# Prediction of activation patterns preceding hallucinations in patients with schizophrenia using machine learning with structured sparsity

Amicie de Pierrefeu, Thomas Fovet, Fouad Hadj-Selem, Tommy Lofstedt, Philippe Ciuciu, Stephanie Lefebvre, Pierre Thomas, Renaud Lopes, Renaud Jardri, Edouard Duchesnay

Received: 5 October 2017 | Revised: 14 December 2017 | Accepted: 2 January 2018

DOI: 10.1002/hbm.23953

RESEARCH ARTICLE

WILEY

# Prediction of activation patterns preceding hallucinations in patients with schizophrenia using machine learning with structured sparsity

Amicie de Pierrefeu[1] | Thomas Fovet[2,3] | Fouad Hadj-Selem[5] | Tommy Löfstedt[6] | Philippe Ciuciu[1,7] | Stephanie Lefebvre[2,3] | Pierre Thomas[2,3] | Renaud Lopes[4,8] | Renaud Jardri[2,3] | Edouard Duchesnay[1]

[1]NeuroSpin, CEA, Paris-Saclay, Gif-sur-Yvette, France

[2]Univ. Lille, CNRS UMR 9193, Laboratoire de Sciences Cognitives et Sciences Affectives (SCALab), PsyCHIC team, Lille F-59000, France

[3]CHU Lille, Pôle de Psychiatrie, Unité CURE, Lille F-59000, France

[4]Imaging Dpt., Neuroradiology unit, CHU Lille, Lille F-59000, France

[5]Energy Transition Institute: VeDeCoM, France

[6]Department of Radiation Sciences, Umeå University, Umeå, Sweden

[7]INRIA, CEA, Parietal team, Univ. Paris-Saclay, France

[8]U1171 - Degenerative and Vascular Cognitive Disorders, Univ. Lille, INSERM, CHU Lille, Lille F-59000, France

**Correspondence**
Renaud Jardri, Univ. Lille, CNRS UMR 9193, Laboratoire de Sciences Cognitives et Sciences Affectives (SCALab), PsyCHIC team, F-59000 Lille, France
Email: renaud.jardri@chru-lille.fr

**Funding information**
Programme Hospitalier de Recherche Clinique, Grant/Award Number: PHRC-N2011 19-02 MULTIMODHAL; Agence Nationale de la Recherche, Grant/Award Number: ANR-16-CE37-0015 INTRUDE

## Abstract

Despite significant progress in the field, the detection of fMRI signal changes during hallucinatory events remains difficult and time-consuming. This article first proposes a machine-learning algorithm to automatically identify resting-state fMRI periods that precede hallucinations versus periods that do not. When applied to whole-brain fMRI data, state-of-the-art classification methods, such as support vector machines (SVM), yield dense solutions that are difficult to interpret. We proposed to extend the existing sparse classification methods by taking the spatial structure of brain images into account with structured sparsity using the total variation penalty. Based on this approach, we obtained reliable classifying performances associated with interpretable predictive patterns, composed of two clearly identifiable clusters in speech-related brain regions. The variation in transition-to-hallucination functional patterns not only from one patient to another but also from one occurrence to the next (e.g., also depending on the sensory modalities involved) appeared to be the major difficulty when developing effective classifiers. Consequently, second, this article aimed to characterize the variability within the prehallucination patterns using an extension of principal component analysis with spatial constraints. The principal components (PCs) and the associated basis patterns shed light on the intrinsic structures of the variability present in the dataset. Such results are promising in the scope of innovative fMRI-guided therapy for drug-resistant hallucinations, such as fMRI-based neurofeedback.

**KEYWORDS**
hallucinations, machine learning, real-time fMRI, resting-state networks, schizophrenia

## 1 | INTRODUCTION

Hallucinations are defined as aberrant perceptions in the absence of causative stimuli. These experiences, especially auditory hallucinations, constitute fundamental features of psychosis (64%–80% lifetime prevalence among schizophrenia-diagnosed patients) and can lead to functional disability and a low quality of life (McCarthy-Jones et al., 2017).

Over the past years, auditory hallucinations have been studied in-depth using brain imaging methods, such as functional and structural magnetic resonance imaging (fMRI and sMRI), to decipher their underlying neural mechanisms. Numerous brain changes have been

  

extensively covered in a wide range of studies in patients suffering from auditory hallucinations (Allen, Larøi, McGuire, & Aleman, 2008; Bohlken, Hugdahl, & Sommer, 2017; Jardri, Pouchet, Pins, & Thomas, 2011; Sommer et al., 2008). Beyond location, the functional dynamics of the neural networks involved in auditory hallucinations have also been studied.

To address this important question, an increasing number of studies have focused on so-called intrinsic connectivity networks (ICN) and their potential role in the onset of hallucinations (Alderson-Day et al., 2016; Northoff & Qin, 2011). ICNs typically reveal interactions among brain regions when the subject is not engaged in any particular task. Frequently reported networks include the default mode network (DMN), the control executive network (CEN), the salience network (SAL), and the sensorimotor network (SMN) (Alderson-Day et al., 2016). Numerous studies have asserted that fluctuations in those ICNs are associated with the onset of hallucination periods. For instance, the emergence of hallucinations correlates with a disengagement of the DMN (Jardri, Thomas, Delmaire, Delion, & Pins, 2013). More recently, stochastic effective connectivity analyses revealed complex interactions among hallucination-related networks, DMN, SAL, and CEN, during the ignition, active phase, and extinction of hallucinatory experiences (Lefebvre et al., 2016).

Despite significant progress in the field, "capturing" the neural correlates of subjective mental events (such as hallucinations) remains a time-consuming task with multiple postprocessing steps and analyses. However, recent progress in machine learning has now paved the way for real-time automatic fMRI decoding of hallucination-related patterns. Such developments may have crucial impacts on the implementation of innovative fMRI-based therapy for drug-resistant hallucinations, such as fMRI-based neurofeedback (Arns et al., 2017; Fovet, Jardri, & Linden, 2015). During fMRI-based neurofeedback, brain activity is measured and fed back in real time to the subject to help her/him progressively achieve voluntary control over her/his own neural activity. Precisely defining strong a priori strategies for choosing the appropriate target brain area/network(s) for fMRI-based protocols appears critical. Interestingly, considering the rapid technical developments of fMRI techniques and the availability of high-performance computing, the pattern classification approach now appears to be one of the potential strategies for fMRI-based neurofeedback sessions.

In this context, the feasibility of fMRI-based neurofeedback relies on robust and reliable classifying performances and on the ability to detect hallucinations sufficiently early to allow the patients the necessary time to modulate their own cerebral activity (Fovet et al., 2016). Rather than detecting hallucinatory events per se, we aim to help patients become aware of the imminence of this experience based on online detection of fMRI signal changes in key networks involved in the ignition of hallucinations. Thus, in this study, we specifically focused on the period preceding the occurrence of a hallucination, that is, the few seconds corresponding to the brain's transition from a resting state to a full hallucinatory state. Interestingly, previous fMRI studies have noted the existence of specific fMRI changes prior to hallucinations (Lennox, Park, Jones, Morris, & Park, 1999; Hoffman, Anderson, Varanko, Gore, & Hampson, 2008; Diederen et al., 2010; Lefebvre et al., 2016).

Among the current machine-learning approaches available for fMRI analysis, multivoxel pattern analysis (MVPA)—a supervised classification method—is gaining recognition for its potential to accurately discriminate between complex cognitive states (Fovet et al., 2016; Haxby, Connolly, & Guntupalli, 2014). MVPA seeks to identify significantly reproducible spatial activity patterns differentiated according to mental states. Extending these methods to the prediction of the phenomena of transition toward hallucinations should provide better insight into the mechanisms of these subjective experiences. Thus, leveraging real-time pattern decoding capabilities and applying them in the case of hallucinations could lay the foundation for potential solutions for affected individuals.

Variations in transition-to-hallucination functional patterns from one patient to another (e.g., due to phenomenological differences) and from one occurrence to the next (e.g., depending on the modalities involved) appears to be the potential major shortcomings in developing an effective classifier. Indeed, such disparities may inexorably lead to a decrease in decoding performances. Therefore, characterizing the variability within the prehallucination patterns across subjects and occurrences is highly desired. Principal component analysis (PCA) is one such unsupervised method that has been successfully applied in the analysis of the variability of a given dataset. The principal components (PCs) and the associated basis patterns shed light on the intrinsic structures of the variability present in a dataset. This unsupervised approach is complementary to the supervised approach described above, as it can help with interpreting the classification performances.

Here, we applied both supervised and unsupervised machine-learning methods to an fMRI dataset collected during hallucinatory episodes. The goal of this article was twofold: (i) to predict the activation patterns preceding hallucinations using a supervised analysis and (ii) to uncover the variability in these activation patterns during the emergence of hallucinations using unsupervised analysis. The goals of these two analyses appear completely complementary in the context of future fMRI-based clinical and therapeutic applications.

## 2 | METHODS

### 2.1 | Participants and experimental paradigms

The population was composed of 37 patients with schizophrenia (DSM-IV-TR criteria, average age = 35.23 years, 10 females/27 males) who were suffering from very frequent multimodal hallucinations (i.e., more than 10 episodes/h). Participants were recruited through the FR2SM network (*Fédération Régionale de Recherche en Santé Mentale*), which groups all the private/public institutions for mental health in the Hauts-de-France region (62% of the participants were hospitalized at the time recruited, 38% received outpatient care). This sample presents a partial overlap with previous works from our team (Lefebvre et al., 2016; Leroy et al., 2017). The clinical characteristics of the recruited subjects are summarized in Table 1. fMRI was acquired at rest. Participants were asked to lie in the scanner in a state of wakeful rest with

**TABLE 1** Clinical characteristics of the recruited samples

| **Age (mean ± *SD*)** | **35.8 ± (9.8) years** |
|---|---|
| Sex | 10 F/27 M |
| CGI (mean ± *SD*) | 4.2 ± (1.6) |
| Dose of antipsychotic treatment (EqOZ) (mg/d) | 42.5 ± (22,4) |
| PANSS (mean ± *SD*) | 82.4 ± (20.3) |
| AHRS (mean ± *SD*) | 26 ± (7) |
| Average number of hallucination episodes per patient | 5.6 |
| Number of patients experiencing hallucinations (by modality) during the fMRI session | |
| *Auditory* | 32 |
| *Visual* | 5 |
| *Tactile* | 7 |
| *Olfactory* | 2 |

*Note*. Abbreviations: AHRS = Auditory Hallucination Rating Scale; CGI = Clinical Global Impressions Scale; EqOZ = Equivalent Olanzapine; PANSS = Positive and Negative Syndrome Scale.

their eyes closed. The subjects experienced an average of 5.6 hallucinatory episodes per scan. The patients' states at different acquisition times were labelled using a semiautomatic difficult procedure, as described in Jardri et al. (2013), Lefebvre et al. (2016), and Leroy et al. (2017) and were assigned to one of the following four categories: transition toward hallucinations (trans), ongoing hallucinations (on), no hallucinations (off), and end of hallucinations (end). This labeling task is a nonstraightforward two-steps strategy; the first step is a data-driven analysis of the fMRI signal using an ICA in the spatial domain. The second step involves the selection of the ICA components associated with possible sensory experiences that occurred while scanning. This pipeline is said to be semiautomatic as it combined the following: (a) an automatic denoising part, based on the classifiers described in De Martino et al. (2007) and (b) a manual and time-consuming part, with the use of an immediate post-fMRI interview conducted with the patient, in which the sensory modalities, number of episodes, and phenomenological features of the experiences were specified.

The study was approved by the local ethical committee (CPP Nord-Ouest France IV), and written informed consent was obtained for each participant enrolled in the study.

## 2.2 | Imaging parameters

The participants underwent an 11-min anatomical $T_1$-weighted 3D multishot turbo-field-echo scan (3 T Philips Achieva X-series, with an 8-elements SENSE head coil). The field-of-view was 256 $mm^2$ with a voxel resolution of 1 mm in all directions. Participants also underwent a blood oxygen level-dependent (BOLD) fMRI session. The parameters of the 3D-PRESTO SENSE sequence were field-of-view 206 × 206 × 153 $mm^3$, TE = 9.6 ms, TR = 19.25 ms, EPI-factor = 15, flip angle = 9°, dynamic scan time = 1000 ms. Because of the multishot nature of the PRESTO sequence, the TR is not equivalent to the scan duration. Each fMRI session consisted of 900 volumes collected for a total acquisition time of 15 min.

## 2.3 | fMRI preprocessing

The anatomical and functional data were preprocessed using SPM12 (WELLCOME, Department of Imaging Neuroscience, London, UK) running on MATLAB R2016a (MathWorks, Inc., Natick, Massachusetts, USA). To control for motion-induced artifacts, the point-to-point head motion was estimated for each subject (Van Dijk, Sabuncu, & Buckner, 2012). Excessive head motion (cumulative translation or rotation >3 mm or 3°) was applied as an exclusion criterion. Applying this filter, one patient was excluded from the analysis. Signal preprocessing consisted of motion correction (realignment of fMRI volumes) and voxel-wise linear detrending. Given that we excluded subjects in whom motion was too influential, we estimated that noise had a contained and, therefore, tolerable impact on the remaining subjects. Moreover, concerning the low-frequency trends in the fMRI signal, we believed that these slow signal intensity drifts did not create excessive artifacts over the signal, given that we were dealing with very short periods of transition. Hence, applying linear detrending was likely sufficient.

Then, we performed coregistration of the individual anatomical T1 images to the functional images and spatial normalization to the Montreal Neurological Institute (MNI) space using DARTEL based on the segmented T1 scans. We did not perform any spatial smoothing step in the preprocessing pipeline. The MNI brain mask was used to restrict voxels considered in the subsequent steps to 67,665 voxels.

## 2.4 | Computation of samples

Prior to training classifiers, the first step involved computing samples from the fMRI signal. The intention was to convert the fMRI signal into vectors of features reflecting the pattern of activity across voxels at a point in time. We opted against creating the samples directly from the fMRI signal. Instead, we created the samples by estimating the activity within each voxel using a linear model. The design of such a model was a crucial part of the learning process. We used a general linear model

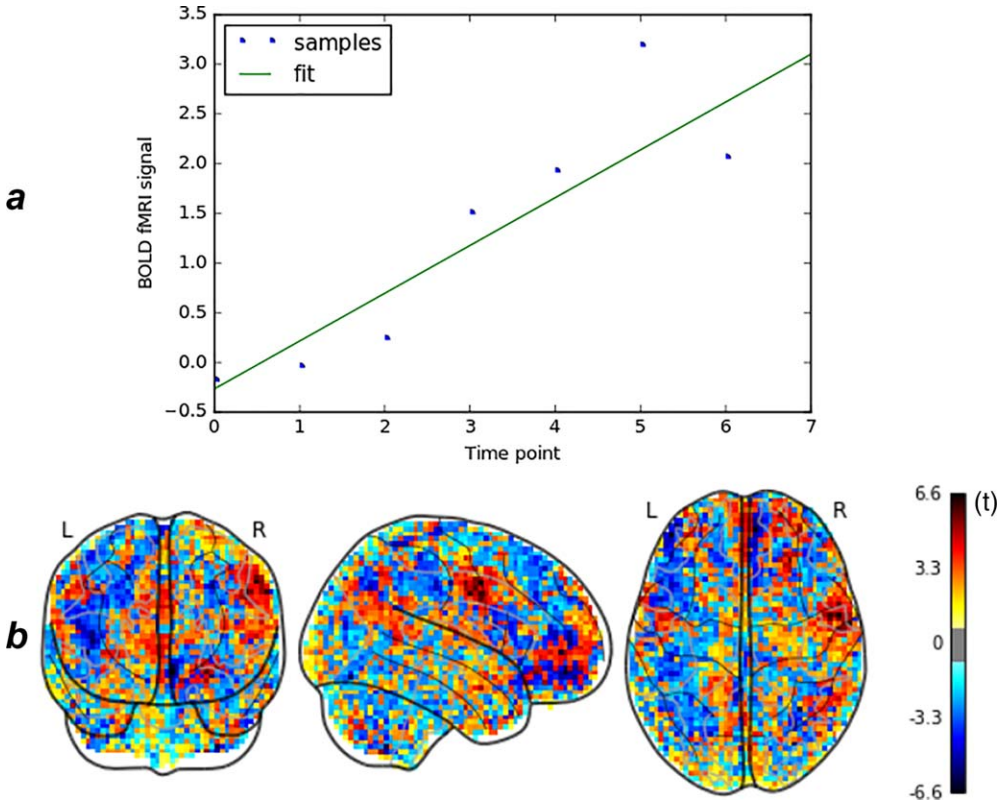


**FIGURE 1** (a) Regression of the fMRI signal time course of a voxel on a linear ramp function (fit is represented in green). (b) Sample created from one set of consecutive prehallucinations scans. The features are the *T*-statistic values associated with the coefficients of the regression in each voxel [Color figure can be viewed at wileyonlinelibrary.com]

(GLM) to estimate the activity within each voxel. From each set of consecutive images within a prehallucination state ("trans" periods) or "off" state, we created one sample. On average, each "trans" or "off" state lasted for 8 consecutive EPI volumes, which appeared sufficient to estimate activity. Based on the GLM, we regressed the fMRI signal time course on a linear ramp function for each set of consecutive volumes. This choice was based on the hypothesis that activation in some regions presents a ramp-like increase during the time preceding the onset of hallucinations.

A sigmoid activation in some regions prior to the occurrence of an hallucination is potentially more realistic than a ramp-like activation. However, to fit a sigmoid function to a set of points, two parameters need to be estimated. Given the fact we only had a limited set of 8 consecutive prehallucination EPI volumes, fitting a sigmoid would have meant leaving only 6 degrees of freedom. Given the arguments above and our wish to reach the highest possible level of robustness, we, thus, chose to use a ramp model in these conditions.

Figure 1a represents the evolution of the signal intensity in one single voxel over the 8 consecutive volumes of a prehallucination period of a subject. In this specific voxel, the signal presents a ramp-like increase during the prehallucination period.

Given that most of the patients hallucinated more than once during the scanning session, we had more samples than patients (376 samples created from 37 patients). The samples that we used as inputs to the machine-learning process were the statistical parametric maps associated with the slope coefficients of the regression (see Figure 1b as an example of one sample). We obtained a dataset of 376 samples: 166 in the resting state (*off* periods) and 210 in the prehallucination state (*trans* periods) with 67,665 features.

## 2.5 | Supervised analysis

All analyses were performed in Python using the scikit-learn toolbox (Pedregosa et al., 2011) and the pylearn-parsimony package (https://github.com/neurospin/pylearn-parsimony). Given the slow, partially manual and interview-intensive nature of the cognitive state labeling pipeline (Jardri et al., 2013), we constructed an algorithm in parallel to detect a transition-to-hallucination state in a real-time, automated fashion exclusively relying on the imaging data. We focused the analysis on the transition toward a hallucination state (trans) with the intention of distinguishing it from the resting-state activity (off).

### 2.5.1 | Classifiers

Learning with hundreds of samples (376) using high-dimensional data ($7 \times 10^4$ voxels) was associated with a high risk of overfitting in the training subjects, leading to poor performances of the independent subjects. Such issues of replicability can be addressed using state-of-the-art regularized learning algorithms.

In this study, we compared two different linear classifiers of binary classification. First, we used a regular linear SVM based on $\ell_2$ (ridge) penalty on the coefficients vectors. The role of the ridge penalty was to shrink the coefficients toward zero to control for the variance of the

fitted coefficients. However, the SVM classifier cannot select significant variables and, rather, tends to produce dense patterns of predictors that are difficult to interpret without arbitrary thresholding. In the context of predictive signature discovery, it is crucial to understand the brain activation patterns that underpin the prediction. We, therefore, sought an approach that selects a reduced number of predictive regions. Feature selection methods, such as recursive feature elimination (RFE) (Guyon, Weston, Barnhill, & Vapnik, 2002), have been used to select a reduced set of predictors (De Martino et al., 2008). However, as they are prone to local minima, these *ad hoc* heuristics tend to be replaced by sparse models based on convex minimization problems that simultaneously optimize the prediction performances while performing the feature selection.

Another solution to obtain a limited number of predictors is the use of $\ell_1$-regularized classifiers (lasso) that produces sparse patterns of predictors by enforcing many voxels to have zero-weights. The combination of ridge and lasso penalties in ElasticNet (Friedman et al., 2010) promotes sparse models while still maintaining the regularization properties of the $\ell_2$ penalty. However, despite the fact that lasso or ElasticNet classifiers have often been advocated as leading to more interpretable models, they generally lead to scattered and unstable weight patterns (Grosenick, Klingenberg, Katovich, Knutson, & Taylor, 2013; Dubois et al., 2014).

Therefore, we proposed to use the benefit of the known structure of brain fMRI images to force the solution to adhere to biological priors, producing more plausible, interpretable solutions. Indeed, MRI data are naturally encoded on a three-dimensional grid. Some voxels are neighbors, whereas others are not. The goal is to obtain a predictive pattern that is both sparse (i.e., limited number of non-null weight), but also structured (i.e., organized into clusters). Therefore, such structured sparsity can be obtained by combining $\ell_1$, $\ell_2$, and total variation (TV) penalties. Such a combination of penalties will enforce the spatial smoothness of the solution while segmenting predictive regions from the background.

Consequently, as a second classifier, we used logistic regression with 3 types of regularization penalties: $\ell_1$, $\ell_2$, and TV (Dubois et al., 2014), which was denoted TV-Elastic-net (TV-Enet). The $\ell_1$ and $\ell_2$ penalties served the purpose of addressing overfitting induced from the MRI data's high-intrinsic dimensionality. The TV penalty also regularized the solution, but its main purpose was to take advantage of the spatial 3D structure. Together, these penalties enabled the generation of a coherent, parsimonious, and interpretable weight map. Moreover, these penalties provided a segmentation of the predictive weight map into spatially contiguous parcels with constant values, which is a highly desirable characteristic in the scope of predictive signature discovery.

#### 2.5.2 | Performance metric, cross-validation, and model selection

Performance was evaluated through a double cross-validation pipeline. The double cross-validation process consists of two nested cross-validation loops. In the outer (external) loop of the double cross-validation, we employed a leave-one-subject-out pipeline where all subjects except one were referred to as the training data, and the remaining subject was used as test data. The test sets were exclusively used for model assessment, whereas the training sets were used in the inner fivefold cross-validation loop for model fitting and model selection. Classifier performances were assessed by computing the balanced accuracy, sensitivity, and specificity with which the test samples were classified. Sensitivity was defined as the ability to identify the transition toward hallucination state (trans), whereas specificity evaluated the ability to identify the resting-state activity (off). The balanced accuracy score was defined as the average of the sensitivity and specificity. We also implemented the receiver operating characteristic (ROC) curve for each classifier, from which the area under the curve (AUC) was computed.

#### 2.5.3 | Result significance

To measure the significance of the prediction scores for both classifiers, we used an exact binomial test while leveraging a paired two-samples *t*-test to compare the decoding performances of the two classifiers.

#### 2.5.4 | Predictive pattern

To analyze the brain regions that drive the prediction, we refitted the model on all samples of the dataset and extracted the associated discriminative weight map. This weight map revealed the spatial patterns that best discriminate the two cognitive states (trans and off). The weights revealed the relative contribution of each voxel to the decision function. Positive weights indicated a positive contribution toward predicting the *trans* state, whereas negative weights signaled a positive contribution toward predicting the *off state*.

### 2.6 | Unsupervised analysis

#### 2.6.1 | Decomposition method

Subsequently, in addition to the supervised analysis, we conducted an extensive analysis of the data using unsupervised machine learning. The goal was to characterize the variability within the prehallucination scans. PCA can extract the significant mode of variation from high-dimensional data. However, its interpretability remains limited. Indeed, the components produced by PCA are often noisy and exhibit no visually meaningful patterns. Nonetheless, our ultimate goal was to understand the variability in the form of intelligible patterns. In this context, we used SPCA-TV (sparse principal component analysis-total variation), which is an extension of regular PCA with $l_1$, $l_2$, and TV penalties on the PCA loadings, promoting the formation of structured sparse components that are relevant in a neuroscientific scope (de Pierrefeu et al., 2017). We hypothesized that the principal components extracted with SPCA-TV could uncover major trends of variability within the prehallucination samples. Thus, the principal components might reveal the existence of subgroups of hallucinations, notably according to the sensory modality involved (e.g., vision, audition, etc.). From the 376 samples, we retained the 210 elements corresponding to the prehallucinations samples. We applied SPCA-TV to these 210 samples and interpreted the resulting principal components.

Additionally, we computed the explained variance of each component yielded by SPCA-TV and investigated whether these components were really capturing a signature of the cognitive process involved in the onset of hallucinations. To do so, we projected each activation map, "off" and "trans" samples, in the basis formed by the principal components and used the subsequent associated scores to decode the

**TABLE 2** The performance of the classifiers

| Classifier | AUC | Acc | Spe | Sen |
|---|---|---|---|---|
| SVM | 0.73 | 0.73 | 0.78 | 0.67 |
| TV-Enet | 0.79 | 0.74 | 0.76 | 0.71 |

*Note*. Abbreviation: AUC = area under the curve.
Prediction accuracies: sensitivity (Sen, recall rate of trans samples), specificity (Spe, recall rate of off samples), and balanced accuracy ((Acc): (Sen + Spe)/2).

mental state of each subject. We used SVM with the same cross-validation pipeline described in the supervised analysis method section.

## 3 | RESULTS

### 3.1 | Supervised analysis

#### 3.1.1 | Classification performances

The classification results are presented in Table 2. Classification of resting state (i.e., *non-hallucination) patterns* (off) versus *transition toward hallucinations patterns* (trans) achieved above chance level decoding performances with both methods. Using the SVM classifier, we obtained an AUC of 0.73 and a balanced accuracy of 0.73, with a specificity of 0.78 and a sensitivity of 0.67. When using the TV-Enet classifier, we obtained an AUC of 0.79 and a balanced accuracy of 0.74, with a specificity of 0.76 and a sensitivity of 0.71. The TV-Enet yielded a significantly increased AUC compared to SVM ($T = 2.87$, $p = .006$).

Since the 37 patients included in this study were suffering from multimodal hallucinations (Table 1), we also evaluated the performance of the prediction of the TV-Enet model on two subsamples, one of which comprised the 32 subjects suffering from auditory hallucinations, among other modalities, and the other comprised the 5 subjects without any auditory hallucinations (Table 3).

For the cohort of patients experiencing auditory hallucinations, we obtained an AUC of 0.80 and a balanced accuracy of 0.75, with a specificity of 0.76 and a sensitivity of 0.73. For the cohort of patients who were not experiencing auditory hallucinations, we obtained decreased prediction performances, namely, an AUC of 0.75, a balanced accuracy of 0.63, with a specificity of 0.74, and a sensitivity of 0.55.

#### 3.1.2 | Predictive weight maps

When using the regular SVM classifier, the relevance of the obtained discriminative weight maps was limited (Figure 2a). The whole brain seemed to contribute to the prediction. It is clinically challenging to interpret the weight map. The TV-Enet classifier yields a more coherent weight map with two defined stable predictive clusters (Figure 2b). The details of these two clusters are described in Table 4.

### 3.2 | Unsupervised analysis

#### 3.2.1 | Relevance of components

The first component explained 2.5% of the variance. The second component explained 1.4% of the variance. The third component explained 0.09% of the variance. The fourth component explained 0.05% of the variance. The prediction of mental states based on the scores associated with each component yielded a significant decoding performance: the classifier was able to distinguish the "trans" samples from "off" samples, with an AUC of 65%, a recall mean of 65%, a sensitivity of 68%, and a specificity of 64%.

#### 3.2.2 | Component weight maps

The components extracted with the SPCA-TV method were of great interest from a clinical point of view (Figure 3). They revealed structured, interpretable patterns of variability within the different prehallucinations periods in our sample. Details regarding the clusters present in each principal component are provided in Table 5.

## 4 | DISCUSSION

Here, we wanted to automate the detection of specific functional patterns preceding hallucination occurrences in participants scanned at rest. First, using supervised analyses, we found evidence of prediction scores with a reliable level of significance. Our prediction of the emergence of hallucinations appeared to be accurate and yielded highly interpretable associated weight maps. Second, using unsupervised analysis, we characterized the variability of the prehallucinations patterns across both occurrences and subjects in the form of intelligible components.

### 4.1 | Supervised analysis

#### 4.1.1 | Decoding performances

The present findings indicated that the two classification algorithms were able to significantly detect the prehallucination patterns in brain activity at rest. Crucially, spatial regularization (TV) combined with the elastic net penalty significantly improved the prediction performances (increased AUC) and provided more balanced specificity and sensitivity. Indeed, traditional SVM naturally tends to allocate the "off" response, which subsequently leads to a good specificity but to a reduced detection rate (sensitivity) of patterns preceding the occurrence of hallucinations.

The studied cohort contained patients who were suffering from complex multimodal hallucinations. Thus, the hallucinations captured during acquisition could have been very heterogeneous not only across subjects but also across occurrences. When evaluating each classifier's performance on the nonauditory hallucinations only, we obtained degraded prediction scores as opposed to the ones obtained with the patients experiencing auditory hallucinations, among other modalities.

**TABLE 3** Prediction performances of the TV-Enet on the subgroup of patients experiencing auditory hallucinations, among other modalities (top row), and on the subgroup of patients who were not experiencing auditory hallucinations

| Presence of auditory hallucinations | AUC | Acc | Spe | Sen |
|---|---|---|---|---|
| Yes | 0.80 | 0.75 | 0.76 | 0.73 |
| No | 0.75 | 0.63 | 0.74 | 0.55 |

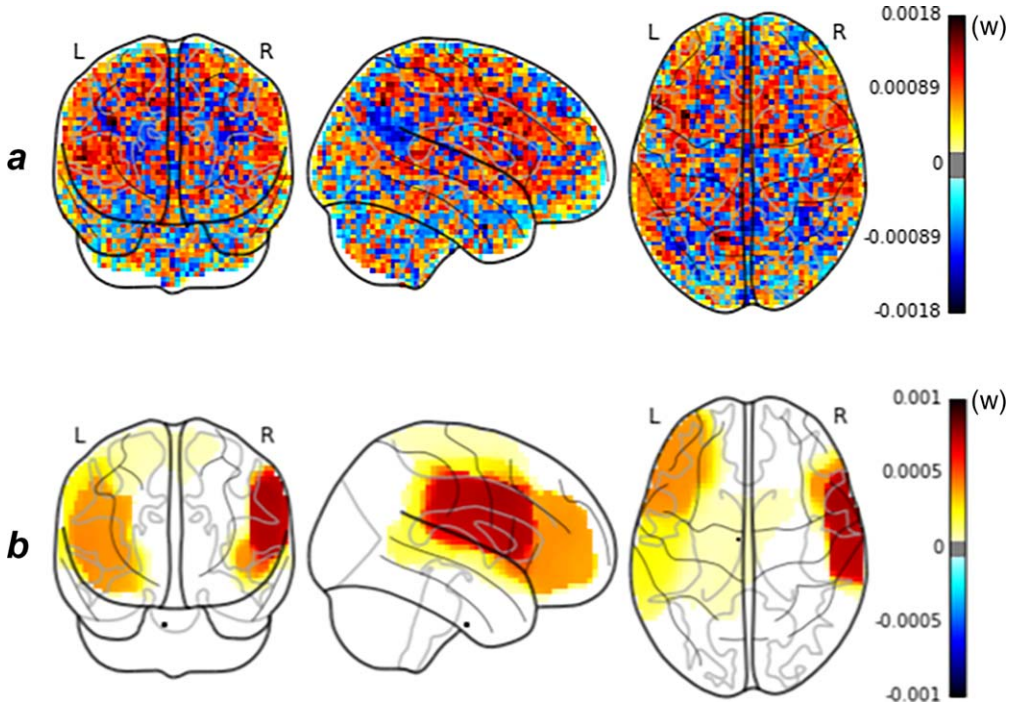


**FIGURE 2** (a) Linear support vector machine (SVM) and (b) TV-Enet predictive weight maps [Color figure can be viewed at wileyonlinelibrary.com]

This finding is to be expected as the learning of the model is conducted on 37 subjects, of whom 32 exhibited auditory experiences. Therefore, our predictive model seemed to be more specific to the prediction of auditory hallucinations than any other modalities.

Considering the above, the intersubject decoding performance that was achieved should be considered reasonably satisfactory.

Furthermore, a comparison to the seminal procedure used for labeling the scans (Jardri et al., 2013) placed our result into perspective. Compared with the procedure that required the incorporation of information from post-fMRI interviews with patients into the labeling process, the proposed machine learning-based method is fully automatic, relying exclusively on the imaging data. Moreover, the learned model can be applied in real-time during data acquisition.

Despite the challenge of gathering so many subjects in an fMRI hallucinations capture dataset ($n = 37$ subjects), we expect that increasing the sample-size may improve performances. We believe that our prediction model can still gain additional useful information from more data. Even if it is difficult to define a clear-cut line for clinical applications, an accuracy of 80% could be considered as acceptable for use in the scope of fMRI-based therapy for drug-resistant hallucinations, such as fMRI-based neurofeedback. The level of 80% stays an arbitrary threshold here, but it is considered satisfactory as detecting $^4/_5$ hallucinations in a clinical setting is already promising.

### 4.1.2 | Predictive weight map interpretation

The predictive maps obtained with the SVM method were dense and difficult to interpret without arbitrary thresholding. Even though the prediction performance was relatively good, a physician will never draw a conclusion from such a black-box model in a clinical setting as presented in Figure 2a. Understanding the brain activation patterns that drive the prediction is crucial. In addition, the predictive map obtained with TV-Enet was considerably more interpretable given that it provided a smooth map composed of two clearly identifiable regions. Interestingly, these regions, especially the speech-related brain regions, were previously shown to be involved in hallucinations (Ćurčić-Blake et al., 2017).

**TABLE 4** Supervised analysis: The clusters in the discriminative weight map

| Clusters | Center in MNI coordinates (x,y,z) | Cluster size (voxels) | Cluster mean weight | Cortical regions involved | Laterality |
|---|---|---|---|---|---|
| 1 | (53,0,15) | 3,541 | $4.1e^{-4}$ | Precentral gyrus, postcentral gyrus, inferior frontal gyrus, central opercular cortex, anterior and posterior supramarginal gyrus, insular cortex, frontal pole, middle frontal gyrus, planum temporale, temporal pole, superior temporal gyrus | Right |
| 2 | (−36,0,28) | 10,134 | $2.0e^{-4}$ | Precentral gyrus, frontal pole, postcentral gyrus, middle frontal gyrus, superior frontal gyrus, insular cortex, frontal orbital cortex, central opercular cortex, inferior frontal gyrus | Left |

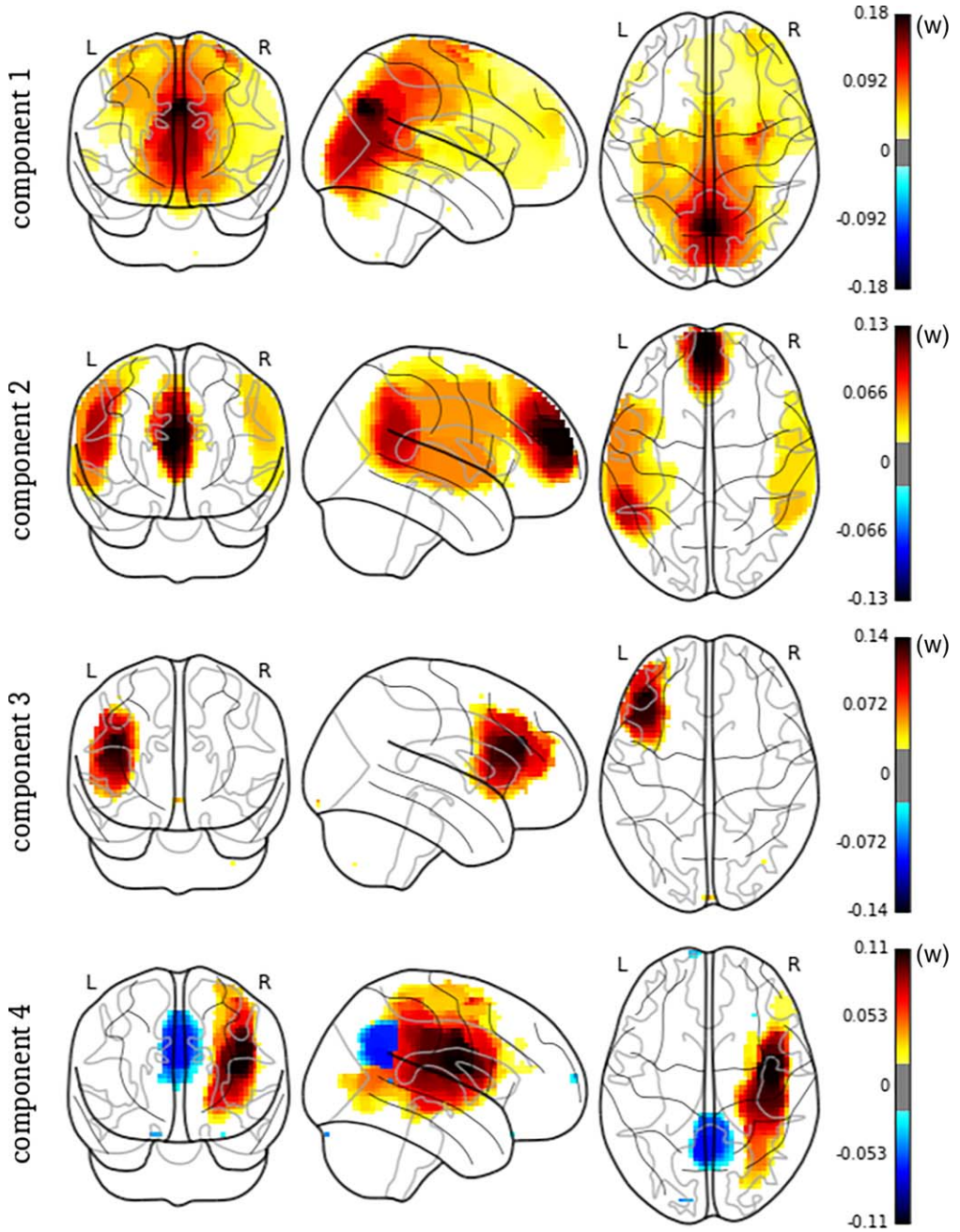


**FIGURE 3** SPCA-TV principal components. Note that the sign is arbitrary [Color figure can be viewed at wileyonlinelibrary.com]

First, the two large, stable predictive fronto-temporal clusters appeared consistent with what we currently know of the networks involved in auditory hallucinations. Indeed, numerous studies have highlighted abnormal resting-state functional connectivity among some temporo-parietal, frontal, and subcortical regions in patients with auditory hallucinations (Alderson-Day, McCarthy-Jones, & Fernyhough, 2015; Allen et al., 2008). Otherwise, patients experiencing auditory hallucinations while in the MRI scanner (in so-called fMRI "capture" studies) demonstrated significantly increased activation in Broca's area, the insula, left middle, and superior temporal gyrus, left inferior parietal lobule, and left hippocampal region (Jardri et al., 2011). Second, the right cluster identified in our study also emphasized the role of the right-sided homologues of the classical speech-related areas (i.e., the right inferior frontal gyrus, right superior temporal, and supramarginal gyrus) in auditory hallucinations, as previously described in the literature. It has been hypothesized that activity in these regions, especially the insula and the right homologue of Broca's area, is associated with the occurrence of auditory hallucinations (Jardri et al., 2011; Sommer et al., 2008), whereas language production in a natural context predominantly activates left-lateralized frontal and temporal language areas. The role of right-sided speech-related areas in the pathophysiology of auditory hallucinations was also mentioned by Mondino, Poulet, Suaud-Chagny, and Brunelin (2016). By neuromodulating a speech-related fronto-parietal network, these authors demonstrated that a

**TABLE 5** Unsupervised analysis: The clusters in the weight maps associated with the first four PCs

| PC | Clusters | Center in MNI coordinates (*x*,*y*,*z*) | Cluster size (voxels) | Cluster mean weight | Cortical regions involved | Laterality |
|---|---|---|---|---|---|---|
| **1** | 1 | (8,−28,27) | 22,002 | 0.05 | Precuneus cortex, posterior cingulate gyrus, precentral gyrus, postcentral gyrus, superior frontal gyrus, frontal pole, lingual gyrus | Right and left |
| **2** | 1 | (−52,−25,28) | 4,249 | 0.05 | Postcentral gyrus, precentral gyrus, anterior and posterior supramarginal gyrus, angular gyrus, middle frontal gyrus, superior temporal gyrus, middle temporal gyrus | Left |
| **2** | 2 | (56,−18,25) | 2,716 | 0.03 | Postcentral gyrus, precentral gyrus, anterior and posterior supramarginal gyrus, angular gyrus, superior temporal gyrus | Right |
| **2** | 3 | (−1,48, 25) | 1,988 | 0.07 | Frontal pole, paracingulate gyrus, anterior cingulate gyrus | Right and left |
| **3** | 1 | (−41,25,1) | 1,857 | 0.08 | Middle frontal gyrus, frontal pole, inferior frontal gyrus, frontal operculum cortex, insular cortex | Left |
| **4** | 1 | (37,−23,26) | 5,022 | 0.05 | Precentral gyrus, postcentral gyrus, middle frontal gyrus, insular cortex, superior parietal lobule, angular gyrus, posterior supramarginal gyrus | Right |
| **4** | 2 | (1,−52,30) | 1,173 | 0.04 | Precuneus cortex, posterior cingulate gyrus | Right and left |

reduction in the resting-state functional connectivity between the left temporo-parietal junction and right inferior frontal areas could be measured, and this reduction was associated with a significant reduction in the severity of the hallucinations.

The high rate of auditory hallucinations in this sample may account for the speech-related regions identified in the predictive map. This explains the fact that these regions are crucial in the prediction process of prehallucinations patterns. Given that 32 of the 37 patients suffered from auditory hallucinations, among other modalities, it is not surprising that such regions previously associated with auditory-verbal hallucinations have been identified as highly predictive. Conversely, since the number of patients suffering from hallucinations in other modalities (visual, tactile, and olfactory) is limited, their weights in the classifier appeared minimal compared to the predictive weights of the auditory hallucinations. Consequently, this explained the degraded prediction performances obtained for the nonauditory hallucinations, as presented in Table 3.

Classification algorithms may ideally benefit from modality-specific training on more restrictive datasets of patients hallucinating in just one sensory modality. However, even if this could be easily performed for voice-hearing, this appears quite challenging for other modalities.

Taken together, these results confirm that adding a penalty to account for the spatial structure of the brain seems relevant in fMRI captures, given that it significantly improves the classifier performance and results in clinically interpretable weight maps.

Here, we demonstrated that supervised classification methods can accurately predict the imminence of a hallucinatory episode. Thus, leveraging real-time pattern decoding capabilities and applying them in the case of hallucinations could lay the foundation for alternative solutions for affected patients in the near future, such as fMRI-based neurofeedback.

## 4.2 | Unsupervised analysis

### 4.2.1 | Relevance of weight maps

The total amount of explained variance was surprisingly low. Indeed, the activation maps of the resting-state fMRI data preceding hallucinations were very noisy, and only a minor part of its variability could be captured.

However, when predicting the mental state of subjects based on the SPCA-TV scores, the decoding accuracy was significant. Naturally, the performance was decreased compared to the performance obtained in the supervised part of this article, which was expected as we were losing some information from the compression of the 67,655 features into 4 scores. However, the fact that we could still significantly distinguish the prehallucination samples from the resting-state samples using those 4 component scores revealed that they made sense and were specifically related to hallucinations. Consequently, although the explained variance was low due to the resting-state nature of the data, the components were relevant and captured the cognitive processes involved in the onset of hallucinations.

### 4.2.2 | Weight map interpretation

The variability in the prehallucination patterns across occurrences and subjects were represented in the form of intelligible components.

The first PC mainly included the weights in the precuneus cortex and the posterior cingulate cortex. The posterior cingulate cortex, which is part of the DMN, is associated with auditory hallucinations (Rotarska-Jagiela et al., 2010). We believe that this component may have captured the visual pathways typically involved in the occurrence of visual hallucinations.

The second PC was composed of one activation cluster in the paracingulate gyrus and the anterior cingulate gyrus and two symmetric bilateral activation clusters in the temporal cortex. This fronto-temporal

component appeared compatible with the processes at the roots of the auditory hallucinations. Interestingly, some processes involved in the occurrence of hallucinations, such as the monitoring of inner speech processes and error detection, are classical functions of the anterior cingulate cortex included in this component (Allen et al., 2008; Mechelli et al., 2007). This second PC yielded regions classically involved in inhibition (paracingulate gyrus and anterior cingulate gyrus) (Allen et al., 2008; Mechelli et al., 2007). The severity of auditory hallucinations has been found to be inversely related to the strength of the functional connectivity between the temporal–parietal junction, the anterior cingulate cortex (ACC), and the amygdala (Vercammen, Knegtering, den Boer, Liemburg, & Aleman, 2010). This ACC dysconnectivity supposedly drove the external misattribution observed during auditory hallucinations (Allen et al., 2007; Mechelli et al., 2007), and might explain global inhibition impairments in the pathophysiology of hallucinations (Jardri et al., 2016), which may account for this feature beyond the schizophrenia-spectrum, as in LSD-induced hallucinations, for instance (Schmidt et al., 2017).

The third PC revealed a cluster in the frontal gyrus and the anterior insula. These regions are important for speech production, encompassing the well-known Broca's area (Small and Hickok 2016) and are involved in auditory hallucinations (Jardri et al., 2011; Sommer et al., 2008).

Finally, the fourth PC included two clusters of opposing signs. In the right hemisphere, there was a large activation cluster that involved the temporo-parietal junction and a deactivation cluster that involved the precuneus cortex and the posterior cingulate gyrus. Interestingly, this PC revealed activation of the brain regions involved in auditory hallucination-related processes and in self-other distinction, such as the right temporo-parietal junction (Jardri et al., 2011; Decety & Lamm, 2007; Plaze et al., 2015), together with a deactivation of key nodes of the DMN, including the posterior cingulate cortex, medial prefrontal cortex, medial temporal cortex, and lateral parietal cortex (Buckner, Andrews-Hanna, & Schacter, 2008). Our results appeared fully compatible with recent fMRI-capture findings demonstrating that aberrant activations of speech-related areas concomitant with hallucinatory experiences follow complex interactions between ICNs, such as the DMN and the CEN (Lefebvre et al., 2016). A disengagement of the DMN during goal-directed behaviors has been seminally evidenced in the resting-state literature (Fox et al., 2005; Lefebvre et al., 2016; Raichle et al., 2001), and similar mechanisms might be involved in hallucinatory occurrences (Jardri et al., 2013; Leroy et al., 2017). Such fluctuations in the ICNs are, thus, thought to be highly involved in the transition from a resting state to an active hallucinatory state.

### 4.3 | Perspectives

In this study, we chose to train a classifier to specifically detect periods preceding the occurrence of hallucinations (i.e., "trans" periods). As mentioned earlier, several studies have demonstrated that this period is potentially associated with specific brain activations. Diederen et al. (2010) demonstrated reduced activity in the left parahippocampal gyrus, the left superior temporal gyrus (STG), the medial frontal gyrus, and the right inferior frontal gyrus (IFG) prior to auditory hallucinations. A study by (Hoffman & Hampson, 2011) also revealed increased activation in the right posterior temporal area compared with its right homologue during the same period. The specific patterns observed in the "trans" period probably corresponded to the triggering mechanisms of the auditory hallucinations, which may have a component in memory (Ćurčić-Blake et al., 2017) and constitute a very interesting target for neurofeedback therapies. Real-time recognition of the "trans" period using the TV-Enet classifier could enable the delivery of visual information (i.e., visual feedback) regarding the imminent onset of hallucinations to the participant during an fMRI-NF session. Such a procedure could help the subject learn effective coping strategies to prevent the occurrence of hallucinations. Similarly, recent effective connectivity findings revealed that the extinction of auditory hallucinations ("end" periods) was associated with a takeover of the frontoparietal CEN (Hoffman & Hampson, 2011; Lefebvre et al., 2016). This finding suggests that the termination of auditory hallucinations is a voluntary process that could benefit from, and be reinforced by, fMRI-NF learning. We believe that such fMRI-NF based on the TV-Enet classifier could reduce the associated distress based on an improvement in the feelings of control and self-efficacy.

One of the major limits of such fMRI-based therapies remains the accessibility and cost of the equipment. It appears fundamental to develop less complex devices as potential second-line treatments for hallucinations, such as near-infrared spectroscopy (NIRS). From this technological transfer perspective, the discriminative maps obtained using the TV-Enet classifier also appear advantageous, given that the identified clusters are cortical regions with activity that are easily measured with NIRS.

## 5 | CONCLUSION

Because the hallucinations were frequently multimodal in the sample of patients recruited for this study, we expected more disparities in the functional patterns associated with their complex hallucinations and the transition toward this state compared with pure auditory experiences. In this context, the significant intersubject decoding performances obtained appeared satisfactory and are promising for future fMRI-based therapy for drug-resistant hallucinations.

### ACKNOWLEDGMENTS

This study was supported by the *Programme Hospitalier de Recherche Clinique* (PHRC-N2011 19-02 MULTIMODHAL awarded to RJ) and the *Agence Nationale de la Recherche* (ANR-16-CE37-0015 INTRUDE awarded to RJ).

### CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest with the content of this article.

## ORCID

*Amicie de Pierrefeu* http://orcid.org/0000-0002-5231-6010
*Thomas Fovet* http://orcid.org/0000-0003-0077-624X
*Tommy Lofstedt* http://orcid.org/0000-0001-7119-7646
*Pierre Thomas* https://orcid.org/0000-0002-8459-7742
*Renaud Lopes* http://orcid.org/0000-0002-2425-2283
*Renaud Jardri* http://orcid.org/0000-0003-4596-1502
*Edouard Duchesnay* http://orcid.org/00000-0002-4073-3490

## REFERENCES

Alderson-Day, B., K., Diederen, C., Fernyhough, J. M., Ford, G., Horga, D. S., Margulies, S., … Jardri, R. (2016). Auditory hallucinations and the brain's resting-state networks: Findings and methodological observations. *Schizophrenia Bulletin*, *42*(5), 1110–1123.

Alderson-Day, B., McCarthy-Jones, S., & Fernyhough, C. (2015). Hearing voices in the resting brain: A review of intrinsic functional connectivity research on auditory verbal hallucinations. *Neuroscience and Biobehavioral Reviews*, *55 (August)*, 78–87.

Allen, P., Amaro, E., Fu, C. H. Y., Williams, S. C. R., Brammer, M. J., Johns, L. C., & McGuire, P. K. (2007). Neural correlates of the misattribution of speech in schizophrenia. *British Journal of Psychiatry*, *190*, 162–169.

Allen, P., Larøi, F., McGuire, P. K., & Aleman, A. (2008). The hallucinating brain: A review of structural and functional neuroimaging studies of hallucinations. *Neuroscience and Biobehavioral Reviews*, *32*(1), 175–191.

Arns, M., J.-M., Batail, S., Bioulac, M., Congedo, C., Daudet, D., Drapier, T., … NExT group. (2017). Neurofeedback: One of today's techniques in psychiatry? *L'Encephale*, *43*(2), 135–145.

Bohlken, M. M., Hugdahl, K., & Sommer, I. E. C. (2017). Auditory verbal hallucinations: Neuroimaging and treatment. *Psychological Medicine*, *47*(2), 199–208.

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network. *Annals of the New York Academy of Sciences*, *1124*(1), 1–38.

Ćurčić-Blake, B., Ford, J. M., Hubl, D., Orlov, N. D., Sommer, I. E., Waters, F., … Aleman, A. (2017). Interaction of language, auditory and memory brain networks in auditory verbal hallucinations. *Progress in Neurobiology*, *148 (January)*, 1–20.

de Pierrefeu, A., Löfstedt, T., Hadj-Selem, F., Dubois, M., Jardri, R., Fovet, T., … Duchesnay, E. (2017). Structured sparse principal components analysis with the TV-elastic net penalty. IEEE Transactions on Medical Imaging, Sep 4. https://doi.org/10.1109/TMI.2017.2749140

Decety, J., & Lamm, C. (2007). The role of the right temporoparietal junction in social interaction: How low-level computational processes contribute to meta-cognition. *The Neuroscientist*, *13*(6), 580–593.

De Martino, F., Gentile, F., Esposito, F., Balsi, M., Di Salle, F., Goebel, R., & Formisano, E. (2007). Classification of fMRI independent components using IC-fingerprints and support vector machine classifiers. *NeuroImage*, *34*(1), 177–194.

De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., & Formisano, E. (2008). Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *NeuroImage*, *43*(1), 44–58.

Diederen, K. M. J., Neggers, S. F. W., Daalman, K., Blom, J. D., Goekoop, R., Kahn, R. S., & Sommer, I. E. C. (2010). Deactivation of the parahippocampal gyrus preceding auditory hallucinations in schizophrenia. *The American Journal of Psychiatry*, *167*(4), 427–435.

Dubois, M., Hadj-Selem, F., Lofstedt, T., Perrot, M., Fischer, C., Frouin, V., & Duchesnay, E. (2014). "Predictive Support Recovery with TV-Elastic Net Penalty and Logistic Regression: An Application to Structural MRI." In 2014 International Workshop on Pattern Recognition in Neuroimaging. https://doi.org/10.1109/prni.2014.6858517.

Fovet, T., Jardri, R., & Linden, D. (2015). Current issues in the use of fMRI-based neurofeedback to relieve psychiatric symptoms. *Current Pharmaceutical Design*, *21*(23), 3384–3394.

Fovet, T., Orlov, N., Dyck, M., Allen, P., Mathiak, K., & Jardri, R. (2016). Translating neurocognitive models of auditory-verbal hallucinations into therapy: Using real-time fMRI-neurofeedback to treat voices. *Frontiers in Psychiatry*, *7 (June)*, 103.

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., & Raichle, M. E. (2005). From the cover: The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences*, *102*(27), 9673–9678.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22.

Grosenick, L., Klingenberg, B., Katovich, K., Knutson, B., & Taylor, J. E. (2013). Interpretable whole-brain prediction analysis with GraphNet. *NeuroImage*, *72 (May)*, 304–321.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*(1), 389–422.

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annual Review of Neuroscience*, *37 (June)*, 435–456.

Hoffman, R. E., Anderson, A. W., Varanko, M., Gore, J. C., & Hampson, M. (2008). Time course of regional brain activation associated with onset of auditory/verbal hallucinations. *The British Journal of Psychiatry*, *193*(5), 424–425.

Hoffman, R. E., & Hampson, M. (2011). Functional connectivity studies of patients with auditory verbal hallucinations. *Frontiers in Human Neuroscience*, *6 (December)*, 6.

Jardri, R., K., Hugdahl, M., Hughes, J., Brunelin, F., Waters, B., Alderson-Day, D., … Denève, S. (2016). Are hallucinations due to an imbalance between excitatory and inhibitory influences on the brain? *Schizophrenia Bulletin*, *42*(5), 1124–1134.

Jardri, R., Pouchet, A., Pins, D., & Thomas, P. (2011). Cortical activations during auditory verbal hallucinations in schizophrenia: A coordinate-based meta-analysis. *The American Journal of Psychiatry*, *168*(1), 73–81.

Jardri, R., Thomas, P., Delmaire, C., Delion, P., & Pins, D. (2013). The neurodynamic organization of modality-dependent hallucinations. *Cerebral Cortex*, *23*(5), 1108–1117.

Lefebvre, S., Demeulemeester, M., Leroy, A., Delmaire, C., Lopes, R., Pins, D., Thomas, P., & Jardri, R. (2016). Network dynamics during the different stages of hallucinations in schizophrenia. *Human Brain Mapping*, *37*(7), 2571–2586.

Lennox, B. R., Park, S. B., Jones, P. B., Morris, P. G., & Park, G. (1999). Spatial and temporal mapping of neural activity associated with auditory hallucinations. *The Lancet*, *353*(9153), 644.

Leroy, A., Foucher, J. R., Pins, D., Delmaire, C., Thomas, P., Roser, M. M., … Jardri, R. (2017). fMRI capture of auditory hallucinations: Validation of the two-steps method. *Human Brain Mapping 38*(10), 4966–4979.

McCarthy-Jones, S., D., Smailes, A., Corvin, M., Gill, D. W., Morris, T. G., Dinan, K. C., … Dudley, R. (2017). Occurrence and co-occurrence of hallucinations by modality in schizophrenia-spectrum disorders. *Psychiatry Research*, *252 (June)*, 154–160.

Mechelli, A., Allen, P., Amaro, E., Jr, Fu, C. H. Y., Williams, S. C. R., Brammer, M. J., Johns, L. C., & McGuire, P. K. (2007). Misattribution

of speech and impaired connectivity in patients with auditory verbal hallucinations. *Human Brain Mapping*, *28*(11), 1213–1222.

Mondino, M., Poulet, E., Suaud-Chagny, M.-F., & Brunelin, J. (2016). Anodal tDCS targeting the left temporo-parietal junction disrupts verbal reality-monitoring. *Neuropsychologia*, *89*, 478–484.

Northoff, G., & Qin, P. (2011). How can the brain's resting state activity generate hallucinations? A 'resting state hypothesis' of auditory verbal hallucinations. *Schizophrenia Research*, *127*(1–3), 202–214.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-Learn: Machine learning in python. *Journal of Machine Learning Research: JMLR*, *12*, 2825–2830.

Plaze, M., Mangin, J.-F., Paillère-Martinot, M.-L., Artiges, E., Olié, J.-P., Krebs, M.-O., Gaillard, R., Martinot, J.-L., & Cachia, A. (2015). Who is talking to me?—Self–other attribution of auditory hallucinations and sulcation of the right temporoparietal junction. *Schizophrenia Research*, *169*(1–3), 95–100.

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., & Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(2), 676–682.

Rotarska-Jagiela, A., van de Ven, V., Oertel-Knöchel, V., Uhlhaas, P. J., Vogeley, K., & Linden, D. E. (2010). Resting-state functional network correlates of psychotic symptoms in schizophrenia. *Schizophrenia Research*, *117*(1), 21–30.

Schmidt, A., Müller, Lenz, F. C., Dolder, P., Schmid, C., Zanchi, Y. D., Lang, U. E., Liechti, M. E., & Borgwardt, S. (2017). Acute LSD effects on response inhibition neural networks. *Psychological Medicine*, October, 1–13.

Small, S. L., & Hickok, G. (2016). The neurobiology of language. *Neurobiology of Language*, Amsterdam: Academic-Press, pp. 3–9.

Sommer, I. E. C., Diederen, K. M. J., Blom, J.-D., Willems, A., Kushan, L., Slotema, K., ... Kahn, R. S. (2008). Auditory verbal hallucinations predominantly activate the right inferior frontal area. *Brain: A Journal of Neurology*, *131*(Pt 12), 3169–3177.

Van Dijk, K. R. A., Sabuncu, M. R., & Buckner, R. L. (2012). The influence of head motion on intrinsic functional connectivity MRI. *NeuroImage*, *59*(1), 431–438.

Vercammen, A., Knegtering, H., den Boer, J. A., Liemburg, E. J., & Aleman, A. (2010). Auditory hallucinations in schizophrenia are associated with reduced functional connectivity of the temporo-parietal area. *Biological Psychiatry*, *67*(10), 912–918.

**How to cite this article:** de Pierrefeu A, Fovet T, Hadj-Selem F, et al. Prediction of activation patterns preceding hallucinations in patients with schizophrenia using machine learning with structured sparsity. *Hum Brain Mapp*. 2018;39:1777–1788. https://doi.org/10.1002/hbm.23953