university of groningen / faculty of science and engineering

Classifier performance for the decoding of decision evidence from intracranial EEG data

Bachelor's Project Thesis

Jos van Goor, s1869051, josvangoor@gmail.com,
Supervisor: Dr. M.K. van Vugt

Abstract: Drift diffusion models have been successfully used to model the brain activity for decision tasks. Drift diffusion models describe the process of accumulating evidence until a certain evidence threshold is reached and a decision is made. In this exploratory research we attempted to find possible ways of classifying decision based intracranial electroencephalography (iEEG) data, and find out more about the time course of the brain's decision process. We used iEEG data from three different tasks, one visual search task, where participants needed to determine the location of a stimulus, and two match-to-sample tasks, where participants needed to compare, or recall images of faces. The participants consisted of people with implanted electrodes due to pharmalogically intractable epilepsy. Initially we tried to train a logistic regression classifier, assessing its performance with 10-fold cross-validation. This yielded results with an accuracy close to chance. In a subsequent attempt we found that feature selection improved the classifiers performance. We also found that support vector machines (SVM) were better able to predict decisions than regularized logistic regression. The SVMs performed at an accuracy consistently above 0.5, increasing towards the moment the participant responds. Finally we used the feature selection and classifier results to get a better understanding of the spatiotemporal evolution of the decision making process through the brain, and to find a pattern of increasing classifier accuracy over time to support the drift diffusion model.

1 Introduction

People make many decisions everyday, big and small, from trivial decisions such as which clothes to wear to a social gathering, to life changing decisions such as choosing what career path to take. Small and large decisions differ in several ways. The process of making small decisions for instance, is a time constrained process. Small decisions can quickly lose their relevance when the decision is not made in time. Big decisions do not, or to a lesser extent, suffer from time constraints. This study focuses on the making of small binary decisions, small decisions with two possible answers.

When modeling how the process of making small binary decisions happens in our brain researchers often use evidence accumulation models (Gold and Shadlen (2007)). Evidence accumulation models are a type of sequential sampling models. The idea behind these kinds of models is that evidence is collected over time, and that each response is represented by some decision boundary. In evidence accumulation models a decision gets made the moment the first decision boundary is reached (Forstmann et al. (2016)).

Currently the most widely used evidence accumulation model is the drift diffusion model (DDM). The DDM uses a single counter which uses collected evidence to move towards the decision thresholds. If evidence supporting one of the decisions is accumulated the counter will move towards that decision's decision boundary. This process keeps going until one of the boundaries is reached, and that decision is made. In our case we expect to see that when a stimulus is first presented no evidence has been collected yet, then as time progresses, more and more evidence is collected to support a certain response. Then, when a decision boundary is reached the response is given. DDMs assume the data that are being collected is noisy, this is a useful feature when working with brain data.

To test the predictions made by the drift diffusion model we need data on the time course of the decision process of the human brain. An increasingly popular method to gather data on these

processes in the human brain is electroencephalography (EEG). EEG is often used to record brain data, and has an excellent temporal resolution, but unfortunately suffers from a poor spatial resolution. Sometimes, when patients undergoing long term invasive monitoring are recruited for research, the opportunity to get intracranial EEG (iEEG) data for research is available. These patients have the recording probes placed directly on and in the brain instead of on the outside of the skull. This invasive placement of the probes has the advantage that these recordings also have a high spatial resolution. Combining this high spatial resolution, with the already high temporal resolution of normal EEG makes that these recordings are considered by many the golden standard for research (Jacobs and Kahana (2010)).

There are always a lot of processes running simultaneously in the brain, such as processes we are not aware of, keeping our body running, but also processes actively processing something we are focused on or busy with. Because there are always multiple things happening recorded brain data can contain a lot of noise, and recording equipment will not only record the processes of interest, but everything that is happening in the brain.

To get a better understanding of what is happening in the brain we can use classification techniques such as generalized logistic models (GLM). In this case we use generalized logistic models to predict the dependent variable, which is the response, on the basis of the predictors, the recorded brain data. GLM have already seen some successes when being used to extract decision signals from decision based brain recordings, and DDMs seem to provide a model that predicts the time course for the accumulation process (van Vugt et al. (2012)).

Other classifiers, such as support vector machines (SVM) have potential to outperform GLM. SVM attempt to map predictor values to vector space so that the different classes are divided as much as possible. New data then get placed into the vector space and predicted to be part of the category where they are placed. SVM are more resilient against noise than GLM, and this robustness make them useful for classifying noisy brain data (Biggio et al. (2011)).

In this research we looked at iEEG data recorded from participants performing binary decision tasks. We attempted to determine whether the time course of these tasks fit the proposed drift diffusion model. To do this we used two classifiers, namely the generalized logistic model, and the state vector machine. We also compared the performance of both classifiers to see if one is more suitable for classifying such data.

2 Methods

In this research we used two different data sets, or experiments. The first experiment consisted of a limited data set with a few participants performing a simple visual search task. The second data set used two more demanding tasks, a perception task and a memory task, and offered data from more participants.

The reason we opted to use a second data set were the disappointing results of the first experiment. This way, we hoped to find out whether the disappointing results came from the fact that the task was too subtle, the data set didn't offer the information required by the classifiers to fit the data, or that we were using the wrong approach.

For both experiments we used several methods to classify the data. We looked at the classifier accuracy over time to see if it fitted the predictions made by the drift diffusion model. The drift diffusion model predicts that the accuracy should increase over time because the brain has accumulated more evidence, and this should be decodable from the brain data.

2.1 Experiment 1

The data used for this experiment were provided by Dr. Jean-Philippe Lachaux from the Lyon Neuroscience Research Center, and were originally used in a study on the suppression of broadband gamma power in the default-mode network (Ossandón et al. (2011)).

2.1.1 Participants

The data were recorded from 14 patients with a mean age of 32 ± 10 of which 11 women, and 3 men, undergoing long term invasive monitoring for intractable epilepsy (Ossandón et al. (2011)). We were provided with data from three of these patients.

2.1.2 Task

The participants were asked to do a search task where each stimulus consisted of an array of 36 letters: 35 letters L and one letter T. These letters were randomly arranged in a six by six square. When the participant had found the location of the T he or she was asked to indicate by a button press whether the T was in the upper half, or in the lower half of the stimulus. There were two search conditions for this task, easy and difficult. For the easy condition the T was gray, and all the L's were black. For the difficult task all the letters were gray. The participants scored 100% accuracy on this task. with an average reaction time of 0.9 seconds for the easy trials, and 1.3 seconds for the difficult trials.

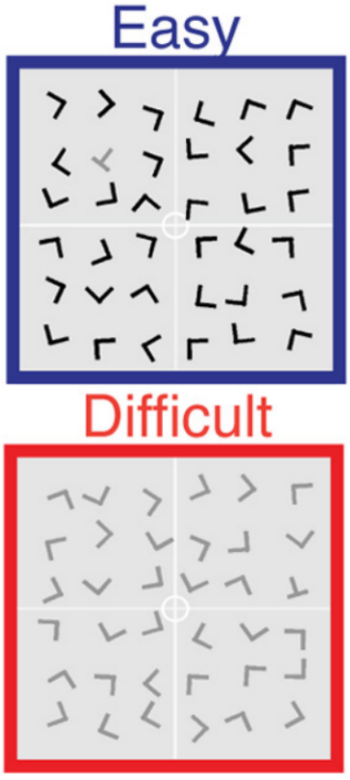

Figure 2.1: Example trials Experiment 1, source: Ossandón et al. (2011)

2.1.3 Recordings

The data were recorded using intracranial electroencephalography at 512 Hz, and were band pass filtered from 0.1 to 200 Hz. Before analyzing the data we performed a z-transform on the data, and divided the trials in 50 vincentized bins starting at the stimulus onset, and ending at the response. Thus the data consisted of slices, a slice consisted of the values for each channel for each trial, for a single time bin.

2.1.4 Analysis

For the analysis we applied two classifiers to the data, a regularized logistic regression model, and a support vector machine. For the regularized logistic regression model we looked at the data with and without feature selection. For the support vector machine we only worked with the feature selected data. Finally we frequency-transformed the data to the high gamma band (80–150 Hz) a frequency previously linked to attentional requirements (Ray et al. (2008)). All these steps are described in more detail below.

First we classified the data with a logistic regression classifier. A classifier was fitted for every slice, then we used 10-fold cross-validation for each classifier to get an accuracy score over time. The model however showed errors because it is only supposed to be used with a low amount of predictors.

Because of this we used feature selection on the data. The feature selection process consisted of splitting the data set into two parts, one for the up response, and one for the down response. Then we compared the channels between the up and down trials for each time bin with a t-test to get a measure of difference. We selected for each slice the top-10 channels based on the highest absolute t-value from the t-test comparisons. We used that selection of 10 channels per slice to fit the logistic regression classifier, and again, used 10-fold cross-validation for each classifier to get an accuracy score over time.

Second, we classified the feature selected data with a support vector machine using a linear kernel. Support vector machines are more resilient against noise, and are well suited to classify noisy brain data (Biggio et al. (2011)). We fitted a support vector machine for each slice, and then used 10-fold cross-validation for each slice to get an accuracy over time.

Finally we used high gamma preprocessing to run an analysis on the high gamma band. The data had to be processed differently for this technique. We

used the EEG Toolbox to split the raw data and perform a wavelet transform to obtain EEG time courses in the 80–150 Hz high gamma band. Finally the transformed data were z-transformed and vincentized as for the previous analysis.

2.2 Experiment 2

The second set of data was provided by Dr. Marieke van Vugt of the University of Groningen. These data were previously used in a research on the tracking of perceptial and memory decisions with intracranial electroencephalography (iEEG) (van Vugt et al. (2017)).

We did not attempt to fit logistic regression for the data of this experiment because this was already done in a previous study (van Vugt et al. (2017)). This study found that the decoded decision evidence showed a dynamic that follows the drift diffusion model, for both memory and perception tasks. The classifier accuracy's were shown to be above chance, but the authors indicated that their classification was far from stellar.

2.2.1 Participants

Participants were recruited from patients undergoing long-term invasive monitoring for pharmacologically intractable epilepsy at Freiburg University hospital (Germany). Sixteen individuals were recruited and participated in the behavioral experiments (van Vugt et al. (2017)).

2.2.2 Task

The participants were asked to perform two sets of tasks, a perceptual and a memory task. Both tasks used the same set of stimuli but varied in the decision process required. The stimuli consisted of synthetic face stimuli created with the Basel Face Model, which allowed for precise manipulation of the similarity of the stimulus. The precise control on the manipulation of the similarity was useful for controlling the difficulty of the task. In the perceptual task two faces, facing outward, were shown. The participants had to determine whether these faces belonged to the same person. In the memory task participants were first shown two faces during a 2000–2075ms jittered study period, followed by a 1000–1150ms jittered blank delay period, after which a probe face was shown. Participants were then asked to indicate whether this face was identical to one of the two faces presented before. The average reaction time for the memory task was around 1.8 seconds, and for the perception task around 2.5 seconds (van Vugt et al. (2017)).

2.2.3 Recordings

The data consisted of intracranial electroenchepalography recordings with a 2000 Hz sampling rate. The data were segmented into trials of 4000 ms, starting 200 ms prior to the onset of the probe stimulus. The data for one of the patients were discarded due to the large number of epileptic spikes. Similarly trials with a kurtosis larger than 15 were removed. A kurtosis larger than 15 is indicative of epileptic spikes. (van Vugt et al. (2017)).

2.2.4 Analysis

The data provided were wavelet transformed to get the time courses in the 4–9 Hz theta band, which previously has been shown to be important for decision making (van Vugt et al. (2012)). The data were also z-transformed and vincentized. The number of bins was chosen such that on average bin duration is approximately 50ms. Both classes in the trials (match and non-match) had an equal number of trials. We considered the data of both tasks to consist of slices. A slice consisted of the values for each channel for each trial for a single time bin.

For our first attempt at classification we fitted a support vector machine (SVM) with a linear kernel function for each slice. We used 10-fold cross-validation to retrieve a classifier error value for the SVM of each slice.

For our second attempt at classification we first used feature selection in an attempt to improve the accuracy of the classifier. The feature selection process consisted of splitting the data set into two parts, one for match, and one for non-match. Then we compared the channels between the match and non-match trials for each time bin with a t-test to get a measure of difference. Next we selected for each slice the top-10 channels based on the highest absolute t-values from the t-test comparisons. We used that selection of 10 channels per bin to fit another SVM, and again performed 10-fold cross-

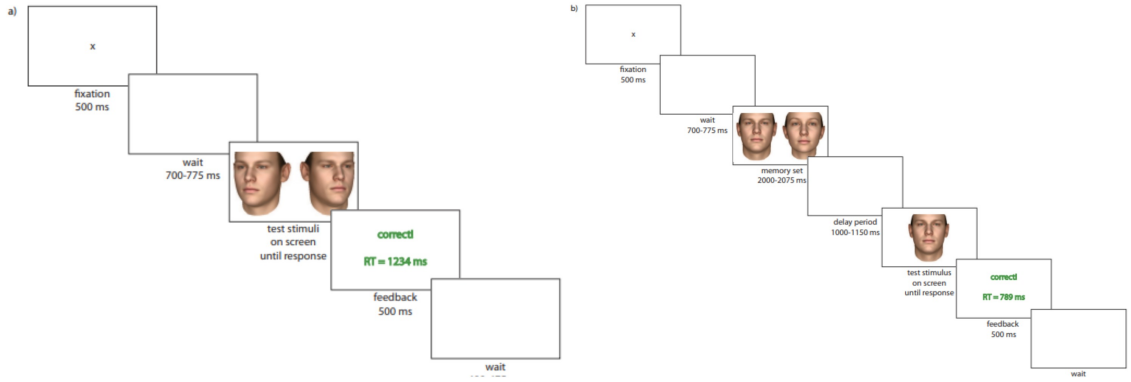

Figure 2.2: Example trials Experiment 2. Left: the perception task. Right: the memory task. source: van Vugt et al. (2017)

validation to retrieve the classifier error for each time bin.

3 Results

3.1 Experiment 1

We initially attempted to classify the iEEG recordings for all channels with a logistic regression model for each slice. As described in the methods the model showed errors because there were too many insignificant channels. This caused the model to be rank deficient, and overparameterized. The results for this analysis can therefore not be trusted and will not be discussed further.

Since classifying with all the data didn't yield the expected results we tried only using the generalized logistic regression model on the channels selected by the feature selection process for each slice. The results of this analysis looked much more promising, as shown in figure 3.1. The classifiers for all three participants managed to get average accuracy's close to 0.6 which is above chance, but failed to show the expected upward trend that would be the result of the evidence accumulation process. Figure 3.1 also shows some downward spikes in accuracy going as low as 0.35, these spikes show classification far below chance, and don't fit the evidence accumulation model or drift diffusion model.

The logistic regression classification of the data after high gamma band processing also failed to show the expected results. We can see in figure 3.2 that the classification accuracy was almost always below chance. The results also failed to show an increase in accuracy over time as predicted by the evidence accumulation model or drift diffusion model.

For a final analysis we classified the data selected by feature with a support vector machine using a linear kernel for every slice. We used 10-fold cross-validation to get a classifier accuracy over time, the results are shown in figure 3.3. The classifier scored an average accuracy of 0.59, and had a maximum of 0.66. These results might look promising but there is no upward trend in the data and moments before the response the model performed the worst. This kind of performance is not in line with the predictions of the drift diffusion model.

3.2 Experiment 2

We will first look at the results of our own classification. Classifying over all the channels and averaging the accuracy's over all test subjects resulted in an accuracy of 0.49, and 0.51 for the perception and memory task respectively. This is an acceptable result seeing that when the stimulus is presented e.g. when the second face is shown for the perception task, or when the two faces are shown for the memory task, there has no information been accumulated yet. Then more information is accumulated. The time course is reflected in the classifiers accuracy, as shown in figure 3.4. Once enough information has been accumulated a decision boundary is reached and a response is given. At this point the classifier shows an accuracy of 0.56, and 0.55 for the perception and memory task respectively. The upwards accuracy slope we observed was consistent

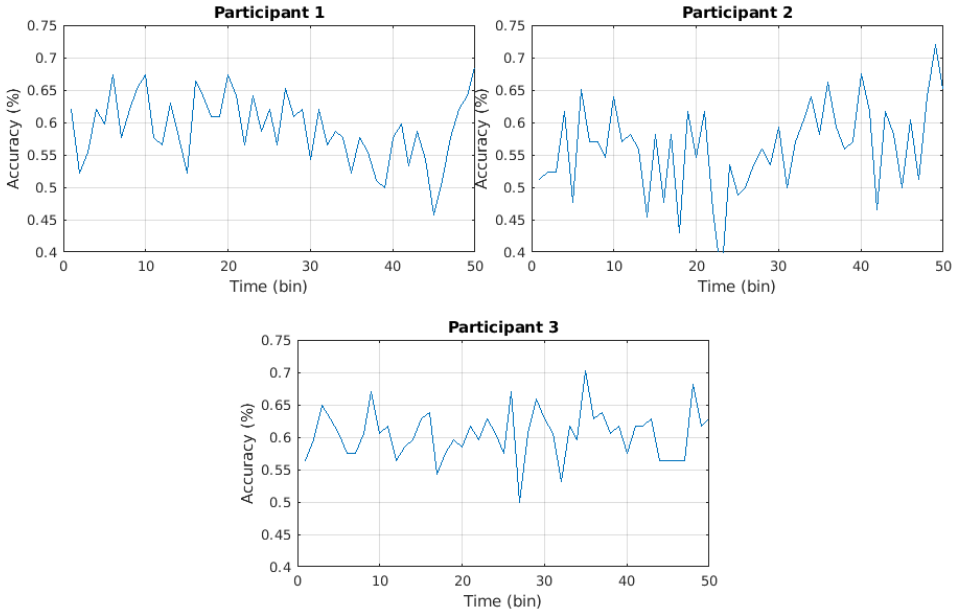

Figure 3.1: Experiment 1: Classification accuracy's for the three participants.

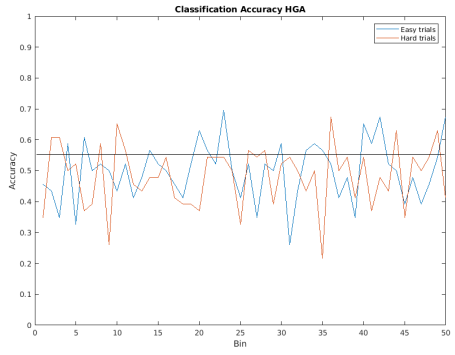

Figure 3.2: Experiment 1: High gamma band classification accuracy

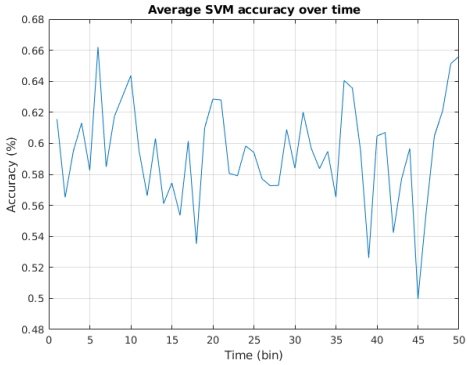

Figure 3.3: Experiment 1: Average SVM accuracy over time

with the predictions of the drift diffusion model.

When classifying after feature selection the accuracy was significantly higher for the memory task ($t(31) = 29.2$, $p = 4.1\text{e}-24$) as well as for the perception task ($t(49) = 40.9$, $p = 1.5\text{e}-39$), vs classifying without feature selection. At the time of the stimulus the average accuracy was 0.55 and 0.56 for the perception task and memory task respectively. Between stimulus and response the accuracy slopes upward and ends up at accuracy's of 0.62 and 0.60 for the perception task and memory task respectively. The time course is shown in figure 3.4.

When we look at our top classifier accuracy for each participant at each task, as shown in figure 3.5, we see that the classifier performance differs between participants. In the previous study by van Vugt et al. (2012) a maximum accuracy of 0.55 was reported for the classification for the patient where

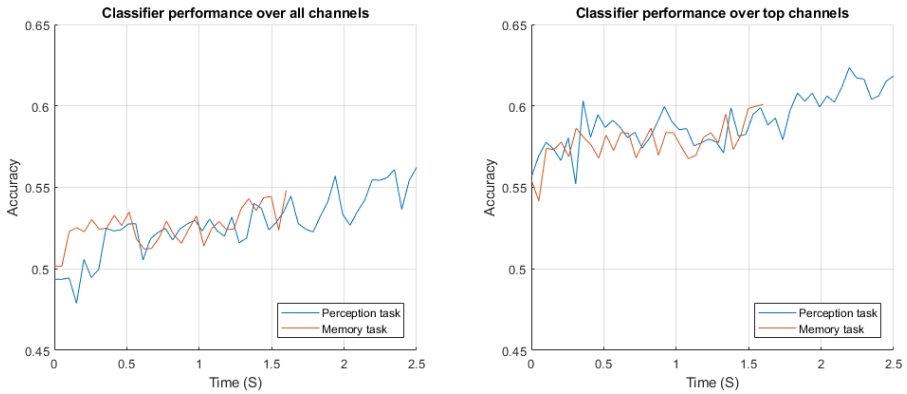

Figure 3.4: Experiment 2: Average SVM accuracy over time. Left: All channels. Right: After feature selection.

the classifier performed the worst, and a maximum accuracy of 0.76 for the patient where the classifier performed best. The maximum accuracy for the participant where our classifier performed worst is 0.59, and the maximum accuracy for the participant where our classifier performed best is 0.79, which shows a slight increase.

4 Discussion

We have explored the possibilities of classifying two data sets using two different classifiers, with and without feature selection. We studied the time course of small binary decisions taking drift diffusion models as our starting point. In line with drift diffusion models we expected to see the classifier accuracy to go up when coming closer to the response because the accumulation of evidence over time should make it easier for the classifiers to distinguish the responses from each other.

4.1 Experiment 1

For the first experiment we failed to produce this upward trend mentioned above. The classification accuracy seems to fluctuate erratically around a constant mean. This was true for both the logistic regression and the support vector machine classification. Even though the classification accuracy is rather consistently above chance we could not find a neural correlate for drift diffusion models in this experiment.

Simply stated, if we had found a neural correlate the data would have shown an upward trend with less deviation, something that we do not observe in the current results. Brain recordings contain a lot of information that is not relevant to the task. The classifier however is not aware of this, and simply looks for correlations which enable it to predict a response. With the amount of recorded channels offered to the classification algorithms in the non-feature selected analysis the chance that the classifier picks only the relevant channels is slim, since those channels are not guaranteed to have the highest correlation.

Inspecting the data on the electrode locations revealed that the probes were not evenly spread over the brain but focused on a certain side, or were placed around a specific part of the brain. This was of course because of the reason the electrodes were implanted in the first place: to discover the source for the patient's epilepsy. The electrodes were mainly placed around the area where the doctors expected to find this source, and were only sparsely placed on other areas of the brain. While this is of course understandable from a medical point of view, it can also be part of the reason why we didn't find neural correlates; we were simply looking at the wrong part of the brain.

If we were looking at the wrong part of the brain, then feature selection wouldn't have done us any good, even if it seemed to improve the classifier's accuracy. The way we did the feature selection, by selecting the channels that showed the largest difference during the decision process, we simply selected

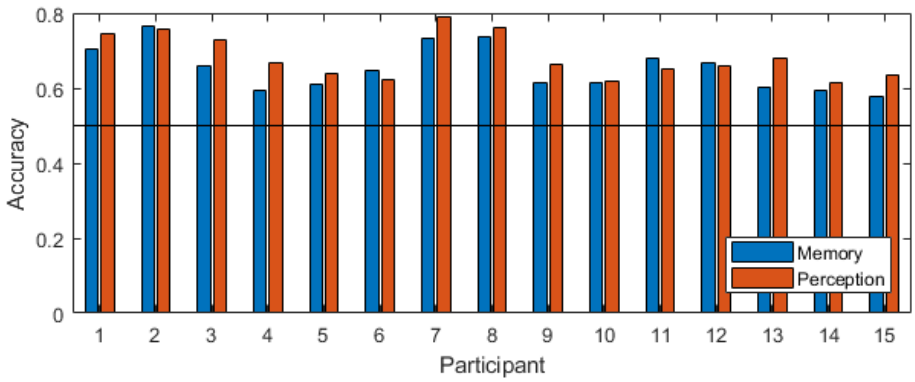

Figure 3.5: Experiment 2: Top accuracy per participant for both tasks.

the strongest correlates. However a strong correlation does not necessarily mean that the channel is related to the decision process at all. This way we could have seen an increase in classifier accuracy even though the results did not tell us anything about the actual process.

Another reason that this classification did not yield the expected results might have been that the task the participants were asked to perform is very subtle, especially for the easy trials. It involved a simple visual search and required no real deliberation. Feature selection might have caused us to select the wrong channels, because in the relevant channels the changes are more subtle that that of noisy channels. Because the task consist mostly of a visual search and requires very little deliberation it might also have been the case that the actual process was too subtle for the classifiers to detect.

4.2 Experiment 2

For the second experiment we looked at data from participants that performed two tasks, a memory task and a perception task. These tasks were more cognitively intensive than the task the participants of experiment 1 were asked to do. We observed results that showed a clear upward trend over time, as predicted by the drift diffusion model. The accuracy over time also showed less deviation, which could indicate that we were looking in the right place. Feature selection also showed a clear improvement in classifier accuracy at every moment of the process, especially for the perception task. Not only did feature selection increase classifier accuracy, it also retained the general shape of the broader analysis. We draw the cautious conclusion that we have found neural correlates of the binary decision process. However there are a few comments to be made.

The mean accuracy of the non-feature selected support vector machine classification starts around chance, or 0.5. This is to be expected because no matter how good the classifier, the brain has only just, or not even, started the decision process. Then when the evidence accumulation process starts gathering more information from the stimulus the classifier's average performance goes up. This is, as already described, the expected trend according to the drift diffusion model (Forstmann et al. (2016)). When we look at the feature selected analysis however, the initial average accuracy is around 0.55. While this initially looks like an improvement over the broader analysis we can ask ourselves where this increase comes from. This higher accuracy stems from a point in time for which we can make the assumption that the brain has not done any work towards evidence accumulation yet. Realizing this might entice us ask the following question: Where does this increase in accuracy come from, if not from the evidence accumulation process. More research is required to answer this question.

We also realize however that one of the problems that occurs in experiment 1 also occurs in this experiment: the probes were not spread evenly around the brain. Because the exact placement of the probes is different for each participant this could also be used to explain the difference in ac-

curacy scores between participants.

4.3 Conclusion

Intracranial electroencephalography is a useful tool to measure brain activity but even though it offers high temporal and spatial resolutions it also has some serious drawbacks that can potentially be of great influence on the results of research. The invasive nature of the monitoring technique, which directly relates to the uneven spread of probes requires researchers using such data to be really aware of what they are, and what they are not looking at and how this can influence their results.

We found some poor results in the first experiment, and some moderately successful results in the second experiment. Logistic regression and support vector machines are both good tools to analyze intracranial EEG data, where the support vector machine model performs a little better, especially when looking at larger amounts of data.

Even though this study has no strong results to present, we can still learn a great deal from it. We have seen some weaknesses of logistic regression and one of the trials, but we have also shown that it is possible to get results that fit the drift diffusion model's predictions. There is however a lot of work to be done.

In this research we primarily focused on the decision process, and we make the assumption that that task starts at the stimulus and ends at or after the response. However, other research has found for a memory task that there are different processing stages between stimulus and response, where the decision task only takes up a small part of the complete time line (Borst and Anderson (2014)). For further research we propose that we first extract the decision part from the entire process and run analysis on that, which might show a stronger correlation with the drift diffusion model's predicted time course.

I would also suggest that future research also look at support vector machine performance when using different kernel functions, because the linear kernel has a higher chance of over-fitting than other kernels. Finding more robust ways of doing feature selection, by for example, filtering out data from brain area's that should not influence the decision process, could also be useful to obtain better results. More research can also be done in looking at how the data change over time. This research used classifiers for each individual time bin with no reference to the other time bins or slices, future studies could attempt to find correlation between the expected time course of evidence accumulation and changes within single channels over time.

References

Biggio, B., Nelson, B., and Laskov, P. (2011). Support vector machines under adversarial label noise. *Journal of Machine Learning Research*, 20:97–112.

Borst, J. P. and Anderson, J. R. (2014). The discovery of processing stages: Analyzing eeg data with hidden semi-markov models. *NeuroImage*, 108:60–73.

Forstmann, B., Ratcliff, R., and Wagenmakers, E.-J. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual Review of Psychology*, 67:641–666.

Gold, J. I. and Shadlen, M. N. (2007). The neral basis of decision making. *Annual review of neuroscience*, 30:535–574.

Jacobs, J. and Kahana, M. J. (2010). Direct brain recordings fuel advances in cognitive electrophysiology. *Trends in Cognitive sciences*, 14:162–171.

Ossandón, T., Jerbi, K., Vidal, J. R., Bayle, D. J., Henaff, M.-A., Jung, J., Minotti, L., Bertrand, O., Kahane, P., and Lachaux, J.-P. (2011). Transient suppression of broadband gamma power in the default-mode network is correlated with task complexity and subject performance. *Journal of Neuroscience*, 41:14521–14530.

Ray, S., Niebut, E., Hsiao, S. S., Sinai, A., and Crone, N. E. (2008). High-frequency gamma activity (80-150 hz) is increased in human cortex during selective attention. *Clinical Neurophysiology*, 119(1):116–133.

van Vugt, M., Brandt, A., and Schulze-Bonhage, A. (2017). Tracking perceptual and memory decisions by decoding brain activity.

van Vugt, M. K., Simen, P., Nystrom, L. E., Holmes, P., and Cohen, J. D. (2012). Eeg oscillations reveal neural correlates of evidence accumulation. *Frontiers in Neuroscience*, 6(106).