaches,” *Schizophrenia research*, vol. 215, pp. 430–438, 2020.

[20] M. N. I. Qureshi, J. Oh, D. Cho, H. J. Jo, and B. Lee, “Multimodal discrimination of schizophrenia using hybrid weighted feature concatenation of brain functional connectivity and anatomical features with an extreme learning machine,” *Frontiers in neuroinformatics*, vol. 11, p. 59, 2017.

[21] P. Orban, C. Dansereau, L. Desbois, V. Mongeau-Pérusse, C.-É. Giguère, H. Nguyen, A. Mendrek, E. Stip, and P. Bellec, “Multisite generalizability of schizophrenia diagnosis classification based on functional brain connectivity,” *Schizophrenia research*, vol. 192, pp. 167–171, 2018.

[22] K. Fukushima and S. Miyake, “Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition,” in *Competition and cooperation in neural nets*, pp. 267–285, Springer, 1982.

[23] “Simple introduction to convolutional neural networks | by matthew stewart, phd researcher | towards data science.” `https://towardsdatascience.com/simple-introduction-to-convolutional-neural-networks-cdf8d3077bac`. (Accessed on 03/17/2021).

[24] T. Gong, Y. Kim, J. Shin, and S.-J. Lee, “Metasense: few-shot adaptation to untrained conditions in deep mobile sensing,” in *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*, pp. 110–123, 2019.

[25] Z. Liu, R. Zhang, Y. Song, and M. Zhang, “When does maml work the best? an empirical study on model-agnostic meta-learning in nlp applications,” *arXiv preprint arXiv:2005.11700*, 2020.

[26] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *International Conference on Machine Learning*, pp. 1126–1135, PMLR, 2017.

[27] M. Bontonou, N. Farrugia, and V. Gripon, “Few-shot learning for decoding brain signals,” *arXiv preprint arXiv:2010.12500*, 2020.

[28] J. D. Cooper, S. Y. S. Han, J. Tomasik, S. Ozcan, N. Rustogi, N. J. van Beveren, F. M. Leweke, and S. Bahn, “Multimodel inference for biomarker development: an application to schizophrenia,” *Translational psychiatry*, vol. 9, no. 1, pp. 1–10, 2019.

[29] R. L. Bluhm, E. A. Osuch, R. A. Lanius, K. Boksman, R. W. Neufeld, J. Théberge, and P. Williamson, “Default mode network connectivity: effects of age, sex, and analytic approach,” *Neuroreport*, vol. 19, no. 8, pp. 887–891, 2008.

[30] J. Faber and L. M. Fonseca, “How sample size influences research outcomes,” *Dental press journal of orthodontics*, vol. 19, no. 4, pp. 27–29, 2014.

[31] K. S. Button, J. P. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. Robinson, and M. R. Munafò, “Power failure: why small sample size undermines the reliability of neuroscience,” *Nature reviews neuroscience*, vol. 14, no. 5, pp. 365–376, 2013.

[32] S.-H. Wang, Y.-D. Lv, Y. Sui, S. Liu, S.-J. Wang, and Y.-D. Zhang, “Alcoholism detection by data augmentation and convolutional neural network with stochastic pooling,” *Journal of medical systems*, vol. 42, no. 1, pp. 1–11, 2018.

[33] J. H. Koskinen, G. L. Robins, and P. E. Pattison, "Analysing exponential random graph (p-star) models with missing data using bayesian data augmentation," *Statistical Methodology*, vol. 7, no. 3, pp. 366–384, 2010.

[34] D. B. Wright, K. London, and A. P. Field, "Using bootstrap estimation and the plug-in principle for clinical psychology data," *Journal of Experimental Psychopathology*, vol. 2, no. 2, pp. 252–270, 2011.

[35] J. M. Mativo and S. Huang, "Prediction of students' academic performance: Adapt a methodology of predictive modeling for a small sample size," in *2014 IEEE Frontiers in Education Conference (FIE) Proceedings*, pp. 1–3, IEEE, 2014.

[36] "ashwinahuja/mamlforpsychiatry: Applying model-agnostic meta learning for deep learning schizophrenia diagnosis." `https://github.com/ashwinahuja/MAMLforPsychiatry`. (Accessed on 03/17/2021).