# Using magnetic resonance imaging to distinguish a healthy brain from a bipolar brain: A transfer learning approach

Martyn, P[1]., McPhilemy, G[2]., Nabulsi, L[2]., Martyn, F.M.[2], Hallahan, B[2]., McDonald, C[2]., Cannon, D.M[2]., Schukat, M[1].

[1] College of Engineering and Informatics, National University of Ireland Galway, H91 TK33 Galway, Ireland
[2] Centre for Neuroimaging & Cognitive Genomics (NICOG), Clinical Neuroimaging Laboratory, NCBES Galway Neuroscience Centre, College of Medicine Nursing and Health Sciences, National University of Ireland Galway, H91 TK33 Galway, Ireland.

**Abstract.** Bipolar Disorder (BD) is a recurrent psychiatric condition characterised by periods of depression and (hypo)mania, it affects more than 1% of the world's population [1]. However, accurate diagnosis can be difficult due to the lack of diagnostic tools available to practitioners. To address this knowledge gap this paper aims to understand how the application of transfer learning, in the context of machine learning techniques, can be used to improve a diagnosis of BD.

Image detection of magnetic resonance images (MRI) was undertaken to identify features of grey matter in BD brains in comparison to healthy controls (HC), which may constitute a biomarker of BD. Additionally, the products of machine learning were investigated for clinical application to efficiently aid in clinical diagnosis by an end user, through a cloud-based application.

The transfer learning model created demonstrated at 88% accuracy the ability to detect features present in the BD brain, not present in controls. Of limitation to this study was the amount of MR images required to train this model. However, this project identifies that it is possible with limited resources to create a model which may prove useful in diagnostic settings in the future.

**Keywords:** Transfer Learning, Bipolar Disorder, Software Engineering

## 1 Introduction

Bipolar disorder (BD) is a mood disorder associated with recurrent shifts in mood alternating between depressive and (hypo)manic episodes as well as disruptions to cognitive function, activity levels, and sleep [1]. These functional alterations have been associated with morphological changes in the grey and white matter structures of the brain [2]. Robust findings of reductions of cortical thickness and volume reduction in subcortical structures, particularly within the inferior frontal gyrus, cingulate gyrus, middle frontal gyrus, hippocampus, amygdala, and thalamus have been demonstrated [3]. However, while these alterations have been replicated in a number of studies, they are not present in all [3]. A number of the aforementioned cortical and subcortical areas are involved in the cortico-limbic system, a circuit responsible for emotional recognition, response, and regulation. Of particular note are the hippocampus, thalamus, amygdala, and cingulate cortex.

### 1.1 Related Work

Machine Learning (ML) through the use of Support Vector Machines (SVM), or Convolutional Neural Networks (CNN) has been used extensively as a method of diagnosing brain disorders through analysing Magnetic Resonance Imaging (MRI). Nunes *et al.* [4] outline an approach using SVM to diagnose BD. A combination of validation techniques was used to examine differences in each method. A meta-analysis of sample-level classifiers was used which generated results of 42.26% and 59.14% for sensitivity and specificity. A Leave-One-Site-Out (LOSO) validation method was also used which generated results of 58.67% for accuracy, while also generating 51.99% for sensitivity and 64.85% for specificity. The final method was an aggregate subject-level validation which generated 65.23% for accuracy, 66.02% for sensitivity and 64.90% for specificity. The LOSO and aggregate validation also generated receiver operating characteristic curve values of 60.92% and 71.49%, respectively.

Iidaka [5] developed a model to diagnose patients with autism spectrum disorder (ASD). The model used a probabilistic neural network (PNN) to classify patients based on certain biomarkers found in resting state functional MRI. The study used a sample size of 312 subjects with ASD and 328 with typical development, taken from the Autism Brain Imaging Data Exchange. The PNN was shown to have an approximately 90% accuracy rate. The study relied upon the notion that "intrinsic connectivity between subdivisions of the brain is altered in patients with ASD compared to controls", [5] which can be analysed in resting state fMRI. The process taken within that study involved a sequence of spatially realigning volumes to the mean volume and temporally realigning the signal within each slice of MRI to that obtained in the middle slice using Whittaker–Shannon (aka sinc) interpolation. These re-sliced volumes were then normalized to the Montreal Neurological Institute space with a voxel size of 3 x 3 x 3mm$^3$. The normalized images were then spatially smoothed using a Gaussian kernel. The model used in the study was a PNN, described as "a PNN (Specht, 1990) is an implementation of the kernel discriminant analysis statistical algorithm, which is organized into a multilayered feed forward neural network to perform classification" [5]. The PNN consists of 4 fully interconnected layers, the input layer, a pattern layer, a summation later and an output layer. The input layer has an equal number of nodes to features and is responsible for distributing input vectors to the pattern layer. The pattern layer takes the inputs and estimates the probability density function, which in the Iidaka study, a Gaussian function was used. The pattern layer outputs to nodes in the summation layer based on the different classes in the model, in this case whether ASD or a control. The output layer then outputs values corresponding to the most appropriate choice from the current data based on the maximum probability or Bayer's rule. The results of the model relied on both Leave One Out Cross Validation (LOOCV) and V-fold cross validation as the validation methods. For LOOCV the accuracy results were 89.4% while the V-fold was 77.2%, 86.9% and 90.3% for 2-fold, 10-fold and 50-fold, respectively. An investigation into confounding factors was also taken into account with similar accuracy values for groups solely related to subjects on-medication, off-medication and subjects who exhibited head movement while the MRI was being taken.

Transfer Learning is the process of taking a pre-trained model and refitting the model's final predictive layer within a CNN in order for it to be redefined to another task. This allows for a much smaller training dataset to be used within the training process. Hon and Khan [6] used a transfer learning method to identify AD by analysing MRI. They used two pre-trained models, the VGG16 CNN and the Inception CNN. They were re-trained using a training set taken from the Open Access Series of Imaging Studies (OASIS) consisting of 416 subjects, made up of both AD and healthy control (HC) subjects. The data was trimmed down to a random selection of 100 AD and 100 HC subjects. The data was then sorted into 32 of the most informative images from the axial plane of the 3D training images, resulting in 6400 images. The average accuracy of these models was 74.12% for the VGG untrained, 92.3% for the trained VGG, and 96.25% for the trained Inception model.

The basic architecture of a CNN is based around a series of stages or layers. As described in Deep Learning by LeCun *et al.*, [7] we see a description of a typical CNN architecture. The initial stages of a CNN architecture are comprised of two types of layers, these are the convolutional layers and pooling layers. The convolutional layers are made up of feature maps. Inside of these, individual nodes connect to local patches of the feature maps from the previous layer through a set of weights called a filter bank. The sum of all these weights are then passed through a non-linearity such as a rectified linear unit (ReLU). A ReLU is an activation function that computes $f(x) = \max(0,x)$. In other terms it thresholds values at 0, outputting 0 when $x < 0$, and outputting a linear function when $x \geq 0$ [8]. Feature maps in a layer share filter banks, while different feature maps use different filter banks. This is due to the fact that, in images, local motifs usually become evident which results in groupings of values in array data. Also, these motifs are invariant to location so they could appear in any area of the image. So, this means that individual units across the scope of image would need to share weights to detect these individual motifs. The pooling layer in a CNN is tasked with merging semantically similar features into one. Typically, this will be done by computing the maximum of a local patch of units in one or multiple feature maps. A CNN will usually be comprised of multiples of the convolutional, non-linearity and pooling layers, followed by more convolutional and fully-connected layers [7].

Training within a CNN happens through a process called backpropagation using stochastic gradient descent. "The backpropagation procedure to compute the gradient of an objective function with respect to the weights of a multilayer stack of modules is nothing more than a practical application of the chain rule for derivatives" [7]. Backpropagation computes the gradients by working backwards from the gradient with respect to the outputs of that module, repeatedly applying this process to propagate gradients though all modules from output to input.

Cross-validation is the process of estimating the accuracy of a prediction model. In Iidaka [5] a description of the Leave One Out Cross validation (LOOCV) method is shown. This method removes one subject's data from the testing dataset and uses the remainder for training a neural network. The removed data is then used to predict the outcome of the removed dataset to analyse the accuracy of the model. The LOOCV validation method is commonly used in MRI based models [9] [10]. Also in Iikada [5] we see a description of another validation method, V-fold cross- validation. With this validation method, a model is built with (V-1)/V proportion of the subjects used. The remaining 1/V is then used to validate the model.

## 2 Methods

The output of this research involved the creation of a transfer learned convolutional neural network (CNN), using BD and HC MRI data, as well as a deployable software application with which to give access to the prediction capabilities of the CNN to an end user. During the course of the transfer learning training, metrics were used to gather data on the accuracy of the trained model. A full software development lifecycle was also employed during the course of the development of the software application. This included design, project management, development, testing and deployment of the application.

The methodology used in the research was quantitative in nature as the essence of the study was to produce new data to show that transfer learning is a viable option in BD diagnostic tooling. The research carried out involved establishing accuracy data from CNN training using transfer learning. This would give an indication as to the veracity of the claim that transfer learning can be used in diagnosing BD using MRI data. The process of creating an application to provide a mechanism for a user to accept the predictive capabilities of the model also produces data, but possibly of a more abstract nature. This would include overall performance of the application in use but also the overall user experience of the application.

### 2.1 Transfer Learning

The transfer learning portion of the project was achieved using the TensorFlow machine learning backend as well as the Keras library. The code for the ML part of the project was written in Python. Keras supports multiple languages but Python is probably the most well supported language for machine learning. This meant it was a good choice for the project. The general design of the data augmentation and model training was taken from a previous study involving transfer learning and MRI prediction of Alzheimer's Disease MRI data [6]. The approach taken in this study led to very high accuracy, so it was deemed a good idea to follow a similar approach. This also mirrored similar methods outlined by Francois Chollet, the author of Keras, in a Keras tutorial [12] describing transfer learning.

A high-level view of the transfer learning process involved a step-by-step process of data acquisition, data augmentation, training and testing. This project took a staged approach to this process, whereby at different times different levels of data were using in training, mainly due to difficulty in accessing a large dataset. As well as this, different levels of data augmentation were used to try to overcome this data problem. In the end, the only real solution was to source more data. The results of this varying dataset show an interesting difference in accuracy. This highlights the importance of the size and quality of a dataset when conducting machine learning experimentation. The different attempts made during this project to utilise the varying size and quality datasets shows this to hold true.

The training process involved heavily leaning on the functions provided by the Keras library, which included functions to ingest data into a format that the TensorFlow engine could then use while training. Other functions involved making predictions using the trained CNN as well as providing the base model for the transfer learning process. The base model used in the project was the VGG16 model, created by K. Simonyan and A. Zisserman of Oxford University for the ImageNet competition [13]. Keras comes with this model as standard, making accessing the model very simple.

Other Python packages were also used during the machine learning portions of the paper. These included the PILLOW and Open-CV packages for image processing, and the Nibabel package for interacting with NII files. Med2image was used for splitting NII files into individual JPEG images, while the DeepBrain package was used for extracting the brain structure from the non-brain parts of the MRI images. These packages allowed for quicker development than rewriting the same functionality from the ground up.

During the early phases of transfer learning research, it was necessary to format the available data into the required directory structures in order for Keras to be able to ingest the image data. This was performed by a combination of Bash and Python scripts. These scripts created the directory structure, converted the MRI data to JPG, augmented the data and saved the JPG images required for later training.

## 2.2 Data

### Data Acquisition

The data used during the training of the CNN came from two different sources. These were the Anatomy Department of NUI, Galway and the OpenNeuro database [14]. Originally it was hoped that the data from NUI, Galway would be sufficient but during the course of training the CNN it became clear that this would not be the case. Some interesting results were elicited from the earlier training sessions using just the NUI, Galway dataset. These results showed the necessity of a sufficiently large data set for machine learning. The OpenNeuro data was sourced to allow for better results from the training.

Some issues did arise from having two separate datasets. One of these issues was that the data was sufficiently different, visually, to necessitate different approaches to data augmentation. The data from OpenNeuro was of a slightly lower quality and of a different image size compared to the NUI, Galway images. This meant that the data augmentation scripts needed to be adjusted for the differently sourced datasets.

One other attempt was made to get more data through the COINS online dataset but the BD MRIs that were present in that dataset were of the sagittal plane and not the axial (transverse) plane like all the other MRI. This was then deemed unusable in the model and thus discarded.

In relation to the NUI Galway dataset, MR images were obtained from 98 individuals aged 16-60 years of age as part of the Clinical Neuroimaging Laboratory Research Programme. All participants provided written informed consent for the relevant studies and ethical approval was obtained from the NUI Galway and Galway University Hospitals Research Ethics Committees. As OpenNeuro was from a public dataset all that was required to make a request to use the data and it was made available. A further set of images was obtained from those used in the Hon & Khan [6] study. These originally came from the Open Access Series of Imaging Studies, a publicly accessible dataset.

#### Datasets

The NUI, Galway data was made up of 42 BD MRI and 56 HC MRI. The OpenNeuro data consisted 49 MRI scans all from individuals with BD. This left a need for extra HC MRI images. To fill this need some of the images from the Hon & Khan [6] study were used to supplement the HC images used in this study. Similar pre-processing and training strategies were utilized in the datasets of this study and the Hon & Khan study. These included similar sizing, cropping, contrast adjustment and brain extraction and so the images used would suit for both studies. The quality of the OpenNeuro dataset was slightly lower than that of the NUI, Galway dataset, with more severe evidence of ghosting and distortion on the images. Most of this was towards the beginning of the MR images and so was usually discarded during the image selection procession.

## 3 Results

### 3.1 Initial Investigation

The training of the CNN can largely be broken into two phases, pre-OpenNeuro data being added and post-OpenNeuro data being added. The results of the initial training sessions before OpenNeuro, essentially showed how the size of the data set was not sufficient to produce any learning in the model. The results showed overfitting in the model with validation accuracy repeatedly returning ~50% and high training accuracy. When the resulting models were used in prediction scenarios the results were always the first class-label entered into the model when compiling, prior to training, showing that the model was constantly picking the first class for whatever it saw every time during training, which explains the 50% accuracy.

The sequence of initial training attempts went as follows and is summarized in Table 1 but the general takeaway from the initial training attempts is that more data was required for the training to be successful. All the of the initial nine attempts at training with the initial showed similar results, hovering around ~50%. Between the attempts variations in the training strategy were employed such as including data augmentation, changing the optimizer during training and adding further layers to the top model of the CNN, or adding weight regularizers to these layers was also employed. None of these strategies was successful at increasing the validation accuracy of the model.

### 3.2 Full Dataset

Following the last attempt at training it was clear that the amount of data was insufficient. This meant that enlarging and retraining on the larger dataset was necessary. Following the inclusion of the OpenNeuro data, a massive improvement in the training results was elicited. After just one round of training the validation accuracy went up to ~75%. This showed a reduction in overfitting. The validation accuracy improved after the first 10-20 epochs and then stayed static hovering around 70-75% accuracy. The validation loss also showed signs of improvement with values varying between 0.8 and 1.2. The validation loss also increased the further into the training the process went.

**Table 1**. Summary of training attempts

| Training Attempt | Validation Accuracy % | Notes |
|---|---|---|
| 1 | ~50 | Initial setup. |
| 2 | ~50-55 | Changed optimiser to Adam. |
| 3-9 | ~50 | Increasing levels of image pre-processing and data augmentation to increase size of dataset. Adjustments to optimisers. |
| 10 | ~50 | Added more layers to CNN top model with added weight regularizers. |
| 11 | ~75 | Combined the NUI, Galway dataset with OpenNeuro dataset. Large jump in validation accuracy. |

Following the inclusion of the OpenNeuro data, another change was made to the data. This was to heighten the contrast of each MRI, in order to increase the detail of the image. Also, a batch normalisation layer was included in the top model. Following these changes and a new round of training of the image data, validation accuracy went up to approximate average of ~77%. This configuration of was then extended to a full 5-fold cross-validation. The 5-fold cross-validation results (Table 2) show the difference in validation accuracy across the training datasets splits. The results from these training can then be calculated to find the average total accuracy of the model of 88% across the 5 folds. Due to some considerable threshing across the different folds, it might have been useful to perform 10-fold cross validation. This is something which could be taken into consideration in further future experimentation.

Tests were performed on the model produced from the final training run and the results were promising. Due to the small dataset, a decision was made to only hold back 3 BD MRIs for separate testing. Several control MRIs were also available as this type of MRI outnumbered the BD MRIs. The results from predicting show that the majority of the selected image slices from the BD MRIs returned Bipolar as the prediction-class, although the prediction probability value was sometimes very low and erratic. This may be the result of some error in the process or the fact the model's probability of it being BD was just deemed low. Some prediction runs returned the opposite results as expected which is not a promising result. This shows some of the downsides of the current model. Some results also returned false positives from control MRIs.

### 3.3 Application Performance

The performance of the application when run over the network on Amazon AWS, is similar to the performance seen when tested locally. The main difference seen in the time from request start to finish is the time taken to upload the MRI from the client to the server. This time is between 5-10 seconds, depending upon the quality of the network at the time. The actual processing time of a request remains the same as a locally tested request at ~15 seconds.

The overall user experience of the application is a largely subjective metric but to due to the simplicity and general stability of the program experienced during testing, the overall user experience is generally good. The major aspects negatively impacting user experience are the wait times during the processing of the MRI, as well as false alarms and general inconsistent hit rates in some results. If these were reduced, user experience would be greatly improved. The majority of the processing time is spent on the image pre-processing and the predicting itself. This is largely due to the number of individual tasks that need to happen during these processes, including I/O and spooling up of TensorFlow backends. These processes take time and as such means that the frontend client may have to wait a long time. These would be interesting areas for future research to improve the quality of the application

## 4 Discussion

The primary outcome from the results of the transfer learning training is that for predicting BD, using transfer learning is at least a semi-viable approach. The literature suggests that certain morphology exists in the brain of BD sufferers distinct from that of a healthy control [16] [17] and from the results seen in this paper it is evident that transfer learning can be used to detect this morphology to a certain degree. From the validation accuracy of 88% with the limited dataset at hand, this shows that with further data and more time spent on improving the quality of the model or using a different base model, higher accuracy could be achieved. This is shown by the results (Table 3) seen in the Hon & Khan [6] study, which served as the basis for this project, the results attained by a similar process were markedly higher. One reason for this higher accuracy could be down to the dataset. The amount of data used by Hon and Khan in their study was slightly larger, by roughly 10%, and their results show accuracy of 92.3% using the same VGG16 model. When they used the Inception V4 model the average accuracy went up to 96.25%. We can see here that if there was a larger dataset of BD MRI then it may be possible to achieve similar levels of accuracy with a model trained using transfer learning to detect BD. Another aspect of their data was homogeneity. The data came from a single source and was very similarly processed prior to ingestion into the model for training whereas the data used in this study came from two sources and needed to be processed in different ways resulting in slightly different images being entered into the model when training.

**Table 2**. Results from Hon & Khan [6] study

| Model | Avg. Acc. (st. dev.) (%) |
|---|---|
| VGG16 (from scratch) | 74.12 (1.55) |
| VGG16 (transfer learning) | 92.3 (2.42) |
| Inception V4 (transfer learning) | 96.25 (1.2) |

Another set of results with which to compare the results attained in this study is from Nunes *et al* [4]. This study used support vector machines rather than a transfer learned CNN for BD classification, and highlights differences between the capabilities of the different classification methods. The results of that study used a Leave-one-site-out (LOSO) cross-validation mechanism returning an accuracy of 58.70% and an aggregate subject-level accuracy of 65.23%. This is a difference of ~30% for LOSO cross-validation and ~23% for aggregate subject-level accuracy compared to the results seen in this paper. This may be due to the heterogenous multisite dataset used in that study. A homogenous dataset may improve accuracy within that single dataset, but increased variation may lead to an overall decline in accuracy as patterns become harder to be learned by the model.

Classification errors between the BD and HC groups during image prediction could be related to certain confounding factors such as medication being taken by the subjects at the time the images were taken. For instance, Lithium use has been associated with increased thickness of the anterior cingulate cortex [18] when taken as a treatment for BD. The cingulate is otherwise thinner in BD, so medication is sometimes used to correct this [19]. This could then lead to errors in classification.

In conclusion, we have shown that transfer learning applied to feature detection in BD does return positive results, while not being totally conclusive in its intended outcome. In terms of future research, the project shows conclusively that further investigation into this area is warranted.

## References

1. Grande, I., Berk, M., Birmaher, B., Vieta, E.: Bipolar disorder. The Lancet. 387, 1561–1572 (2016). https://doi.org/10.1016/S0140-6736(15)00241-X.
2. Hibar, D.P., Westlye, L.T., van Erp, T.G.M., Rasmussen, J., Leonardo, C.D., Faskowitz, J., Haukvik, U.K., Hartberg, C.B., Doan, N.T., Agartz, I., Dale, A.M., Gruber, O., Krämer, B., Trost, S., Liberg, B., Abé, C., Ekman, C.J., Ingvar, M., Landén, M., Fears, S.C., Freimer, N.B., Bearden, C.E., the Costa Rica/Colombia Consortium for Genetic Investigation of Bipolar Endophenotypes, Sprooten, E., Glahn, D.C., Pearlson, G.D., Emsell, L., Kenney, J., Scanlon, C., McDonald, C., Cannon, D.M., Almeida, J., Versace, A., Caseras, X., Lawrence, N.S., Phillips, M.L., Dima, D., Delvecchio, G., Frangou, S., Satterthwaite, T.D., Wolf, D., Houenou, J., Henry, C., Malt, U.F., Bøen, E., Elvsåshagen, T., Young, A.H., Lloyd, A.J., Goodwin, G.M., Mackay, C.E., Bourne, C., Bilderbeck, A., Abramovic, L., Boks, M.P., van Haren, N.E.M., Ophoff, R.A., Kahn, R.S., Bauer, M., Pfennig, A., Alda, M., Hajek, T., Mwangi, B., Soares, J.C., Nickson, T., Dimitrova, R., Sussmann, J.E., Hagenaars, S., Whalley, H.C., McIntosh, A.M., Thompson, P.M., Andreassen, O.A.: Subcortical volumetric abnormalities in bipolar disorder. Molecular Psychiatry. 21, 1710 (2016).
3. Hajek, T., McIntyre, R., Alda, M.: Bipolar disorders, type 2 diabetes mellitus, and the brain: Current Opinion in Psychiatry. 29, 1–6 (2016). https://doi.org/10.1097/YCO.0000000000000215.
4. Nunes, A., Schnack, H.G., Ching, C.R.K., Agartz, I., Akudjedu, T.N., Alda, M., Alnæs, D., Alonso-Lana, S., Bauer, J., Baune, B.T., Bøen, E., Bonnin, C. del M., Busatto, G.F., Canales-Rodríguez, E.J., Cannon, D.M., Caseras, X., Chaim-Avancini, T.M., Dannlowski, U., Díaz-Zuluaga, A.M., Dietsche, B., Doan, N.T., Duchesnay, E., Elvsåshagen, T., Emden, D., Eyler, L.T., Fatjó-Vilas, M., Favre, P., Foley, S.F., Fullerton, J.M., Glahn, D.C., Goikolea, J.M., Grotegerd, D., Hahn, T., Henry, C., Hibar, D.P., Houenou, J., Howells, F.M., Jahanshad, N., Kaufmann, T., Kenney, J., Kircher, T.T.J., Krug, A., Lagerberg, T.V., Lenroot, R.K., López-Jaramillo, C., Machado-Vieira, R., Malt, U.F., McDonald, C., Mitchell, P.B., Mwangi, B., Nabulsi, L., Opel, N., Overs, B.J., Pineda-Zapata, J.A., Pomarol-Clotet, E., Redlich, R., Roberts, G., Rosa, P.G., Salvador, R., Satterthwaite, T.D., Soares, J.C., Stein, D.J., Temmingh, H.S., Trappenberg, T., Uhlmann, A., Haren, N.E.M. van, Vieta, E., Westlye, L.T., Wolf, D.H., Yüksel,

D., Zanetti, M.V., Andreassen, O.A., Thompson, P.M., Hajek, T.: Using structural MRI to identify bipolar disorders – 13 site machine learning study in 3020 individuals from the ENIGMA Bipolar Disorders Working Group. Molecular Psychiatry. 1 (2018). https://doi.org/10.1038/s41380-018-0228-9.

5. Iidaka, T.: Resting state functional magnetic resonance imaging and neural network classified autism and control. Cortex. 63, 55–67 (2015). https://doi.org/10.1016/j.cortex.2014.08.011.
6. Hon, M., Khan, N.M.: Towards Alzheimer's disease classification through transfer learning. In: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 1166–1169. IEEE, Kansas City, MO (2017). https://doi.org/10.1109/BIBM.2017.8217822.
7. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature. 521, 436–444 (2015). https://doi.org/10.1038/nature14539.
8. Agarap, A.F.: Deep Learning using Rectified Linear Units (ReLU). arXiv:1803.08375 [cs, stat]. (2018).
9. Fu, C.H.Y., Mourao-Miranda, J., Costafreda, S.G., Khanna, A., Marquand, A.F., Williams, S.C.R., Brammer, M.J.: Pattern Classification of Sad Facial Processing: Toward the Development of Neurobiological Markers in Depression. Biological Psychiatry. 63, 656–662 (2008). https://doi.org/10.1016/j.biopsych.2007.08.020.
10. Zeng, L.-L., Shen, H., Liu, L., Wang, L., Li, B., Fang, P., Zhou, Z., Li, Y., Hu, D.: Identifying major depression using whole-brain functional connectivity: a multivariate pattern analysis. Brain. 135, 1498–1507 (2012). https://doi.org/10.1093/brain/aws059.
11. Sabahi, F.: Secure Virtualization for Cloud Environment Using Hypervisor-based Technology. IJMLC. 39–45 (2012). https://doi.org/10.7763/IJMLC.2012.V2.87.
12. Chollet, F.: Building powerful image classification models using very little data, https://blog.keras.io/building-powerful-image-classification-models-using-very-little-data.html, last accessed 2019/08/01.
13. VGG16 - Convolutional Network for Classification and Detection, https://neurohive.io/en/popular-networks/vgg16/, last accessed 2019/07/23.
14. OpenNeuro, https://openneuro.org/, last accessed 2019/07/28.
15. Brownlee, J.: How to Use Weight Decay to Reduce Overfitting of Neural Network in Keras, https://machinelearningmastery.com/how-to-reduce-overfitting-in-deep-learning-with-weight-regularization/, last accessed 2019/08/21.
16. Hajek, T., Cullis, J., Novak, T., Kopecek, M., Blagdon, R., Propper, L., Stopkova, P., Duffy, A., Hoschl, C., Uher, R., Paus, T., Young, L.T., Alda, M.: Brain Structural Signature of Familial Predisposition for Bipolar Disorder: Replicable Evidence For Involvement of the Right Inferior Frontal Gyrus. Biological Psychiatry. 73, 144–152 (2013). https://doi.org/10.1016/j.biopsych.2012.06.015.
17. Hibar, D.P., Doan, N.T., Jahanshad, N., Cheung, J.W., Ching, C.R.K., Versace, A., Bilderbeck, A.C., Uhlmann, A., Mwangi, B., Krämer, B., Overs, B., Hartberg, C.B., Abé, C., Dima, D., Grotegerd, D., Sprooten, E., Bøen, E., Jimenez, E., Howells, F.M., Delvecchio, G., Temmingh, H., Starke, J., Almeida, J.R.C., Goikolea, J.M., Houenou, J., Beard, L.M., Rauer, L., Abramovic, L., Bonnin, M.,

Ponteduro, M.F., Keil, M., Rive, M.M., Yao, N., Yalin, N., Najt, P., Rosa, P.G., Redlich, R., Trost, S., Hagenaars, S., Fears, S.C., Alonso-Lana, S., van Erp, T.G.M., Nickson, T., Chaim-Avancini, T.M., Meier, T.B., Elvsåshagen, T., Haukvik, U.K., Lee, W.H., Schene, A.H., Lloyd, A.J., Young, A.H., Nugent, A., Dale, A.M., Pfennig, A., McIntosh, A.M., Lafer, B., Baune, B.T., Ekman, C.J., Zarate, C.A., Bearden, C.E., Henry, C., Simhandl, C., McDonald, C., Bourne, C., Stein, D.J., Wolf, D.H., Cannon, D.M., Glahn, D.C., Veltman, D.J., Pomarol-Clotet, E., Vieta, E., Canales-Rodriguez, E.J., Nery, F.G., Duran, F.L.S., Busatto, G.F., Roberts, G., Pearlson, G.D., Goodwin, G.M., Kugel, H., Whalley, H.C., Ruhe, H.G., Soares, J.C., Fullerton, J.M., Rybakowski, J.K., Savitz, J., Chaim, K.T., Fatjó-Vilas, M., Soeiro-de-Souza, M.G., Boks, M.P., Zanetti, M.V., Otaduy, M.C.G., Schaufelberger, M.S., Alda, M., Ingvar, M., Phillips, M.L., Kempton, M.J., Bauer, M., Landén, M., Lawrence, N.S., van Haren, N.E.M., Horn, N.R., Freimer, N.B., Gruber, O., Schofield, P.R., Mitchell, P.B., Kahn, R.S., Lenroot, R., Machado-Vieira, R., Ophoff, R.A., Sarró, S., Frangou, S., Satterthwaite, T.D., Hajek, T., Dannlowski, U., Malt, U.F., Arolt, V., Gattaz, W.F., Drevets, W.C., Caseras, X., Agartz, I., Thompson, P.M., Andreassen, O.A.: Cortical abnormalities in bipolar disorder: an MRI analysis of 6503 individuals from the ENIGMA Bipolar Disorder Working Group. Mol Psychiatry. 23, 932–942 (2018). https://doi.org/10.1038/mp.2017.73.

18. Emsell, L., McDonald, C.: The structural neuroimaging of bipolar disorder. International Review of Psychiatry. 21, 297–313 (2009). https://doi.org/10.1080/09540260902962081.
19. Machado-Vieira, R., Manji, H.K., Zarate Jr, C.A.: The role of lithium in the treatment of bipolar disorder: convergent evidence for neurotrophic effects as a unifying hypothesis. Bipolar Disorders. 11, 92–109 (2009). https://doi.org/10.1111/j.1399-5618.2009.00714.x.