PAPER

Comparison of logistic regression, support vector machines, and deep learning classifiers for predicting memory encoding success using human intracranial EEG recordings

To cite this article: Akshay Arora *et al* 2018 *J. Neural Eng.* **15** 066028

Recent citations

- Spike detection and sorting with deep learning
  Melinda Rácz *et al*

View the article online for updates and enhancements.

Comparison of logistic regression, support vector machines, and deep learning classifiers for predicting memory encoding success using human intracranial EEG recordings

Akshay Arora[1], Jui-Jui Lin[1], Alec Gasperian[1], Joseph Maldjian[2], Joel Stein[3], Michael Kahana[4] and Bradley Lega[1,5,6]

[1] Department of Neurological Surgery, University of Texas—Southwestern Medical Center, Dallas, TX 75390, United States of America
[2] Department of Radiology, University of Texas—Southwestern Medical Center, Dallas, TX 75390, United States of America
[3] Department of Radiology, University of Pennsylvania, Philadelphia, PA, 19104, United States of America
[4] Department of Psychology, University of Pennsylvania, Philadelphia, PA, 19104, United States of America

E-mail: bradlega@gmail.com

Abstract

Objective. We sought to test the performance of three strategies for binary classification (logistic regression, support vector machines, and deep learning) for the problem of predicting successful episodic memory encoding using direct brain recordings obtained from human stereo EEG subjects. We also sought to test the impact of applying t-distributed stochastic neighbor embedding (tSNE) for unsupervised dimensionality reduction, as well as testing the effect of reducing input features to a core set of memory relevant brain areas. This work builds upon published efforts to develop a closed-loop stimulation device to improve memory performance. *Approach*. We used a unique data set consisting of 30 stereo EEG patients with electrodes implanted into a core set of five common brain regions (along with other areas) who performed the free recall episodic memory task as brain activity was recorded. Using three different machine learning strategies, we trained classifiers to predict successful versus unsuccessful memory encoding and compared the difference in classifier performance (as measured by the AUC) at the subject level and in aggregate across modalities. We report the impact of feature reduction on the classifiers, including reducing the number of input brain regions, frequency bands, and the impact of tSNE. *Results*. Deep learning classifiers outperformed both support vector machines (SVM) and logistic regression (LR). *A priori* selection of core brain regions also improved classifier performance for LR and SVM models, especially when combined with tSNE. *Significance*. We report for the first time a direct comparison among traditional and deep learning methods of binary classification to the problem of predicting successful memory encoding using human brain electrophysiological data. Our findings will inform the design of brain machine interface devices to affect memory processing.

[5] UT—Southwestern, Neurological Surgery MS 8855, 5323 Harry Hines Blvd, Dallas, TX 75390, United States of America
[6] Author to whom any correspondence should be addressed.

S Supplementary material for this article is available online

(Some figures may appear in colour only in the online journal)

Introduction

Machine learning classifiers have seen broad application in the field of cognitive neuroscience. Prominent examples include multivoxel pattern analysis in fMRI (Stelzer *et al* 2013), and for control of EEG-based brain–machine interface devices (Hortal *et al* 2015). Most recently, our group has published data that utilizes binary classification of intracranial EEG (iEEG) signal to control a brain machine interface device designed to improve memory performance in humans (Ezzyat *et al* 2017, 2018). This responsive neuromodulation strategy is predicated on the ability to decode whether the brain is in a state that is favorable or unfavorable for successful memory encoding, as stimulation must only be delivered when an individual is *unlikely* to remember a given item. These published analyses utilize logistic regression for classification. While preliminary results are promising, the effectiveness of closed-loop stimulation strategies to modify memory performance depends on the effectiveness of the classifier used to control stimulation. As such, even minor improvements in performance of a classifier can potentially lead to clinically meaningful improvements in device performance. Further, as the design of a brain machine interface device to affect memory performance matures, important questions will need to be resolved regarding the optimal number of brain regions (and which regions) will need to be sampled from for effective behavioral improvement. This is a critical question, as presumably a clinical device will not have available the same density of brain electrode coverage employed in (short term) clinical seizure mapping investigations. Ultimately, a core set of critical locations will have to be selected (perhaps on a case by case basis). In support of these important questions, we examined two alternative binary classification strategies for predicting successful memory encoding using intracranial EEG signal, namely support vector machines and recurrent neural networks, comparing performance to that achieved using a logistic regression classifier. We also wanted to compare the results achieved using a classifier built with all available brain data versus performance for which a subset of brain regions were included, with the goal of addressing the problem of the optimal BCI design given practical limitations on which brain areas can be sampled from.

We wanted to test the utility of support vector machines because these have been used for pattern classification for EEG signal in several studies (Kaper *et al* 2004, Chandaka *et al* 2009, Kumar *et al* 2014, Li *et al* 2014), showing good performance and robustness across many subjects at the expense of greater computational intensity during classifier training compared to logistic regression. They offer a non-linear method. Previous applications of polynomial SVM to

BCi-related binary classification have shown performance improvement between five and 15% relative to linear classification methods (Kaper *et al* 2004, Xu *et al* 2004, Schlögl *et al* 2005, Lotte *et al* 2007). EEG-based BCI devices typically offer high data dimensionality (multiple channels, frequencies, and time series samples, interactions among these variables) relative to the number of observations. In this environment, SVMs are an appealing method because of relative insensitivity to overtraining due to embedded regularization. They are (theoretically) less sensitive to the 'peaking phenomenon' by which increasing dimensionality reduces classifier performance. We also wanted to test the performance of recurrent neural networks (RNNs) to this problem. RNNs offer theoretical advantages over traditional machine learning approaches for the modeling of timeseries data (Hermans and Schrauwen 2013), as they are able to learn complexities in the time domain without susceptibility to over classification. RNNs include an additional synapse or weight matrix to connect a synapse to its own inputs from a previous time step. The ability to incorporate previous information discovery allows the network to learn complex sequences as in speech for example Graves *et al* (2013), although limitations in capabilities of traditional RNNs necessitated the development of LSTM cell strategies (long short term memory cell) (Pascanu *et al* 2013). The LSTM cell is essentially a memory block which has an input gate, a forget gate and an output gate (Sutskever *et al* 2014). RNNs without LSTM cells suffer from a vanishing gradient problem which inhibits them from learning long-term dependencies of data (Sutskever *et al* 2014).

In iEEG data, implantation arrays used for intracranial EEG are custom tailored to the hypothesized epileptic brain region, so the resulting data sets can be quite heterogenous from subject to subject in terms of electrode coverage. This is a problem for generalizing results when comparing different machine learning strategies. Stereo EEG investigations for suspected temporal lobe epilepsy however often rely on a core set of common brain regions to localize seizure patterns; as such it provides a common set of brain regions across subjects that permit valid comparison of model performance at the subject level (Serletis *et al* 2014). We used a unique data set consisting of 30 stereo EEG patients in whom electrodes had been inserted into a common set of brain regions associated with mnemonic processing as part of the standard seizure mapping protocol at our institution. This data set consisted of 15 subjects with electrodes in the dominant and 15 subjects with electrodes in the non-dominant hemisphere. *A priori*, we also planned to generate models using all data from each individual as well as more restrictive training data focused only on the five brain regions common to all subjects in our dataset. The latter models would be the most generalizable across

subjects, since all subjects would contribute equivalent data. Along with the main goal of comparing different machine learning strategies, we also wanted to examine the effects of applying t-stochastic neighbor embedding (tSNE) dimensionality reduction to our data (Maaten and Hinton 2008, Gisbrecht *et al* 2015). T-SNE has previously been employed for dimensionality reduction purposes in EEG timeseries data, with the goal of improving classifier performance by reducing the co-linearity problem that can affect logistic regression and other machine learning methods (Birjandtalab *et al* 2016). We hypothesized that the addition of tSNE would have the greatest impact on the highest dimensionality data and would most improve classifier performance for logistic regression, although this was not what we found. Finally, we wanted to use this unique dataset to explore the impact of connectivity information (coherence) on model performance. This necessitates the use of a trial-by-trial estimate of connectivity calculated across a consistent time window, so we focused on coherence in lower frequencies for which significant connectivity does not vary as much across the time series (compared to gamma range connectivity). Theta frequency connectivity modulation has been observed during the encoding and retrieval of episodic memories throughout the brain (Burke *et al* 2013).

Methods

Electrode locations

Thirty patients at UT Southwestern Medical Center undergoing iEEG monitoring as part of clinical treatment for intractable epilepsy were recruited to participate in this study. A subset of these individuals contributed data to the previous responsive stimulation publications (Ezzyat *et al* 2018). Data from these subjects (as well as the full data set from the Restoring Active Memory program) is available by request at the following URL: http://memory.psych.upenn.edu/Request_RAM_Public_Data_access. Localization of electrodes was determined by expert neuroradiology review following robotically-assisted insertion. Electrodes in seizure onset zones were eliminated as part of standard processing pipeline based on epileptology review. Electrodes were either AdTech (19 participants, model RD10R) or PMT (13 participants, model 2102) stereo EEG depth electrodes with cylindrical contacts spaced between 3 and 8 mm apart, depending upon the depth of the brain target. No grid or strip electrodes were used. A typical implantation array for a given patient included between 13 and 17 electrodes, each of which included from eight to 14 contacts arrayed linearly along the shaft of the electrode. All electrodes not observed to be within brain parenchyma were removed. Across these 30 subjects, electrodes were implanted in regions including the orbitofrontal cortex, fusiform gyrus, anterior cingulate, insula, amygdala, supramarginal and angular gyrus, superior parietal lobule, mesial and lateral pre-frontal cortex, and ventro-lateral pre frontal cortex, along with five core areas common to all subjects: hippocampus, superior parietal lobule (SPL), posterior cingulate gyrus, lateral temporal cortex (at coronal plane just posterior to amygdalo-hippocampal sulcus), and inferior parietal lobule

(angular or supramarginal gyrus). This core set of five brain regions were conserved across subjects because they include temporal and extra-temporal areas routinely sampled from to characterize temporal lobe epilepsy. These core regions are also well-characterized participants in memory encoding and retrieval networks, so it provided a well-suited dataset with implicit *a priori* focus on memory-relevant brain regions to test the idea that supervised dimensionality reduction could improve classifier performance.

Behavioral task

All participants performed sessions of a verbal free recall task wherein they were visually presented words on a computer screen from a predetermined pool of common nouns. Details of this task are described in numerous previous publications, but are included here for clarity (Sederberg *et al* 2003, Lega *et al* 2011). The memory task consisted of a series of lists of 12 memory items; each testing session included either 12 or 25 such lists. Each word was presented for 1.6 s, separated by a blank screen of 300–500 ms with (100 ms of random jitter around 400 ms). The period analyzed for each item was 1800 ms, consisting of the 1600 ms period during which items were on the screen and 200 ms afterward. Each list was followed by a 30 s period of simple arithmetic distractors in the form of 'A + B + C = ??' in effort to deter working memory rehearsal. Subjects were instructed to answer as many of these as possible. Finally, participants were given a 30 s recall period and instructed to recite as many words as possible, in any order, from the previously presented list. Successfully encoded items were those that were recalled during this retrieval period, only from the list immediately preceding (from other lists were considered intrusions and not scored as correctly encoded). For this analysis we did not include signal from the retrieval period (when individuals are recalling the items) in the prediction models. Recorded iEEG from participants' sessions were time-locked with word presentations. Participants contributed from three of these sessions, where a session consisted of 25 presentations of unique lists. Subjects contributed an average of 864 words each with an average of 23% correctly recalled per subject (some items thrown out due to task interruptions, etc). Figure 1 shows a schematic of the recall task.

iEEG processing

We used a standard subsequent memory effect paradigm as the substrate for pattern classification. Encoding events were divided into successful and unsuccessful groups. Each event window consisted of time series data of 1800 ms duration, with 500 ms buffer to avoid edge effects, averaged into six time windows using a bipolar re-referencing scheme. Power was obtained by downsampling the raw EEG recorded on a clinical Nihon–Koden system from 1 kHz to 250 Hz, then utilizing morlet wavelets (wave number of 6) to extract power from 2 to 100 Hz that was averaged into six log-spaced frequency bands (delta: 2.5–5 Hz, theta: 4–9 Hz, alpha: 9–16 Hz, beta: 16–25 Hz, gamma: 40–65 Hz, and high gamma: 65–100
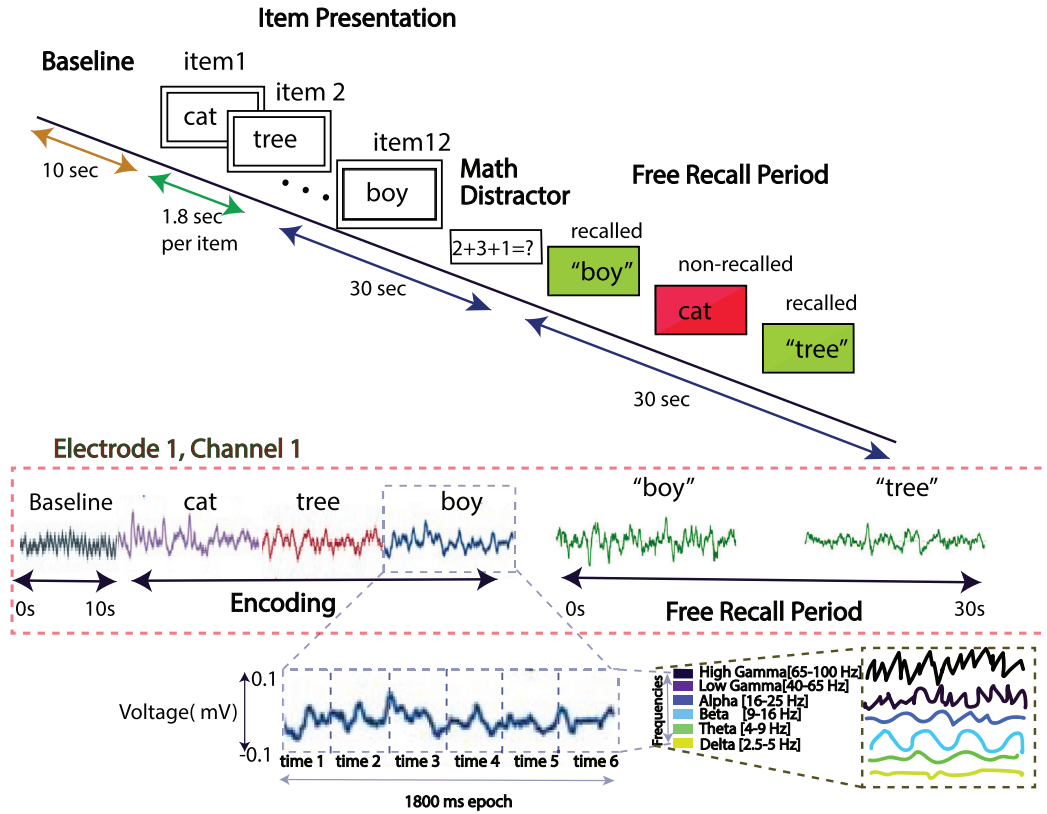
Figure 1. Schematic of free recall task. Goal of classifier is to use brain activity recorded at the time of item presentation to classify which words would be successfully remembered during the free recall period. Local field potential examples for the encoding epoch divided into six time bins and six frequency bands after normalization using baseline period.

Hz). We used a 2.5 Hz lower boundary for the delta band based upon our own observations that oscillations in this frequency range exhibit memory-related properties in the core encoding network (Lega *et al* 2011). A kurtosis algorithm was applied (threshold of 4) to exclude encoding events with signal artifact (Sederberg *et al* 2003). Finally, intracranial power was averaged for all the electrode channels in a region for each event and Z-transformed to normalize across subjects relative to pre-stimulus baseline. We sought to assess the impact of connectivity data on the performance of our classifiers. With the hippocampus as a seed region, we quantified connectivity with each of the other four regions in our 'core area' dataset. Coherence was calculated using the Chronux package (Bokil *et al* 2010), focused on a broad low frequency range (3–10 Hz) based upon previous publications for memory-relevant theta frequency connectivity (Watrous *et al* 2013, Burke *et al* 2014b). Results were compared with classifiers trained using only power data.

Dimensionality reduction

Input matrices were events (~800 word items) × time (nine bins) × frequency (six bins) × regions (up to 22). We wanted to test the effects of new dimensionality reduction techniques in our work. We used t-distributed stochastic neighbor embedding (tSNE). tSNE provides a way to reduce high-dimensionality data and aid in visualization of embedded features. This is particularly suitable for high-dimensional

EEG data that lie on several different, but related, low-dimensional time series manifolds of several trials. The algorithm positions the data points optimally in the projection map by defining the similarity between N data points in the high-dimensional space X and the low dimensional space Y by measuring the the pairwise similarity between data points as shown below.

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N} \quad (1)$$

$$where \quad p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)} \quad (2)$$

for space X and we calculate a similar conditional probability in the low dimensional space Y, where y_i and y_j are the counterparts of x_i and x_j from the higher dimension. Thus we model the similarity of map point y_j to y_i using the following equation:

$$q_{ij} = ((1 + \|y_i - y_j\|^2)Z)^{-1} \quad (3)$$

$$where \quad Z = \sum_{k=1}^{N} \sum_{k \neq l}^{N} (1 + \|y_k - y_l\|^2)^{-1}. \quad (4)$$

If similarity for these points y_i and y_j in the lower dimension perfectly match the similarity for the high dimensional datapoints x_i and x_j, then the ratio of the conditional probabilities would be 1. The overall goal of t-SNE is to obtain a

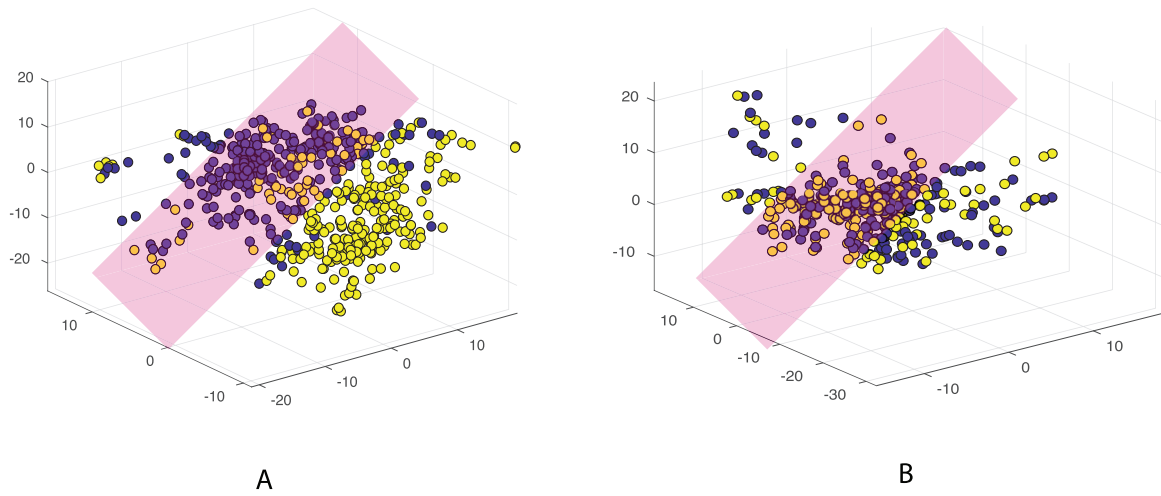## t-SNE Visualization of subject level data

A

B

Figure 2. Examples of tSNE data visualization. Examples in which the application of tSNE appeared to improve classification success for a subject (A) and one for which tSNE did not seem to aid classification (B).

low-dimensional representation that minimizes the difference between $p_{i|j}$ and $q_{i|j}$. The σ_i parameter is computed for each point x_i in the dataset so that the effective number of its neighbors corresponds to the fixed parameter called μ (perplexity) shown as below:

$$\mu = 2^{-\sum_j^N p_{j|i} log_2 p_{j|i}}. \qquad (5)$$

The joint probability Q is computed using a Student's t-distribution kernel with one degree of freedom. P is the joint probability. The heavy tails of this t-distribution generate a larger separation of dissimilar points than a Gaussian distribution. t-SNE aims to continuously position these data points in space Y by minimizing the cost function C that is measure of (Kullback–Leibler) divergence between probability distributions:

$$C(P, Q) = KL(P||Q) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}}. \qquad (6)$$

In this respect, t-SNE gradient has the advantage of strongly repelling data points which have small pairwise distances in space Y and tries to solve the crowding problem. In this work, we are dealing with over 2000 dimensions in the time series data. The high computational complexity introduced by employing t-SNE on our dataset requires that we make use of the Barnes–Hut t-SNE (Van Der Maaten 2013) an evolution of the t-SNE algorithm that introduces different approximations to reduce the computational cost. When computing the t-SNE embedding, we projected our dataset into three-dimensional space to visualize it (shown in figure 2), however specific dimensionality reduction was set to 100 for the high-dimensional (all brain location) data and 30 dimensions for the reduced dataset (selected brain regions). We set the Barnes–Hut parameter σ to 0.5 for all analyses. We used a perplexity value of 30 and empirically examined separation achieved with values from 5 to 50. We initialized the gradient descent algorithm with 1000 iterations and observed the error during separation.

We applied the t-SNE algorithm separately to the input data matrix of each subject across all dimensions (electrode location, frequency band, time domain). Therefore the resulting (reduced) matrix that resulted from t-SNE was bespoke for each individual, permitting differences in relevant information across regions from individuals with slightly different implant arrays. The utility of essentially unsupervised dimesionality reduction (t-SNE applied across all available input data) could then be compared to the utility of reduction in the number of input brain regions inherent in the models generated using the core set of five brain regions described above.

Synthetic minority oversampling technique (SMOTE)

We used synthetic minority oversampling technique to handle a class imbalance problem between recalled and non-recalled events (Chawla *et al* 2002). In the free recall task (and similar strategies to test episodic memory), individuals will always remember fewer items than they will recall. SMOTE is an established a strategy for addressing this issue, which can impact model performance (Xue and Hall 2015). This algorithm is used to generate synthetic samples of data by using the feature space of the dataset. The minority class (successful encoding) is over-sampled by taking each minority class sample and augmenting the synthetic samples in the direction of the k minority class of their nearest neighbors. We choose k on the basis of the amount of the desired over-sampling. Our implementation used three nearest neighbors ($k = 3$). As an important note, we performed SMOTE only on the training set and not on the testing set in order to maintain complete isolation for the testing set.

Pattern classifiers

We used three different methods of pattern classification for this analysis. Inputs to all classifiers in the core models were

Table 1. Grid search parameters for SVM model.

C	0.1	1	2	3	5	10
Gamma	1.0×10^{-08}	1.0×10^{-06}	1.0×10^{-05}	1.0×10^{-04}	1.0×10^{-03}	1.0×10^{-02}
kernel	Gaussian (RBF)	Linear	Quadratic	Cubic		

Table 2. RNN grid search parameters.

Batch size	10	15	30	40	45	50	60
Epochs	100	200	400	800	1000	1500	2000
Optimizer	adam	rmsprop	dadelta	adagrad	momentum	nag	sgd
Learning rate	0	0.01	0.03	0.001	0.003	0.0001	0.0003
Dropout fraction	0.1	0.15	0.2	0.25	0.3	0.35	0.40
Hidden layers	1	2	3	4			

oscillatory power averaged into six time bins. A logistic regression model was designed using parameters drawn from those in previously published work by our group, using $N - 1$ cross validation. Holdout data was an entire testing session (third session) with the classifier trained on two sessions (leave one *session* out cross validation), consistent with previously published methods (Ezzyat *et al* 2017). We believe this validation method is critical for testing classifier performance. Trained classifiers were fed the held out session data and classification accuracy was assessed using area under the receiver operating characteristic curve (AUC), a standard measure of a classifier's ability to generate true positives while avoiding false positives. The entire confusion matrix was also calculated. Following our published methods, for the logistic regression model we employed a binomial distribution (with link:logit) with C parameters optimized through grid search (Ezzyat *et al* 2017). Differences compared to this publication were in the number of time bins used and a standard set of lambda values for regularization (0.1, 0.01, 0.001).

SVM computes a hyperplane which finds the largest gap between the successful and failed recall classes. Points may lie on the edges of this 'gap', but not inside. The distance between these outside hyperplanes can be maximized by finding the greatest distance between edge cases with respect to the normal vector $\vec{w}$ (defined as the distance between the two edge hyperplanes). This works perfectly for linearly separable data, but fails when data is not necessarily so, such as in the case of many dimensions. As our data is highly complex, we utilized several kernel based SVM combinations (e.g. quadratic, Gaussian) in order to transform data into space that can more easily be separated with a hyperplane. The C and gamma parameters of the SVM were tuned. Incorporating a cost function allows for adjustment of alpha and beta error levels; applying kernels like quadratic or gaussian alters the shape of the hyperplane for more optimal fitting. These kernels are represented by these following equations:

$$K(\vec{x}_i, \vec{x}_j) = exp(-\gamma \|\vec{x}_i - \vec{x}_j\|)$$
$$K(\vec{x}_i, \vec{x}_j) = (p + \vec{x}_i \cdot \vec{x}_j)^q. \tag{7}$$

The parameters of the SVM classifier such as the scale of the Gaussian kernel and the degree of the polynomial kernel can be changed to improve model fit. These were optimized using a grid search method. The C and gamma parameters in SVM are modeled by the RBF (radial basis function) kernel itself. The C parameter trades off misclassification of training examples against simplicity of the decision surface: a low C makes the decision surface smooth. The gamma parameter defines the influence of training examples on the model. The C-gamma grid is shown in table 1; parameters were selected using only the training and not testing data set to avoid overfitting.

For the RNN analysis we vectorized the samples for input into the neural network and dispersed them into many batches during training. The one-hot encoding technique (binary representation of categorical data) was adopted for labels, consistent with previous methods for EEG analysis (Palaniappan and Mandic 2007). Weights and biases of the neural network were initialized by randomly choosing values from a truncated normal distribution. We selected hyperparameters (hidden layers, training iterations, learning rate, batch size, and dropout probability). As described above, we used a long-short term cell (LSTM) framework with cells enveloped in a dropout wrapper. This aids in regularization of the model by testing the output against random loss of hidden layer neurons (preventing overfitting). Finally, the addition of L2-regularization improved our model's resistance to overfitting, and, in concert with the dropout algorithm (Hinton *et al* 2012, Srivastava *et al* 2014), resulted in exceptional robustness and reliability of the classifier. We tested the classifier with session-wise cross-validation (as for SVM and LR models) and logged the measured AUC and accuracy over the defined epochs. Grid search parameters are shown in table 2.

Our deep learning model was implemented and trained using Google's Tensorflow2 Deep Learning Framework and monitored continuously using Tensorboard (www.tensorflow.org). The entire model was trained with the help of an NVIDIA GPU running on a Google Cloud GPU instance. We credit this platform with improving the learning rate and capacity of our model (NVIDIA Tesla K80). We plotted the time required to train an optimized model for each type of classifier. LR and SVM models were trained on a single 384 GB (quad-core) node running RedHat Enterprise Linux 6, and the SLURM job scheduling software, while RNN models were trained on the GPU nodes described above. This difference in the computing platform for each is the reason training times were shorter
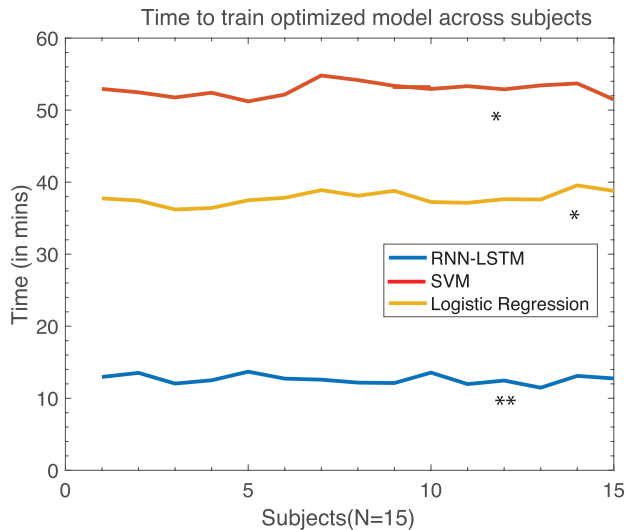
Figure 3. Time needed to train optimized model across classifier type. * denotes training on single quad-core node, while ** indicates training on GPU nodes employing Nvidia Tesla K80 cards.

Table 3. SVM versus logistic regression.

	SVM	LR	Corrected p value	Subject number
Select regions with tSNE	0.68	0.60	0.005	15/15
All regions with tSNE	0.64	0.58	0.024	13/15
Select regions without tSNE	0.63	0.59	0.013	14/15
All regions, without tSNE	0.64	0.56	0.009	14/15
Non-dominant, select regions, with tSNE	0.67	0.59	0.020	15/15

for RNN models in spite of greater computational intensity, shown in figure 3.

Statistical comparisons

We focused our analysis on a comparison of AUC values at the subject level and across subjects. Because a unique model was generated for each individual subject, we calculated an AUC value per subject and are able to compare the fraction of subjects exhibiting greater or smaller AUC for one model or the other (for example SVM versus LR). We also report the results of a Wilcoxon signed-rank test for comparison of distributions of AUC values across subjects between modalities. For the recurrent feature elimination analysis (across brain regions) we used the Kruskal–Wallis test. P values were subjected to FDR correction using the Benjamini and Hochberg method (Benjamini and Hochberg 1995)

Results

SVM versus LR with and without tSNE dimensionality reduction

Results of the SVM and LR comparison is presented in table 3. In the dominant hemisphere, SVM significantly outperformed LR both in magnitude of mean AUC (p value for comparison presented in table) and in number of subjects in whom the SVM model provided better classification. This held true with and without the application of tSNE and whether all regions or only select regions were included in the classification model. We directly tested the impact of including all regions in a given patient implantation in the SVM and LR models. The classifiers performed significantly worse compared to models developed using only select regions (corrected $p = 0.012$ for SVM and corrected $p = 0.011$ for LR,

signed rank test comparing AUC distributions for select versus all regions). Full subject data, including F1 plots, are shown for the best-performing model (select regions with tSNE) in figure 4. When we removed t-SNE dimensionality reduction across all regions, classification was slightly worse for LR (AUC = 0.56), although tSNE had the largest benefit for prediction in the SVM model with selected regions. We discuss ways to optimize the benefits of tSNE in the *Discussion*. This finding regarding the benefit of regional selectivity has important implications for the construction of classifiers for predicting stimulation as well as the design of clinical devices building upon our previous findings.

SVM versus LR in data from the non-dominant hemisphere

We applied the identical methods to a separate cohort of 15 subjects who had stereo EEG implantation in the language non-dominant hemisphere. In all subjects, this was the right hemisphere as proven by preoperative fMRI or Wada testing. This was a separate patient cohort, providing evidence of robustness for our methods. For the *a priori* planned comparison of LR versus SVM using signal obtained from selected brain regions with dimensionality reduction (the best performing model), the mean AUC was 0.67 for SVM versus 0.59 for logistic regression (corrected $p = 0.020$). Results for all subjects are summarized in figure 5.

SVM classifier performance with and without connectivity information

We wanted to examine the effect of adding connectivity information to the prediction model to determine if this information would improve performance of the classifier. We employed spectral coherence because this offers a trial by trial estimate of connectivity during memory encoding that is not available with the phase locking statistic or other measurements of phase–phase synchrony (Lachaux *et al* 1999, Bokil *et al* 2010). Compared to an SVM model that did not include connectivity information, across 15 subjects, the addition of spectral coherence information among five brain regions did not improve classifier performance significantly, either with our without the addition of tSNE to the data (corrected $p = 0.331$). The classification performance with and without connectivity information is summarized for all subjects in figure 6.
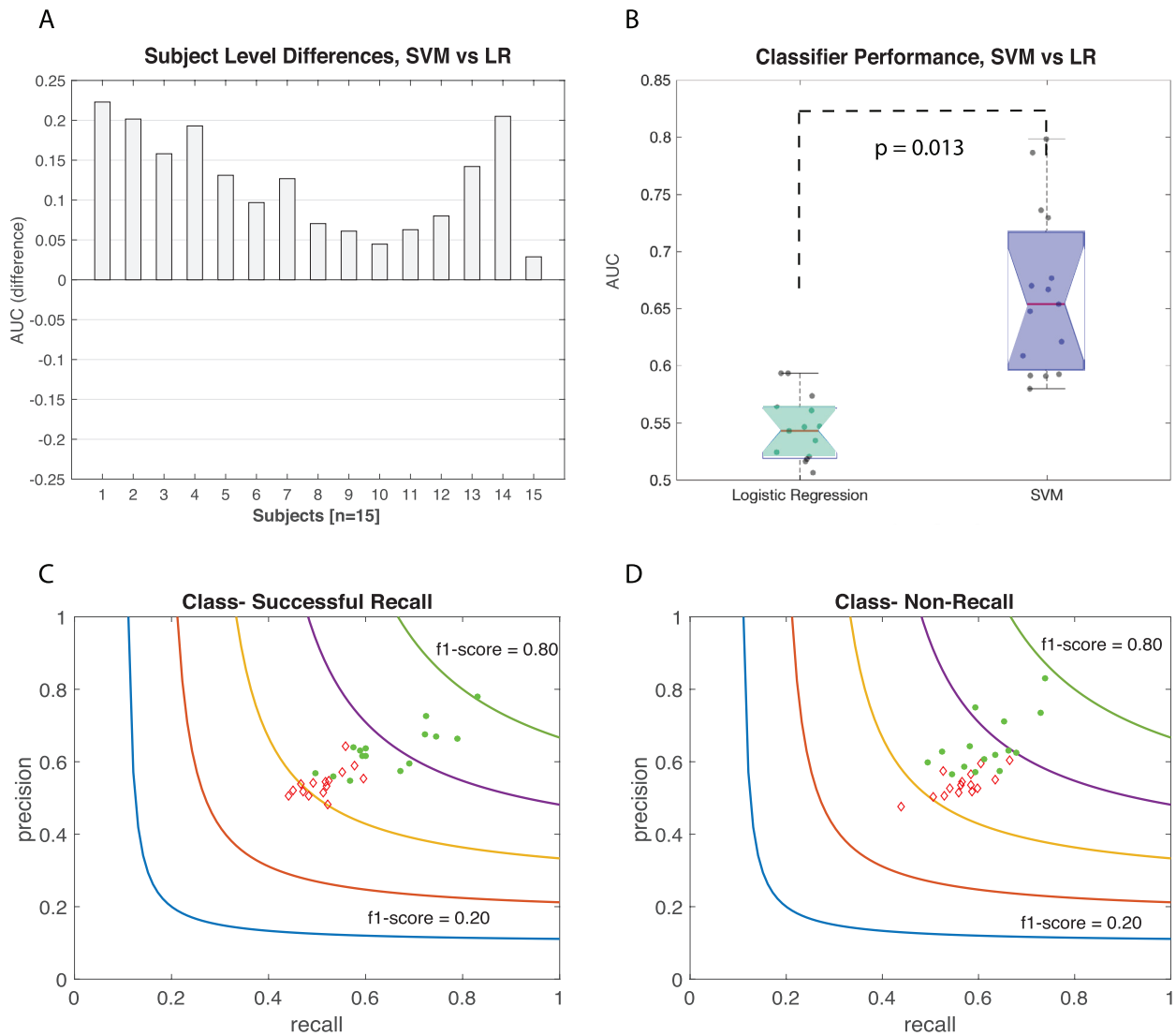
Figure 4. SVM versus LR classifier performance (with tSNE, selected regions). (A) AUC difference for each subject (SVM-LR). (B) Notched boxplots showing distribution of AUC values for SVM versus LR using select brain regions. FDR corrected p values shown. (C) F1 score for the recall and (D) non-recall classes for both SVM and LR models. LR results in red, SVM in green.

SVM classifier performance with reduced frequency band information

We tested the effect of reducing the number of input frequencies on the performance of the SVM-based classifier (analogous to reduction in the number of input regions). We selected the theta and high gamma bands for this model based upon existing human data suggesting that power changes in connectivity information in these frequencies predict successful encoding (Burke *et al* 2014b). Results are shown in figure 7.

SVM versus LSTM RNN model

We next sought to compare the results of the best-performing SVM based classifier (reduced brain regions, with tSNE) with a recurrent neural network model based upon LSTM cells. We designed and implemented a recurrent neural network prediction model via TensorFlow software and investigated its performance (Abadi *et al* 2016). Across 15 subjects, the

RNN classifier achieved a mean AUC of 0.722 as compared to mean AUC of 0.68 in SVM, which was significantly different (corrected $p = 0.013$). RNN based classification was superior in 12 of 15 subjects. Results of the comparison are shown in figure 8.

We observed that RNNs did not improve classification accuracy for three of the subjects. We examined the correlation between AUC and memory recall performance (fraction of successfully recalled events), looking to assess the impact of class imbalance on recall performance. We observed that for all three methods (LR, SVM, and RNNs), the models performed better as subjects recalled a larger fraction of items, with a significant correlation between recall fraction and classifier performance ($\rho = 0.52, 0.51, 0.62$ for RNN, LR and SVM respectively). Interestingly, SVM models were the most sensitive to recall fraction but also achieved the best magnitude of classification for those subjects who were better performers. We observed that RNN models were less sensitive to poor memory performance in subjects (and the associated
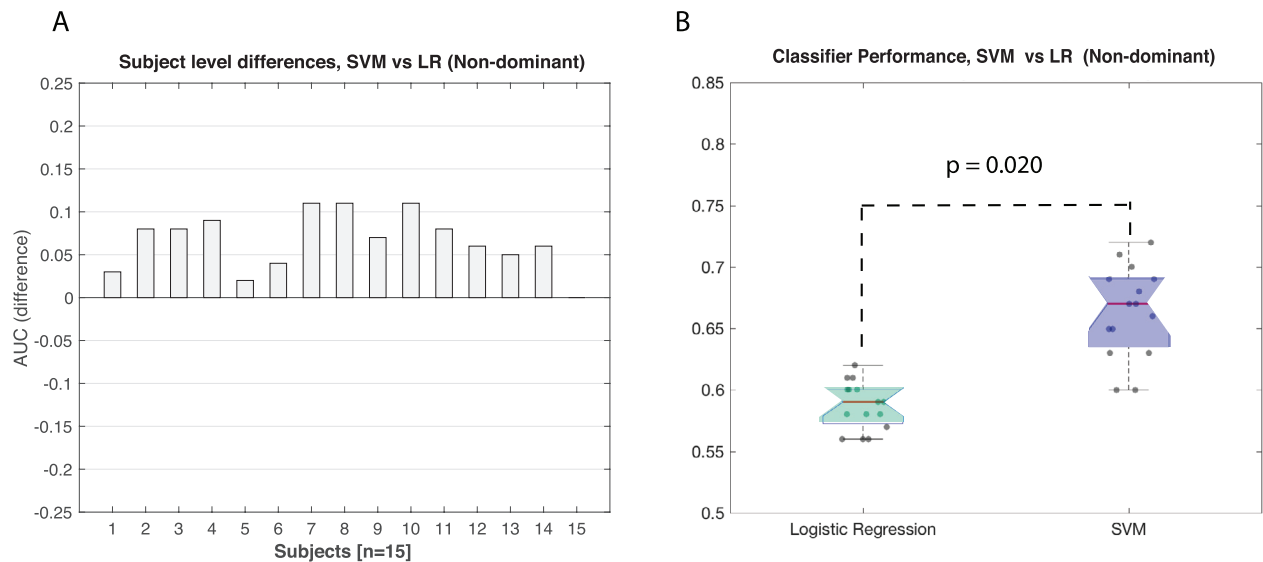
A

B

Figure 5. Right hemisphere SVM versus LR. (A) AUC difference for each subject (SVM-LR). (B) Mean AUC difference.
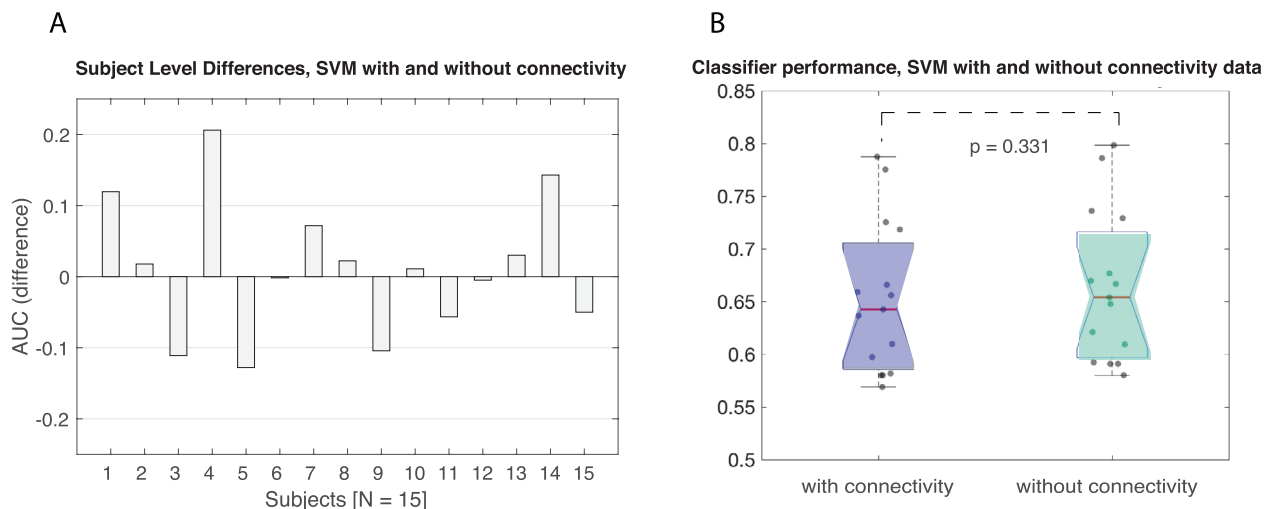
A

B

Figure 6. Effect of adding connectivity information (theta frequency coherence) on SVM classifier performance. (A) AUC difference for SVM classifier with and without connectivity information for each subject with tSNE applied to data. (B) Mean AUC difference; FDR corrected p value.

greater class imbalance). Results are shown in figure 9; we use the best-performing SVM and LR models for this comparison.

Using machine learning to model memory loss after seizure surgery

Our unique data set, including subjects that had electrodes in matched brain regions, allowed us to use the machine learning classifier in a new way. Because a core set of five brain regions was shared across all subjects we could examine the effect on overall classifier performance by leaving one brain region out of the predictive model and examining the effect on the AUC to demonstrate the feasibility of this method. We believe this is a novel way of quantifying how much a given brain region is contributing to the overall process of episodic memory that we believe in future can be applied to individual

subjects. It is based on recurrent feature elimination (RFE) described in the genetics literature (Guyon *et al* 2002). To test this idea, we trained the SVM classifier using input data from five common regions for each subject and then on four of the five regions (leaving one region out of the training set for the model), comparing the effects on the overall predictive accuracy. By depriving the model of the information for a given brain region, the resulting deficit in classification accuracy we believe provides some estimate of the role of a given region within an memory network. Results are shown in figure 10, showing the mean AUC difference across regions. We tested this effect using a repeated measures non-parametric model with a primary factor of region, suggesting that leaving out the hippocampus had a relatively greater impact on classification than other regions (corrected $p = 0.008$, Kruskal–Wallis test). This finding we believe validates this method conceptually,
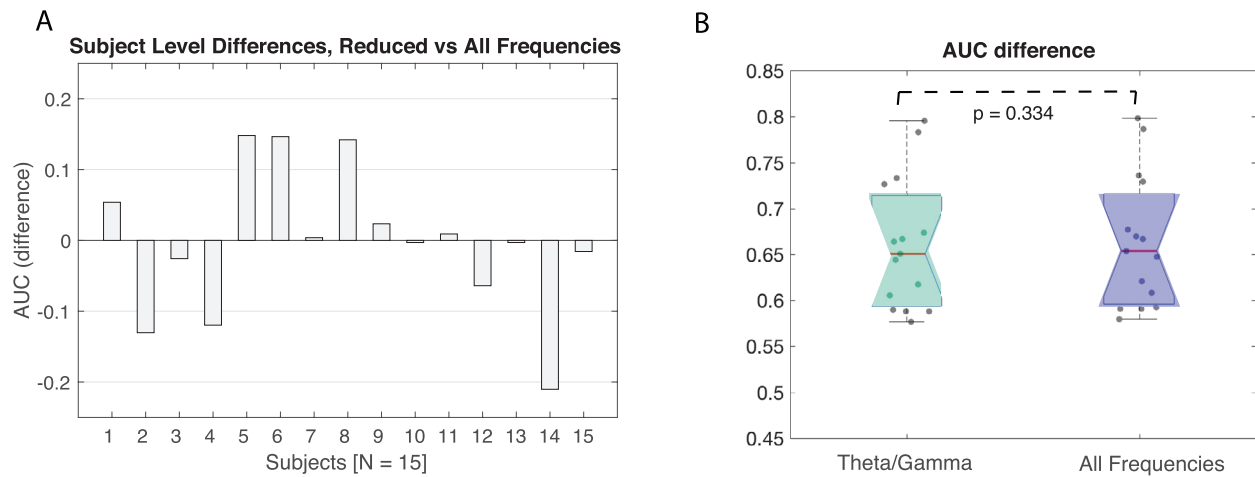
Figure 7. Effect of frequency band reduction (theta and gamma bands exclusively used) on classifier performance. (A) AUC difference for each subject for best performing SVM model (reduced regions, with tSNE) with only theta/gamma bands included (left) and all frequency bands (right). (B) Mean AUC difference with no significant difference across subjects in model performance; FDR corrected p value.
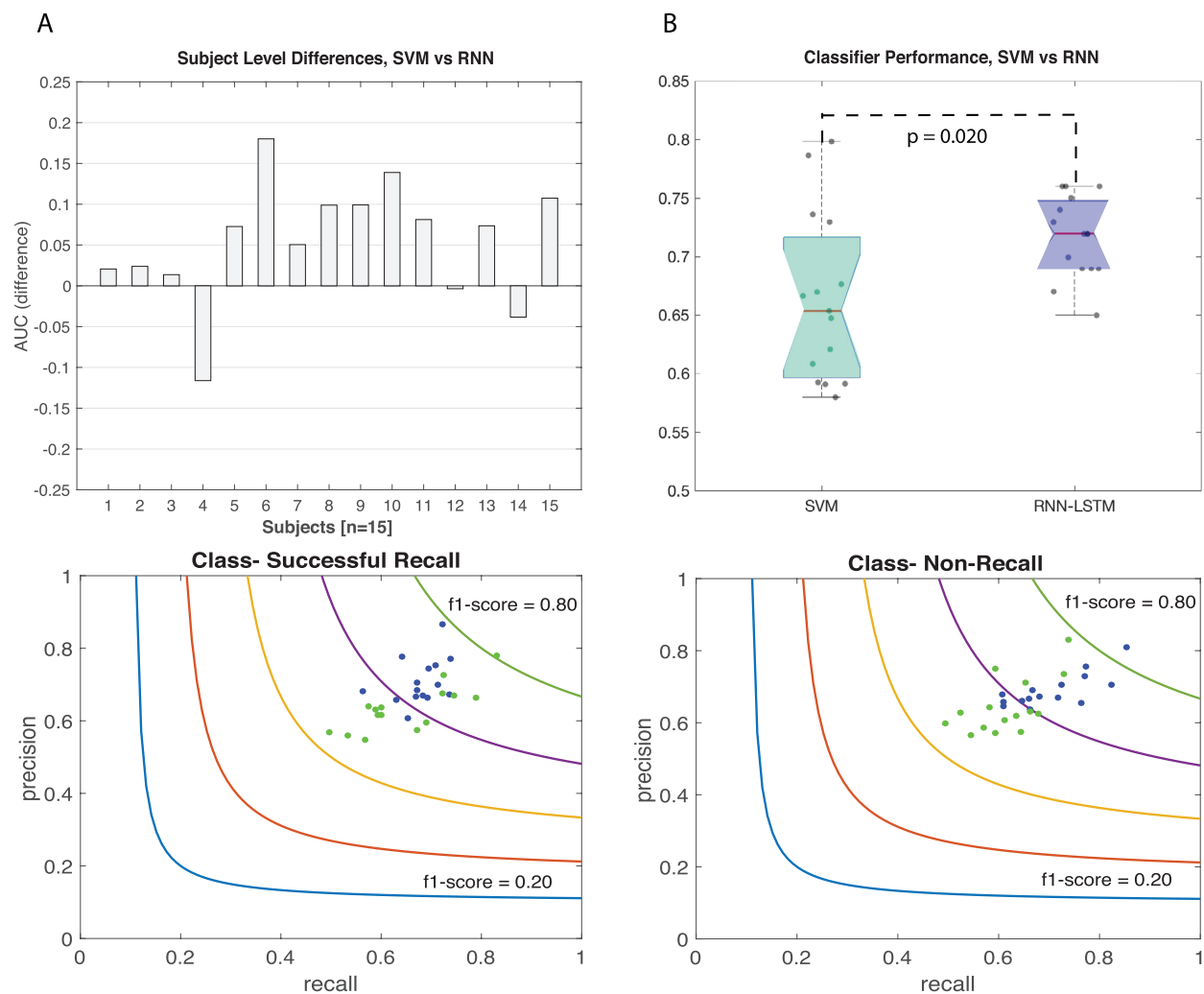

Figure 8. Direct comparison of best-performing SVM model versus RNN. (A) AUC difference for each subject (RNN-SVM). (B) Mean AUC difference. (C) F1 score for the recall and (D) non-recall classes for both SVM and RNN models.
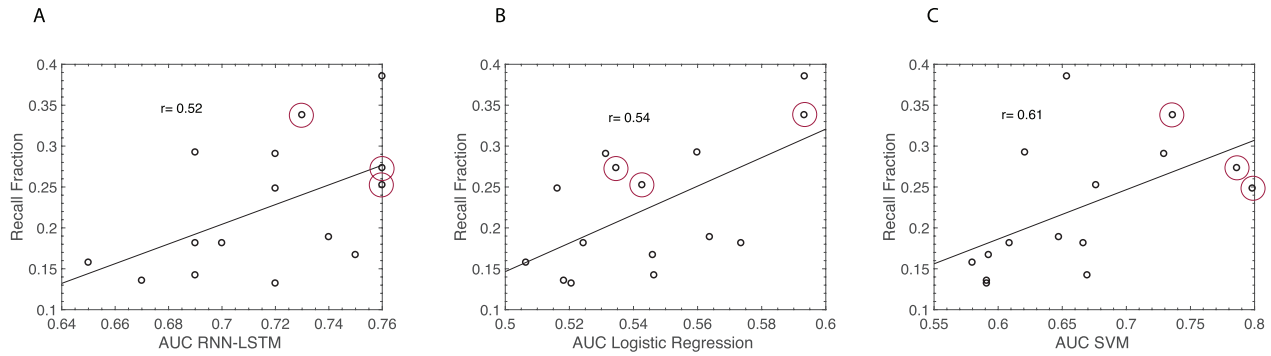
Figure 9. Observed correlation between subject recall performance and classifier performance. (A) RNN-LSTM model. (B) LR model. (C) SVM model. Circles drawn around subjects in whom SVM outperformed the RNN model.
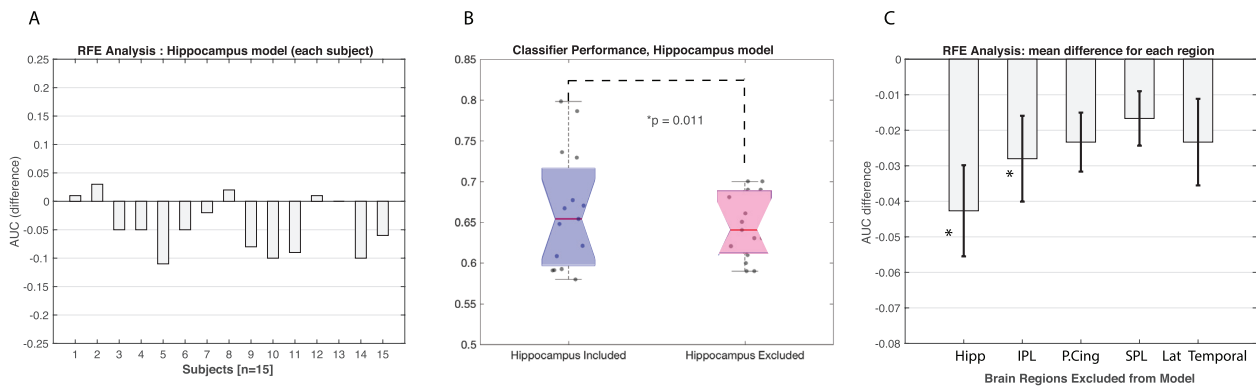

Figure 10. Results of recurrent feature elimination. (A) AUC difference for each subject for hippocampal model (full model—hippocampus excluded). (B) AUC distributions for full and hippocampal excluded SVM models. (C) Mean AUC difference for each of five select regions excluded from model (hippocampus, posterior cingulate, superior parietal lobule, inferior parietal lobule, lateral temporal cortex). Largest impact on AUC observed with leaving hippocampus out of model, least for SPL. Significant reduction in AUC observed for the hippocampus and IPL, corrected p values 0.02 and 0.05, respectively.

as one would predict hippocampal information would be relatively more valuable to predicting memory performance across subjects. While these results are promising, we discuss the limitations of this approach in the *discussion* section.

Discussion

Support vector machines outperformed logistic regression classifiers in nearly all subjects

Logistic regression algorithms can be sensitive to noise and can lead to erroneous classifications, although penalized classifiers are less so. The noise in timeseries EEG data led us to believe that SVM would outperform logistic regression, as we observed. The relatively modest effect sizes for subsequent memory underlie this noise issue (Lega *et al* 2011). Logistic regression also suffers from a multi-collinearity problem; this occurs when predictor variables are highly correlated leading to instability in their associated coefficients due to inflation of the standard errors. This can affect classifier performance because logistic regression is based on the strong assumption that independent variables and their log odds have a linear relationship between them, the violation of which can lead to misclassifications. SVM does a better job of modeling nonlinear relationships by using nonlinear kernels that separate the data using a hyperplane (Lotte *et al* 2007). With optimization of

the kernel, SVM can handle large feature space more effectively than logistic regression taking a geometric rather than probabilistic approach to classification. For all of these reasons, our results are consistent with *a priori* expectations of superior performance for SVM for classification of EEG timeseries data, which shows improvement on the order of 6% in other EEG classification problems (Garrett *et al* 2003). We may derive additional optimization benefit by using a multiple kernel approach rather than optimization of individual kernels on a subject basis as in our method (Li *et al* 2014). An important point of emphasis is that for the select region-based models, input data is fairly uniform across subjects. The consistent performance across subjects with uniform input array is a strong argument for the generalizability of our findings. One caveat for the LR comparison however is that our model here incorporated a greater number of time bins than in our previously published data, as we have observed temporal dynamics in memory effects that we wanted to use to improve model performance (Burke *et al* 2014a, Lin *et al* 2017). This additional dimensionality may have adversely impact the LR method.

Using selected regions improves classification accuracy

A principal goal of our analysis is to inform future design of BCI devices for memory. A specific challenge for memory

devices is that the number of regions that can be sampled from in a given individual is necessarily limited, even more so as the technology moves beyond 'opportunistic' recordings in epilepsy patients and towards the design of an implant explicitly for people with memory problems. The diminished benefit of adding additional input electrode contacts has been reported previously in EEG based BCI classification efforts (Kaper *et al* 2004). From a practical standpoint, it may be possible to sample from a relatively limited number of well-chosen brain areas and achieve good classification and device performance. The selection of these regions may in the future be guided by non-invasive information such as fMRI-based connectivity analysis or network control theory (Gu *et al* 2015). These regions could be selected based on known anatomical circuitry (such as the mesial temporal limbic network that characterized sampling in our study) or could be bespoke based upon individual patient factors. This remains an active area of investigation.

Connectivity information and frequency bands

The issue of optimizing data inputs to a classifier model based on *a priori* considerations is directly related to the question of adding connectivity information. Connectivity information led to improved classifier performance in around half of individuals but was not significant overall. We have not tested the addition of multiple different connectivity parameters, and the most commonly used phase synchrony parameter (phase locking statistic) must be calculated across trials and therefore cannot be used for classifier training and the type of analysis we performed here (Lachaux *et al* 1999). Certainly, other trial by trial estimates of phase synchrony and other estimates of coupling such as cross frequency relationships could be included in future models. We may observe that the reduction in highly correlated features offered by techniques such as tSNE has a greater impact on model performance with connectivity and power information included (if specifically tuned, see below). Or, more traditional machine learning approaches such as SVM may be incapable of handling this added information and the inclusion of connectivity data will require RNNs. Overall, this observation we believe informs BCI device design by helping identify critical features for *a priori* dimensionality reduction (focusing on within-site power memory effects). We also observed that reducing the frequency bands included in our (reduced dimension) SVM model did not strongly affect classifier performance overall, indicating that the critical information for classification occurs in the theta and gamma bands across subjects (although not in every subject). This is not surprising given previous observations in human subjects demonstrating the importance of activity in these frequency ranges for predicting successful memory encoding (Burke *et al* 2014a). However this also represents something that can be optimized on a subject by subject basis.

t-SNE improved performance in the logistic regression algorithm by 1%–5%

We had initially anticipated that dimensionality reduction would provide the greatest benefit for logistic regression

classifiers given the issues mentioned above with sensitivity to multicollinearity. However in practice, the greatest magnitude of impact of tSNE was for SVM classifiers operating with data from selected regions. The magnitude of performance improvement we observed was in line with our expectations derived from the literature (Garrett *et al* 2003, Gisbrecht *et al* 2015). The importance of this finding is that it implies an optimized model requires both *a priori* rationally selected feature reduction and 'hypothesis neutral' feature reduction provided by algorithms such as tSNE. tSNE may ultimately prove more valuable for the higher dimensionality dataset when we specifically tune its input parameters for these data, although for the purposes of comparison for this analysis we kept them identical. tSNE may also be of use as a visualization tool for rapidly determining which features should be included in a model (*a priori* reduction of input data) prior to model training, as our plots in figure 2 reflect. The dataset available to us for testing rational dimensionality reduction across subjects (the five core regions included in our selected regions model) was essentially based on convenience: these regions were those most highly conserved across subjects. The design of a commercial brain–machine interface device will require careful *a priori* selection of brain regions based upon the application of our RFE method to rank the importance of brain regions towards classification across a large dataset that includes significant representation (in terms of electrode coverage) in a broad number of brain areas. Further, the application of pre-implantation fMRI may be a means to identify relevant brain regions, but this requires an analysis in which *the same individuals* participate in both an fMRI memory paradigm as well as memory testing following electrode implantation using a similar memory task. Both of these efforts are ongoing.

RNNs outperform SVM in 75% of subjects

Recurrent neural networks were designed for sequential learning, and effective classification using timeseries data is a core achievement of RNNs (Hinton *et al* 2012, Graves *et al* 2013, Hermans and Schrauwen 2013). Our models required 1000 iterations to reach convergence per subject employing four layers of LSTM cells. In the future, we intend to further optimize classification using the RNN method by providing additional timeseries resolution (greater number of time bins used for this work). RNNs with LSTM should be able to accommodate the expanded dimensionality of the data without over fitting. They also open the possibility of modeling list behavior within the classification algorithm. Memory from stimuli that are sequences of items contains temporal architecture such as the primacy and recency effects. Including this temporal list information may improve pattern classification. This is because it has been reported that the underlying oscillatory patterns reflect the temporal architecture observed in behavioral data (Serruya *et al* 2012). While our RNN model outperformed both SVM and LR methods, the magnitude of the benefit in terms of classification was relatively modest compared to SVM models. Further refinement of the particular features of the RNN model may improve this

result in the future. Improvements in classification may be achievable using a different deep learning approach based on convolutional neural networks, as has been recently published (Bashivan *et al* 2015). The insight of this approach was to preserve spatial information of the recording (surface EEG) electrodes when constructing the neural network, drawing upon image recognition strategies. The high classification accuracy this publication reports suggest that such spatial information preservation may be an effective strategy at improving overall performance, although stark differences in the nature of the memory task (working memory load) versus episodic memory encoding mean that the approach would have to be tested carefully in our data. Another factor that reduces our reported classification accuracy is the use of session wise cross-validation, by which classifier performance is always tested on data from an experimental session different from those used in the training set (typically a separate day of the experiment). For a brain machine interface device, session wise cross validation is critical since we have observed that EEG patterns associated with memory encoding success can change across sessions and the identification of features that do not vary across sessions is necessary. We believe that demonstrating good performance across multiple data acquisition days is necessary to establish stable features to guide responsive stimulation (Ezzyat *et al* 2018).

Class imbalance in memory classification

In episodic memory paradigms such as free recall, there is necessarily a class imbalance problem related to the fact that individuals forget more items than they remember. This cognitively intensive task elicits strong memory-related effects however, making them well-suited to exploring and understanding how the brain processes memories. Importantly, this class imbalance situation actually approximates our daily experience in which a single presentation of a memory item is generally insufficient to elicit robust encoding. This in turn requires execution of a strategy to match classes; for this purpose we employed SMOTE, having chosen to over-sample the minority class rather than the opposite (Chawla *et al* 2002). Updated strategies of class balancing may be a strategy for improvement of machine learning classifiers in memory paradigms. We observed that RNN models were less sensitive to class imbalance than either SVN or LR models in our data, but that SVM models may perform slightly better for subjects who are high performers (with classes that are nearly matched already).

Subject versus population level classification

The properties of recurrent neural networks suggest that they are well-suited to employing a different strategy for predicting memory encoding success using EEG timeseries data (Graves *et al* 2013). Specifically, they may allow us to utilize multiple subject data to develop a cross subject classifier with good performance on the individual subject level. Concatenating trials and time series across multiple subjects will greatly increase the dimensionality of the data but may also allow the classifier to learn relatively invariant features of memory performance across individuals. The data set we employed with relatively consistent sampling across brain regions via the stereo EEG method is well-suited to this strategy as subject data is more uniform. The appeal of a classifier trained across multiple subject data is that it would permit custom modulation strategies that do not require individualized training via invasive intracranial electrodes. Such a model could be updated with noninvasive patient specific information from fMRI or DTI acquisitions. The application of a model trained on general features across many subjects would obviate the need for individualized testing. This remains an ongoing area of investigation, utilizing a large subject pool with electrodes across multiple brain regions Ezzyat *et al* (2018). This approach will likely need to be combined with systematic testing of optimum electrode locations as described above.

Other concerns for the design of a brain machine interface device for memory

Development of a brain machine interface device to alter memory performance will require optimization of several factors. The principal goal of this analysis was to examine the utility of alternative methods of binary classification for one of these factors (performance of the classifier of brain recording information), but we have also examined strategies for optimizing other factors including which brain regions are included in a classifier (related to dimensionality reduction). Another concern for such a putative device is whether stimulation will alter normal cognitive processing in other domains, such as working memory, mathematical processing, or spatial navigation. This will require that a classifier not only distinguishes between successful and unsuccessful encoding situations but also between memory encoding and these other cognitive states. Limiting the application of stimulation to epochs only when the brain is actively attempting to encode new memories should reduce its impact on other domains and allow normal homeostatic processes to occur. This will require a classifier to distinguish between brain states and limit stimulation overall. Limiting the needed number of input channels also can reduce the side effect profile of a brain machine interface device, because it would entail less risk of brain injury due to electrode implantation. For this reason, investigators must determine the minimum number of recording locations that produce effective classification. There is a related question, as to whether a brain machine interface device would perform better if designed to enhance memory performance for specific memory items, taking advantage of distinct representational maps and associated EEG features for reactivation (Edelman 1993, Belal *et al* 2018). We recently tested whether incorporating semantic information would alter classifier output using a version of the free recall task in which a portion of memory items come from semantic categories while most of the items do not (categorized free recall). We observed that, with data across whole brain, logistic regression classifiers performed equally well for both semantically-related and unrelated items (Weidemann *et al* 2018). However, it is possible that classification accuracy or BCI device performance can be

improved with preservation of semantic information; this may depend upon inclusion of brain regions (such as lateral temporal cortex) that provide more specific semantic features of a memory representation. The role of item-specific information in device design remains an active area of investigation.

Predicting memory loss after temporal lobe surgery

We believe that our preliminary analysis can form the foundation of a strategy to predict memory deficits after temporal lobe epilepsy. Using recurrent feature elimination, the difference in classifier performance between the model incorporating all available brain regions versus a model with a single region removed from the training set provide some estimation of the impact of surgery on memory behavior. Certainly this strategy will need prospective validation, but the use of EEG for this purpose offers advantages over cross-subject models based on language lateralization scores that currently offer the best results (Sidhu *et al* 2015). As discussed above, our experimental protocol prioritizes brain signals associated with multiple different semantic representations (those that are invariant across individual memory items but are similar in that they were successfully recalled) at the expense of information tied to specific memory features (the representation of 'king' versus 'cat' for example). With temporal lobectomy, there is undoubtedly an impact on memory ability for general episodic abilities and for specific representations (especially anomia following resection of the dominant temporal lobe). A comprehensive prediction of the effects of surgery will need to include both types of information to offer credible recommendations to a patient. While the addition of connectivity information did not significantly improve classifier performance in our analysis across subjects, this information may still prove critical for estimation of memory decline after surgery by modeling the impact of removal of a single brain region that participates within an overall memory network.

Conclusion

We tested logistic regression, support vector machine, and deep learning approaches to predicting memory encoding success using iEEG recordings. In line with results from other BCI applications, we observed significant improvement using both SVM and RNN strategies, and further that a model based on select brain regions may outperform higher dimensionality models. We applied recurrent feature elimination to model the effects of hippocampectomy on episodic memory and tested the application of a recently devised dimensionality reduction strategy (tSNE). These findings can inform new strategies of responsive stimulation paradigms for episodic memory and other applications.

Acknowledgments

ORCID iDs

Bradley Lega https://orcid.org/0000-0003-3634-406X

Supplementary Material

Supplementary information can be found online at stacks.iop.org/JNE/15/066028/mmedia.

References

Abadi M *et al* 2016 Tensorflow: large-scale machine learning on heterogeneous distributed systems (arXiv:1603.04467)

Bashivan P, Rish I, Yeasin M and Codella N 2015 Learning representations from EEG with deep recurrent-convolutional neural networks (arXiv:1511.06448)

Belal S, Cousins J, El-Deredy W, Parkes L, Schneider J, Tsujimura H, Zoumpoulaki A, Perapoch M, Santamaria L and Lewis P 2018 Identification of memory reactivation during sleep by EEG classification *NeuroImage* **176** 203–14

Benjamini Y and Hochberg Y 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing *J. R. Stat. Soc.* B **57** 289–300

Birjandtalab J, Pouyan M B and Nourani M 2016 Nonlinear dimension reduction for EEG-based epileptic seizure detection *IEEE-EMBS Int. Conf. on Biomedical and Health Informatics* (IEEE) pp 595–8

Bokil H, Andrews P, Kulkarni J E, Mehta S and Mitra P P 2010 Chronux: a platform for analyzing neural signals *J. Neurosci. Methods* **192** 146–51

Burke J F, Long N M, Zaghloul K A, Sharan A D, Sperling M R and Kahana M J 2014a Human intracranial high-frequency activity maps episodic memory formation in space and time *NeuroImage* **85** 834–43

Burke J F, Sharan A D, Sperling M R, Ramayya A G, Evans J J, Healey M K, Beck E N, Davis K A, Lucas T H and Kahana M J 2014b Theta and high-frequency activity mark spontaneous recall of episodic memories *J. Neurosci.* **34** 11355–65

Burke J F, Zaghloul K A, Jacobs J, Williams R B, Sperling M R, Sharan A D and Kahana M J 2013 Synchronous and asynchronous theta and gamma activity during episodic memory formation *J. Neurosci.* **33** 292–304

Chandaka S, Chatterjee A and Munshi S 2009 Cross-correlation aided support vector machine classifier for classification of EEG signals *Expert Syst. Appl.* **36** 1329–36

Chawla N V, Bowyer K W, Hall L O and Kegelmeyer W P 2002 Smote: synthetic minority over-sampling technique *J. Artif. Intell. Res.* **16** 321–57

Edelman G M 1993 Selection and re-entrant signalling in higher brain function *Neuron* **10** 1–20

Ezzyat Y *et al* 2017 Direct brain stimulation modulates encoding states and memory performance in humans *Curr. Biol.* **27** 1251–8

Ezzyat Y *et al* 2018 Closed-loop stimulation of temporal cortex rescues functional networks and improves memory *Nat. Commun.* **9** 365

Garrett D, Peterson D A, Anderson C W and Thaut M H 2003 Comparison of linear, nonlinear, and feature selection methods for EEG signal classification *IEEE Trans. Neural Syst. Rehabil. Eng.* **11** 141–4

Gisbrecht A, Schulz A and Hammer B 2015 Parametric nonlinear dimensionality reduction using kernel t-SNE *Neurocomputing* **147** 71–82

Graves A, Mohamed A R and Hinton G 2013 Speech recognition with deep recurrent neural networks *IEEE Int. Conf. on Acoustics, Speech and Signal Processing* (IEEE) pp 6645–9

Gu S *et al* 2015 Controllability of structural brain networks *Nat. Commun.* **6** 8414

Guyon I, Weston J, Barnhill S and Vapnik V 2002 Gene selection for cancer classification using support vector machines *Mach. Learn.* **46** 389–422

Hermans M and Schrauwen B 2013 Training and analysing deep recurrent neural networks *Advances in Neural Information Processing Systems* pp 190–8

Hinton G *et al* 2012 Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups *IEEE Signal Process. Mag.* **29** 82–97

Hortal E, Planelles D, Costa A, Iáñez E, Úbeda A, Azorín J M and Fernández E 2015 SVM-based brain–machine interface for controlling a robot arm through four mental tasks *Neurocomputing* **151** 116–21

Kaper M, Meinicke P, Grossekathoefer U, Lingner T and Ritter H 2004 BCI competition 2003-data set IIb: support vector machines for the p300 speller paradigm *IEEE Trans. Biomed. Eng.* **51** 1073–6

Kumar Y, Dewal M and Anand R 2014 Epileptic seizure detection using DWT based fuzzy approximate entropy and support vector machine *Neurocomputing* **133** 271–9

Lachaux J P, Rodriguez E, Martinerie J and Varela F J 1999 Measuring phase synchrony in brain signals *Hum. Brain Mapp.* **8** 194–208

Lega B, Jacobs J and Kahana M 2011 Human hippocampal theta oscillations and the formation of episodic memories *Hippocampus* **22** 748–61

Li X, Chen X, Yan Y, Wei W and Wang Z J 2014 Classification of eeg signals using a multiple kernel learning support vector machine *Sensors* **14** 12784–802

Lin J J, Rugg M, Das S, Stein J, Rizzuto D, Kahana M and Lega B 2017 Theta band power increases in the posterior hippocampus predict successful episodic memory encoding in humans *Hippocampus* **27** 1040–53

Lotte F, Congedo M, Lécuyer A, Lamarche F and Arnaldi B 2007 A review of classification algorithms for EEG-based brain–computer interfaces *J. Neural Eng.* **4** R1

Maaten L V D and Hinton G 2008 Visualizing data using t-SNE *J. Mach. Learn. Res.* **9** 2579–605

Palaniappan R and Mandic D P 2007 Biometrics from brain electrical activity: a machine learning approach *IEEE Trans. Pattern Anal. Mach. Intell.* **29** 738–42

Pascanu R, Mikolov T and Bengio Y 2013 On the difficulty of training recurrent neural networks *Int. Conf. on Machine Learning* pp 1310–8

Schlögl A, Lee F, Bischof H and Pfurtscheller G 2005 Characterization of four-class motor imagery EEG data for the BCI-competition 2005 *J. Neural Eng.* **2** L14

Sederberg P B, Kahana M J, Howard M W, Donner E J and Madsen J R 2003 Theta and gamma oscillations during encoding predict subsequent recall *J. Neurosci.* **23** 10809–14

Serletis D, Bulacio J, Bingaman W, Najm I and González-Martínez J 2014 The stereotactic approach for mapping epileptic networks: a prospective study of 200 patients: clinical article *J. Neurosurg.* **121** 1239–46

Serruya M D, Sederberg P B and Kahana M J 2012 Power shifts track serial position and modulate encoding in human episodic memory *Cerebral Cortex* **24** 403–13

Sidhu M K, Stretton J, Winston G P, Symms M, Thompson P J, Koepp M J and Duncan J S 2015 Memory fMRI predicts verbal memory decline after anterior temporal lobe resection *Neurology* **84** 1512–9

Srivastava N, Hinton G, Krizhevsky A, Sutskever I and Salakhutdinov R 2014 Dropout: a simple way to prevent neural networks from overfitting *J. Mach. Learn. Res.* **15** 1929–58

Stelzer J, Chen Y and Turner R 2013 Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control *NeuroImage* **65** 69–82

Sutskever I, Vinyals O and Le Q V 2014 Sequence to sequence learning with neural networks *Advances in Neural Information Processing Systems* pp 3104–12

Van Der Maaten L 2013 Barnes–Hut-SNE (arXiv:1301.3342)

Watrous A J, Tandon N, Conner C R, Pieters T and Ekstrom A D 2013 Frequency-specific network connectivity increases underlie accurate spatiotemporal memory retrieval *Nat. Neurosci.* **16** 349–56

Weidemann C T *et al* 2018 Neural activity reveals interactions between episodic and semantic memory systems during retrieval *J. Exp. Psychol.* (accepted)

Xu W, Guan C, Siong C E, Ranganatha S, Thulasidas M and Wu J 2004 High accuracy classification of EEG signal *Proc. 17th Int. Conf. on Pattern Recognition* vol 2 (IEEE) pp 391–4

Xue J H and Hall P 2015 Why does rebalancing class-unbalanced data improve AUC for linear discriminant analysis? *IEEE Trans. Pattern Anal. Mach. Intell.* **37** 1109–12