A Topological Study of Brain Ageing

Ameer Hassan Saadat-Yazdi

4th Year Project Report
Artificial Intelligence and Mathematics
School of Informatics
University of Edinburgh

2021

Abstract

Topological data analysis (TDA) is a newly emerging field of data science which captures information about the shape of data using the mathematical theory of topology. TDA has been shown to be a powerful tool to engineer interpretable features for machine learning. This report shows that persistent homology, a tool of TDA, can be used to quantitatively study the qualitative patterns of neurodegeneration due to ageing. In particular, the report demonstrates that the features obtained from perstent homology give meaningful features which can be used to train a regression model to predict age with an R^2 score of 66%. It further goes on to demonstrate that these features can also be used to detect neurodegenrative diseases like Alzheimer's disease and constructs a model which achieves an F-1 score of 62%. Evidence is presented to suggest that the topological classifier may perform better than classifiers built on convolutional neural networks.

Acknowledgements

I am foremost grateful to my supervisor Rik Sarkar for the time he spent in guiding the writing of this report and the insights he shared with me throughout the semester. I extend this gratitude to Rayna Andreeva for her help and encouragement when I was feeling uncertain about my work. This project would have been an impossible feat were it not for their help.

I thank all my friends who allowed me to drone on for hours about my project and faked interest in my project as I explained the mathematical details of topology to them. Were it not for those hours I would not have been able to translate my thoughts into coherent words much less be able to compose them into this report.

Finally, I give thanks to my family for supporting me during these turbulent times and showing faith in my capacity to produce a piece of work I am proud of.

Contents

Chapter 1

Introduction

This report reveals that TDA is able to capture and summarise many qualitative biomarkers of ageing and provides an intuitive representation of the shape of the brain. These representations can be used in traditional machine learning pipelines to build models of brain ageing and/or disease with relatively high accuracy. These models, being trained on topological features are more interpretable than previously used techniques, leading to safer and better understood tools for clinical applications.

1.1 Problem

Researchers have identified several qualitative biomarkers with which to measure ageing given the physical structure of the brain. It is well known, for example, that there is a strong correlation between the thinning of the cerebral cortex and age [17]. Ageing is also seen to coincide with the emergence of lesions and general atrophy [39] of the interior of the brain. Although there are many qualitative markers of ageing, there is no obvious way to measure these changes quantitatively.

Being able to determine the age of the patient solely based on their brain anatomy can be useful for many reasons. For example, it allows us to better understand the physical changes that characterise normal ageing and then identify anomalies which indicate structural changes due to illnesses like Alzheimer's.

This report provides a method for translating qualitative observations of brain ageing into quantitative measures using persistent homology, a tool to capture topological information. These features allow us to measure the statistical relationship of these changes with age. It further investigates how these new insights can help with identifying and diagnosing Alzheimer's disease. The main contribution of this work is the identification of a novel biomarker of ageing which encodes the qualitative markers discussed above.

Given that this is a novel approach there is little previous work to fall back on. The choice of how to compute topological features, for example, had to be informed by reviewing the literature on the effects of ageing on the brain. Another challenge was to determine what exactly the features computed capture about the structure of the brain which again relied on an understanding of the existing biomarkers of ageing.

1.2 Why persistent homology?

Topology is a field of mathematics that seeks to characterise and differentiate between shapes. Whereas geometry is focused on computing local information such as curvature and distance, topology attempts to understand how global properties such as connectedness and the number of holes can differentiate objects. This lends itself quite well to cases of image analysis where the geometry of images can substantially differ while qualitative properties are similar. This is particularly useful in brain imaging, where the structure of the brain can vary tremendously. Topological features overlook geometric variation and can identify global differences in shape.

Persistent homology is one method that can be used to obtain topological information from data by studying its holes. Given that these methods are built on strong theoretical foundations we know exactly what information is being captured by the features derived from persistent homology. This means that, using persistent homology, one can obtain set of highly interpretable features which can be used to understand precisely what information models trained on these features capture about the data. This is of particular importance in clinical applications where explainability of models is highly valued.

1.3 Topological effects of ageing

Let us explore why topological insights are useful in studying changes in the brain. The images, in section 1.1, were chosen to demonstrate the changes that occur during ageing. In the young brain we see that the grey matter consists more or less of a single contiguous region with small isolated segments. In the older brain the grey matter is broken up into several disconnected regions, whereas the younger brain is mostly intact. In topological terms, we say that the younger brain has fewer connected components, whereas the older brain contains many. We also observe that the grey matter in the younger brain encloses a number of small white regions whereas the older brain has many small pockets of white matter. These pockets correlate to holes in topological terms and are the main features persistent homology seeks to capture.

Similarly, the white matter of the young brain is more or less fully connected while the older brain is broken up by the grey matter. Thus older brains should show a larger number connected components in the white matter.

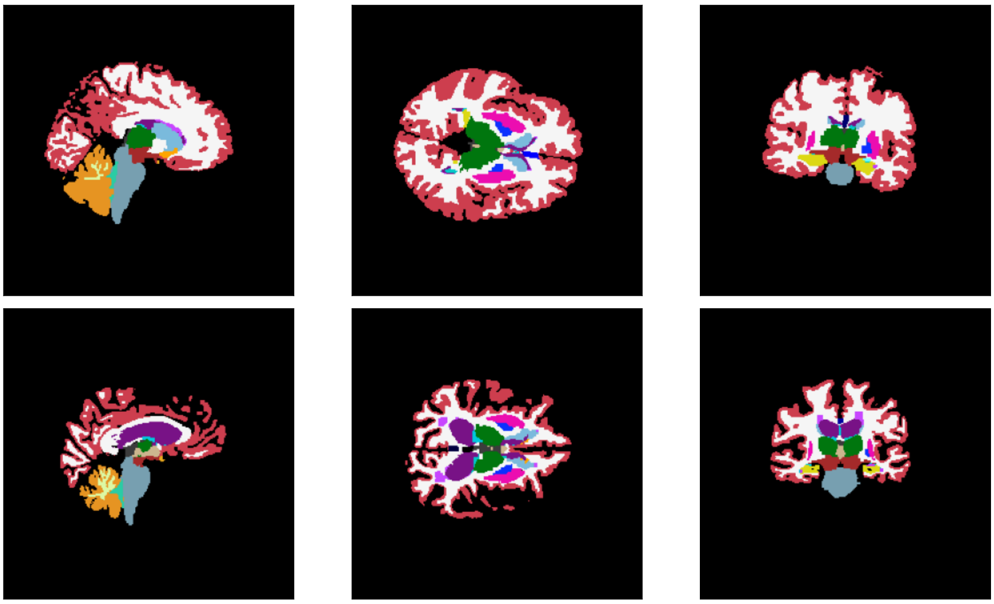


Figure 1.1: Brain of a healthy 48 year old (above) and an 83 year old (below). Grey matter is shown in red and orange. The white matter consists of everything else.

This should provide a general intuition as to why persistent homology was chosen to study these changes. It captures, very naturally, various phenomena which are associated with growing older and combines the various signs of deterioration together simply by measuring holes.

1.4 Summary of approach

The approach taken in this study consists of three separate pipelines. Three sets of features are computed on the brain, one derived from the outer layer (grey matter), one derived from the inner from the internal structure (white matter) and a final one acquired from a point cloud reconstruction of the outer surface.

The homology is computed differently over the three segments. In the grey matter it is computed in such a way that it captures changes in density and thickness and the white matter features indicate changes in pixel intensity (section 3.3). Homology in the point cloud is more complex, it captures information about curvature, density and connectedness simultaneously.

The three sets of features are combined in a final random forest model to predict ageing. The model indicates which features are most strongly correlated with ageing allowing for a better interpretation of what the topology tells us about age. These insights are then used to study the topological changes caused by Alzheimer's.

1.5 Overview of results

I have performed a series of experiments which show a significant relationship between brain topology and ageing. In particular I show that the various signs of deterioration can be linked to the increase in 0 and 1 dimensional holes in both the white and grey matter, as well as in the cerebral cortex (section 4.2). The final model, described in section 5.2, achieves an R^2 score of 0.66, demonstrating a strong correlation between topological changes and ageing.

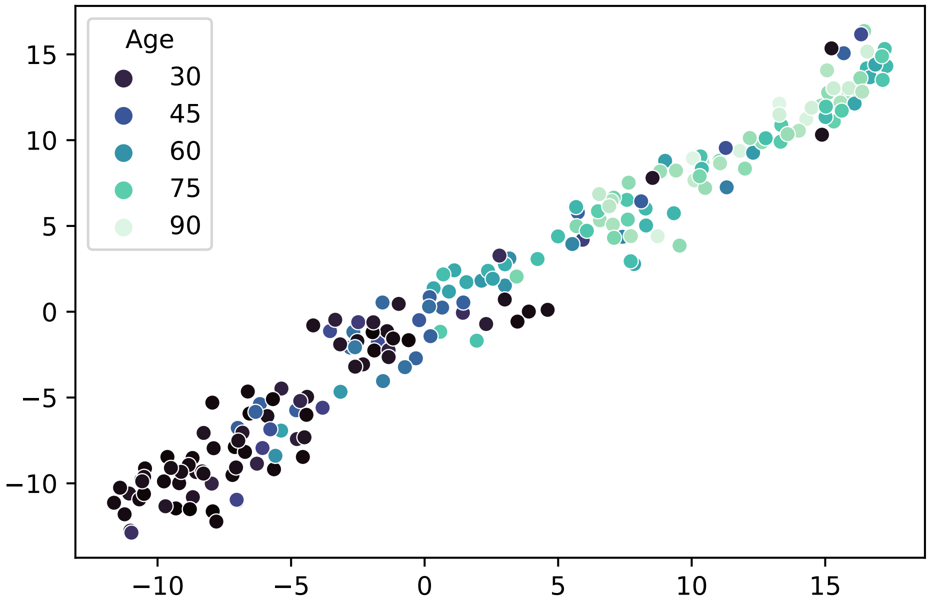


Figure 1.2: TSNE plot of grey matter features shows that the features extracted naturally cluster scans by age. Topological features are able to intrinsically capture changes that occur with age.

I also show that persistent homology alone can capture a lot of qualitative information about the structure of the brain. I show that persistent homology can be used to measure the thickness of the cerebral cortex (section 4.1), as well as the degree of folding, called gyrification, of the outer surface (section 5.1). This provides a unified approach to obtain measurements of various parts of the brain which alternatively would have to be performed using time intensive and specialised algorithms.

Finally, in chapter 6, these findings are used to study scans of patients with Alzheimer's. The markers of Alzheimer's are tangled up with the features of ageing as explored in section 6.1. A solution to this problem is identified in section 6.2 producing a final model with an F1 score of 0.53 compared to a baseline score of 0.31, table 6.3.

To summarise this report contributes the following to the existing literature:

- A method to compute informative topological features from structural MRI scans

- Discussion of how changes in persistent homology correlates with known biomarkers of ageing
- A novel efficient method to measure gyrification and cortical thickness
- A high precision and interpretable model to predict Alzheimer's disease

1.6 Report Outline

The report is structured as follows:

- Chapter 2 gives a brief and intuitive introduction to the mathematical and neurological terminology and concepts used throughout the paper. The mathematical introduction is supplemented with appendix A which contains a more rigorous introduction to homology.
- Chapter 3 describes, and provides justification for, the various preprocessing steps which produced the set of features used in later chapters.
- The features computed in the previous chapter are then used in chapter 4 to predict age based on the topology of 3D MRI images. The models are evaluated and explored to identify what topology says about internal brain structure.
- Similarly, chapter 5 performs a similar exploration, this time by studying the topology of the outer surface of the brain.
- Having studied how age affects the topology of the various parts of the brain, 6 proposes and evaluates a model to detect Alzheimer's.
- The final chapter provides an outline of some of the main limitations of the report, suggestions for further work and a summary of the points discussed in previous chapters.

Chapter 2

Background

In this chapter, I wish to give an intuitive explanation of the features that topological data analysis (TDA). Specifically the aim of this chapter is to understand the basic notion of topology and how it is measured by persistent homology. For a more formal introduction to homology see appendix A. I will also give a brief summary of neuroanatomy introducing some physical characteristics of the brain and some useful notions that will be used throughout the report.

2.1 Persistent Homology

Topology is a branch of mathematics that captures the shape of objects based on global properties. In contrast to geometry, which concerns itself mostly with the local behaviour of shapes e.g. curvature, length, angle etc... topology is better suited to the study of global properties such as connectedness and compactness. Essentially, from the lens of topology, two objects, A and B, are considered identical if there exist continuous maps $A \leftrightarrows B$ which transform A to B and B to A.

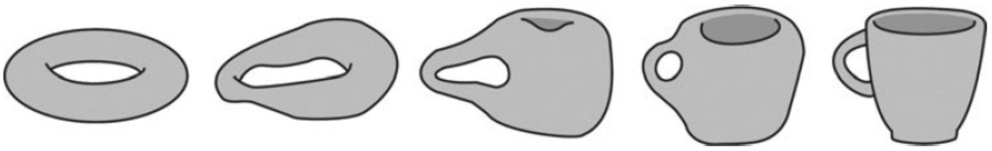
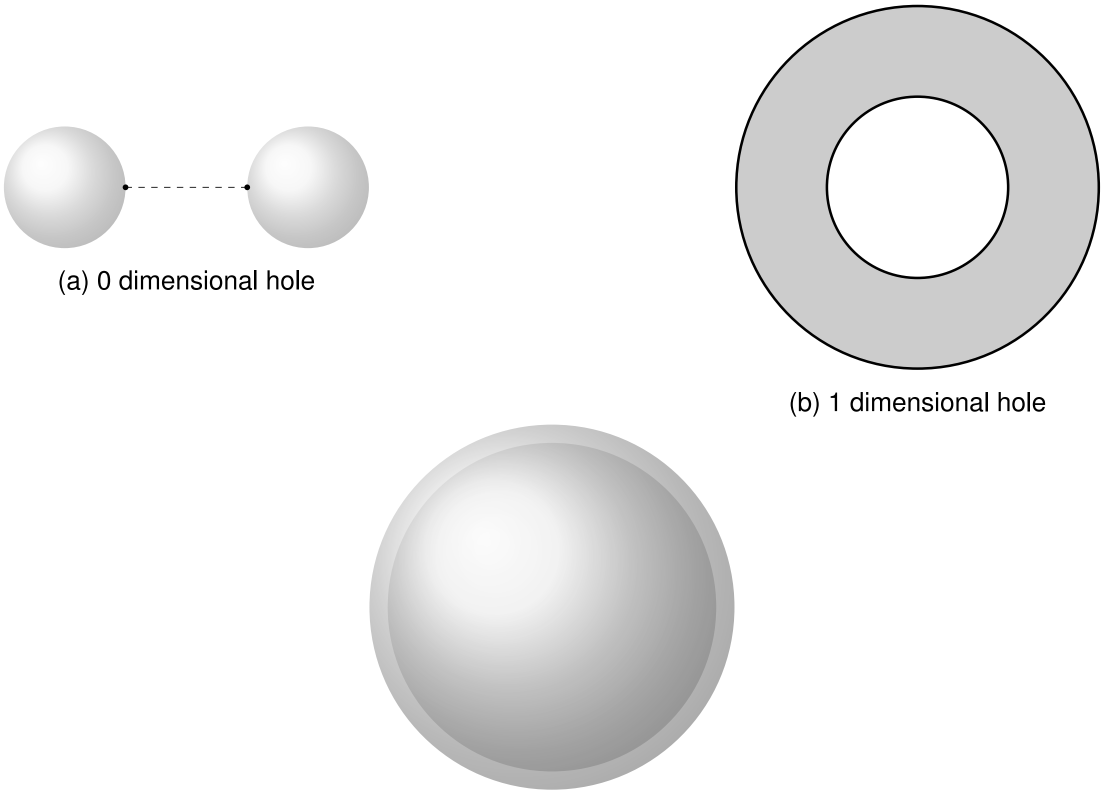


Figure 2.1: Continuous transformation of a coffee cup into a doughnut and vice versa. This is an example of two objects with the same topology.

Just as how length and volume are tools to measure geometry, there exist several tools that can be adopted in order to measure topology. One tool that is particularly useful in data analysis is homology. Homology is a measure of how many holes a shape has and generalises the notion of a flat hole to arbitrary dimensions.[23] Homology defines a zero dimensional hole as the gap between connected components. A one dimensional hole is what one would intuitively think of as a hole; a disk shaped gap on the surface of an object. A two dimensional hole is a void, a missing region competely enclosed by the object e.g. the interior of a balloon.



(a) 0 dimensional hole

(b) 1 dimensional hole

(c) 2 dimensional hole

Figure 2.2: Examples of holes in varying dimensions. A zero dimensional hole is the gap between connected components, a one dimensional hole is a disk shaped gap and a two dimensional hole is the void enclosed by a hollow sphere.

In order to compute the homology of an image we need to represent its topology. This is normally done by representing the image as a cubical complex, a set of cubes glued together like building blocks, see A. A limitation of this representation is that a cubical complex can only capture shape and is not able to represent pixel intensities. This means that one must choose which pixels should be included in the complex and which should be discarded. This choice is usually done by discarding all pixels with intensity below a certain threshold. However, different choices of threshold values will give different cubical complexes with drastically different topology. Figure 2.3 shows how holes appear and disappear at different threshold values. In addition to this, given that topology does not discriminate between sizes, a small one pixel hole would be the same as a hole spanning a much larger region. This means that the homology computed in this way is highly sensitive to small perturbations in the data or choices of filtration value.

The key insight which led to the inception of topological data analysis was to construct multiple complexes $\mathcal{C}_i$ by varying the threshold value such that,

$$\mathcal{C}_0 \subset \mathcal{C}_1 \subset \ldots \subset \mathcal{C}_k.$$

This sequence is called a filtration on C and the indices, i, in C_i are the filtration values. The holes that persist across many subsets in the filtration can be considered to give a truer representation of the topology underlying the data. Homology with filtrations is called persistent homology.[9]

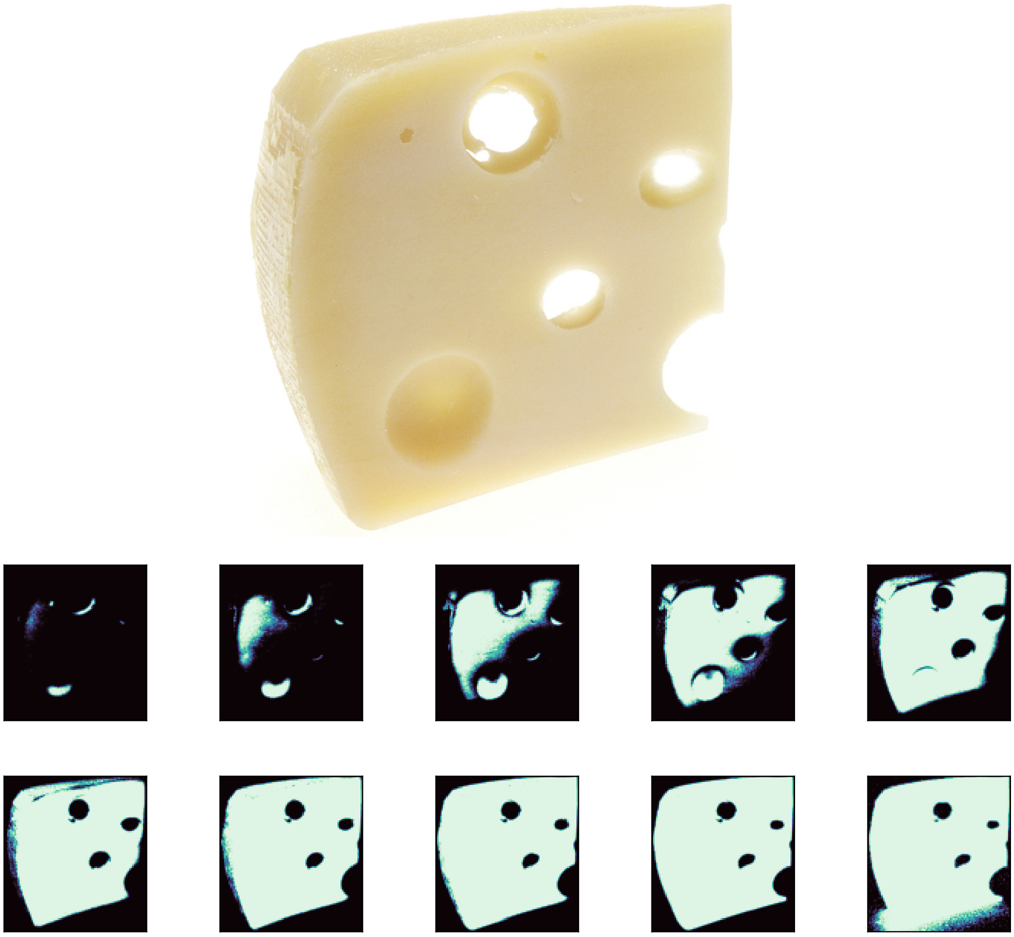


Figure 2.3: A block of cheese with 3 one dimensional holes. The filtration shown filters out all pixels above a certain intensity. As the threshold is increased the number of holes changes, however, the three holes persist for over half the set of filtrations.

Figure 2.3 shows how changing the threshold value can give rise to holes being generated or vanishing. We say that the filtration value at which a hole appears is its 'birth' and the value at which it vanishes is its 'death'. And so the persistent homology of a filtered complex can be represented by the set

$$\{(birth(h), death(h)) : h \in H_n\}$$

where H_n is the the set of n-dimensional holes in the complex. This means that this set of pairs gives all of the information captured by persistent homology. These pairs can be plotted in two dimensions and visualised as persistence diagrams, 2.4.

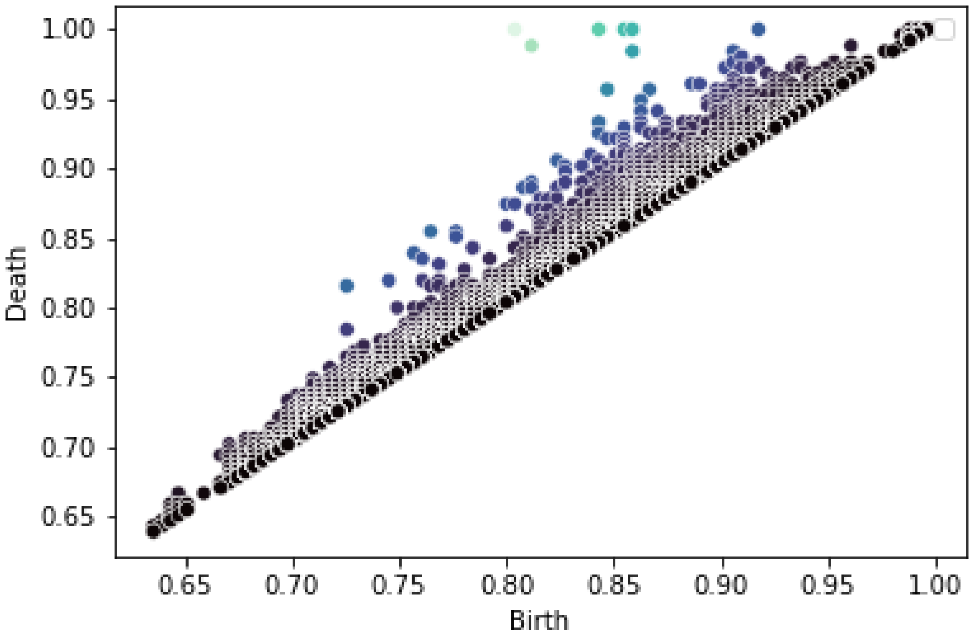


Figure 2.4: Persistence diagram of the one dimensional homology of 2.3. There are many holes close to the diagonal which emerge and die instantly, these are due to noise. However, there are a few points corresponding to the actual holes in the image in bright green. The points further from the diagonal correspond to holes which persist longer.

Images, including 3D images, come with a natural topological structure obtained by drawing edges around the pixels. It is not so clear how to construct a topological structure given a finite set of discrete points in space. The traditional approach is to construct the Vietoris-Rips complex on the set of points X. It is constructed as follows,

$$VR_r(X) = \{A \subseteq X | \forall x, y \in A, d(x,y) < r\}.$$

This definition says that given a value r and a set of points, we include every subset A of X where all the points have distance, d, less than r. A subset here is seen as an edge or face in the complex. We can easily induce a filtration on the Vietoris-Rips complex by increasing the value of r from 0 to infinity, see 2.5.

You can visualise the Vietoris-Rips complex by imagining a set of points in 3D space growing uniformly at a constant rate. Whenever two of the points touch their centres are joined by an edge. When three points intersect at a point, the triangle that their edges form is filled in. This generalises to higher dimensions as well but in this report, we will remain in the first two.

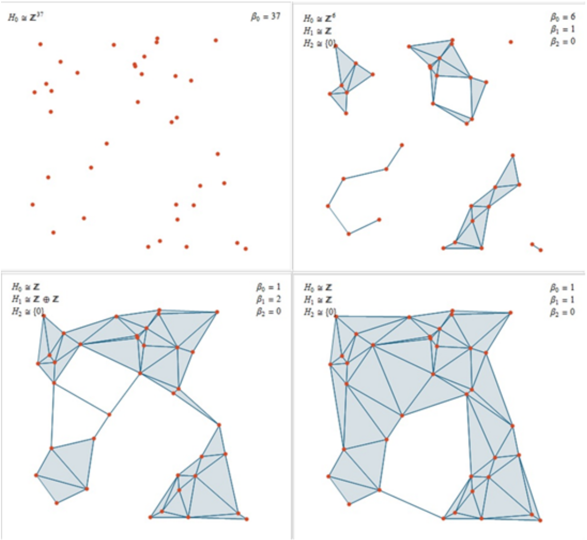


Figure 2.5: Vietoris-Rips filtration showing how the simplicial complex changes as r is increased.[11]

2.2 Basic Neuroanatomy

In this paper I discuss the relevance of a number of neuronal regions which play a significant role in detecting brain ageing. At a very basic level the brain can be separated into two main components, white and grey matter. The former consists of long chains of neurons surrounded by a fatty layer, which gives this region its colour. White matter plays no direct role in processing information, instead its main function is to pass information across regions of grey matter. Damage to the white matter leads to slower cognition due to a reduction of neural pathways. Gray matter is primarily located on the outer layer of the brain in the region known as the cerebral cortex. This region is where the majority of neural processing takes place. This region appears darker in the scans due to the lack of an insulation layer associated with the interior of the brain.[36]

The shape of the cerebral cortex is a useful indicator of both ageing and disease. The thickness, measured at each point between the outer surface of the brain and the interior side of the cortex for example acts as a strong indicator of neurodegeneration. In order to increase its surface area, the cerebral cortex is folded in on itself. This is called gyrification and the amount of gyrification is usually measured using the gyrification index which measures the ratio between the surface area of the cortex and the area of the convex hull surrounding the brain[45].

2.3 Related Works

Topological data analysis has seen a wide range of applications in biomedical sciences and has been used in a variety of settings to discover hidden structures in healthcare related data. In medical imaging, persistent homology has been used to identify brain tumors [38], improve image segmentation [8], distinguish between cancerous and healthy cells and map the functional structure of the brain. Beyond its applications

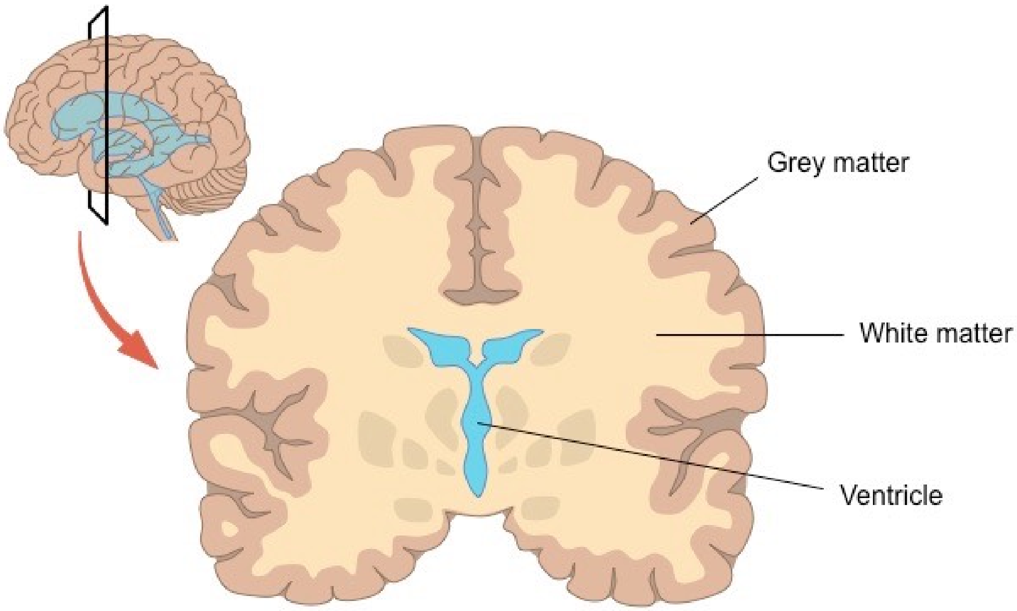


Figure 2.6: Cross section of the brain showing difference between grey and white matter.[1]

to imaging, topological methods have been used to uncover new subgroups of diabetic patients [29], identify subgroups of breast cancer patients with high survival rates [34] and characterise the structure of DNA. Appendix C gives a brief overview of some of these applications and the techniques they use.

As far as I have been able to determine there is no existing literature which seeks to apply persistent homology directly to structural MRI scans. Nonetheless, there have been a number of papers which have applied similar techniques to those used here in order to study various neurological phenomena. In one instance, researchers obtained 3D representations of brain artery networks in the brain; using persistent homology, the authors identified a correlation of 0.52 between topological metrics and brain age [3]. Another work applies topological methods to functional MRIs, these are MRI images taken in rapid succession which capture brain activity. Here the authors show that changes in the topology of connectivity networks can be associated with ADHD [20]. The majority of the literature follows the approach of the latter paper, seeking to use topological methods to understand networks of brain connectivity [1].

Statistical analysis of MRI scans and brain regions have been done predominantly using voxel and surface based morphometry. These methods measure statistically significant changes of volumes across different regions of the brain over the population of interest. The resulting measurements can then be used to identify which sub-population a patient belongs to. Although these methods are useful in their transparency and interpretability, they suffer greatly from confounding variables such as head movement and individual variations in grey matter folding making interpretations of results difficult.

[1] Image from `https://ib.bioninja.com.au/options/option-a-neurobiology-and/a2-the-human-brain/brain-matter.html`

[41], [24]

Neural models have been developed to identify diseases [26], [2] and predict brain age [10], [13]. These models are significantly more accurate than statistical methods but lack the interpretability required for clinical diagnosis. It has been shown however, that age regression models can be used as reliable indicators of premature deterioration of cognitive function. [25]

The work presented in this report adds to the current literature by providing a method to compute a set of highly interpretable and robust features which can be used to predict age, classify illnesses and better understand changes within the structure of the brain. The topological methods not only allow for transparency but also incorporate scale, translation and rotation invariance which overcomes the issues seen in voxel based morphometry and even neural networks.

Chapter 3

Methods

Before settling on the research question discussed in chapter 1 other applications of TDA to biomedical data were also explored. I attempted to detect congenital heart defects from CT scans of hearts, identify arrhythmia in ECG signals B, understand genetic evolution of viruses and studied a number of possible applications to fMRI data. The results of initial experiments in these domains gave mixed results and age prediction was identified as the most promising line of investigation for this project.

The persistent homology pipeline generally consists of several steps:

1. Construction of filtered complexes from data
2. Generation of persistence diagrams
3. Preprocessing and vectorisation of the diagrams
4. Machine learning

Steps 2-4 require a number of design choices, these were made by running experiments of half of the training set, to save on computational runtime, this subset was sampled to preserve the distribution of ages. In order to validate our models I train on 75% of this subset and report errors on the remaining 25%.

3.1 Data aquisition, preprocessing and software used

The dataset used in this study, OASIS-3, was provided by Oasis [27]. The dataset is comprised of a longitudinal neuroimaging study the dataset is longitudinal consisting of multiple images taken of around 450 patients over a period of 10 years. In order to simplify modelling, two images of the same subject taken at different times are considered to be independent. Of the entire dataset, only 888 images were used to reduce memory requirements the spread of subjects is shown in 3.1. Preliminary analysis to judge the promise of the approach was performed on the OASIS-1 dataset [30] a significantly smaller dataset with lower resolution images.

To begin with the data was drawn from other sources with no preprocessing, I found early on that preprocessing MRI scans took more time and expertise than was available

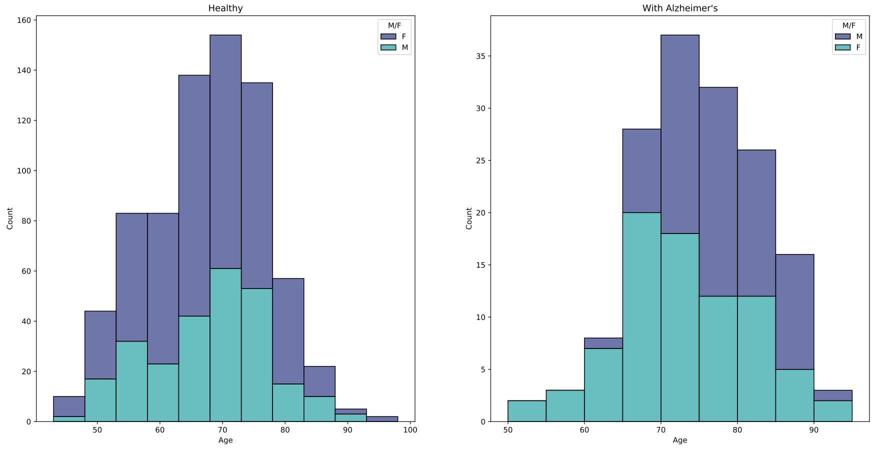


Figure 3.1: Age distribution and sex of subjects in the dataset.

for the project which is why I made the final choice to use the Oasis-3 dataset. The dataset was chosen for the availability of precomputed FreeSurfer [14] files which include skull-stripped scans, brain segmentations [16], cortical thickness measurements [15] and 3D reconstructions. This allowed for a much more streamlined analysis of the images, and meant that I could relate the results directly to the precomputed measurements of brain volumes and thicknesses.

In addition to FreeSurfer, several python packages were used to aid in the analysis. For the computation of persistent homology, i.e. computing persistence diagrams and features, giotto-tda was used [42]. In order to manipulate brain imaging data and load FreeSurfer files, nibabel was employed [6]. Machine learning models and other data manipulations were performed with scikit-learn [35].

3.2 Baseline

The baseline models are trained on volume measurements computed by the FreeSurfer pipeline. This serves as a strong baseline as these measurements are known to change reliably with both age [43] and disease [21]. In particular, these measurements combined with cortical thickness serve as the main diagnostic markers for Alzheimer's disease. The volumes measured are of the following regions:

- Intracranial
- Left/right-hand cortex
- Total cortex
- Subcortical gray matter
- Supratentorial
- Cortical white matter

- Left/right-hand cortical white matter

3.3 Image based pipeline

This section describes how topological features are extracted from the 3D MRI volumes. I rely mainly on empirical results to make choices between various design options in the pipeline.

Our data is given as a set of 3D MRI scans and associated segmentation files which allow us to extract specific regions of the brain. Given that the white and grey matter of the brain are affected differently by ageing, we use the FreeSurfer 'aseg.mgz' files to extract two copies of the skull-stripped brain 'brain.mgz' in which we remove the white matter and grey matter respectively.

In order to compute the persistence of our complexes, we require an appropriate filtration function which derives intuitive and statistically meaningful features. A number of possible functions from which we made our choice are outlined in [19] Given that the lesions in the white matter appear as a variation in pixel intensity of the raw scans, we simply filter on those values. Preliminary experiments performed on the white matter using other filtrations methods gave rise to features with no correlation to ageing, confirming our choice of function. On the other hand, the changes we seek in the grey matter are structural and we require a filtration function which captures local changes in the geometry. One appropriate choice is density, which associates to each pixel the number of active pixels within a given radius. The radius of the density function, is an additional hyperparameter which we set to 3, the largest value which still gives a manageable runtime. Figure 1.2 confirms that this is an appropriate choice.

For a number of reasons, it is difficult to perform statistical inferences directly on persistence diagrams. Instead, it is common in the literature to compute a vector representation of these diagrams. A zoo of possible methods are available to do this. There is close to no theoretical basis to help identify the most appropriate representation, instead an empirical approach was chosen, see tables 3.1. The results point towards choosing the Betti curve representation. The Betti curve divides the range of filtration values into a number of bins, and counts the number of features in each bin. This is the only persistence representation which gives equal weight to every hole, regardless of its 'lifetime'[1].

[1] The range of filtration values over which a hole lasts, $lifetime = birth - death$

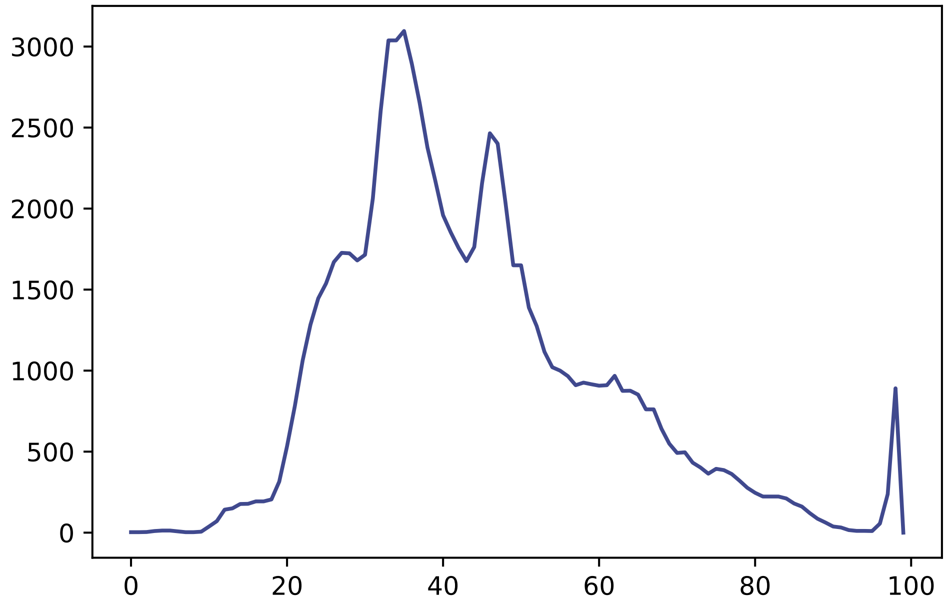


Figure 3.2: Betti curve of 2.3. This captures the density of points as you move along the diagonal of 2.4. The y-axis gives the counts for how many features are seen in each bin.

Representation	R^2 score
Betti Curve	0.3369
Persistence Landscape	0.1447
Heat Kernel	0.2773
Persistence Image	0.2918
Silouhette	0.0763
Euler Characteristic	**0.3967**

(a) Whole brain

Representation	R^2 score
Betti Curve	**0.4798**
Persistence Landscape	0.0359
Heat Kernel	0.4789
Persistence Image	0.4490
Silouhette	0.2574
Euler Characteristic	0.4406

(b) Grey matter

Table 3.1: Comparison of persistence representations for image based approach. Also, provides evidence that separating out grey and white matter will provide better results.

3.4 Point cloud pipeline

FreeSurfer's 3D reconstructions provide point clouds which can be used to construct a simplicial complex via the Vietoris-Rips filtration. See 5.3 for examples. In order to save on computational time and memory, I used an approximation of Vietoris-Rips, namely the weak alpha filtration which is known to produce similar results in low dimensions. Given the size of the point clouds it was infeasible to compute the complexes directly and so the points were downsampled with the VoxelGrid filter in order to obtain more manageable complexes. Note that even with downsampling, the second dimension was too large to compute and led to overflow errors. Thus I was forced to restrict my attention to the first two dimensions.

The choice of persistence representation was performed as above with results given in table 3.2

Representation	R^2 **score**
Betti Curve	**0.5160**
Persistence Landscape	0.1256
Heat Kernel	0.4064
Persistence Image	0.4144
Silouhette	0.4958
Euler Characteristic	0.4354

Table 3.2: Comparison of persistence representations for point cloud based approach.

Chapter 4

Cortical and Subcortical Topology

Here I train several random forest models and identify that topological features explain up to 51% of the variance in age. I was unable to perform a direct comparison of this technique with neural methods as the neural models were unable to learn given the size of our dataset. The best neural model was achieved using a variant of [26] which only learned a constant function i.e. an R^2 score of 0. I believe that this is due to the size of the training set being significantly smaller than that used in the literature. However, due to time and memory restrictions the use of larger datasets was not possible.

4.1 Cortical Thickness

A simple experiment demonstrates the ability of persistent homology to distinguish between rings of varying thickness, if we consider the brain a sphere then we should observe a direct correspondence between the persistence of H_2 features and the thickness of the outer membrane of the brain.

There are several caveats to using persistent homology as a direct measure of cortical thickness. Firstly, the maximum thickness able to be measured is bounded by the choice of radius of the density filtration. Secondly, a hole persists as long as the ring bounding it remains unbroken. If there is a thinner region of this hole that vanishes quicker than the rest, the persistence of the entire ring is bounded by the thickness of its thinnest section. Given the complexity of the structure of the cortex however, the folds it contains give rise to a large number of rings of varying thickness, which when combined give a general impression of the average thickness of the cerebral cortex. A random forest model trained solely on the grey matter features to predict average cortical thickness [1] achieved an R^2 score of 0.64.

[1] Computed from the FreeSurfer files: $\sum \frac{(lh.thickness \cdot lh.surfarea)+(rh.thickness \cdot rh.surfarea)}{lh.surfarea+rh.surfarea}$

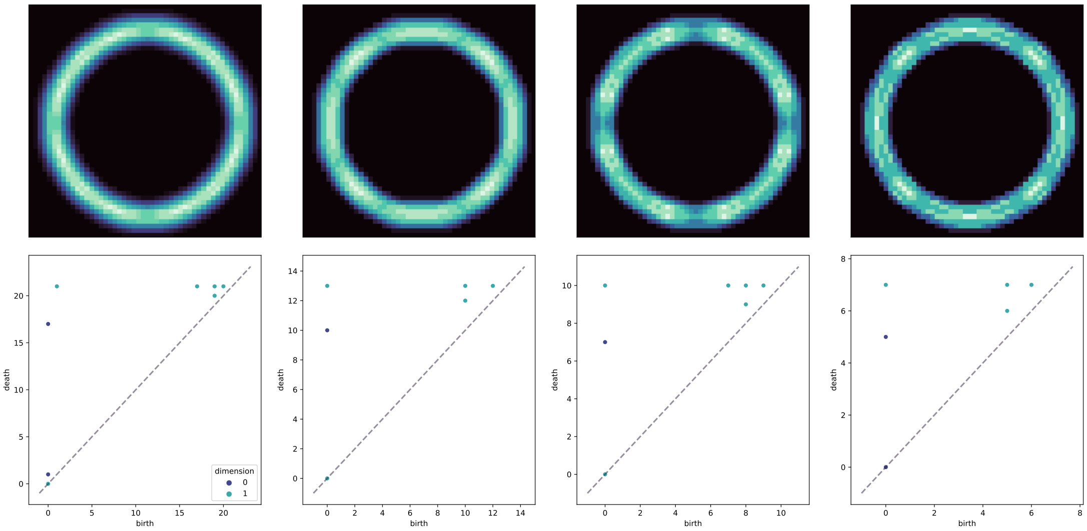


Figure 4.1: Rings of increasing thickness smoothed out with a density filtration demonstrate that persistent homology captures thickness; as the rings become thicker we observe the leftmost 1-dimensional hole feature dies sooner.

4.2 Interpreting the features

Before we train the model, we can investigate how the features we computed vary with age. Here I seek to understand how topology reflects known changes experienced during ageing.

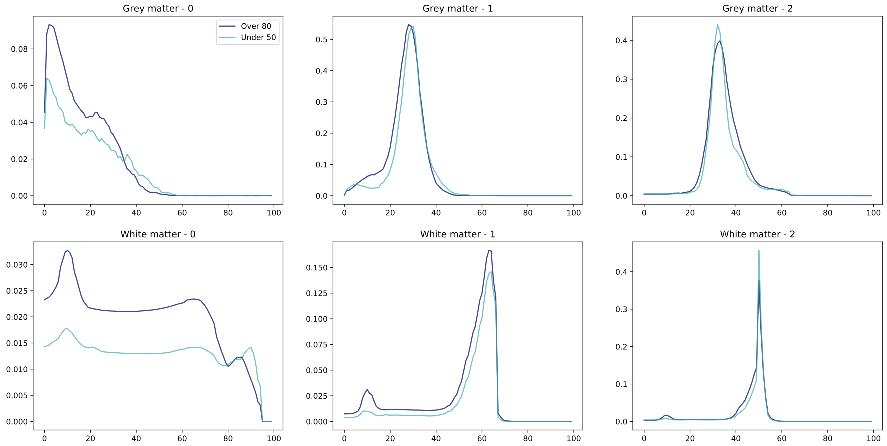


Figure 4.2: Comparison of average Betti curves for grey matter in young vs. old subjects. Vertical axis shows the Betti number and horizontal axis gives the filtration number. The 0 and 1 dimensional Betti numbers are consistently lower in younger subjects than older ones, proving that ageing increases the number of holes in grey matter.

Figure 4.2 shows that the first two dimensions contain more holes in older subjects, this can be identified with the deterioration of the cerebral cortex in the grey matter and increase in lesions in the white matter. However, in the second dimension the peaks are lower, possibly because the increase in gaps in lower dimensions breaks down the boundaries of voids.

By zooming in on the correlations between individual grey matter features we see a similar phenomenon as described above (figure 4.3). In the first two dimensions of grey matter features we see a strong positive spike followed by a negative one as one increases the filtration value. If we refer back to 1.1, we notice that the deterioration of the cerebral cortex caused by ageing gives rise to many disconnected spots (zero dimensional holes) in the image as well as to small patches of white matter fully enveloped by grey matter (one dimensional holes). This is the first spike. If we further increase the filtration value these spots vanish since they are thinner and less dense giving lower Betti numbers. The third dimension displays a kind of bimodal correlation, this could possibly correspond to different sub-populations with varying patterns of ageing.

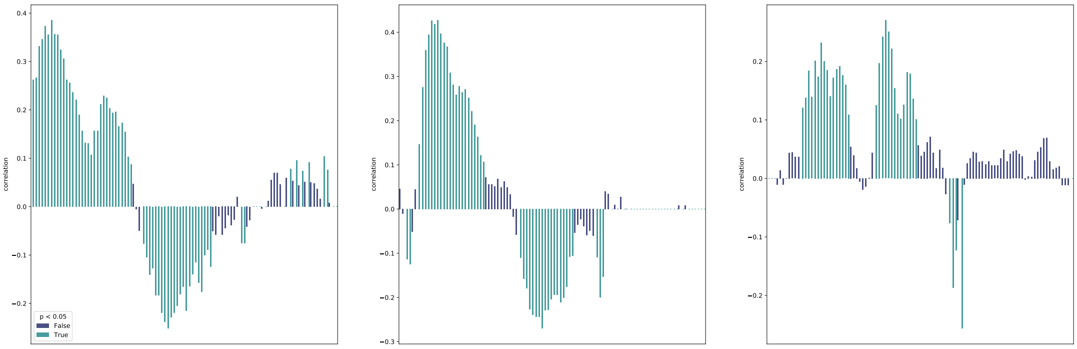


Figure 4.3: Correlations of features in grey matter with age. The p value indicates the likelihood that the correlations are spurious.

The white matter displays different behaviour, possibly due to our choice of filtration value. We see that the presence of holes in the first two dimensions is consistently positively correlated with ageing. The number of gaps and lesions in the white matter are high at almost every filtration level. Particularly in the second dimension we see that the presence of holes and discolourations are robust indicators of ageing.

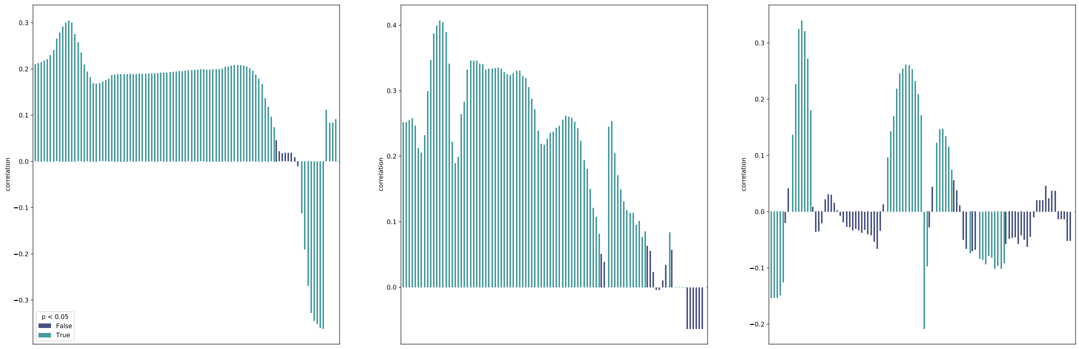


Figure 4.4: Correlations of features in white matter with age.

4.3 Understanding the model

The model trained on both white and grey matter features performs the best with an R^2 score of 0.51 (see 5.1). In order for these results to be useful we need to be able to interpret the random forest models, this is not necessarily straightforward given the nature of random forests. We can leverage the fact that random forest model predictions depend only on a small number of high importance features [4]. So, in order to understand our models, it suffices to understand only its behaviour given the most important features.

One approach to interpret random forest regression models is to plot partial dependencies for the most important features. This is done by choosing a feature x and sampling n datapoints from our testing data. We then vary x from 0 to 1 and plot the model's prediction. Averaging over all n samples gives a general idea of how the model uses x to make predictions.

Figure 4.5 gives a good indication the model behaves as we would expect, in general an increase in Betti number is associated with a higher age. However, in certain cases a negative correlation is seen. One possible reason for this in the case of grey matter features is that the overall density of the brain is much lower at higher ages, and so much of the brain would disappear at high filtration values. If this were true we would expect a greater number of holes to be seen early on which vanish sooner than those in younger subjects. In effect, this would produce Betti curves with a steeper decline and higher peak, which is precisely what we see in the grey matter curves of figure 4.2. Figure 4.5 indicates that the most informative grey matter features are obtained at low filtration values in dimension 1. Particularly, between 10-20 of 4.2 we see a distinct difference between older and younger brains, as younger brains have a larger number of 1 dimensional holes at lower filtration values.

Further insights can be drawn by investigating the distribution of importances across regions and dimensions. We see that although white matter features appear to be the most important, overall, the model is mostly influenced by grey matter features in

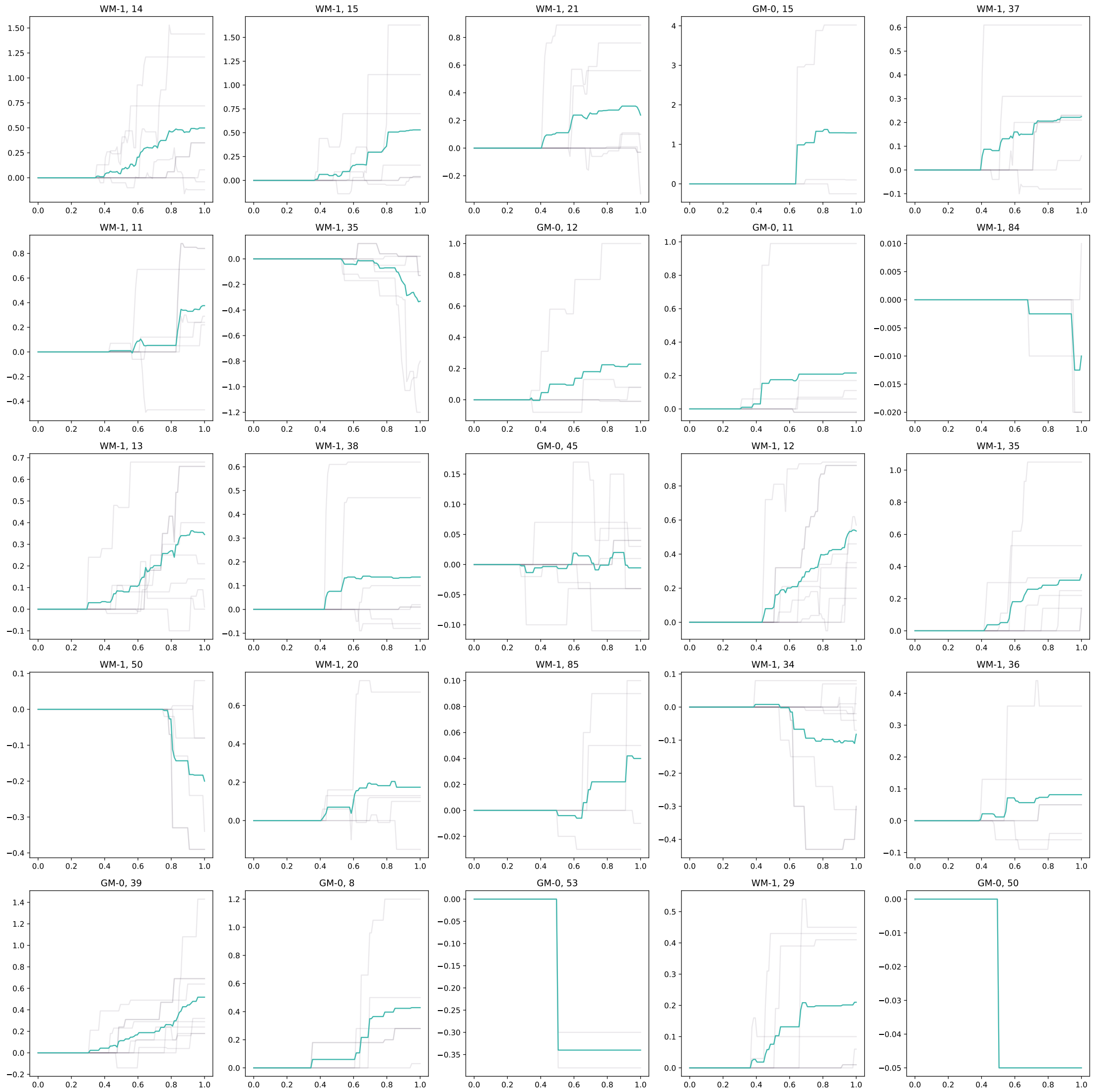


Figure 4.5: Partial dependency plots for top 25 features showing how the change of a single feature affects the model's predictions. The mean over the samples is shown in green. Vertical axis represents deviation of model prediction from model's prediction at $x = 0$. The titles indicate which feature is being plotted, (grey or white matter)-dimension, index of betti curve.

the 1st dimension. These features correspond, as was showed before, to measures of thickness in the cerebral cortex.

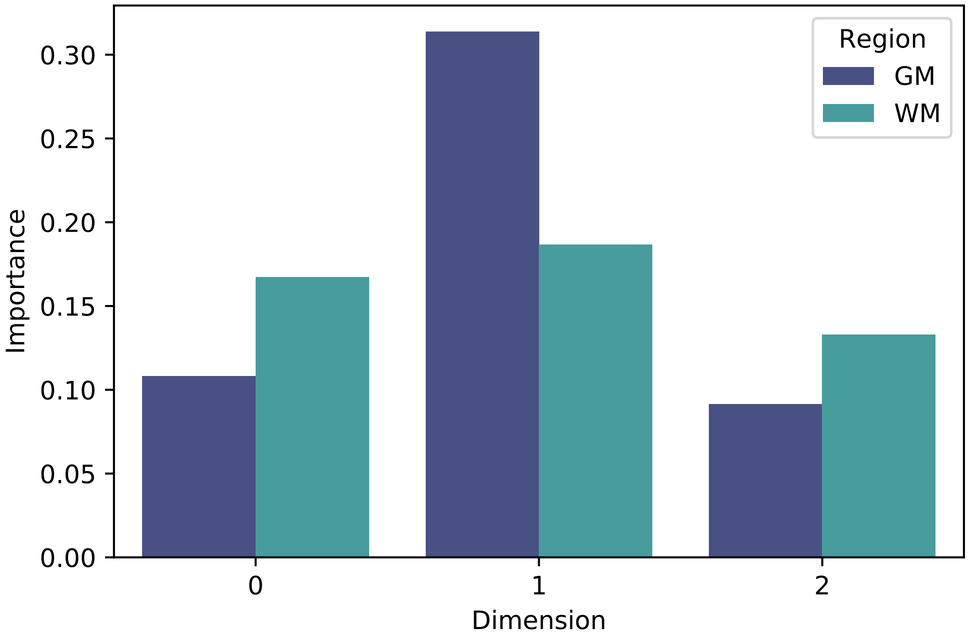


Figure 4.6: Histogram showing the sum of feature importances across different regions of the input data.

4.4 Points of failure

There are several interesting points to raise about the failure cases in the model. Firstly, the model appears to perform better on males than on females. The mean absolute error for male subjects is 4.72 ± 0.08 while for females the number is 5.23 ± 0.08. This is interesting since the dataset contains significantly more females than males, 85% more to be exact. Figure 4.7 indicates that the model is skewed by age as well, performing significantly better on older populations than younger ones. Upon deeper investigation, it appears as though many subjects demonstrate signs of premature deterioration of grey matter. It seems that females experience the greatest variation in deterioration with some showing no signs of ageing at over 70 and others showing signs as early as 52.

Let us examine two instances of failure, one in which the model severely overestimated the age, 4.8, and one in which the model severely underestimates, 4.9. The former image indicates that the model has identified the features we expect, the brain shows heavy deterioration particularly in the posterior of the cerebellum. The latter on the other hand shows a fully healthy brain with little signs of anatomical deterioration. So, the model is clearly picking up on the features we hope it to identify, however it may be that topological information may not be sufficient to fully capture the variety ageing patterns seen in healthy adults.

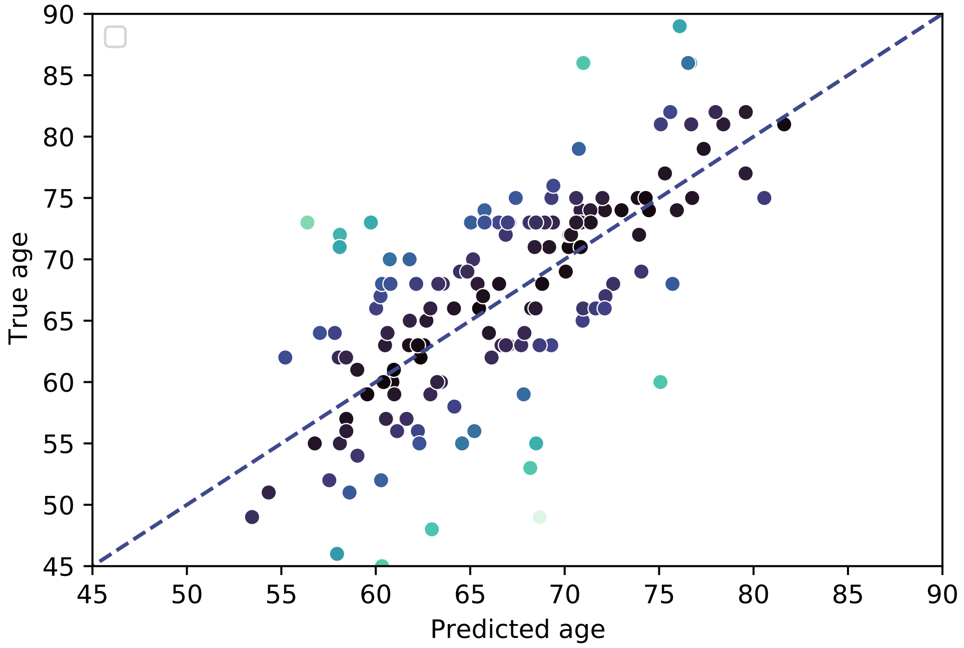


Figure 4.7: Predictions of the model vs true age with points coloured by distance from the diagonal. The model makes more mistakes with under 60s than with older subjects.

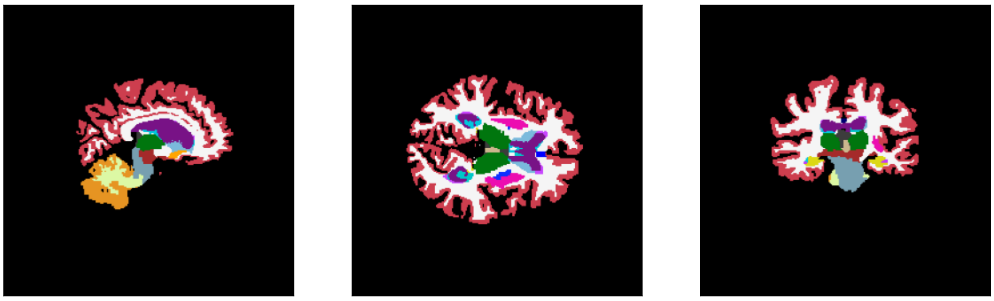


Figure 4.8: Brain of a 53 year old male, labelled as 68 by the model. Subject shows heavy reduction in cerebral cortex and white matter volume.

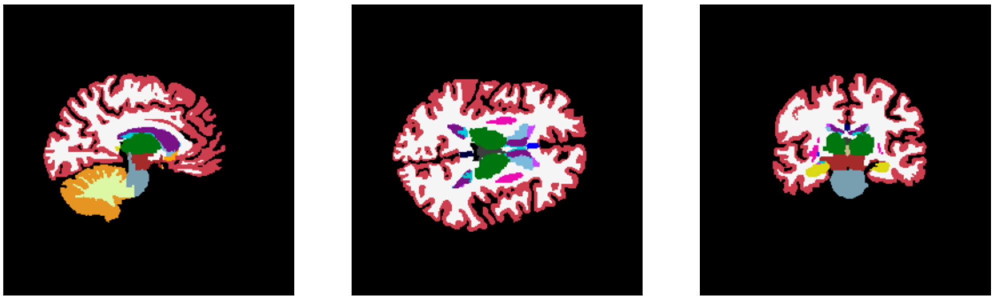


Figure 4.9: Brain of a 73 year old female, labelled as 57 by the model. Subject shows healthy brain, little signs of ageing.

Another possible reason for the model's inaccuracy could lie with the data itself. There are several examples in the dataset which suffer from issues in the segmentation. Although the FreeSurfer files were filtered to retain only well segmented images, it is impossible to avoid issues completely. Badly segmented images lead to chunks of missing grey matter for which the model overestimates the age. Although there are only a handful of cases where this issue is sever, this does introduce a bottleneck to the accuracy on this task.

Chapter 5

Surface Topology

This chapter proceeds in a similar vein to the previous one, exploring how the point cloud reconstructions of the brain's surface changes with age. Although the results here may seem similar to the previous chapter it is important to note that they represent disparate properties of the brain. The previous method identifies changes in volume and density of brain the brain while the point cloud describes the geometry of the outer structure of the brain in very high resolution. Thus we would expect persistent homology in this setting to identify how the curvature [7] and surface area of the outer layer of the brain changes with age.

5.1 Evaluating the point cloud model

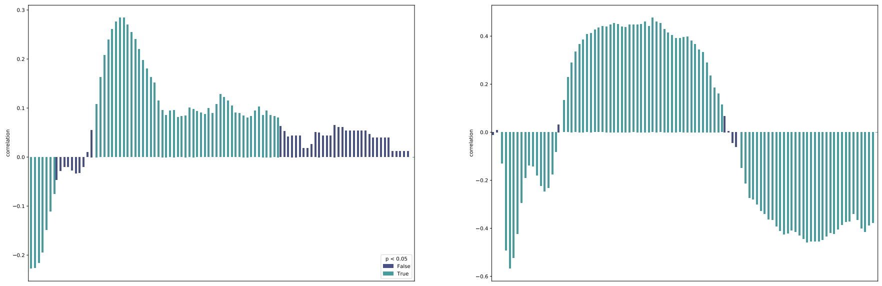


Figure 5.1: Correlations of features in point cloud homology. Zero dimensional on the left and 1 dimensional on the right.

The early filtrations of 5.1 which count the number of points in the point cloud (zero dimensional features) and their proximity (one dimensional features) indicates that at low filtrations younger brains contain a much larger number of points densely packed together. These points quickly combine to form connected simplices as they are densely packed. However, in older brains, these points are less dense and take longer to connect and die giving rise to many disconnected regions at mid to high filtrations. This is

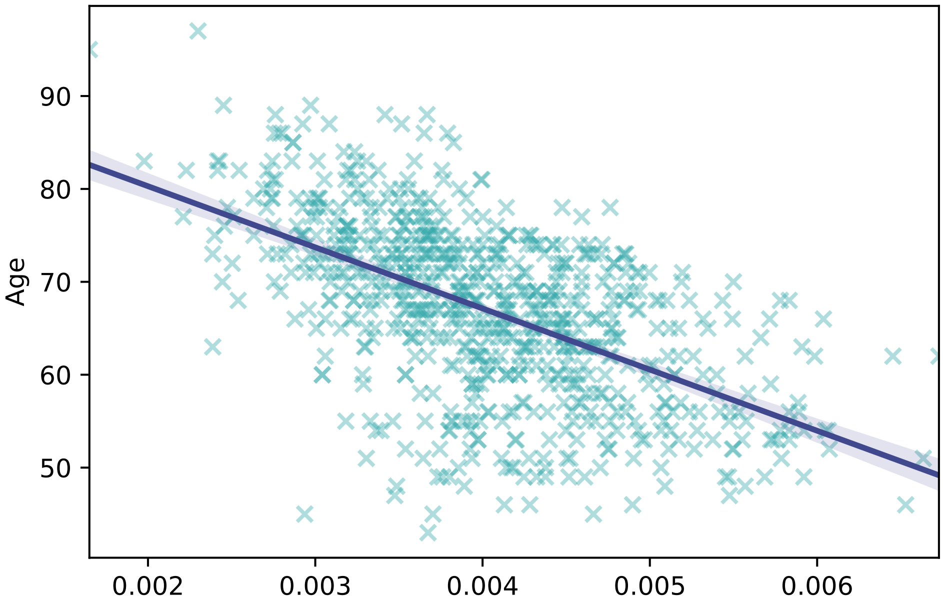


Figure 5.2: Plot showing how γ varies with age.

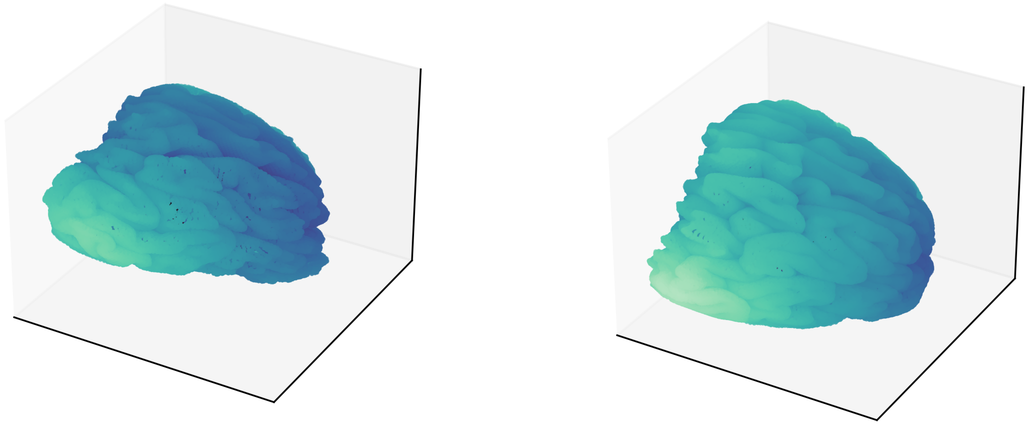
reflected by the one dimensional features which indicate an increase in holes as these simplices slowly connect around gaps. At high filtration values the Betti numbers correlate negatively, indicating that younger brains have a significantly richer structure at higher scales. This could be due to the in the gyrification (folds of the outer layer) which is a known biomarker of ageing [40]. At index 4 of the Betti curve in dimension 1, the magnitude of correlation is at its highest, with $\rho = -0.57$, 5.2. I will call this feature γ for gyrification.

Upon constructing the model we notice that the model takes advantage of this correlation, giving this feature a much higher relative importance than the rest.

5.2 Constructing a combined model

The analysis in previous sections suggests that the image and point cloud features capture different information about subjects that will be useful in determining age. And so, it is natural to ask if combining the two would enhance the model's predictions. I trained a random forest model on the grey matter, white matter and point cloud. In order to keep comparisons fair I dropped the third dimension which keeps the input dimension identical to the previous model's. The final results are summarised in table 5.1.

Although the combined model performs significantly better than the previous ones, the model continues to suffer from similar biases. The model continues to greatly mispredict the ages of subjects under 50, see 5.4. The bias in sexes is reversed, this is surprising as left hemisphere changes are known to correlate more strongly with



(a) Low γ corresponds to smaller folds appearing in the cortex, the surface is packed densely with folds. This is the brain of a 94 year old.

(b) High γ corresponds to larger and fewer folds. The figure shown is of a 42 year old person.

Figure 5.3: Left hemisphere of brains showing differences in γ.

Model	R^2 **score**	**MAE**
Baseline	0.46	5.43
White matter	0.44	5.82
Grey matter	0.48	5.54
White + grey matter	0.51	5.48
Point cloud	0.53	5.15
All combined	**0.66**	**4.47**

Table 5.1: Test set model performances for various feature sets.

male ageing [22]. Men achieve a MAE of 5.55 ± 0.05 while women score lower at 5.09 ± 0.04.

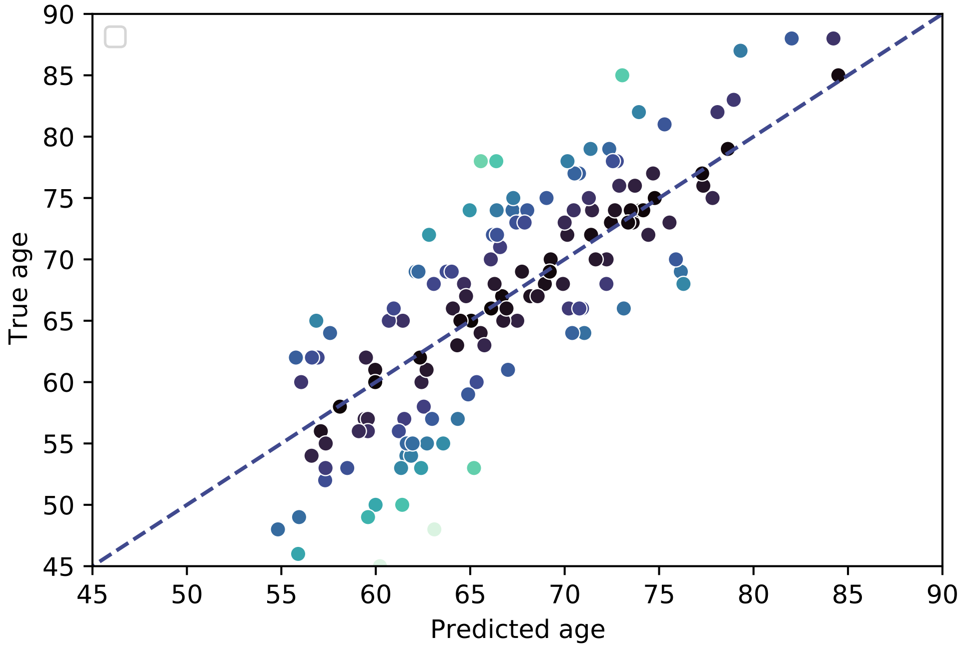


Figure 5.4: Combined model predictions vs. true age.

Chapter 6

Detecting Alzheimer's

The changes in the brain due to Alzheimer's are quite similar to the changes seen during ageing. The difference being that the deterioration is normally more exaggerated. Given that our model appears to be able to identify certain biomarkers of ageing, we expect that the same features should enable us to identify if a patient has Alzheimer's. By removing the effect of ageing from the features, I was able to achieve a 72% improvement upon the baseline model's F1 score, a random forest model trained on volumetric features (3.2), and classify the disease with relatively high precision. The model trained on topological features may perhaps also be an improvement on the state of the art results in [44] on the same dataset, yet I am cautious in emphasising this comparison since the models in this report are only trained on a portion of the data used by the authors of the previous paper.

Topological SVM	3D CNN	Dartel SVM
0.74	0.71	0.71

Table 6.1: Test set balanced accuracies for the best model trained in this paper (topological svm) vs. state of the art results in [44]. Note that results may be skewed as the topological model was trained and tested on a subset of the data used by the latter two models.

6.1 Diagnosis without age dependency

Features	F1 score	Balanced acc.
Grey matter	0.00	0.40
White matter	0.13	0.51
Point cloud	0.43	0.75
Combined	0.30	0.76
Baseline	0.36	0.67

Table 6.2: Validation set performace of random forest models with oversampling.

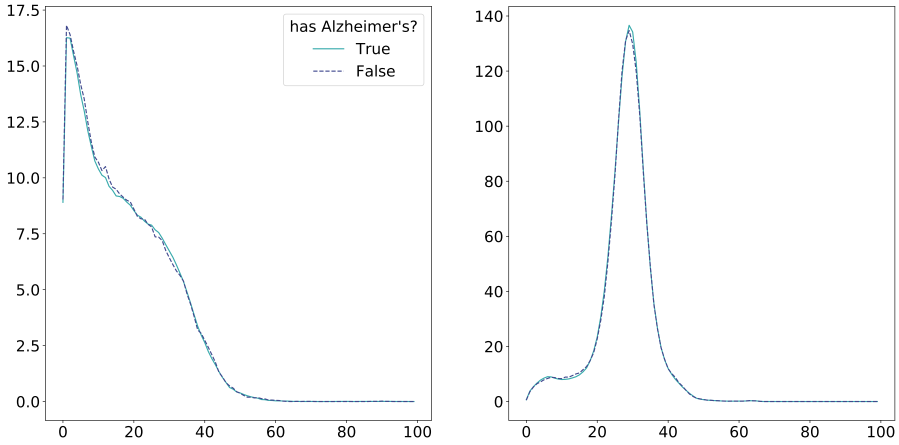


Figure 6.1: Average Betti curves of patients with and without Alzheimer's.

A naïve attempt to train a random forest classifier gives very poor results, 6.2. The accuracy of the models can be slightly improved by oversampling the positive cases of Alzheimer's in the training set to add 100 additional cases. This is surprising as the literature suggests that Alzheimer's disease has a distinct effect on the topology of the brain.

Previous works strongly indicate that changes in the grey matter, particularly of cortical thickness, are robust indicators of Alzheimers [28]. This does not seem to be reflected by the results which indicate that the model is unable to determine a relationship between grey matter topology and disease. This is likely because changes in cortical thickness are also strongly related to ageing, thus without incorporating a patient's age into the model it is impossible to distinguish between these phenomena. Additionally, figure 6.1 indicates that the grey matter Betti curves of patients with and without Alzheimer's are virtually identical. Again, this is possibly due to the fact that the thinning of the cerebral cortex experienced during ageing is confused with Alzheimer's.

The point cloud representation outperforms all others, we also see that its precision is improved by the oversampling. Looking at feature importances it appears as though γ is the most important feature of the point cloud model, 6.2. recall that γ is a measure of the folding of the cerebral cortex, this feature was shown to correlate strongly with the patient's age, possibly supplying the model with enough information to adjust for changes in age.

One of the main indicators of Alzheimer's are the dilation of the ventricles in the brain[33]. On an MRI scan this appears as the expansion of the large dark regions in the centre of the brain, see 6.3. The issue with topological features is that they are indifferent to changes in size and so this marker is not well identified by persistent homology. The atrophy of the hippocampus and white matter also serve as diagnostic features of MRI scans [18], but this is also true for ageing. This makes it difficult for the

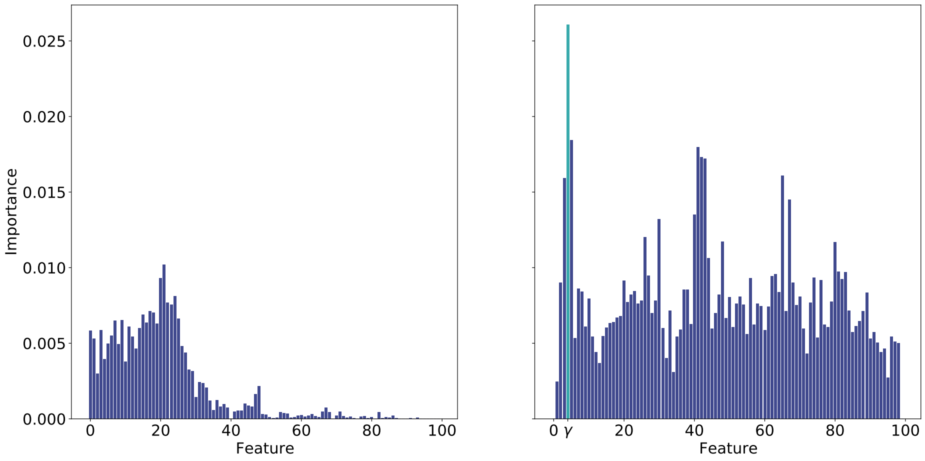


Figure 6.2: Feature importances of the point cloud model with γ highlighted.

model to be able to distinguish between the patterns caused by ageing and dementia. Therefore, white matter features are likely inappropriate for this task.

6.2 Disentangling Ageing from Alzheimer's Disease

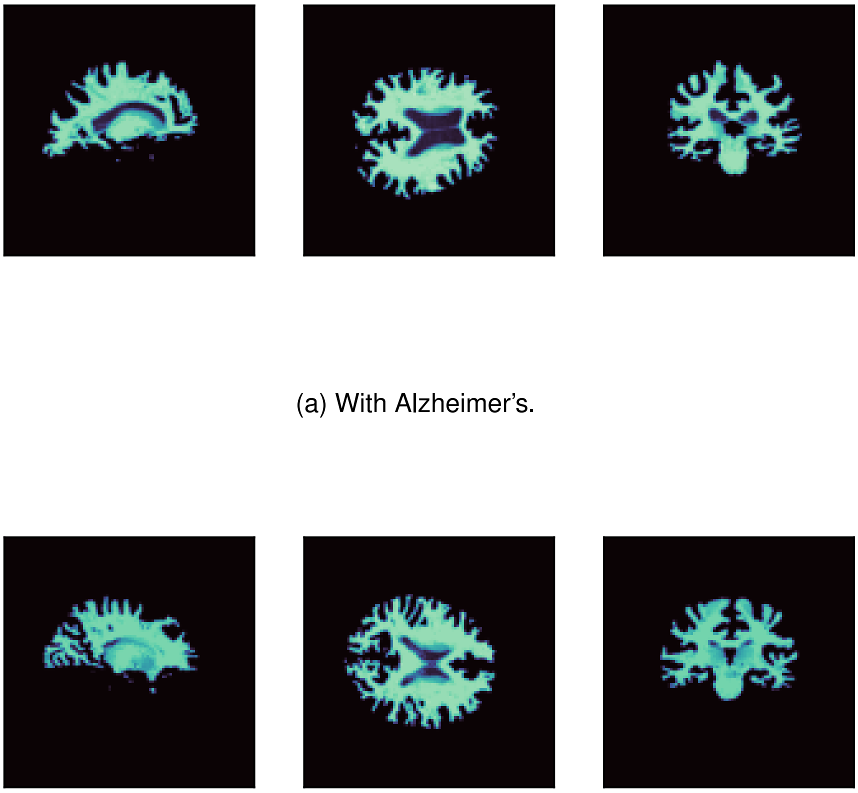
In order to separate ageing and Alzheimer's simply adding the age of subjects as an additional feature was found to not improve the results. Instead I trained a model to predict the expected Betti curve given the patient's age. Then by comparing the expected Betti curve with the true curve the model can measure how far the topology of the brain differs from a healthy brain.

Figure 4.5, suggests that the relationship between individual features and age is well modelled by a linear function. So, the expected features were modelled by linear regression. The regression model was trained on 200 samples from healthy patients from training set. A support vector classifier was then trained on the remaining dataset, taking the difference between the expected features and the true features as input. Again oversampling of Alzheimer's patients was found to drastically improve results.

Features	**F1 score**	**Balanced acc.**
Grey matter	**0.62**	**0.74**
White matter	0.53	0.64
Point cloud	0.44	0.57
Combined	0.47	0.59
Baseline	0.20	0.47

Table 6.3: SVM classifier performance on test set given disentangled features.

The results, 6.3, show a large improvement in performance, both in terms of accuracy and F1 score. In particular, both grey and white matter features now behave more like



(a) With Alzheimer's.

(b) Without Alzheimer's.

Figure 6.3: White matter of two patients with and without Alzheimer's. The size of the central holes is a strong indicator for the presence of Alzheimer's which causes these holes to appear much larger.

one would expect. Clearly, compensating for age is essential for identifying disease. However, this introduces a novel problem. The accuracy of the models are limited by the accuracy of the model of ageing. For example, the baseline features are not well modelled by a linear function and so compensating for age with a linear regression actually worsens the final result.

Chapter 7

Conclusion

The aim of this chapter is to critically analyse the methods used in the previous sections and identify some potential limitations. It suggests some possible solutions that could be explored in future work as well as some fundamental limitations of persistent homology as a measurement of brain ageing. Also mentioned are several tangential lines of inquiry which emerged in the process of writing this report. This chapter also serves to summarise the main findings of this project and their relationship with the goals outlined in the introduction.

7.1 Conceptual limitations

Topology is mainly useful in its ability to abstract away local structure. However, this also leads to a loss of information which can have negative impacts when attempting to measure degeneration at the local level. This means that topological data analysis will be limited by the extent to which neurodegeneration acts globally on the shape of the brain. In contrast to this, voxel-based morphometry is able to measure minute changes in a person's brain over multiple clinical visits allowing for a better understanding of changes at the subject level.

In section 4.4, figures 4.8 and 4.9 indicate that it is not uncommon for ones brain to deviate from the expected pattern. Therefore, an accurate model would need to incorporate information about the patient's lifestyle, genetics and health records. Furthermore, the dataset consists of subjects from age 40 onwards, it is not clear whether persistent homology would be able to capture the more subtle changes seen in the brains of younger adults.

The discussions in previous chapters have also touched upon certain biases in the classification accuracy towards certain subpopulations in the dataset. For example, section 4.4 identifies a statistically significant bias in the ability of the model to certain genders. Figure 4.7 shows that the model is also biased towards younger subjects, where the images obtained from under 55 year olds are consistently over estimated. This may perhaps be due to the lack of subjects in the younger age group in the training data but may also suggest a deeper problem; that the methods used may be ill-equipped to identify the age of someone who has yet to experience more widespread deterioration of the brain which begins to arise at a later stage.

Another limitation is the reliance of this project on precomputed FreeSurfer files. The computational cost and expert knowledge required to preprocess images using FreeSurfer invoke issues applying these methods to other datasets. The FreeSurfer pipeline is prone to errors and requires human input in order to rectify mistakes and also requires several hours to process the images for a single subject. Alternatively, one could make use of a more minimal method to segment grey and white matter regions and discard the point cloud based pipeline entirely at a cost to predictive accuracy.

Finally, the model we use here is trained on purely topological information. Without altering the models it is not immediately clear how one could incorporate additional information into the model. For example, the dataset used contains images from different modalities which emphasise different kinds of details. Ideally, we would like to inform the model of the imaging modalities that were used to generate the final freesurfer files so that it can identify these differences.

7.2 Future work and suggested improvements

The experiments in section 4.1 are overly simplistic. Although they indicate a strong correlation between grey matter topology and cortical thickness the precise nature of this relationship is not well explored. Additional study and experimentation would be need to understand exactly which features in the Betti curve correspond to thickness and how. Similarly, section 5.1 hints that γ may correlate directly with the gyrification index but additional analysis would be required to verify this relationship statistically.

There is also the question of how the models constructed throughout the report compare with state-of-the-art methods. Given that only a subset of the OASIS-3 dataset was used a direct comparison is difficult to make. It is necessary to compare how to persistent homology models compare to convolutional neural networks and voxel-based morphometry in order to draw more conclusive evidence as to whether or not persistent homology provides better markers of ageing. This was attempted during this project but, due to the size of the dataset, the neural network was impossible to train effectively. Additionally, the methods described in this paper provide only a proof of concept and further hyperparameter tuning would likely improve results further.

The emerging methods of persistent local homology provide an alternative direction. Persistent local homology allows the computation of homology at the local level combining all of the data to give insights about both local and global structure. Given that ageing causes deterioration differently in different regions, perhaps these methods will give more appropriate features which can emphasise the topology of more relevant parts of the brain. There is also a multidimensional generalisation of persistent homology which can be used to simultaneously compute the homology over multiple filtrations of the data. This would allow us to combine for example the intensity and density filtrations and compute the whole brain topology in one go. However, the tools available for these computations and the theoretical backing for these methods are limited.

While chapter 6 focused on how topological features can be used to detect Alzheimer's, the final model was not properly analysed in order to determine what exactly was being detected. Although the model scores are high, it still fails to detect 40% of Alzheimer's cases which makes it unsuitable for clinical application. In order to understand why this is, an analysis of both the regression and classification models would be required which the time-frame for this report did not permit.

Previous applications of persistent homology to MRI scans have predominantly focused on functional scans which capture brain activity in addition to the brain's structure. Therefore, given the results here, it would be natural to expect drastic changes in the topology of brain networks as the brain ages, perhaps even providing novel insights into these changes. There is also a gap in the literature concerning the topological changes in brain activity associated with neurodegenerative diseases such as Alzheimer's disorder which could additionally be explored.

Beyond the issues investigate in this report, there are further potential directions for the application of TDA methods to biomedical problems. While my preliminary experiments gave unimpressive results, I believe that with better data preprocessing and more sophisticated parameter choices new insights can emerge. Appendix B outlines one of the more promising directions, suggesting a new application of persistent homology which can be used to detect abnormalities in heartbeats.

7.3 Final remarks

This report has demonstrated that persistent homology can indeed be used as a marker for brain ageing. I have provided a methodology, with justification from existing literature, in order to compute topological features that correspond to brain ageing. Namely, I argue that white matter and grey matter filtrations should be computed separately to emphasise the different changes and suggest that a density filtration on grey matter images can be used a measurement of cortical thickness. By introducing surface topology, which measures gyrification among other properties, to the regression models, the accuracy is improved.

This relationship between topology and age is then leveraged to obtain a classifier of Alzheimer's which outperforms the baseline. This was done by negating the effects of ageing from the topological features in order to obtain a measure of the deviation of brain topology from a healthy population. This difference then acts as a strong indicator for the presence of abnormal neurodegenration linked to Alzheimer's disease.

This report has proven that studying ageing from the lens of topology can bring new understanding, at the very least by providing robust and efficient computational techniques to replace existing methods. The methods and results outlined in this report only scratch the surface of what can be learned by studying the structural topology of the brain. With the rapid development of new theory and methods to compute persistent homology it is clear that there is a vast world of unexplored techniques that can shed new light on the ageing process.

Appendix A

A Formal Introduction to Homology

A.1 Cubical Complexes

The computation of homology requires us to define a topogical structure in which faces and boundaries are clearly defined. Given the nature of the data (3D voxel images) the notion of a cubical complexes lends itself as a natural way to do this. [46]

Definition A.1.1 (Elementary interval)**.** *An elementary interval is an interval of the form* $[n, n+1]$ *or* $[n, n]$ *with* $n \in \mathbb{Z}$*. We say that the latter interval is degenerate.*

Definition A.1.2 (Elementary cube)**.** *An elementary cube is a product of elementary intervals* $C = I_1 \times I_2 \times \ldots \times I_n$*.*

Thus a cube in 3D $[x, x+1] \times [y, y+1] \times [z, z+1]$, represents a voxel in 3D at (x, y, z).

Definition A.1.3 (Boundary of a cube)**.** *We define the boundary,* ∂_n*, of an n-dimensional cube to be the sum* $\partial C = (\partial I_1 \times I_2 \times \ldots I_n) + (I_1 \times \partial I_2 \times \ldots I_n) + \ldots + (I_1 \times I_2 \times \ldots \partial I_n)$*. The boundary of an elementary interval is the sum of it's boundary values* $[n+1, n+1] - [n, n]$ *in the non-degenerate case and* 0 *otherwise.*

Definition A.1.4 (Cubical complex)**.** *A cubical complex* $\mathfrak{C}$ *is a collection of cubes that is closed under taking the boundary i.e. for every cube* $C \in \mathfrak{C}$*,* $\partial(C) \in \mathfrak{C}$*.*

A.2 Homology

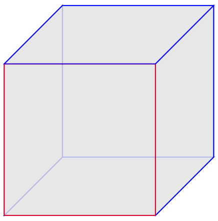
We can now proceed to give the formal definition of homology,

Definition A.2.1 (Chain Complex)**.** *A chain group* C_n *is the free abelian group generated by the set of n-dimensional cubes in the complex. The chain complex of a cubical complex is the sequence of chain groups and boundary operators given by*

$$C_n \xrightarrow{\partial_n} C_{n-1} \xrightarrow{\partial_{n-1}} \ldots \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} 0$$

With the boundary operator extended linearly to the chain groups.

The goal of homology is to identify the holes of the cubical complex. Holes here are defined as closed loops (cycles) that are not boundaries of other objects. For example consider the ring of edges surrounding one face of a cube.



The boundary of this ring is the sum of the boundaries of the edges which all cancel out to give 0. This gives a conveniant definition of a cycle as the subgroup of elements in the chain group that vanish under the boundary operator. Given that the red edges surround a face, they do not produce a hole. However, if the hole was removed, these edges would not bound any cube in our complex and so would surround a hole. This notion can be made rigorous as follows,

Definition A.2.2 (Homology)**.** *It can be shown that the boundary operator is a homomorphism between chain groups. It can also be shown that the $Im(\partial_{n+1}) \subseteq \ker(\partial_n)$, intuitively, all boundaries must be cycles. So, in taking the group of cycles $\ker(\partial_n)$ and factoring out the cycles which act as boundaries to higher dimensional cubes in the complex, $Im(\partial_{n+1})$, we are left with a group, $H_n(\mathfrak{C}) = \ker(\partial_n)/Im(\partial_{n+1})$, containing all the n-dimensional holes of the cubical complex.*

In this paper and in most application of TDA in general, we are note concerned with the internal structure of the homology groups. Usually, it is only necessary to count the number of holes in the data.

Definition A.2.3 (Betti Number)**.** *The nth Betti number β_n is the number of n dimensional holes in homology. $\beta_n = rank(H_n)$.*

In order for homology to be a meaningful measure it needs to be invariant under transformations which preserve topological structure. These kinds of transformations are formalised by the notion of homotopy.

Definition A.2.4 (Homotopy)**.** *Let X, Y be topological spaces, for example these spaces could be cubical complexes. Then, a continuous map $H : X \times [0,1] \to Y$ with $H(x,0) = f(x)$ and $H(x,1) = g(x)$ is a homotopy between f and g. Two spaces are said to be homotopy equivalent if there exist two continuous maps $f : X \to Y$, $g : Y \to X$ such that for each of the compositions $g \circ f$ and $f \circ g$ there exists homotopies to the identity function.*

What this means is that two a homotopy from f to g is a continuous map that changes f to g as we move along the interval $[0,1]$. A homotopy equivalence between X and Y says that we can go from X to Y and back and have something that topological resembles our starting point. One can think of this as Y being similar enough to X in that we can turn the former into the latter by stretching or squishing but without making any cuts or gluing together of different sections.

Appendix B

Detecting Anomalies in Heartbeats

Persistent homology is able to detect and classify anomalies in ECG signals measuring heart beats [12]. It was shown that by including a Betti curve representing the topology of the signal, a neural network was able to achieve close to perfect accuracy in detecting arrhythmia. Inspired by this work and that of [5] which discusses how persistent homology can be used to detect perodicity in signals I attempted to investigate whether an alternative method could be used to detect arrhythmia using a more lightweight and interpretable method. Here I use the MIT BIH dataset [32] provided by [31].

B.1 Topology and Time Series Data

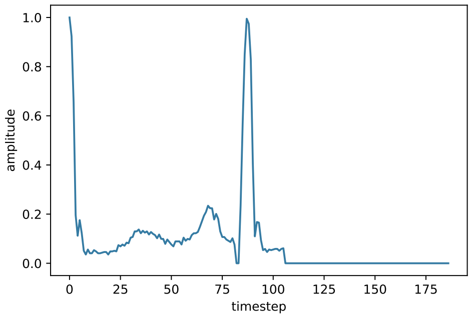
Firstly, let me introduce how persistent homology can be used to study time series data. Consider a 1-dimensional function $f(t)$ over time. The line drawn by this function can be embedded in a d-dimensional space using what is known as sliding window embedding. This embedding is defined as follows

$$TD_{\pi,\tau}(f) = \begin{bmatrix} f(t) \\ f(t+\tau) \\ f(t+2\tau) \\ \vdots \\ f(t+(d-1)\tau) \end{bmatrix}.$$

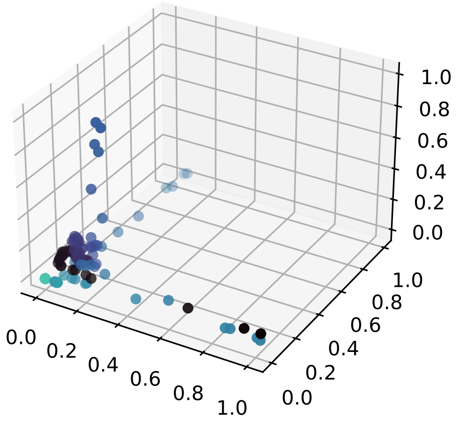
Where τ is the periodicity of the embedding. What is interesting about this embedding is that if a signal contains a period, a repetition of the same pattern at a later point, the embedding forms a closed loop in the embedding space. And so, by computing the homology of the shape produced by the embedding, one can classify the patterns present in the shape.

B.2 Detecting arrhythmia

By the above logic, if one were to compute the embedding of a person with no underlying conditions, one would expect loops in the resulting embedding signfiying a



(a) Raw signal corresponding to a heartbeat.



(b) A three dimensional sliding window embedding.

Figure B.1: A normal heartbeat and its embedding.

steady periodic pattern. An abnormal pattern associated with heart condition leads to erratic heart beats that would give rise to breaks in the periodicity of the signal. The embedding of such a signal would then give rise to no closed loops.

Arrhythmia is the clinical term for a number of various anomalous patterns of heartbeats. Figure B.2 indicates that the homology of the classes should provide sufficient information in the detection and classification of abnormal patterns.

B.3 Results and Discussion

To compute the persistent homology of the embedding I used the Vietoris-Rips filtration to obtain persistence diagrams. These diagrams were then transformed into Betti curves which served as input features to a support vector classifier. The classifier's accuracy was only slightly better than a trivial classifier; labelling all the images as normal achieves an accuracy of 82.7% while the trained classifier model achieves a testing accuracy of 85.0%.

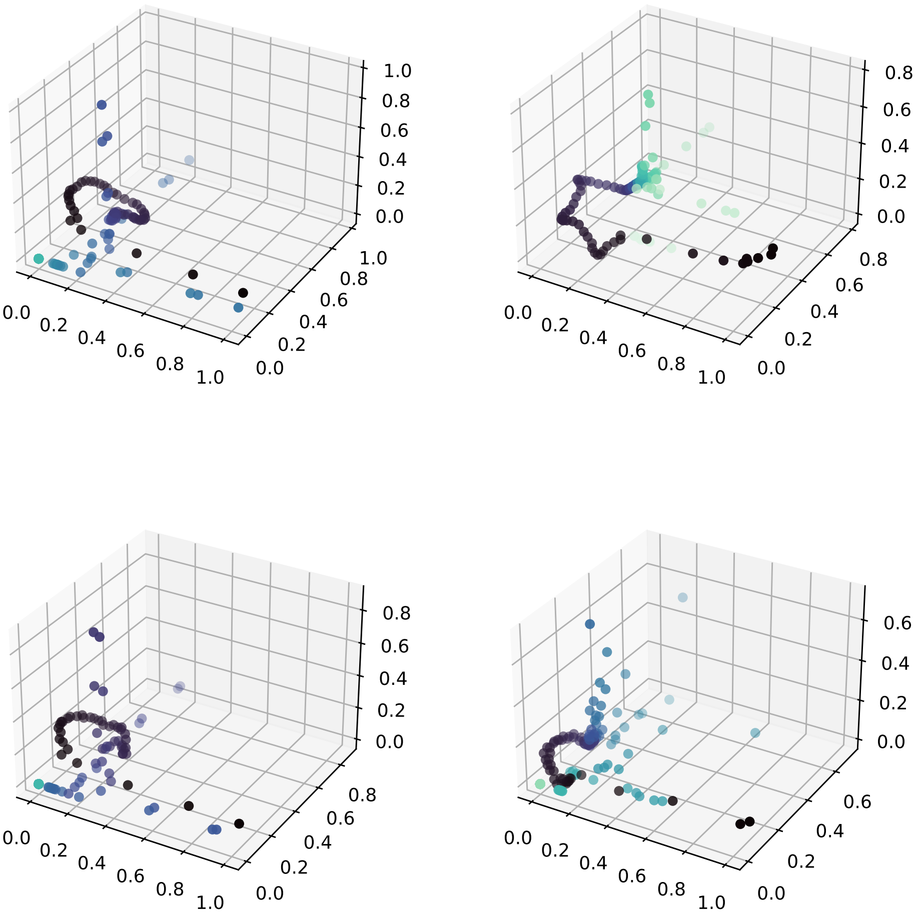


Figure B.2: Example embeddings taken from each of the four arrhythmia classes showing how they are characterised by distinct topological shapes.

There could be a number of reasons due to which the classifier performs badly, one obvious issue is that the preprocessed data divides a sequence of heartbeats from a single patient into one beat slices. In reality, it would be necessary to compute the embedding of a series of beats from a patient and examine whether breaks in periodicity result in topological changes. In order to to properly apply TDA to this problem, one would have to consider the an appropriate way to process the data to achieve the desired results.

Another consideration is that the parameters were chosen fairly arbitrarily. While it seems that topology can give some inidication of arrhythmia, the choices embedding dimension and τ produces vastly different results. Perhaps by choosing these values more carefully it would be possibly to obtain more useful representations of the data.

Appendix C

Other Applications of TDA

Given the breadth of applications and methods used in topological data analysis, I wish to outline some of its most significant contributions to biomedical data analysis. Since the aim of this chapter is to give a broader view of TDA, I present papers which describe approaches to TDA which diverge from those used in this report and provide new perspectives on the field.

C.1 Genomics

Given a set of gene sequencing data from a species over time, one of the main problems in genomics is to deduce the corresponding the phylogenetic network. This network describes how the genetic structure of a species evolves and describes the ancestry of genetic traits. There are two kinds of evolution which contribute to the structure of these networks, vertical and horizontal. Vertical evolution is caused mutations in the genetic information passed from parent to child during asexual reproduction. Many organisms, including bacteria, also have the ability to transmit genetic information laterally to another member of the same species resulting in horizontal evolution.

Most traditional techniques to deduce evolutionary patterns are only able to capture vertical evolution, and so produce branching tree-like data structures. Figure C.1[1] shows that horizontal evolution instead causes distinct branches to merge, making tree based algorithms unsuitable. This is an issue since genetic resistence to drugs in bacteria is often seen to be developed by horizontal evolution instead. Since horizontal evolution gives rise to cycles in the evolutionary network methods that are able to detect these kinds of 'holes' are extremely useful. The idea detailed in [37] is to use persistent homology to detect the presence of cycles i.e. holes in these networks. The notion of persistence is extremely useful here in that it allows geneticists to detect horizontal evolution at various scales. This allows them to answer the question of whether or not horizontal evolution has occured and also how distant the branches were before they merged.

[1] Image taken from https://isabellewinder.com/hybrid-zones-reticulation-and-the-extended-synthesis/

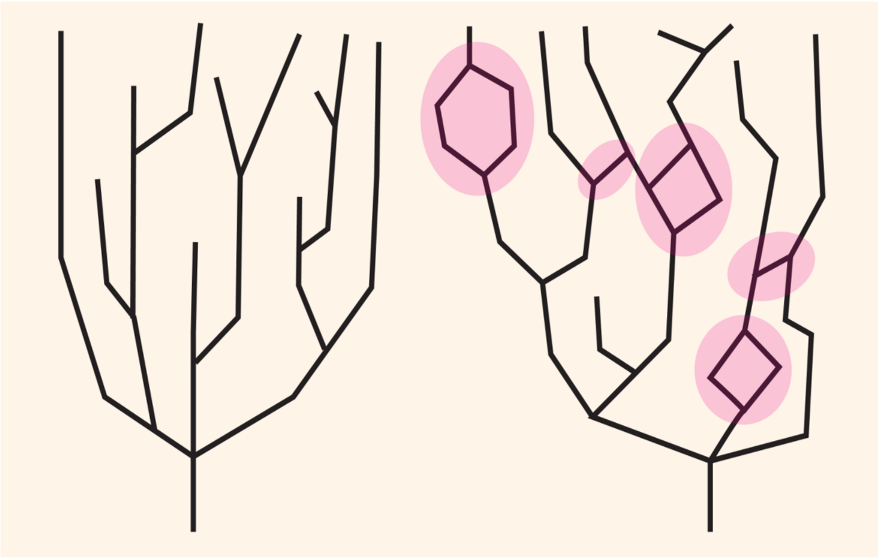


Figure C.1: A phylogenetic tree representing vertical evolution (left) and a network showing horizontal evolution (right).

C.2 Tumor detection

One issue with using persistent homology to derive features from data is that persistence diagrams are not uniquely represented by the data, meaning that some information is lost. This is an issue when attempting to use persistent homology to classify distinct objects since we wish to be able to separate them solely based on their persistence diagrams. There exists however a solution, Euler characteristic transform (ECT). ECT transforms data into a topological representation which preserves all of its information, meaning that you can go back and forth between the two representations. What this means is that the data can be transformed by ECT into a new space equipped with a distance measure which can directly compute how topologically similar two samples are.

The authors of [38] show how ECT can be used to detect Glioblastoma, a form of brain tumor with distinctive topological structure, in MRI images of the brain. The authors show that the ECT has a number of nice properties, it allows for a metric to compute topological similarity between two objects and also structured enough to compute probability distributions. By training a Gaussian regression model on the ECT of Glioblastoma images they find that their method outperforms state of the art techniques. Their method also provides a confidence score which is essential when using these techniques in clinical settings.

C.3 The Mapper algorithm

Biomedical data analysis often makes uses of dimensionality reduction techniques in order to gain new understanding of data. Many of the existing dimensionality reduction techniques make use of topology to some extent in order to preserve the structure of the data in lower dimensions. One technique that has emerged as a direct consequence of persistent homology is known as the Mapper algorithm and leverages the same notion of filtration used to compute persistence. The goal of the mapper algorithm is to produce a graph from a discrete set of data points. This graph can then be embedded in 2- or 3-D space allowing for intuitive data exploration while also clustering data points based on similarity.

Mapper has been used in several breakthrough papers to discover new subgroups of patients with unique properties. By converting high dimensional data into mapper graphs (figure C.2), the authors of [29] where able to first identify coarse grained subgroups of patients. By refining the clustering they found 3 distinct clusters of diabetes patients with distinct clinical properties. These 3 clusters were then identified with unique genes that they found to correspond with the symptoms experienced by patients in the same cluster.

A similar approach was applied to breast cancer patients to understand why certain groups survive the disease compared to others [34]. The researchers again visualised their patient data with the mapper algorithm and found that the graphs produced two distinct chains of nodes. One chain contained patients with 100% survival and where able to identify the genetic sequence that led to this resilience.

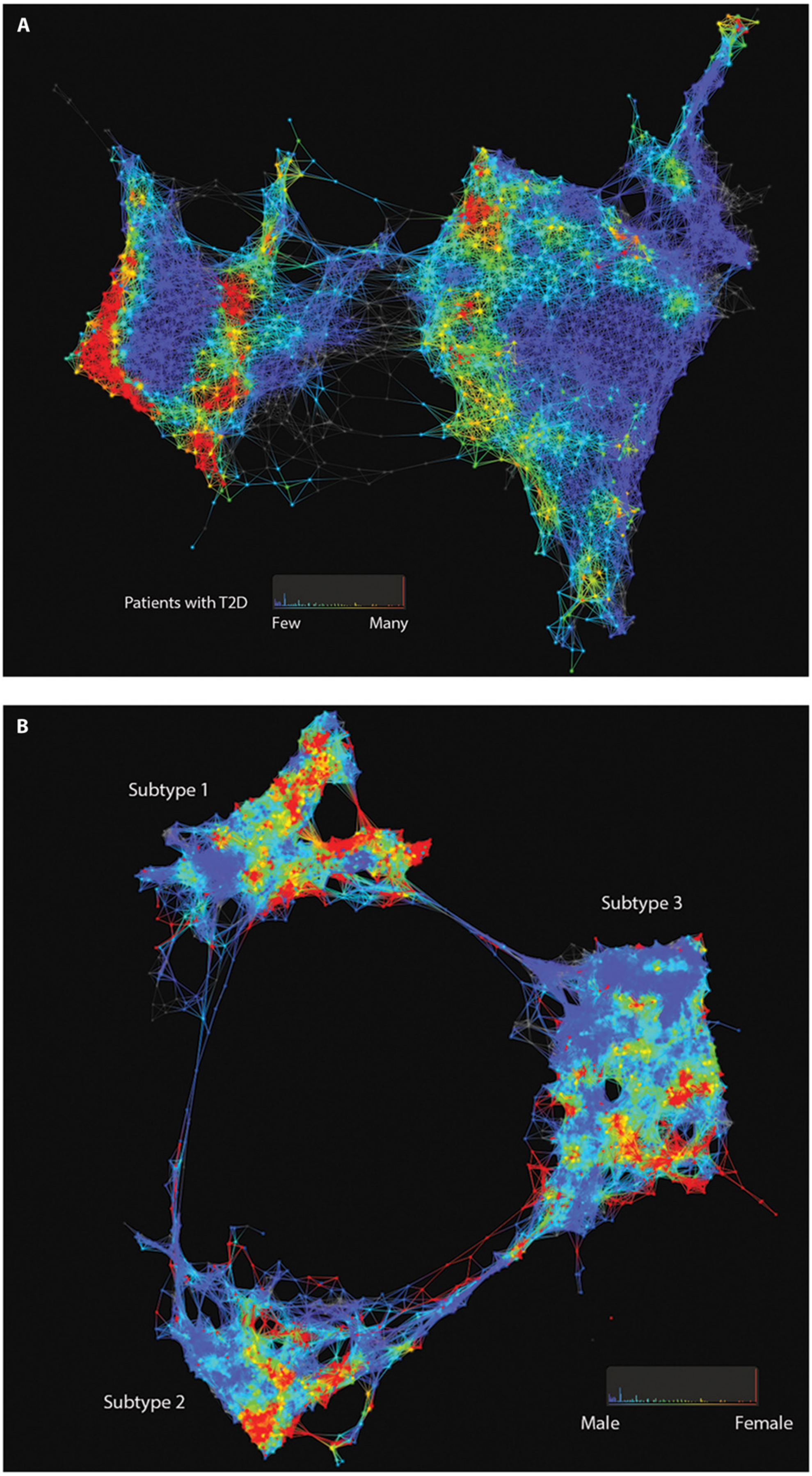


Figure C.2: Example of mapper graphs showing how data can be clustered into distinct groups.

Bibliography

[1] K. L. Anderson, J. S. Anderson, S. Palande, and B. Wang. Topological Data Analysis of Functional MRI Connectivity in Time and Space Domains. ISSN 0302-9743. doi: 10.1007/978-3-030-00755-3_8.

[2] S. Basaia, F. Agosta, L. Wagner, E. Canu, G. Magnani, R. Santangelo, and M. Filippi. Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. 21:101645. ISSN 2213-1582. doi: 10.1016/j.nicl.2018.101645.

[3] P. Bendich, J. S. Marron, E. Miller, A. Pieloch, and S. Skwerer. Persistent Homology Analysis of Brain Artery Trees. 10:198–218, 2016. ISSN 1932-6157. doi: 10.1214/15-aoas886.

[4] G. Biau. Analysis of a Random Forests Model. *J. Mach. Learn. Res.*, 13(null): 1063–1095, Apr. 2012. ISSN 1532-4435.

[5] C. Bresten and J.-H. Jung. Detection of gravitational waves using topological data analysis and convolutional neural network: An improved approach, 2019.

[6] M. Brett, C. J. Markiewicz, M. Hanke, M.-A. Côté, B. Cipollini, P. McCarthy, D. Jarecka, C. P. Cheng, Y. O. Halchenko, M. Cottaar, E. Larson, S. Ghosh, D. Wassermann, S. Gerhard, G. R. Lee, H.-T. Wang, E. Kastman, J. Kaczmarzyk, R. Guidotti, O. Duek, J. Daniel, A. Rokem, C. Madison, B. Moloney, F. C. Morency, M. Goncalves, R. Markello, C. Riddell, C. Burns, J. Millman, A. Gramfort, J. Leppäkangas, A. Sólon, J. J. van den Bosch, R. D. Vincent, H. Braun, K. Subramaniam, K. J. Gorgolewski, P. R. Raamana, J. Klug, B. N. Nichols, E. M. Baker, S. Hayashi, B. Pinsard, C. Haselgrove, M. Hymers, O. Esteban, S. Koudoro, F. Pérez-García, N. N. Oosterhof, B. Amirbekian, I. Nimmo-Smith, L. Nguyen, S. Reddigari, S. St-Jean, E. Panfilov, E. Garyfallidis, G. Varoquaux, J. H. Legarreta, K. S. Hahn, O. P. Hinds, B. Fauber, J.-B. Poline, J. Stutters, K. Jordan, M. Cieslak, M. E. Moreno, V. Haenel, Y. Schwartz, Z. Baratz, B. C. Darwin, B. Thirion, C. Gauthier, D. Papadopoulos Orfanos, I. Solovey, I. Gonzalez, J. Palasubramaniam, J. Lecher, K. Leinweber, K. Raktivan, M. Calábková, P. Fischer, P. Gervais, S. Gadde, T. Ballinger, T. Roos, V. R. Reddam, and freec84. nipy/nibabel: 3.2.1, Nov. 2020. URL `https://doi.org/10.5281/zenodo.4295521`.

[7] P. Bubenik, M. Hull, D. Patel, and B. Whittle. Persistent homology detects cur-

vature. *CoRR*, abs/1905.13196, 2019. URL http://arxiv.org/abs/1905.13196.

[8] N. Byrne, J. R. Clough, G. Montana, and A. P. King. A Persistent Homology-based Topological Loss Function for Multi-class CNN Segmentation of Cardiac MRI. 2020.

[9] G. Carlsson. Topology and Data. 2009.

[10] J. H. Cole, R. P. Poudel, D. Tsagkrasoulis, M. W. Caan, C. Steves, T. D. Spector, and G. Montana. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker.

[11] J. Cuenca Jiménez. Persistent homology for defect detection in non-destructive evaluation of materials. *The e-Journal of Nondestructive Testing*, 21, 01 2016.

[12] M. Dindin, Y. Umeda, and F. Chazal. Topological data analysis for arrhythmia detection through modular neural networks. *CoRR*, abs/1906.05795, 2019. URL http://arxiv.org/abs/1906.05795.

[13] X. Feng, Z. C. Lipton, J. Yang, S. A. Small, and F. A. Provenzano. Estimating brain age based on a healthy population with deep learning and structural mri.

[14] B. Fischl. Freesurfer. 62:774–781, 2012. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2012.01.021. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3685476/.

[15] B. Fischl and A. Dale. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proc Natl Acad Sci*, 97:11050–11055, 2000. ISSN 0027-8424. doi: 10.1162/jocn.2007.19.9.1498.

[16] B. Fischl, D. H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. van der Kouwe, R. Killiany, D. Kennedy, S. Klaveness, A. Montillo, N. Makris, B. Rosen, and A. M. Dale. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. 33:341–355, 2002. ISSN 0896-6273. doi: 10.1016/s0896-6273(02)00569-x.

[17] A. M. Fjell, L. T. Westlye, I. Amlien, T. Espeseth, I. Reinvang, N. Raz, I. Agartz, D. H. Salat, D. N. Greve, B. Fischl, A. M. Dale, and K. B. Walhovd. High consistency of regional cortical thinning in aging across multiple samples. *Cerebral Cortex*, 19(9):2001–2012, jan 2009. doi: 10.1093/cercor/bhn232.

[18] G. B. Frisoni, N. C. Fox, C. R. Jack, P. Scheltens, and P. M. Thompson. The Clinical Use of Structural MRI in Alzheimer Disease. 6(2):67–77. doi: 10.1038/nrneurol.2009.215.

[19] A. Garin and G. Tauzin. A Topological "Reading" Lesson: Classification of MNIST using TDA.

[20] Z. Gracia-Tabuenca, J. C. Díaz-Patiño, I. Arelio, and S. Alcauter. Topological Data Analysis Reveals Robust Alterations in the Whole-Brain and Frontal Lobe Functional Connectomes in Attention-Deficit/Hyperactivity Disorder. 7: ENEURO.0543–19.2020. ISSN 2373-2822. doi: 10.1523/eneuro.0543-19.2020.

[21] Y. Gupta, K. H. Lee, K. Y. Choi, J. J. Lee, B. C. Kim, G. R. Kwon, and and. Early diagnosis of alzheimer's disease using combined features from voxel-based morphometry and cortical, subcortical, and hippocampus regions of mri t1 brain images. 14:e0222446, 2019. ISSN 1932-6203. doi: 10.1371/journal.pone.0222446.

[22] R. C. Gur, P. D. Mozley, S. M. Resnick, G. L. Gottlieb, M. Kohn, R. Zimmerman, G. Herman, S. Atlas, R. Grossman, and D. Berretta. Gender differences in age effect on brain atrophy measured by magnetic resonance imaging. *Proc Natl Acad Sci U S A*, 88(7):2845–2849, Apr 1991.

[23] A. Hatcher. *Algebraic Topology*. Cambridge University Press. ISBN 9780521791601.

[24] C. Hutton, B. Draganski, J. Ashburner, and N. Weiskopf. A comparison between voxel-based cortical thickness and voxel-based morphometry in normal aging. *NeuroImage*, 48(2):371–380, nov 2009. doi: 10.1016/j.neuroimage.2009.06.043.

[25] B. A. Jonsson, G. Bjornsdottir, T. E. Thorgeirsson, L. M. Ellingsen, G. B. Walters, D. F. Gudbjartsson, H. Stefansson, K. Stefansson, and M. O. Ulfarsson. Brain age prediction using deep learning uncovers associated sequence variants. *Nature Communications*, 10(1), nov 2019. doi: https://doi.org/10.1038/s41467-019-13163-9.

[26] S. Korolev, A. Safiullin, M. Belyaev, and Y. Dodonova. Residual and plain convolutional neural networks for 3d brain mri classification, 2017.

[27] P. J. LaMontagne, T. L. S. Benzinger, J. C. Morris, S. Keefe, R. Hornbeck, C. Xiong, E. Grant, J. Hassenstab, K. Moulder, A. G. Vlassenko, M. E. Raichle, C. Cruchaga, and D. Marcus. OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease. doi: 10.1101/2019.12.13.19014902.

[28] J. P. Lerch, J. Pruessner, A. P. Zijdenbos, D. L. Collins, S. J. Teipel, H. Hampel, and A. C. Evans. Automated cortical thickness measurements from mri can accurately separate alzheimer's patients from normal elderly controls. 29:23–30, 2006. ISSN 0197-4580. doi: 10.1007/978-3-322-89521-9_13.

[29] L. Li, W. Y. Cheng, B. S. Glicksberg, O. Gottesman, R. Tamler, R. Chen, E. P. Bottinger, and J. T. Dudley. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci Transl Med*, 7(311):311ra174, Oct 2015.

[30] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner. Open access series of imaging studies (oasis): Cross-sectional mri data in young, middle aged, nondemented, and demented older adults. 19:1498–1507, 2007. ISSN 0898-929X. doi: 10.1162/jocn.2007.19.9.1498.

[31] G. B. Moody and R. G. Mark. Mit-bih arrhythmia database. doi: 10.1109/cic.1990.144205.

[32] G. B. Moody, R. G. Mark, and A. L. Goldberger. Physionet: a research resource for studies of complex physiologic and biomedical signals.

[33] S. M. Nestor, R. Rupsingh, M. Borrie, M. Smith, V. Accomazzi, J. L. Wells, J. Fogarty, and R. B. and. Ventricular enlargement as a possible measure of alzheimer's disease progression validated using the alzheimer's disease neuroimaging initiative database. *Brain*, 131:2443–2454, 2008. ISSN 0006-8950. doi: 10.1093/brain/awn146.

[34] M. Nicolau, A. J. Levine, and G. Carlsson. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proc Natl Acad Sci U S A*, 108(17):7265–7270, Apr 2011.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[36] D. Purves, G. J. Augustine, D. Fitzpatrick, W. C. Hall, A.-S. LaMantia, and L. E. White, editors. *Neuroscience*, volume 424. Sinauer, Sunderland, Mass., fifth edition.. edition, 2012. ISBN 9780878936953. doi: 10.1016/j.neuroscience.2019.12.012.

[37] R. Rabadán and A. J. Blumberg. *Topological Data Analysis for Genomics and Evolution*. Number field,. doi: 10.1017/9781316671665.

[38] M. Rucco, L. Falsetti, and G. Viticchi. Towards personalized diagnosis of glioblastoma in fluid-attenuated inversion recovery (flair) by topological interpretable machine learning, 2020.

[39] P. Scheltens, D. Leys, F. Barkhof, D. Huglo, H. C. Weinstein, P. Vermersch, M. Kuiper, M. Steinling, E. C. Wolters, and J. Valk. Atrophy of Medial Temporal Lobes on Mri in "probable" Alzheimer's Disease and Normal Ageing: Diagnostic Value and Neuropsychological Correlates. 55(10):967–972. ISSN 0022-3050. doi: 10.1136/jnnp.55.10.967.

[40] J. Soares, B. Cao, B. Mwangi, I. Passos, M.-J. Wu, Z. Keser, G. Zunta-Soares, D. Xu, and K. Hasan. Lifespan gyrification trajectories of human brain in healthy individuals and patients with major psychiatric disorders. 81:S233, 2017. ISSN 0006-3223. doi: 10.1007/978-3-322-89521-9_13.

[41] Y. Taki, R. Goto, A. Evans, A. Zijdenbos, P. Neelin, J. Lerch, K. Sato, S. Ono, S. Kinomura, M. Nakagawa, M. Sugiura, J. Watanabe, R. Kawashima, and H. Fukuda. Voxel-based morphometry of human brain with age and cerebrovascular risk factors. *Neurobiology of Aging*, 25(4):455–463, apr 2004. doi: 10.1016/j.neurobiolaging.2003.09.002.

[42] G. Tauzin, U. Lupo, L. Tunstall, J. B. Pérez, M. Caorsi, A. Medina-Mardones, A. Dassatti, and K. Hess. giotto-tda: A topological data analysis toolkit for machine learning and data exploration, 2020.

[43] K. B. Walhovd, A. M. Fjell, I. Reinvang, A. Lundervold, A. M. Dale, D. E. Eilertsen, B. T. Quinn, D. Salat, N. Makris, and B. Fischl. Effects of age on

volumes of cortex, white matter and subcortical structures. 26:1261–1270, 2005. ISSN 0197-4580. doi: 10.1016/j.neurobiolaging.2005.05.020.

[44] J. Wen, E. Thibeau-Sutre, M. Diaz-Melo, J. Samper-González, A. Routier, S. Bottani, D. Dormont, S. Durrleman, N. Burgos, and O. Colliot. Convolutional neural networks for classification ofalzheimer's disease: Overview and reproducible evaluation. 63:101694, 2020. ISSN 1361-8415. doi: 10.1016/j.media.2020.101694.

[45] K. Zilles, E. Armstrong, A. Schleicher, and H.-J. Kretschmann. The human pattern of gyrification in the cerebral cortex. 179:173–179, 1988. ISSN 0340-2061. doi: 10.1007/bf00304699.

[46] D. Ziou and M. Allili. Generating cubical complexes from image data and computation of the euler number. *Pattern Recognition*, 35(12):2833 – 2839, 2002. ISSN 0031-3203. doi: https://doi.org/10.1016/S0031-3203(01)00238-2. URL `http://www.sciencedirect.com/science/article/pii/S0031320301002382`. Pattern Recognition in Information Systems.