Predicting Emotions from Brain Data Using Various Machine Learning Models

Flore Barde, Manon Béchaz, Laura Reimoser
CS-433 Project 2, EPFL

***Abstract*—A fundamental goal of affective neuroscience is to understand how emotions are represented in the brain. To this end, different machine learning methods were used to map functional magnetic resonance imaging (fMRI) data to associated emotions. The Random Forest Classifier has proven to be the most accurate, achieving an accuracy of 0.33%. In parallel, the machine learning techniques have allowed us to investigate the importance of different brain regions in the prediction of emotions. From a neuroscience perspective, the results are coherent.**

I. Introduction

Understanding how emotions are represented in the brain is a central aim of affective neuroscience. In fact, despite decades of neuroimaging research, it's still unclear how exactly they appear in the brain. For instance, whether they are represented as discrete categories or as points in a continuous dimensional space is still a topic of contention [1].

Machine learning methods can provide powerful tools to map physical measurements to categories, in that particular case, emotions. It can also be used to assess the importance of each brain region in the prediction of emotions [2]. For that reason, the goal of this work was to use these techniques to address emotion representation in the brain based on 13 categories (e.g., "Fear," "Sadness," and "Happiness") and to build a model capable of predicting these emotions.

The rest of this work is structured in the following way: Section II gives an overview of the data and its preprocessing. Section III briefly explains the models that were applied and how the parameters for those models were selected. Section IV presents and discusses the obtained results, and Section V provides a short conclusion.

II. Data and Preprocessing

A. Data acquisition

The functional MRI of 30 participants (including 18 female, with average age 25.91) has been measured to obtain their blood-oxygen-level-dependent responses while they watched 14 short films. Scanning was performed on a 3 Tesla, Siemens trio system. The films were selected from the LIRIS database [3] based on their rich and diverse emotional content. In a previous experiment, an independent sample continuously annotated the films for 13 discrete emotions.

B. Preprocessing

a) Preprocessing of brain images: Using the Schaefer 400 brain atlas [4] and the Harvard-Oxford atlas as distributed with FSL (FMRIB Software Library) [5], obtained brain images were averaged over 400 (resp. 48) brain regions. These regions were identified as functionally relatively homogenous, allowing for data reduction (from over 20000 data points per time point to 400, resp. 48). Two datasets with 400 and 48 features, respectively, were thus obtained.

b) Preprocessing of annotations files: From the previous study, a ground truth emotion time course was calculated which represent the average emotion rating of at least three independent raters, after standardization using z-scores (mean = 0, std = 1). A predominant emotion was assigned to each time point using a given threshold. For time points where no emotion label reached this threshold the label 'Neutral' was assigned. The distribution of the different emotions obtained for a 0.5 threshold is depicted in Figure 1.

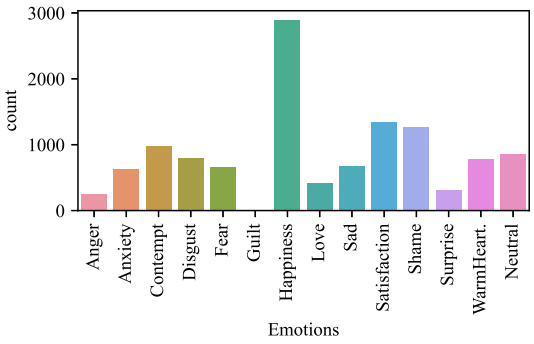

Fig. 1. Emotions present in the dataset, for a threshold of 0.5.

c) Averaging over time: To smooth out the physiological data, both the brain and emotional data were averaged over time. Since emotion annotations were collected at 1/1s, and fMRI data at 1/1.3s, both were averaged over 13 seconds. Due to the relatively slow process of hemodynamic response in the brain (which appears as increased levels of oxygen in brain areas just after neuronal activity), a delay of 6 seconds between the brain and emotion data has also been introduced.

d) Concatenation: After the brain and emotion data were preprocessed, they were merged into a large matrix, containing the brain data for the individual films in blocks and the corresponding emotional data for the films. In total, 11'841 data points were obtained for the dataset with 48 features, and 14'928 for the one with 400 features.

e) Feature engineering: In order to improve the performance of the machine learning models, all feature were normalized to zero mean and unit variance. Other features

engineering methods have been tested (removing correlated features and features with low variance, oversampling of the minority emotion classes, etc), but they did not prove to be helpful and were therefore not retained.

f) Splitting of the data: Finally, as it is usually done in machine learning, datasets were split into training and testing sets. In this particular case, however, the splitting method plays an essential role. Indeed, randomly splitting the data lead to models being trained and tested on very similar data, coming, for example, from the same participant, watching the same film, with just a few seconds difference between the point used for the training and the one for the test. This lead to very good results for the models, but made them unable to anticipate the emotions of new individuals.

A first way to overcome this issue is to simply "truncate" the concatenated matrix after a certain percentage, e.g. to keep the first 90% of the matrix and use it as training data and use the last 10% for testing. However, this has the disadvantage that the model is not trained at all on certain films or certain participants.

The second and most commonly used way to separate the data here is to split each film according to a certain ratio. For example, if a train-test ratio of 90-10 is desired, 90% of the brain and related emotion data of a single film (independently of the subjects) is put into the training set and the other 10% into the test set. This ensures that the model is trained with each film and participant.

III. Methodology

A. Methods

Different methods were tested in the course of this work:

a) Fully Connected Neural Network: An artificial neural network in general is built following the example of natural neural networks. It consists of variously connected neurons. Typically, the neurons are arranged in different layers. The first layer is the input layer, and the last one is the output layer. In between, there are an arbitrary number of hidden layers. The term Fully Connected Neural Network (FCNN) is used when all neurons in one layer are connected to all neurons in the next layer.

In our case, A FCNN with two hidden layers with 6 and 14 nodes, respectively, was used.

b) Convolutional Neural Network: The structure of a classical Convolutional Neural Network (CNN) consists of one or more convolutional layers followed by a pooling layer. In principle, this unit can repeat itself as often as necessary. The final layer is a fully meshed layer. In contrast to FCNN, not all neurons from one layer are connected to all neurons in the next layer in a CNN. The training of a CNN is generally supervised [6]. The CNN used here is presented in Figure 5 of the Appendix. Both NN were implemented using the `tensorflow.keras` library.

c) Decision Tree Classifier: The Decision Tree Classifier (DTC) is a multistage classification method. Multistage, in this case, means that a complex decision is split into several simpler decisions. The solution derived from the simpler decisions then corresponds, in the best case, to the solution of the complex problem. A decision tree generally consists of a root node that represents all possible classes of the classification problem. For decision-making, there exist a number of interior nodes and a number of terminal nodes. The terminal nodes represent the final decision [7], [8]. `sklear DecisionTreeClassifier` was used in this work to implement DTC.

d) Random Forest: A Random Forest is a classification and regression procedure consisting of several individual decision trees. For a classification, each tree in the forest takes a decision on a class. The class with the most votes determines the final classification. Both unweighted and weighted decision-making are possible [9]. `sklearn RandomForestClassifier` was used.

e) Support-Vector Machine: Support-Vector Machines (SVM) divide a set of objects into classes in such a way that the largest possible area around the class boundaries remains free of objects; new objects are mapped and assigned to these classes. SVMs are considered to be very robust among methods based on statistical learning frameworks [10]. To implement this algorithm, `SVC` method from the `sklearn` library was used.

f) K-Nearest-Neighbor-Algorithm: The k-Nearest-Neighbor-Algorithm (kNN) is a classification procedure in which a class is assigned by taking into account the k nearest neighbours. This is a so-called lazy learning algorithm, as the learning consists of simply storing the training examples. Normalising the data can increase the accuracy of this algorithm [11]. `sklearn KNeighborsClassifier` was used in course of this work.

B. Parameter Selection

The best parameters for the different models were found using a grid search and cross validation approach. More specifically, the method `GridSearchCV` from the library `scikit-learn` was used on specified parameters such as the activation function, the batch size, etc.

C. Libraries

Apart from the library previously named `Numpy, pandas, tensorflow` and `sklearn` were used throughout this work. `seaborn` and `nilearn` were also used for vizualization purposes.

IV. Results

A. Influence of the Threshold

To begin with, the threshold's influence on the "Neutral" emotion on the model's accuracy was studied. Tables I and II report the obtained accuracies of the 6 studied methods for thresholds ranging from 0 to 1. The corresponding neutral percentages are also indicated. It can be noticed that until a threshold of 0.7, the models' accuracies are not impacted by the addition of a neutral

	Threshold	0.0	0.1	0.2	0.3	0.4	**0.5**	0.6	0.7	0.8	0.9	1.0
	Neutral percentage [%]	0	0.322	1.447	1.768	5.627	**8.595**	13.773	28.108	36.127	45.123	58.628
Model's accuracy [%]	Convolutional Neural Network	11.481	26.094	7.318	26.094	13.896	**10.786**	11.289	24.037	33.126	41.138	59.626
	Decision Tree Classifier	26.524	26.524	26.524	26.524	22.458	**22.531**	21.431	25.974	35.589	47.644	59.626
	Fully Connected Neural Network	26.453	25.759	25.304	25.233	25.376	**25.376**	20.712	25.974	35.589	47.644	59.626
	K Nearest Neighbours	26.787	26.764	26.620	26.572	26.381	**26.142**	24.539	29.251	36.378	47.644	59.626
	Random Forest Classifier	28.151	27.959	29.107	29.012	26.741	**26.741**	25.401	29.061	38.077	49.055	60.679
	Support Vector Machine	19.421	19.564	19.205	19.373	19.421	**19.277**	20.402	19.373	21.693	22.171	26.285

TABLE I
Accuracy of the models with respect to the threshold used for the neutral state. The data with the 400 brain regions are used here.

	Threshold	0.0	0.1	0.2	0.3	0.4	**0.5**	0.6	0.7	0.8	0.9	1.0
	Neutral percentage [%]	0	0.322	1.447	1.768	5.627	**8.595**	13.773	28.108	36.127	45.123	56.628
Model's accuracy [%]	Convolutional Neural Network	5.458	11.873	14.602	17.024	18.014	**15.455**	23.575	23.132	32.855	47.082	37.598
	Decision Tree Classifier	26.475	26.475	26.475	26.475	26.475	**26.475**	25.895	27.732	35.483	47.082	59.365
	Fully Connected Neural Network	26.475	26.475	26.475	26.475	26.475	**26.475**	25.112	26.066	35.482	47.082	59.365
	K Nearest Neighbours	29.001	29.001	29.001	29.001	28.317	**27.431**	27.567	30.979	37.495	47.014	59.365
	Random Forest Classifier	33.640	32.992	33.641	34.391	33.572	**33.061**	33.163	34.732	40.736	48.857	60.423
	Support Vector Machine	28.283	28.727	29.512	29.887	28.317	**26.646**	27.567	32.207	35.448	45.547	57.898

TABLE II
Accuracy of the models with respect to the threshold used for the neutral state. The data with the 48 brain regions are used here.

emotion. However, after this value, the success rate starts to increase and is almost always equal to the percentage of neutral. This is due to the fact that neutrality becomes predominant over all other emotions. Thus, the models manage to predict it every time but can not find any other emotion. Therefore, a trade-off needs to be found between a reasonable number of neutrals and the ability of the models to predict more than just one emotion. If not stated otherwise, a threshold of 0.5 will be used throughout the rest of this report.

B. Influence of the Parcellation

As stated earlier, two datasets have been created, resulting from the average of the brain images over 400 and 48 regions, respectively. As shown in Tables I and II, averaging over 48 regions produces better results, with an accuracy reaching 33% against 26.7% when averaging over 400 brain regions. In fact, using a coarser parcellation may reduce noise, even tho some data is lost, which makes the data more suitable for machine learning applications.

C. Assessing the Models' Quality

When assessing the quality of the machine learning methods used, some important differences can be noted. First, a significant and surprising difference is observed between the FCNN and the CNN, the latter performing noticeably worse than the other methods. Then, except for the Random Forest Classifier that achieve somewhat better accuracy, all methods are comparable.

Moreover, the performance achieved by the models implemented is coherent with previously published works on the subject. Indeed, previous models had accuracies [1] ranging from 26% to 47% [2], [12], [1]. For most of the methods used in this paper, the obtained performances are within this range.

[1]Note that these values were obtained using different number of emotions, which has to be taken into account when comparing the accuracies.

D. Influence of the Model on the Predicted Emotions

So far, only the overall accuracy of the models has been taken into account. It is, however, interesting to also consider the performance of the models, emotion by emotion. Figure 2 summarises all these results. The precision, computed using the corresponding metric from the `scikit-learn` library, is given for each emotion and model considered, using the 48-regions dataset. Precision refers to the number of true positives divided by the total number of positive predictions [13]:

$$p = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}}},$$

with N_{TP} the number of true positives and N_{FP} the number of false positives. Clearly, precision is not the only metric assessing the quality of a model, but due to space constraints in this work, only precision is presented.

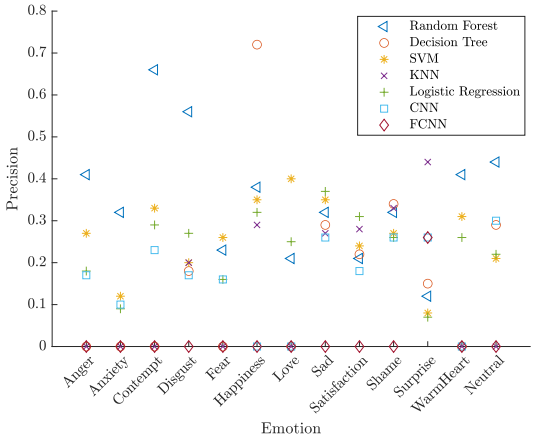

Fig. 2. Scatter plot showing how precisely different models predict emotions. The data with the 48 brain regions are used here, as well as a threshold of 0.5.

The good performance of the Random Forest Classifier is again underlined in this Figure. In comparison, while achieving

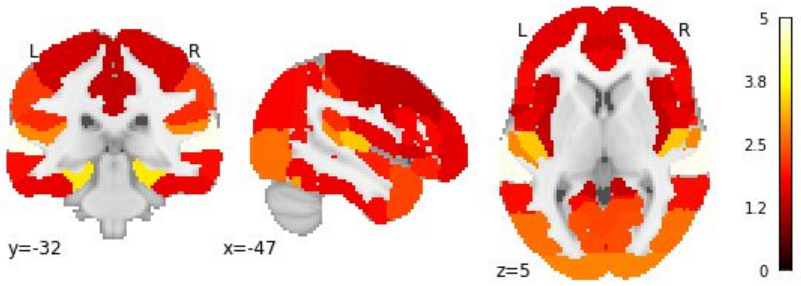

Fig. 3. Brain map presenting the importance of the different brain regions in the prediction of emotions.

the highest precision for 'Happiness', the Decision Tree Classifier scores poorly for other emotions. Similarly, the predictions of the two neural networks (FCNN and CNN) are not very good. The FCNN, in particular, only shows a value that deviates from 0 for the emotion 'Surprise'. Overall, the results are consistent with the accuracies given in Table II.

Secondly, considering more carefully the predictions provided by the Random Forest Classifiers, the confusion matrix provided in Figure 4 was obtained.

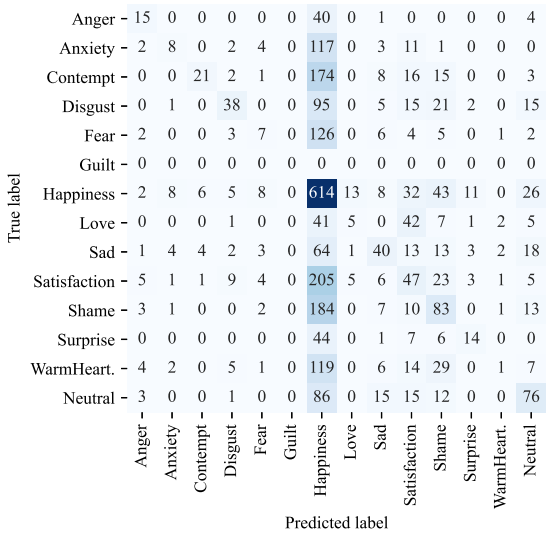

Fig. 4. Confusion Matrix for the Random Forest Classifier. The 48-brain regions dataset is used, along with a threshold of 0.5.

The first thing that stands out is the very large number of 'Happiness' predictions, regardless of the true emotion. This is probably due to the large number of dataset points labeled as 'Happiness'. In fact, this emotion represents more than 24% of labels present in the dataset. This imbalanced class distribution leads to bias towards the majority class, i.e. 'Happiness'. A way to get around this problem is to over-sample, i.e. add copies of instances from the under-represented class. In this case, however, no improvement has been noticed. Under-sampling the majority class has not been considered due to the already limited amount of data available.

Apart from the predominance of 'Happiness' predictions, all other emotions, when not predicted as happiness, are mostly predicted correctly.

E. Feature importance

Computing the feature importance of each of the 48 brain regions allows us to determine how useful they are at predicting emotions. Figure 3 presents a visualisation of the importance of the different brain regions in the prediction of emotions. As a general rule it appears that areas related to visual and auditory processing have stronger predictive value to distinguish emotions. This may be related to the nature of the stimuli being audio-visual films.

Specifically, the five most important areas were the superior Temporal Gyrus, the Parahippocampal Gyrus, the Heschl'sch Gyrus and the Fusiform Cortex. These areas are related to auditory and language processing, social cognition, visual memory and face processing. Thus, it seems coherent for them to play an important role in distinguishing emotions, particularly in the context of watching films. It is somewhat surprising that areas in the frontal lobe that are associated with higher cognitive processing such as decision making, reasoning and planning appeared as relatively unimportant. Finally, the somato-motor cortex had the lowest feature importance, which is coherent as it is related to sensory perception and movement, and is thus not emotion related. This is also consistent with the context of the experiment. Indeed, in the fMRI, participants are not moving or sensing anything, resulting in a minimal variance in these regions.

V. Conclusion

To conclude, machine learning techniques have allowed us to predict emotions from brain images with a maximal accuracy of 33%. Although this result remains small, this accuracy is still much larger than the one expected from random classification (considering 14 emotional states, the accuracy in that case would be 1/14, i.e. 7%), and is in the range of previous attempts [2], [12], [1]. Of the different classifiers in use, Random Forest proved to be the most efficient.

Moreover, from a neuroscience perspective, the results were coherent, and outlined reasonable brain regions as important features for the prediction of emotions.

VI. Acknowledgment

We thank Elenor Morgenroth for her precious supervising and help throughout this work. Also, we would like to thank the Medical Image Processing Lab (MIP:Lab) for hosting our project. Finally, we would like to thank Ilaria Ricchi for her help concerning the machine learning part.

Link to public repository:

https://github.com/CS-433/ml-project-2-flm2-0

References

[1] P. A. Kragel and K. S. LaBar, "Multivariate neural biomarkers of emotional states are categorically distinct," *Social cognitive and affective neuroscience*, vol. 10, no. 11, pp. 1437–1448, 2015.

[2] H. Saarimäki, E. Glerean, D. Smirnov, H. Mynttinen, I. P. Jääskeläinen, M. Sams, and L. Nummenmaa, "Classification of emotion categories based on functional connectivity patterns of the human brain," vol. 247, p. 118800. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811921010715

[3] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen, "Liris-accede: A video database for affective content analysis," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 43–55, 2015.

[4] A. Schaefer, R. Kong, E. M. Gordon, T. O. Laumann, X.-N. Zuo, A. J. Holmes, S. B. Eickhoff, and B. T. T. Yeo, "Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI," *Cerebral Cortex*, vol. 28, no. 9, pp. 3095–3114, 07 2017. [Online]. Available: https://doi.org/10.1093/cercor/bhx179

[5] C. for Morphometric Analysis (CMA), "Harvard-oxford atlas, distributed with the fmrib software library (fsl)." [Online]. Available: https://scalablebrainatlas.incf.org/human/HOA06#howtocite

[6] C. M. Bishop, *Pattern recognition and machine learning*, ser. Computer science. New York, NY: Springer, 2006. [Online]. Available: http://swbplus.bsz-bw.de/bsz250316129cov.htm

[7] S. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.

[8] P. H. Swain and H. Hauska, "The decision tree classifier: Design and potential," *IEEE Transactions on Geoscience Electronics*, vol. 15, no. 3, pp. 142–147, 1977.

[9] F. Livingston, "Implementation of breiman's random forest machine learning algorithm," *Machine Learning Journal Paper.*, vol. Fall 2005. [Online]. Available: http://datajobstest.com/data-science-repo/Random-Forest-[Frederick-Livingston].pdf

[10] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[11] Z. Deng, X. Zhu, D. Cheng, M. Zong, and S. Zhang, "Efficient k nn classification algorithm for big data," *Neurocomputing*, vol. 195, pp. 143–148, 2016.

[12] B. Azari, C. Westlin, A. B. Satpute, J. B. Hutchinson, P. A. Kragel, K. Hoemann, Z. Khan, J. B. Wormwood, K. S. Quigley, D. Erdogmus, J. Dy, D. H. Brooks, and L. F. Barrett, "Comparing supervised and unsupervised approaches to emotion categorization in the human brain, body, and subjective experience," vol. 10, no. 1, p. 20284. [Online]. Available: https://doi.org/10.1038/s41598-020-77117-8

[13] scikit learn, "sklearn.metrics.precision_score," 16.12.2021. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html

Appendix

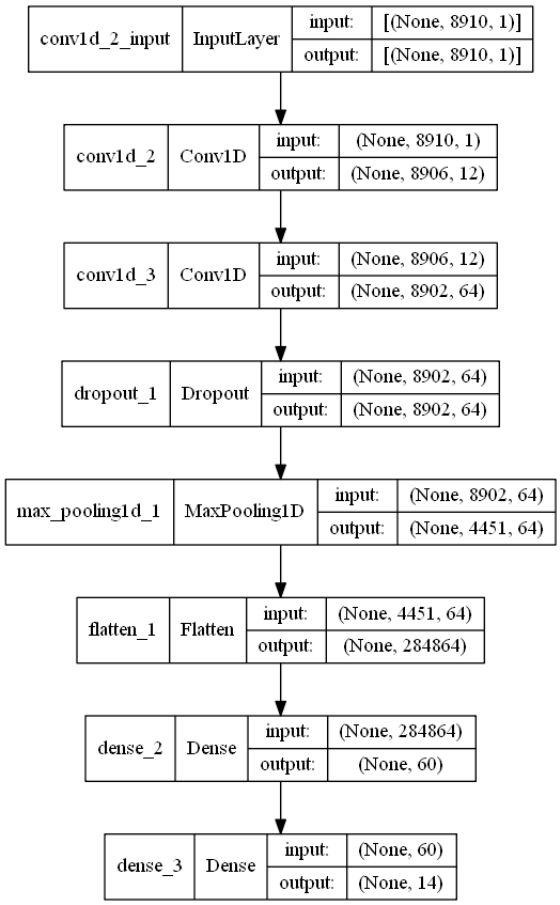

Fig. 5. Convolutional Neural Network architecture used.