Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Proceedings (IEEE Int Conf Bioinformatics Biomed). Author manuscript; available in PMC 2023 May 10.

Published in final edited form as:

Proceedings (IEEE Int Conf Bioinformatics Biomed). 2022 December ; 2022: 3101–3108. doi:10.1109/bibm55620.2022.9994989.

Determining and Validating Population Differences in Magnetic Resonance Angiography Using Sparse Representation

Steve Mendoza[1], **Fabien Scalzo**[2], **Aichi Chien**[1,*]

[1]Department of Radiological Science, David Geffen School of Medicine at UCLA

[2]Department of Computer Science, Seaver College of Natural Sciences, Pepperdine University, Malibu, United State of America

Abstract

Goal: Identifying population differences can serve as an insightful tool for diagnostic radiology. To do so, a reliable preprocessing framework and data representation are vital.

Methods: We build a machine learning model to visualize gender differences in the circle of Willis (CoW), an integral part of the brain's vasculature. We start with a dataset of 570 individuals and process them for analysis using 389 for the final analysis.

Results: We find statistical differences between male and female patients in one image plane and visualize where they are. We can see differences between the right and left-hand sides of the brain confirmed using Support Vector Machines (SVM).

Conclusion: This process can be applied to detect population variations in the vasculature automatically.

Significance: It can guide debugging and inferring complex machine learning algorithms such as SVM and deep learning models.

Keywords

Brain Vasculature; Magnetic Resonance Imaging; Cerebral Angiography; Circle of Willis; Gender Difference; Morphometry; Machine Learning; Medical Imaging

Index Terms—

Morphometry; Machine Learning; Medical Imaging

I. INTRODUCTION

Determining medical imaging patterns across populations could be useful for medical imaging data interpretation and diagnosis. A typical study would involve a statistical

*Corresponding author Aichi Chien, PhD, Department of Radiological Science, David Geffen School of Medicine at UCLA, 10833 LeConte Ave, Box 957350 Los Angeles, CA 90095. aichi@ucla.edu.

Conflicts of interest/Competing
All authors do not have conflicts of interest nor competing interest.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

analysis of a few hundred cases to find high anatomical variance areas. Finding and validating these high variance regions is known as morphometry. The conventional approach to solve this problem would be to consider two populations and see which voxels contribute the highest variance between different subjects, also known as voxel-based morphometry (VBM). In this methodology, one acquires images from several patients and registers them to each other using an anatomical atlas. VBM has been most often applied to MRI images to identify changes in neurologically healthy patients and those with various psychiatric disorders (Bookstein, 2001). However, it is almost impossible to distinguish population differences from differences in registration or other image processing details, such as dynamic range, without proper normalization (Thirion, 2006). A few papers (Bookstein, 2001) have mentioned some considerations for interpreting brain morphometry data. Among them are the registration and the normalization of the brain data. Several groups use different algorithms for normalization and registration that influence the results. (Thirion, 2006) Due to its simplicity, it is one of the most popular methods for determining differences among populations. There is also a well-established platform for determining statistical significance for any brain differences found. The first accepted method is the Gaussian linear method to determine statistical significance. (Friston, 1995) The individual voxels are individual variables, and the Gaussian linear field tries to find corresponding voxels that are different in two respective populations. However, if the data is not well registered, the statistical pattern would reflect misregistration rather than anatomical differences. Aside from these caveats, the VBM approach is used for many studies since it has a straightforward statistical interpretation of the results and has several clinical use cases.

A. Introduction to Multi-Variate Pattern Analysis (MVPA)

The second class of population-based analysis is multi-variate pattern analysis (MVPA) (Norman, 2006). MVPA is comparatively recent, of which the prime example is the SVM (Vapnick, 1994), which was prominent after 2006 (Norman, 2006). The main goal of MVPA is to treat the data in a multi-variate fashion where interactions and combinations of pixels or features provide the classification power. As opposed to VBM, in MVPA, the pixels are statistically dependent. In contrast, for univariate analysis, one voxel or pixel uniquely determines the statistical significance, with p values representing the statistical significance of voxels.

However, for MVPA, model interpretation is under debate, with most attempts utilizing classification performance and AUC curves as a proxy for statistical rigor. (Allefield, 2016) In particular, an SVM model is adequate if it shows performance better than that of random permutation of labels. This interpretation of classification performance has drawbacks (Allefield, 2016) since the classification error has no direct connection with population prevalence. In principle, the classification accuracy between two groups could be due to a few outliers, but there is no way to tell if that is the case with the standard statistical tests done with SVM. The key is to have a minimum threshold of the frequency of these patterns. For instance, using their method, we can infer the minimum population prevalence. Using the standard interpretation of the SVM cross-validation score, rare subjects in a population could contribute to the cross-validation results.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Alternatively, there have been other groups who, in response to these criticisms, have combined the standard cross-validation testing with confound controls. (Snoek, 2019) The confound control aims to quantify how common a certain confound is in a population and try to control for it, thereby signaling other features that could be useful for classification. In (Snoek 2019), they focus on gender classification using structural MRI and control for the larger brain size of males. While not implementing their specific confound regression technique, we consider their ideas when interpreting our results. In MRI imaging, some populations show more motion than others that could impact the studies of MVPA results, with statistically significant effects on brain segmentation results based on motion results. (Alexander-Bloch, 2016). We aim to control for registration confounds between males and females and find statistically significant populations while controlling for registration and subtle head motion differences.

A recent extension of VBM is pattern-based morphometry or PBM. The essence of the technique is to use sparsity applied to subtraction images between populations (Ganokar, 2011). The theoretical basis for using sparsity in brain imaging was that brain networks are sparse. (Ramenazi, 2015). Among the first applications using Dictionary learning is image compression (M. Ahalon, 2006). The dictionary method implementation uses a wavelet transform of the image data; this wavelet transforms are the dictionary that would allow a compressed representation of the image, allowing for faithful preservation while at the same time allowing for compression. The general mathematical intuition is that it is a generalization of the k-means clustering algorithm, which for a certain k, the idea is to split the data so that k different centers are found such that any one point in the dataset is no further than a value epsilon from the closest k. In that sense, we can think of the dictionary learning problem as having a set of basis vectors k. These k basis vectors can reconstruct any training data example with an error less than a value epsilon.

Pattern-Based Morphometry (PBM) tries to answer the same questions as VBM, with a more global data representation. The main difference in the PBM approach is in the sparsity constraint. This sparsity constraint encourages the algorithm to generate global samples present in several samples rather than highly local and noisy voxel-based representations. Specifically, we will search for a set of 6 subtraction images to represent the differences between two populations. Our definition of subtraction images is that we take one case minus all the population members of another population. The user then selects a parameter k, the number of images to compare each image in the other population. The key metric is to take the Euclidean distance between images, and the algorithm considers the closest k images for each image. These difference images then represent differences among different populations via image subtraction. The goal of PBM is to compress the representation of all these images into a sparse dictionary showing the most salient differences between the populations. The number of different images to consider is a hyperparameter, known as the number of atoms. The user specifies the number of atoms for the algorithm to find, converging to the resultant atoms that best represent dictionaries by minimizing the distinction between the set of difference images and the closest dictionary steps to each case. We formalize the above steps by constructing two distinct groups, S defined as

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$S = \{S_1, \ldots, S_n\}$ and group 2 with $\{Z_1, \ldots, Z_r\} \subset Z$, where r is the subject of neighbors in our case six. The definition of the difference vectors is:

$$D_{ij} = S_i - Z_i \; \forall i \in \{1, \ldots, n\} \; and \; \forall j \in \{1, \ldots, r\}. \tag{1}$$

From the equation above, the order matters; the definition of group 1 or group 2 will influence the patterns. In the following subsection, we then use Equation 1 to describe our specific dictionary learning approach.

B. Introduction/Review of Dictionary Learning

We review principal mathematical equations and visualize the results for a brief overview of dictionary learning theory and motivations. In the first paper describing PBM, the authors described differences in brains between healthy and Alzheimer's patients. We will discuss the mathematical details discussed in (Aharon 2006), (Ganokar 2011). The main set of equations in the paper is as follows.

$$\begin{aligned} &\underset{B,C}{\text{minimize}} && \|X - BC\|_F^2 \\ &\text{subject to} && \forall i, \; \|c_i\| 0 \; \leq \; T \end{aligned} \tag{2}$$

The above is a minimization problem: a basis matrix B, the dictionary atoms, and C, a sparse matrix used to rank the dictionary atoms. X is the original dataset.

$$\begin{aligned} \forall i \; \in \; \{1, 2, \; \ldots n\} &\underset{c_i}{\text{minimize}} && \|x_i - Bc_i\|_2^2 \\ &\text{subject to} && \forall i, \; \|c_i\| 0 \; \leq \; T \end{aligned} \tag{3}$$

Equation 3 is used to solve (2). The algorithm initializes basis vectors, then searches for the sparest representation.

$$E_k = X - \sum_{j \neq k}^{k} b_j c_j \tag{4}$$

The equation below updates the basis vectors B_k using the error metric as shown in the previous equation.

$$\underset{B_k}{\text{minimize}} \quad \|E_k - \; b_k c_k\| \tag{5}$$

A few groups have implemented variations of the same algorithm used in this paper in different contexts. In medical imaging, these groups have focused on fMRI and MRI imaging. To the best of our knowledge, there has not been a detailed sparsity-based study of MRA imaging. However, at least one group is working on gender-based differences in the vasculature, using SVM and graph-based kernel methods (Kwitt, 2013).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

While there has been considerable scholarship on dictionary learning and in the visualization of morphometry differences, to our knowledge, relatively few works have analyzed these patterns in the machine learning context. There is a body of literature since the original formulation of the PBM principle aims to analyze the multi-variate patterns. These multi-variate patterns are more challenging to interpret than the more traditional linear analysis and statistical methods. While researchers apply MVPA to MR images of the brain, we have not seen this analysis applied to the visualization of MRA vasculature. We apply some statistical methods to dictionary learning and attempt to find correlations between the patterns and the statistical significance. We also put these results into context with further data processing and confound analyses. Our goal is to find morphologically and anatomically relevant features. To do so, we aim to find statistical significance. Once we have statistical significance, we also aim to do confound analysis and eliminate data-taking confounds that may appear as statistically relevant features. Our first step in the paper is to verify that our preprocessing method yields good results based on a statistical atlas and registration. The second step is to find statistically significant regions through permutation testing and training an SVM classifier. The last step is to control using confound analysis, where we try to reproduce the dictionary results using some registration transform. We try to find regions that may show the anatomical differences between genders through all these steps.

II. Methods

For registration, it is necessary to have a proper template. Our template is a statistical atlas optimized for the non-linear nature of registration described in detail in (Moches 2019); small vessels need non-linear transformations, such as the SyN method described in (Avants 2009), necessitating an atlas with high resolution containing minute details.

A. Code and Data Sourcing

Much of the code stemmed from a GitHub repository known as pypbm or python pattern-based morphometry. The original code uses Python 2 and, in our project, was rewritten to make it compatible with Python 3. The primary purpose of the python code was to call ANTS registrations routines built on the machine running the registration algorithm itself. We plan to release the parts of the code used for this project on a private GitHub upon request after publication.

Our MATLAB routines use open-source code based on the K-SVD algorithm, as described on the GitHub page. We also have made our codes for the normalization and display of the dictionary learning results. We use MATLAB 2018b for visualization and plotting. The main parts of the code gathered from others were from the GitHub repository; the rest utilized built-in MATLAB functions and implementations.

There are some differences between the GitHub pipeline and our pipeline. While we use the same ANTS registration algorithm with the default parameters described in the project, we use an MRA atlas (Mouches, 2019) instead of the standard pipeline of using an MR image as an intermediary and then applying registration transformation to the MRA data. We also

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

use an algorithm to segment the data, using (Jerman 2016), of which there is a MATLAB implementation, and we use their default values for our segmentation.

All our MRA data used in this study comes from published resources and is free to use provided one uses the proper citations and attributions.

B. Dataset Description

Our registration template (Mouches, 2019) includes data from 3 different institutions and has a high resolution of 0.5mmx0.5mmx0.8mm. By comparison, earlier works that have done a similar analysis have conducted the study based on a 1mm x 1mm x 1.75mm space. The earlier technique transformed the MRA images into the MRI space of lower resolution for the MRI template. The multiple scanners control for multiple acquisition conditions, making our method applicable to many scanners. Table 1 shows the MRA acquisition parameters for the institutions used in generating the atlas. The multi-center analysis is essential since an algorithm well-tuned to the parameters of one institution may not generalize to a different institution's data. Therefore, it is better to have a robust data processing and normalization pipeline for better inference. The dataset used to create the registration template is the same dataset we used for our analysis. The parameters include echo time (TE), relaxation time (TR), and field strength (T).

The group used the standard protocol for registering MRA images to generate the statistical atlas. Their protocol consists of 4 main steps. The first step is to normalize the MRA images using a bright and sharp contrast from the dataset. The selected case normalizes the images to control the different institutions' data. Within this first step, they also do a form of histogram matching. This histogram matching algorithm splits the 3D image into 100 regions and normalizes the histogram. Histogram matching is a standard normalization algorithm used in many registration pipelines. The second and third steps consist of segmenting the cases. The first step segments the vessel in a centerline representation. The second uses a distance transform based on a centerline to obtain a radius representation. The last step is to register these processed images to each reference MRI image using rigid registration and non-rigid registration to an MRI atlas.

The registration template helps interpret our registration results. It is also essential since the atlas components can help us with the results, as shown in Figure 1. Our atlas shows the raw grayscale values and is a 32-bit type image. We do not use the entire brain for the analysis since the brain periphery has smaller and harder to analyze vessels. Analyzing and processing a more minor brain region speeds analysis as the registration process is computationally expensive. One entire brain takes 200 minutes (about three and a half hours) to register using one CPU, but our subsection takes 20 minutes.

We use the registration template to define the image space we analyze. Due to the computational cost of complete image analysis, we decided to focus on the Circle of Willis crop since the circle occupies a small area. The volume dimensions are 150x150x61 in voxel space. For the image representation, we show both coronal and axial pictures.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

C. Pipeline description

We must check that the image quality is adequate for dictionary generation. In the original paper that described the statistical atlas, the authors describe how some volumes could not be adequately segmented, hence were not used in the final atlas. So, in accordance, we use a quantitative quality control mechanism omitting inadequate quality images or significant outliers that obscure our analysis. To do so, we will use the concept of the dice coefficient, which shows a comparison between different segmentation techniques and uses the dice coefficient measured against a manually segmented case and an algorithmic segmentation of the same manually segmented case. We do not have access to ground truth for each patient, so our reference will be a case that is the average of the dataset, so it is a template. While an imperfect benchmark, it is an adequate screening tool for poorly registered or segmented cases.

After the segmentation check, we will set a threshold to select the valid cases. Literature using these datasets often describes the original dataset and mentions the number of data excluded but often gives qualitative reasons for exclusion. The literature includes the statistical atlas, in which the authors used 544 cases out of 700. Our analysis starts with 570 from the same statistical atlas; using our threshold of 0.65 dice coefficient, we end up with 400 for the final analysis; eleven of these had no gender leaving 389 cases for the final analysis. The statistical atlas paper mentions that segmentation was the number one reason for case exclusion, motivating our work's quantitative registration measure. For our case, the registration algorithm converged to a result registered to the template for all, but some data had faulty segmentation. To generate the dice coefficient, we take an average image of the normalized image, segment it, then compare the resulting segmented image with the individual cases. The dice coefficient measures the deviation from the average image, which is considered the accurate segmentation and provides for a quantitative discriminator.

We do the segmentation in 3 dimensions using an algorithm to detect various tubular structures such as aneurysms (Jerman, 2016). The segmentation results are then cast into two dimensions to compare cases. The segmentation verifies image quality with better image quality corresponding to better data. However, all the dictionary learning comparisons use grayscale images since the corresponding atoms are smoother than those generated using binarized images.

For the dictionary atoms, we used the grayscale images but normalized using the algorithm shown below:

Since the images came from different centers with varying signal strengths, we needed to account for the biases for each center. A specific group, the IHOP, a psychiatry center, showed images brighter than average. Using the grayscale images gives smoother atoms than using binary images.

Figure 2 shows our workflow. The flowchart has two parts. The upper left concerns processing and details for each case. The lower right describes the processing steps for the group. We check each case individually, and if the registration completes, we then calculate the Dice coefficient. There are some cases where the registration fails due to noisy imaging

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

or a low contrast image, so in these cases, we apply an image contrast manually or apply the non-local means filter. The lower right concerns the steps taken for the SVM-based classification. These are after the images meet the criteria outlined in the upper left.

D. SVM and Statistical Testing

To calculate the statistical significance of dictionary learning, we trained 1000 permutations of labels of dictionary learning. To determine the importance of the dictionary pattern, we take one iteration of the dictionary learning and compare it with the permutations of the dictionary learning. The p-value for each pixel is as follows:

$$\hat{p} = (\#\{pixel_i > pixel^*\} + 1)/(P + 1) \tag{6}$$

Using this calculation per pixel, we then get a map of the p-value for each pixel. We then threshold this to only show p values of 0.05 or lower as the results.

E. SVM Background

In this work, we use the C-SVM implementation as implemented in MATLAB 2018, called fitcsvm. We use the segmentation setup described beforehand to analyze the SVM as binary images. This default mode creates a ten-fold cross-validation setup and returns the average validation error. Nine folds are used for training in this implementation, while the remaining fold serves as a validation dataset. As described in the Results section, we repeat this process 1000 times and compute each experimental setup's mean and standard deviation. We pick the value of 1000 since another study using SVM for gender classification used 1000 to compare various SVM kernels for gender classification. For use in the SVM algorithm, we flatten the images into vectors. The fitcsvm algorithm solves the standard SVM binary classification linear program as described below:

$$f(x) = x'\beta + b \tag{7}$$

In the equation above, x is the data, flattened to a 1-dimensional vector, β is a vector of coefficients that define the separating hyperplane, and b is the bias term.

The goal is to minimize the linear program:

$$0.5||\beta||^2 + C\sum \xi i \tag{8}$$

Above, C is the cost function, which can be user-defined. However, we did not include a customized cost function, and ξ is the slack variable to accommodate data points further from the separating hyperplane. Fewer slack variables usually result in more robust models.

1) SVM Significance Analysis—To determine if our results using SVM have statistical significance, we will compare the output of the permutation test with the default labels. As described in (Kwitt 2013), the permutation test estimates the p-value for multi-variate analysis. The equation described in their paper for the empirically derived p-value is:

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\hat{p} = (\#\{e_i < e^*\} + 1)/(P + 1) \quad (9)$$

The expression e^* is the error rate with the proper labels. The e_i is the error rate of each one of the 1000 permutations. The number of permutations that are less than the error rate of the classification error of the SVM model with the correct labels determines the statistical significance. The intuition is that when training an SVM model, these cross-validation results could be a result of chance, so we shuffle the labels of 0 and 1 and see if the shuffled labels are different from the proper labels. Many studies using SVM use the notion of cross-validation as statistical evidence of valid classification, so it is a black box. If there is overlap between the two classes, then the results of the SVM learning are not statistically significant and would not lead to null hypothesis rejection. For a significance value of 0.05 or lower, we would need to have at least 1000 trials. This approach is distinct from the univariate statistical method from voxel-based morphometry. The purpose of the SVM is to see if there is a classification difference between the left and right-hand sides. Figure 3 shows the input images for the left- and right-hand crops. To do so, we will use theanova1functionasimplemented 2018.

III. Results

A. Introduction and Data Validation

To test the validity of the dictionary atoms, we will use an SVM classifier and first verify a statistically significant classification. We have not tuned the SVM classification; the default parameters yielded statistically substantial results independently. The first test compares regular data and random permutations of the same dataset. We will use binary images in the SVM classifier; since this machine learning algorithm performs best with normalized or binary data where all values in a vector are between 0 and 1. We use a linear SVM using a standard kernel with default parameters as a MATLAB library, called fitcsvm. (MATLAB 2018) We discuss the coronal and axial results separately then consider their differences. The key is to gain intuitions behind the saliency maps and identify differences between regions. Before showing the SVM results, we will deliver the output of the dictionary learning to show how we could select crops to guide our analysis further. While the first part is qualitative, it is a valuable tool for further interpretation and analysis. This tool intends to show where differences exist among populations and suggest further research. Our section aims to show how dictionary learning and SVM coincide, offering the utility of our approach and suggesting how to use these results for further analysis.

We first start with the Dice Coefficient means and standard deviation for males and females. Females show a more significant dice coefficient than males, implying more variation in the segmentation in males than in females, and fewer overall male cases were used compared to female patients; more female subjects met our criteria, the consequences of which could also come about later. This more significant variation in male brains could impact our morphometry results. The average Dice coefficient of the cases we chose for our final analysis is 0.70 with a standard deviation of 0.03.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

B. Axial View

1) Visual Representation of Dictionary Patterns—As we describe in Figure 4, the values of dictionary learning can guide us to interpret the images. We hypothesize that these regions should show the most variation further from the mean values or values with low variability. We test this hypothesis through permutation testing of both SVM and dictionary learning and using the SVM classification results. To more clearly show where the variation lies and standardize the dictionary learning analysis, we focus on the average dictionary atom; we take the six images than average them. We show the results of this standardization in Figure 5.

2) SVM and Classification Results—We start our SVM analysis with the axial viewpoint. The first results concern random permutations, where we shuffle the labels and do validation tests 1000 times. The results show a significant difference between them, with an error rate of 48 percent with the permutations, random chance, and a 38 percent error rate using the original labels. The split between male and female subjects is 42 percent and 58 percent. So, we can see that the random permutations give results consistent with random chance. The error rate is 10 percent lower with the proper labels, comparable to earlier SVM work classifying males and females (Kwitt, 2013). Figure 5, and the bottom two rows of Table 2, show the left and the right-hand sides of the axial viewpoint. As discussed earlier in this section, we can visually observe differences between the left and right MCA. We want to quantify these differences by considering only one side or the other by considering classification strengths. Thus, we define a binary mask, blanking part of the hemisphere. Right and left classifications mean that only the right and left-hand sides are intact; the other side is blank. Using this logic, as seen in Figure 5, we see some differences in classification in the right and left-hand sides. The error rate is lower using the right-hand side than only the left-hand side. The error rate difference means that the right-hand side has more discriminatory power than the left-hand side.

This finding makes intuitive sense when considering the dictionary patterns shown previously. As seen in Figure 5, the right-hand side is brighter than the left-hand side. According to the dictionary learning scheme, the right-hand side shows more salient regions. The SVM results also support that notion, lending credibility to the dictionary approach for finding salient areas. We quantify the statistical significance in Table 5, which shows the ANOVA results with 10, 100, and 1000 cases. The results are not significant until we consider 100 groups, which coincides with SVM results in another paper that performed 1000 SVM permutations using cross-validation. These results do not disprove the null hypothesis that there is no statistical significance between them; the ANOVA shows how many samples we would need to show differences between the means of these two cases. Also, ten permutations would not be enough to determine statistical significance since more data is necessary for a permutation sense to determine if outcomes can arise by chance. The results with 10, 100, and 1000 cases show how we would need to simulate 1000 permutations for statistically significant results. Table 4 shows the calculated p-values for the instances of the entire image, the left, and right crops. For all three cases, we follow the definition of the p-value as described in the Methods section. There is a statistically different value of the proper labels and the random permutations, but the left-hand side shows

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

less statistical significance. The results give more quantitative validation to the dictionary method of finding statistically significant saliency regions.

C. Coronal View

To complement our analysis of the axial viewpoint, we now consider the coronal view, where the ICA is lower in the figure, as shown in Figure 7. We see that the left and right are symmetric, so their classification accuracies should be closer to each other than when considering the axial viewpoint. We show the analytical details in the following section, which corroborates the analysis shown in Figure 6. Another interesting observation is that we see differences in the ICA angle, as shown at the bottom of the images in Figure 6.

1) SVM and Classification Results—Compared to the axial view, the coronal view shows less accuracy in classification. However, the estimated classification accuracy is above chance. The random labels permutation test accuracy is like the axial view; the permutation test is a valid control demonstrating that the classification results are different concerning the different viewpoints. In particular, the left-side only standard deviation is higher than the right side, meaning that the statistical difference is low for both hemispheres. The left and right-hand sides for the SVM classification do not show a significant difference; the left and right-hand sides do not show striking visual contrasts. Table 5 shows the statistical analysis of the SVM results. The F score is less for the cases of 100 and 1000 cases, as compared to the axial view. Lastly, we consider the estimated p-value of the SVM classification. Table 7 shows the calculated p-values. We see there is no statistically significant difference between random permutations on the coronal viewpoint when considering the left- and right-hand sides individually.

D. Dictionary Validation

The next step in our analysis is to see the variations in the dictionary as we rerun the same dictionary and try to reproduce the same patterns. We chose 50 iterations for training the dictionary since the RMSE showed a slight change after 20, so 50 iterations are sufficient. In practice, the dictionary learning algorithm converges when the RMSE shows no difference between subsequent iterations, and the atoms are no longer updated. Although we have run the same algorithm and obtained the same RMSE, the results vary since dictionary training is stochastic. However, when rerunning the algorithm, we find the same patterns. We also create a statistical map to determine the statistical significance of our dictionary. The inputs will be the Female-Male and Male-Female averaged dictionary atoms discussed earlier. We use the average dictionary atom as defined in Figure 6. When implementing similar MVPA analysis, the literature observes some variations with the runs, requiring preprocessing with a gaussian kernel to smooth the results (Allefield, 2016). We use the notion of permutation testing as described when we implemented the SVM analysis. Recall that we find statistically significant classification results when considering the axial viewpoint but not with the coronal view. In our case, for permutation testing, we train the dictionary with 1000 random permutations of the training labels and compare the two vectors element by element. While this is a simple concept, it is not mathematically rigorous as there are variations in training the dictionary, and these can only really be estimates of statistical significance. Permutation testing is the only general known way

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

for determining the dictionary learning significance (Gaonkar, 2013). In addition, this test avoids the problem of using classification accuracy as a means for determining statistical significance. In dictionary permutation testing, we estimate the feature prevalence directly using the permutation concept, but the dictionary coefficient weights do not suffer the limitations of cross-validated accuracy.

Figure 7 shows the results of the dictionary permutation testing. The Female-Male comparison, in red, shows more significance than the Male-Female comparison shown in blue. Figure 7 also supports the SVM results between the left and right-hand sides. The right-hand side had more discriminatory features than the left-hand side. So, two pieces of permutation testing show the asymmetries. However, an important thing to note is that we have not quantified the effect of dictionary learning. The higher values indicate the increased presence of a vessel in one location versus another, but it does not show how often this effect is. The only thing we have done is to show that globally, there are differences within populations and that the information content globally is different within different regions of the image. An alternative is to implement population-based inference that would indicate quantitatively how prevalent an effect is in the population (Allefield, 2016).

We also compare the cases by transforming the average dictionary atoms into z scores. We take the average measure of the dictionary atoms and subtract the mean. The absolute value will give the variance of the method.

IV. Discussion

Compared to other MVPA methods mentioned and described in the paper, our dictionary learning method allows for direct visualization of anatomy provided one implements a proper normalization scheme. Compared to SVM visualization techniques, this technique also allows for a rough quantitative visualization of how significant differences are between groups. While univariate methods only show a small number of statistically significant voxels, our technique can identify substantial portions of anatomy where there are differences between the two groups. The visualization and the permutation testing align as the higher contrast regions as shown using the visualization map correspond to the most statistically significant parts when performing dictionary learning permutation testing. It complements techniques such as SVM, and we show that also SVM performance correlates with the contrast map we have generated.

However, there are some considerations that we must highlight. The most important is that there is variation in the dictionary learning maps, each run of the algorithm yields different results, but qualitatively they are all similar. Although there is no exact one-to-one correspondence between the dictionary learning and the SVM classification map, there is enough evidence to suggest that dictionary learning can highlight the salient parts of the imaging. The dictionary learning can distinguish between the foreground and background elements. The only regions showing a high z score are regions where there was population variation. However, more work is needed to better correspond between dictionary learning and the SVM. It is a simple procedure to train dictionary learning, and it would give some clues as to where the salient differences are. Some studies suggest that deep learning and

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

other machine learning algorithms may rely on subtle training biases based on the training set and acquisition parameters (Albadawy, 2018). For instance, a brain segmentation deep learning model performed worse when used in an institution different from the training institution. Having a way to visualize decisions using a simple algorithm can help debug and interpret complex models.

We believe the cause of the patterns comes from subtle motion differences between genders. Studies have investigated the motion differences between the genders and have found that, on average, males show more motion. The motion differences study also corroborates our findings regarding the dice coefficient; male cases show less successful segmentation results than females and offer more movement. (Alexander-Bloch, 2016), Our morphometry pipeline relies on registration and segmentation, so we study literature related to morphometry in MRI imaging. The literature also does address the problem of segmentation and motion and implications for morphometry. Our results suggest subtle registration confounds are statistically significant and can form a basis for classification even though they may not reflect anatomical differences between the two groups. We plan to investigate motion correction and confound regression as previously outlined in this paper for further analysis. We have found some evidence of subtle registration confounds. We hope this technique can serve as a mathematical way to check for various preprocessing confounds and hopefully lead to robust machine learning algorithms. The use case for this algorithm is as a first step before training an SVM or deep learning model, and to see confounds and hopefully regress them out or counterbalance them before resuming training.

V. Conclusion

In this work, we have taken a dictionary learning algorithm and used it to facilitate machine learning interpretation, trying to determine classification statistical significance using as few features as possible. While we do show that we have determined statistical significance using a few regions, we would still need to do a rigorous analysis of subtle motion artifacts to conclusively report on population differences due to anatomy. Nevertheless, it is important to observe biases in preprocessing and these biases in our opinion can greatly influence machine learning performance, as alluded to in our results. As deep learning methods are powerful but can be opaque, it would be interesting to compare features that deep learning uses and those found using dictionary learning. Also, this method can also serve as a pre-processing check and reveal subtle data features that need additional testing, so that other machine learning algorithms such as deep learning can deal with cleaner data.

Funding

This research was supported in part by NIH NHLBI R01HL152270

This research was supported in part by NIH NHLBI R01HL152270

VI. References

1). AlBadawy Ehab A. et al. “Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing.” Medical Physics 45 3(2018): 1150–1158.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

2). Alexander-Bloch Aaron et al. “Subtle in-scanner motion biases automated measurement of brain anatomy from in vivo MRI.” Human brain mapping 37 7(2016): 2385–2397.

3). Allefeld C, Haynes JD, 2014. Searchlight-based multi-voxel pattern analysis of fMRI by cross-validated MANOVA. Neuroimage 89, 345–357. [PubMed: 24296330]

4). Allefeld Carsten et al. “Valid population inference for information-based imaging: From the second-level t-test to prevalence inference.” Neuroimage 141 (2016): 378–392.

5). Aylward Stephen R. et al. “Spatial Graphs for Intra-cranial Vascular Network Characterization, Generation, and Discrimination.” Medical image computing and computer-assisted intervention: MICCAI … International Conference on Medical Image Computing and Computer-Assisted Intervention (2005): 59–66. [PubMed: 16685829]

6). Avants BB, et al. "Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of the elderly and neurodegenerative brain.” Medical Image Analysis 12 1(2008): 26–41.

7). Bullitt E et al. “Vessel tortuosity and brain tumor malignancy: a blinded study.” Academic Radiology 1232–40 2005 [PubMed: 16179200]

8). Bookstein Fred L.. “Voxel-Based Morphometry Should Not Be Used with Imperfectly Registered Images.” NeuroImage 14 6(2001): 1454–1462.

9). Friston K, Holmes A, Worsley K, Poline JP, Frith C, Frackowiak R, 1995. Statistical parametric maps in functional imaging: a general linear approach. Hum. Brain Mapp 2, 189–210.

10). Gaonkar B, Davatzikos C “Analytic estimation of statistical significance maps for support vector machine-based multi-variate image analysis and classification.” NeuroImage 78 (2013): 270–283.

11). Gaonkar B, Pohl K, & Davatzikos C (2011). Pattern-based morphometry. Medical image computing and computer-assisted intervention: MICCAI … International Conference on Medical Image Computing and Computer-Assisted Intervention, 14(Pt 2), 459–466. [PubMed: 21995061]

12). Hebart Martin N, et al. “Deconstructing multi-variate decoding for the study of brain function.” Neuroimage 180 (2018): 4–18.

13). Haxby J, Gobbini M, Furey M, Ishai A, Schouten J, Pietrini P, 2001. Distributed and overlapping representations of faces and objects in the ventral temporal cortex. Science 293, 2425–2430 [PubMed: 11577229]

14). Ashburner J, et al. “Voxel-Based Morphometry.” Encyclopedia of Neuroscience (2009): 471–477.

15). Jerman T, Pernu? F, Likar B, & ??iclin, ?. (2016). Enhancement of Vascular Structures in 3D and 2D Angiographic Images. IEEE Transactions on Medical Imaging, 35(9), 2107–2118. [PubMed: 27076353]

16). Kwitt RK… "Python Pattern Based Morphometry" https://github.com/KitwareMedical/TubeTK-pypbm. (2013).

17). Kwitt Roland et al. “Studying cerebral vasculature using structure proximity and graph kernels.” Medical image computing and computer-assisted intervention: MICCAI … International Conference on Medical Image Computing and Computer-Assisted Intervention 16. Pt 2(2013): 534–541.

18). Aharon M, et al. “K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation.” IEEE Transactions on Signal Processing 54 11(2006): 4311–4322.

19). MATLAB, version 9.5.0.1067069 (R2018b) Update 4.. The MathWorks Inc, 2018.

20). Mouches Pauline et al. “A statistical atlas of cerebral arteries generated using multi-center MRA datasets from healthy subjects.” Scientific Data 6 1(2019): 29.

21). Norman K, Polyn S, Detre G, Haxby J, 2006. Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. Trends Cogn. Sci 10, 424–430. [PubMed: 16899397]

22). Ramezani Mahdi et al. “Joint sparse representation of brain activity patterns in multi-task fMRI data.” IEEE Transactions on Medical Imaging 34 1(2014): 2–12.

23). Snoek, Lukas et al. “How to control for confounds in decoding analyses of neuroimaging data.” Neuroimage 184 (2019): 741–760.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

24). Thirion B, Flandin G, Pinel P, Roche A, Ciuciu P, Poline JB, 2006. Dealing with the shortcomings of spatial normalization: Multi-subject parcellation of fMRI datasets. Hum. Brain Mapp 27, 678–693. [PubMed: 16281292]

25). Whitwell Jennifer L. “Voxel-Based Morphometry: An Automated Technique for Assessing Structural Changes in the Brain.” Journal of Neuroscience 29 31(2009): 9661–9664.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

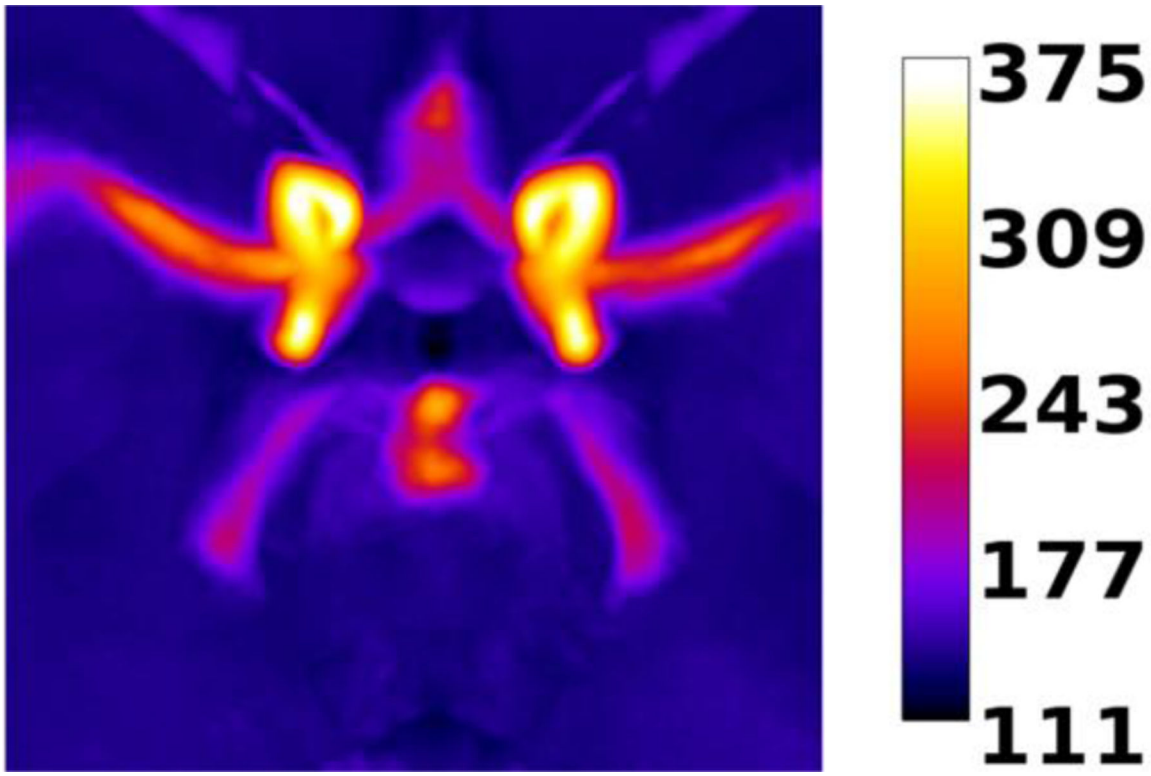

Figure 1:
Statistical Atlas used in the registration. Shown is the original 32-bit image used in registration. The grayscale scale is the raw scale unaltered from the publication. The area represents the region of interest we are interested in analyzing. This section of the brain contains the main arteries, highlighting the critical clinical parts of the brain. Of note is the varying arterial intensities. In our intensity-based registration algorithm, the brighter image regions carry more weight. The bright areas correspond to the Interior Carotid Arteries (ICA). The lower intensities correspond to the Posterior circulating arteries and Anterior circulating arteries.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

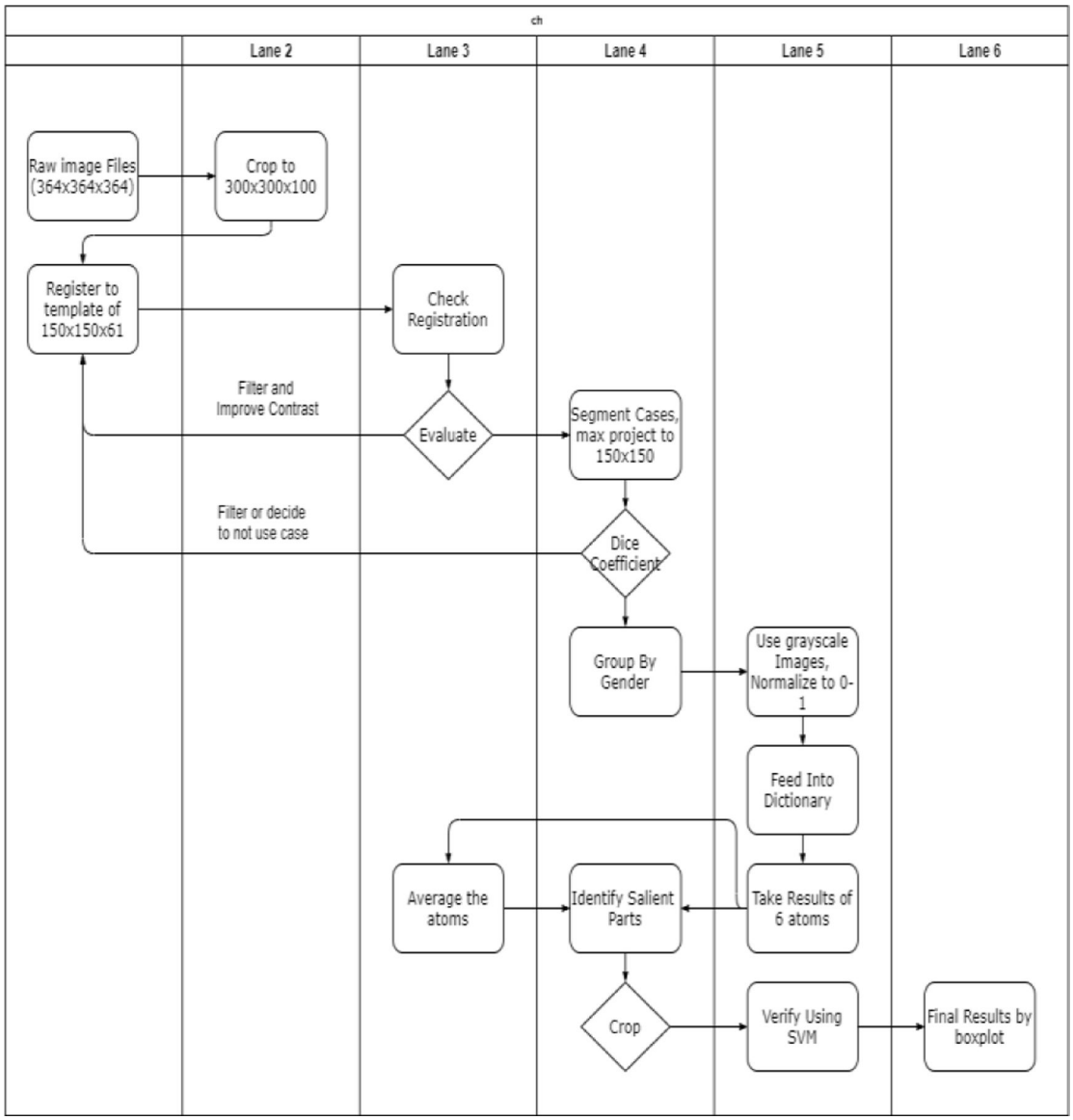

Figure 2:

Flowchart for the image processing pipeline for dictionary learning and SVM. The upper left side corresponds to steps taken to process individual images. For instance, we run the volumes through the registration again if it fails after applying a denoising operation. Afterward, once all the volumes were registered, the next step was segmentation. The segmentation process filters cases that may have some faulty segmentation. We cast the segmented 3-dimensional case into 2-dimensions for the dice coefficient analysis. After selecting, we construct dictionary atoms for a specific axis, either axial or coronal; then, we find the salient regions. We analyze the left and right hemispheres using SVM analysis based on those salient regions.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

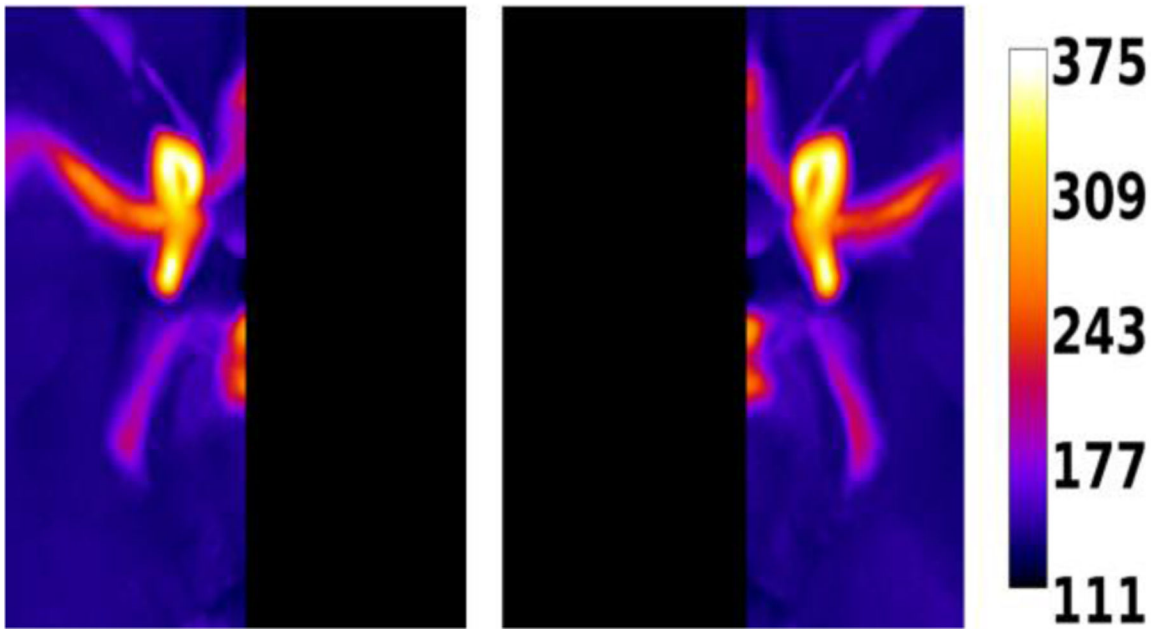

Figure 3:

Visual description of the crop to determine statistically significant differences between the left- and right-hand sides. In the results section, the left- and right-hand sides will correspond to the left- and right-hand crops, as shown in this figure. The image crop uses the same statistical atlas as Figure 2. We take the output of the registered images, segment them, and then crop them as shown in the figure for the final analysis.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

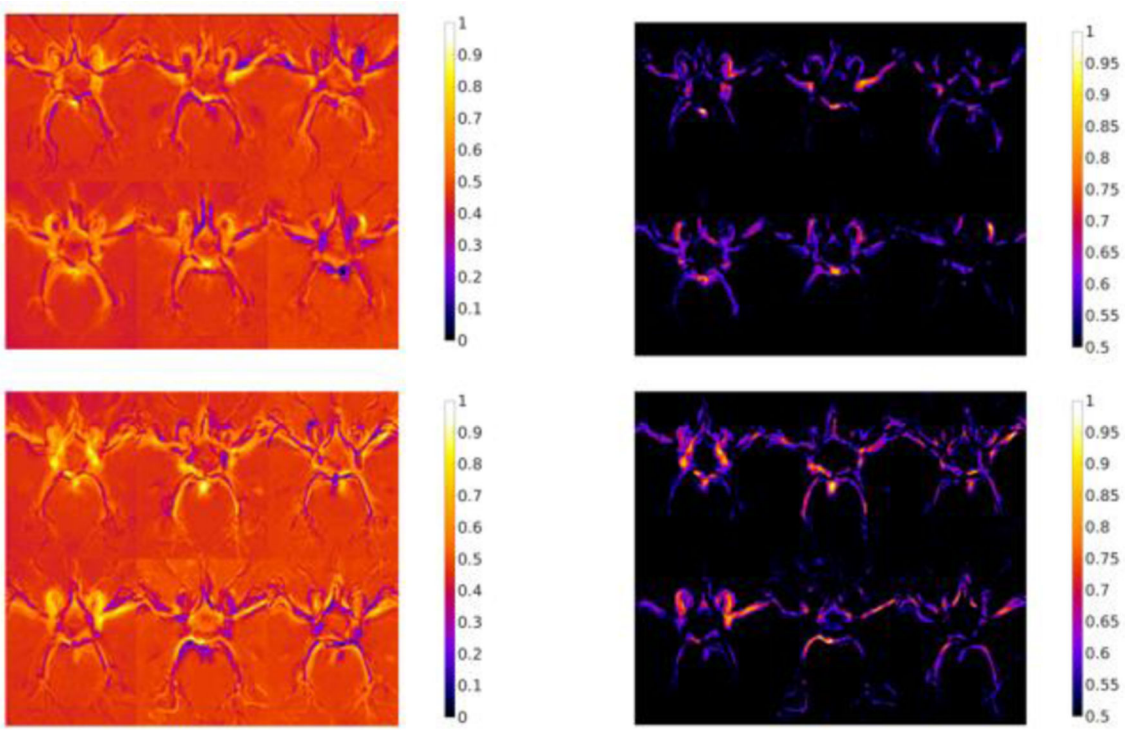

Figure 4:

The six dictionary patterns for the axial viewpoint. We have the dictionary patterns on the top when running the algorithm on (Male-Female) difference images. On the bottom, we have the reverse (Female-Male) difference images. On the left-hand side, we have the raw outputs of the dictionary learning. When performing dictionary learning, the prominence or significance of the feature scales from the distance of the mean. The probability scales with the relative intensity display. We threshold the image to observe the most prominent features.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

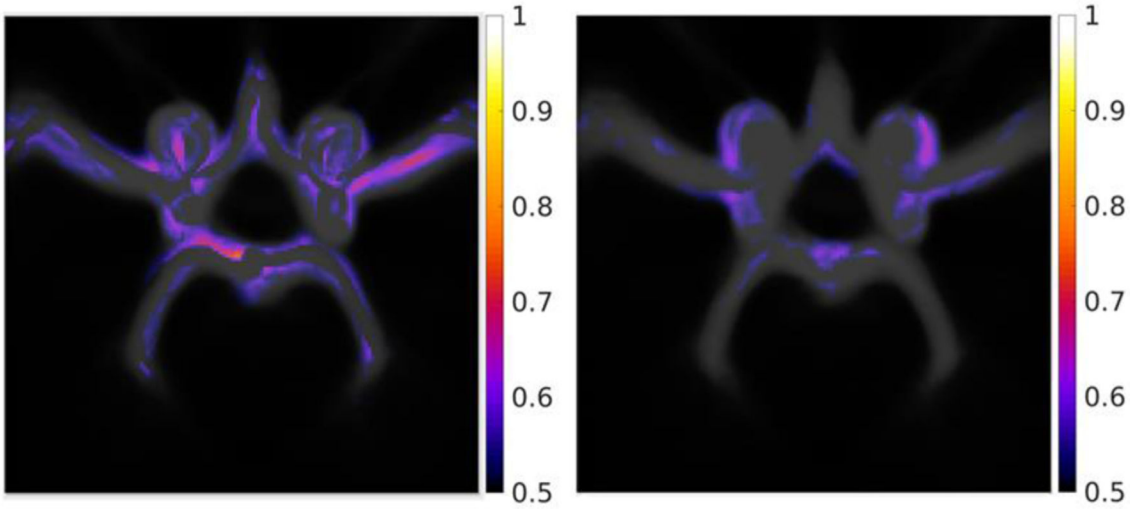

Figure 5:

Average Dictionary Patterns. We have the average dictionary pattern of (Female-Male) on the left-hand side, and on the right, we have (Male-Female). We see that there is more variation on the right-hand side of the image than on the left. We then use this figure to justify our experimental setup and test our hypothesis. The gray background represents the average image of the case calculated, taking the average of both populations' instances.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

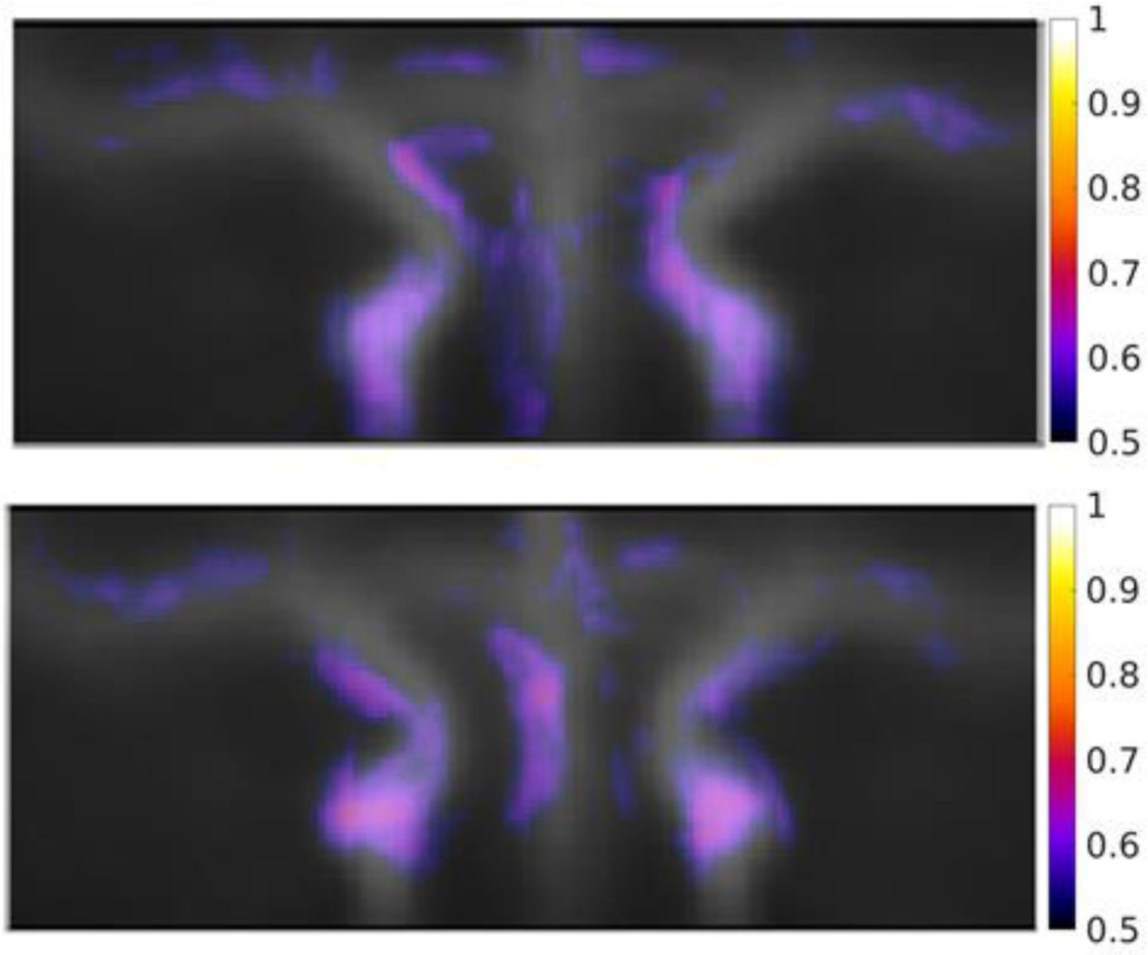

Figure 6:

The coronal viewpoint follows the same logic used to generate Figure 6. The top of the image shows the (Female-Male) cases and the bottom shows (Male-Female) difference images. In contrast to figure 6, we do not see a clear difference between left and right images, and we perform the analysis between the left and right images as a comparison.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

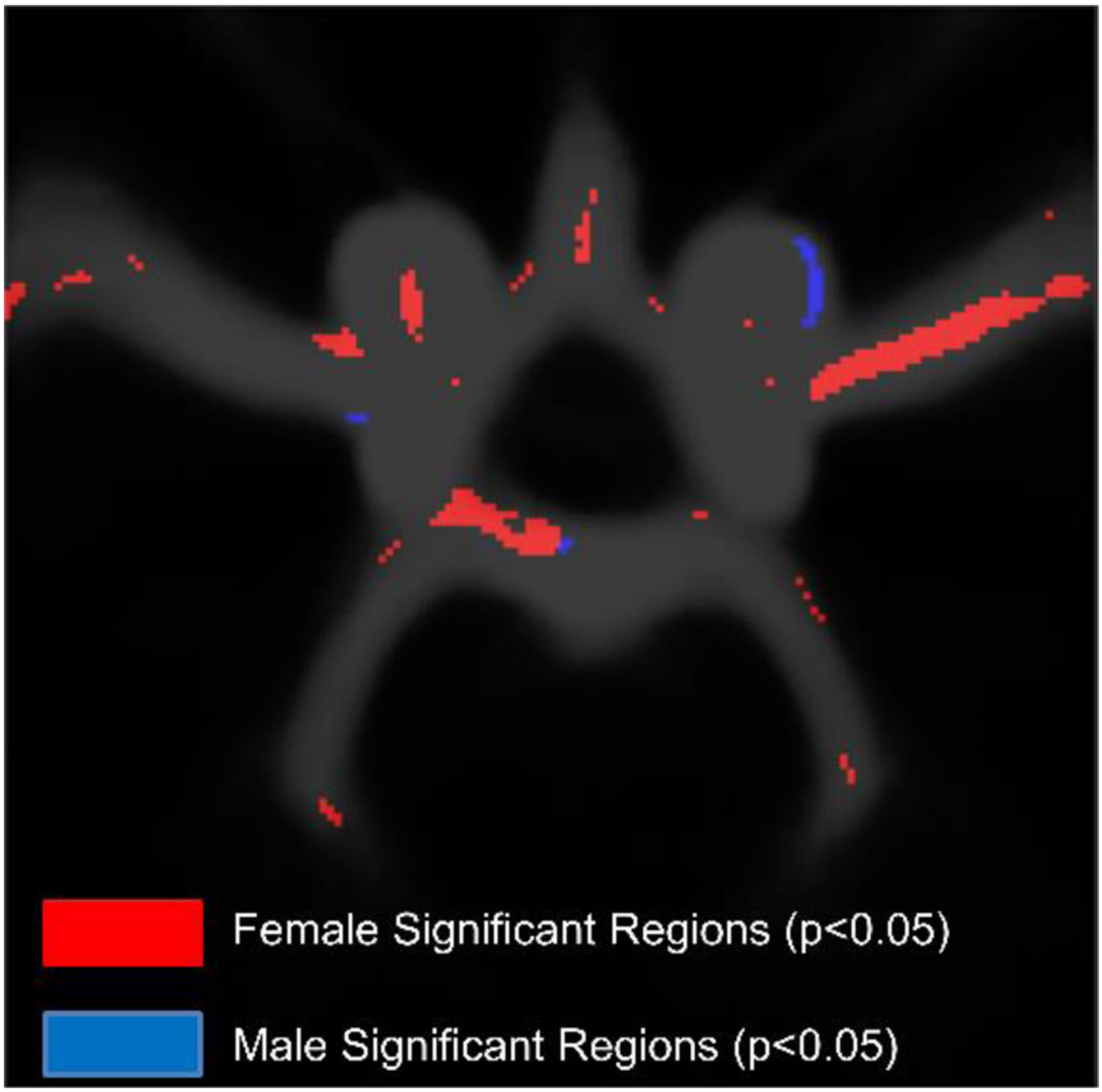

Figure 7:
Dictionary Permutation Testing. Most of the significant regions correspond to the bright areas shown earlier in Figure 6. The Female-Male subtraction images show more outstanding statistically significant parts than the opposite. We also see differences between the left and right, with the right-hand side defined beforehand, having more significant statistically substantial regions.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1:

MRA acquisition parameters for the centers used in the statistical atlas generation. TE stands for echo time, TR for repetition time. Field strength corresponds to the MRA strength.

Dataset	Field Strength	TE	TR
	T	ms	ms
Hammersmith Hospital	3	5.7	16.7
Guys Hospital	1.5	6.9	20
NC Dataset	3	3.5	35

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2:

Results of SVM Permutation with Axial View. C.I. = Confidence Interval

Condition	Mean Error Rate	Standard Deviation	95 % ± C.I.
	%	%	%
Correct Labels	37.01	1.31	0.08
Random Labels	47.41	3.07	0.19
Left Side	40.73	1.49	0.09
Right Side	39.31	1.46	0.09

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 3:

Statistical Analysis for the Axial Dimension data. The test performed is the ANOVA test, with different sample sizes of 10,100 and 1000 samples sampled randomly from each distribution.

Sample Size	P-Value	F	MS
10	0.32	1.04	0.0032
100	6.36e-10	42.26	0.01
1000	1.60e-92	462.9	0.1

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 4:

Estimated P-values for the axial dimension comparison between the left and the right-hand sides. The data in the table results from the SVM permutation test described earlier.

Crop taken	P-Value
Entire Image	0.0009
Left	0.006
Right	0.002

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 5:

Results of SVM Permutation with Coronal View. C.I= Confidence Interval

Condition	Mean Error Rate %	Standard Deviation ± %	95 % C.I. ± %
Correct Labels	41.66	1.33	0.08
Random Labels	47.76	3.13	0.194
Left Side	44.33	1.54	0.09
Right Side	43.55	1.36	0.09

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 6:

Anova analysis increasing the sample sizes analyzed for statistical analysis. To achieve statistical significance, we need to study at least 100 trained permutated models.

Sample Size	P-Value	F	MS
10	0.1369	2.39	0.00053
100	1.41e-05	19.82	0.005
1000	4.85e-43	198.45	0.045

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 7:

Estimated P-values for the coronal dimension comparison between the left and the right-hand sides. The takeaway is that compared to the axial viewpoint, using only the left or right crop would not lead to statistically significant classification when using SVM. There is some overlap between the random label permutation and correct label performance.

Crop taken	P-Value
Whole Image	0.009
Left	0.12
Right	0.06