UNIVERSIDADE D
COIMBRA

Daniel Andrade Pinho Agostinho

Differential diagnosis of Alzheimer's Disease based on multimodal imaging data (MRI, PIB and DTI)

Dissertation to obtain a Master's degree in Biomedical Engineering,
supervised by Prof. Francisco José Santiago Fernandes Amado Caramelo, PhD, and Prof. Miguel Sá Sousa Castelo Branco, PhD,
and presented to the Physics Department of Faculty of Sciences and Technology of the University of Coimbra

September 2019

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D COIMBRA

Differential diagnosis of Alzheimer's disease based on multimodal imaging data (MRI, PIB, DTI)

Dissertation submitted to obtain the degree of Master in Biomedical Engineering in the Specialty of Neurosciences.

Diagnóstico diferencial da doença de Alzheimer com base em dados de imagem multimodal (MRI, PET e DTI)

Author

Daniel Andrade Pinho Agostinho

Supervisors

Prof. Francisco José Santiago Fernandes Amado Caramelo, PhD

Prof. Miguel Sá Sousa Castelo Branco, PhD

Jury

President	Prof. Joel Perdiz Arrais, PhD
Supervisor	Prof. Francisco José Santiago Fernandes Amado Caramelo, PhD
Expert	João Valente Duarte, PhD

Colaboração Institucional

Laboratório de Bioestatística e Informática Médica- FMUC.

Instituto de Ciências Nucleares Aplicadas à Saúde.

Coimbra, September, 2019

“Learning never exhausts the mind.”

-Leonardo da Vinci

Agradecimentos

O alcançar desta etapa não teria sido possível sem a colaboração, auxílio, carinho, dedicação e sacrifícios por parte de várias pessoas ao longo de todo o meu percurso académico. Por esta mesma razão, não quero deixar passar esta oportunidade para agradecer a todos aqueles que, direta ou indiretamente, contribuíram para o meu sucesso e a minha chegada até aqui.

Em primeiro lugar gostaria de agradecer aos meus pais e irmão, estes que são os pilares principais da minha vida e que mais sacrificaram para tornar possível não só o meu ingresso no ensino superior, mas como a sua conclusão. Sem eles, e sem o seu sacrifício, nada disto seria possível e gostava de dedicar este trabalho especialmente a eles.

Gostava de deixar um carinhoso agradecimento também aos meus amigos que me acompanharam nestes últimos 5 anos. Um obrigado por toda a ajuda a nível académico, mas uma ovação por todos os momentos partilhados num contexto extra académico que permitiram o meu desenvolvimento como pessoa. O meu de hoje só existe graças à vossa contribuição.

Não poderia também deixar de agradecer aos meus orientadores por me terem acompanhado no desenvolvimento e escrita deste trabalho e por toda a sua disponibilidade para me ajudar a qualquer momento.

Por fim, um especial agradecimento à minha afilhada Joana que me ajudou imenso na escrita com melhor clareza deste trabalho. Sem ela, a leitura deste trabalho seria um processo bem mais complicado.

Resumo

A doença de Alzheimer (DA) é a mais comum forma de demência humana e, de momento, não possui um critério de diagnóstico bem definido, embora, neste contexto, se dê ênfase ao uso de biomarcadores, provenientes de neuroimagens. Técnicas de imagem como a ressonância magnética (RM), tomografia de emissão de positrões e imagem de tensor de difusão são usadas, quer separadamente, quer numa abordagem multimodal, no estudo e na classificação da DA. As abordagens multimodais concentram-se maioritariamente no uso de biomarcadores baseados na ressonância magnética, que são posteriormente combinados com outro tipo de neuroimagem ou dados biológicos.

Nesta tese, propomos analisar os efeitos da combinação dos dados provenientes das 3 modalidades de imagem, bem como estudar a informação complementar que é dada através da combinação de cada modalidade. Para alcançar este objetivo começámos por criar classificadores base, um para cada modalidade de imagem, e depois examinamos os efeitos da sua combinação, usando técnicas de ensemble.

Os resultados obtidos mostram que a combinação das 3 modalidades de imagem melhora a performance geral dos classificadores base (exatidão-98%, sensibilidade-99%, especificidade-97%), mas não apresentam uma melhoria significativa em relação ao uso da combinação de apenas MRI+PIB (exatidão-98%, sensibilidade-99%, especificidade -98%) ou MRI+DTI (exatidão-97%, sensibilidade-94%, especificidade-99%). Mais ainda, a combinação de PIB+DTI (exatidão-91%, sensibilidade-93%, especificidade-90%) não mostrou qualquer melhoria em relação aos classificadores base, o que sugere uma falta de informação complementar entre estas duas modalidades de imagem.

Estas descobertas podem representar benefícios clínicos não apenas para as instituições, reduzindo custos, mas também para o bem-estar do paciente, reduzindo o desconforto causado pelo longo tempo de aquisição de imagens PET e removendo a necessidade de exposição a radiação ionizante.

Palavras-chave: Doença de Alzheimer (DA), Abordagem Multimodal, Machine Learning, Classificação, Diagnostico, Ensemble Learning.

Abstract

Alzheimer's disease (AD) is the most common form of dementia in humans and currently it does not have a defined diagnostic criterion, although, nowadays, it emphasizes the use of neuroimaging biomarkers. Magnetic resonance imaging (MRI), positron emission tomography (PET) and diffusion tensor imaging (DTI) are used, either alone or in multimodal approaches, in the study and classification of AD. Today's multimodal approaches focus on the use of MRI related biomarkers, as a base, and then combining them with other type of imaging or biological data.

Here we propose to analyse the effects of the combination of the data from the three imaging modalities, as well as study the complementary information, provided from the combination of each modality. To achieve this goal, we start by creating base classifiers, one for each different imaging modality, and then examine the combination effects, using ensemble techniques.

The results show that the combination of all three imaging modalities improves the general performance of the base classifiers (accuracy-98%, sensitivity-99%, specificity-97%), however, it did not show a significant improvement over the use of the combination of just MRI+PIB (accuracy-98%, sensitivity-99%, specificity-98%) or MRI+DTI (accuracy-97%, sensitivity-94%, specificity-99%). Furthermore, the combination of PIB+DTI (accuracy-91%, sensitivity-93%, specificity-90%) did not show any improvement over the base classifiers, suggesting a lack of complementary information between the two imaging modalities.

These findings could represent clinical benefits not only for the institutions, by reducing costs, but also for the patient's wellbeing by reducing the discomfort caused by the lengthy acquisition time of the PET scans and by removing the need of exposure to ionizing radiation.

Keywords Alzheimer's Disease (AD), Multimodal Approach, Machine Learning, Classification, Diagnosis, Ensemble Learning

Index

FIGURE INDEX

TABLE INDEX

TERMS AND ABBREVIATIONS

AD – Alzheimer's Disease

CN – Cognitively Normal

MRI – Magnetic Resonance Imaging

PET – Positron Emission Tomography

PIB – Pittsburgh Compound-B

DTI – Diffusion Tensor Imaging

MCI – Mild Cognitive Impairment

NFTs – Neurofibrillary Tangles

Aβ – Amyloid Beta

PS – Presenilin

APP – Amyloid Precursor Protein

GM – Grey Matter

WM – White Matter

CSF – Cerebrospinal Fluid

ML – Machine Learning

SVM – Support Vector Machine

LR – Logistic Regression

ROI – Region of Interest

SUVR – Standard Uptake Value Ratio

FBM – Filter Based Method

EBM – Embedded Based Method

1. INTRODUCTION

The past few years prompted a significant improvement of living conditions and thus a substantial extension of human lifespan, which ultimately propelled the increasing occurrence of neurodegenerative diseases[1].

Alzheimer's disease (AD) is one of these conditions [1–3].

Considered to be the most prevalent form of human dementia, AD affects the entirety of the brain, hence, its primary characteristic is the progressive loss of memory, more noticeable concerning episodic memories which are formed in the hippocampus[1,2].

To aid patients suffering from Alzheimer's disease, it is crucial to mitigate its progression, which is primarily attained by providing an early and definitive diagnosis of the condition [3,4].

In this context, the use of medical imaging, such as Positron Emission Tomography (PET), Magnetic resonance (MRI) and Diffusion Tensor (DTI), has been of utmost importance in AD's diagnosis and study, considering that it allows technicians and neurologists to identify lesions in the brain, either by analysing MR images or by seeing, *in-vivo*, radiotracer labelling of neurofibril plaques of beta-amyloid, through PET images[1,2].

However, despite the significant help provided by these imaging techniques, a definitive diagnosis is only possible post-mortem[2,5,6].

This restriction ultimately led to advances in the field of computer science, which included the creation of new and more precise tools, capable of diagnosing AD, even in its prodromal state.

Within these successful advances we come upon machine learning, frequently used in the generation of classifiers that are able to distinguish between normal patients and patients with AD [7,8]. In turn, the classifiers can be obtained by using only one type of neuroimaging (PET or MRI), or using a multimodal approach, that combines two different types of neuroimaging (PET combined with MRI) [9].

Undoubtedly, the use of classifiers in AD has been incredibly remarkable and it has provided numerous of acceptable and precise results in the diagnosis of the condition, but science is in a constant development[9]. Therefore, to improve the accuracy of the obtained results even further, data from other methods can be added: DTI, Diffusion Tensor Imaging.

However, this situation brings up other issues that should be carefully considered, such as the high magnitude of correlations within the collected data, leading to excessive interdependence, as well as the debatable application of the three classifiers in AD diagnosis. Thus, it is essential to study and analyse if the combination of the methods does, in fact, improve the accuracy of the diagnosis, or if it is overall unnecessary, and the risk of exposing patients to ionizing radiation (in the case of PET imaging) and put them through extreme discomfort is not worth it.

The fundamental ambition of this thesis is to investigate the use of three different classifiers, based on three types of imaging modalities, and in due time, to evaluate and compare each of their performances.

Moreover, a new classifier ensemble will be created, through a multimodal approach combining all of the classifiers, and later assessed, to estimate the significance of the improvement in the AD diagnosis, as well as the complementarity between the different data.

It is arguable that this improvement might not add much significance in the diagnosis. The question remains whether the complementarity between the data is an advantageous tool. These issues are relevant to assess whether diagnosis can be achieved while sparing participants from discomfort and increased number of unnecessary examinations.

Lastly, this thesis will be organized into 5 chapters. In Chapter 2 we will present a review of the current literature about the thesis theme. Chapter 3 explains the general workflow of the study. Chapter 4 shows the results from the study. Finally, Chapter 5 exposes the main conclusions drawn from the study. Furthermore, an Appendix Chapter is also included, and all the complementary data can be found there.

2. LITERATURE REVIEW

Literature review was done following a systematic revision approach. Firstly, we start by searching the PubMed library (https://www.ncbi.nlm.nih.gov/pubmed/) for all relevant publication with the following search phrase:

(("Alzheimer Disease"[Mesh]) AND ("Diagnostic Imaging"[Mesh] OR "Positron-Emission Tomography"[Mesh] OR "Magnetic Resonance Imaging"[Mesh] OR “Diffusion Tensor Imaging” OR "Multimodal Imaging"[Mesh]) AND ("Machine Learning"[Mesh] OR “Multi-Kernel learning” OR "Classification"[Mesh]) AND ("Diagnosis"[Mesh]))

The search results show 168 papers from which 134 were excluded due to a variety of reasons:

- 21 were not classification problems;
- 12 were classification between Alzheimer’s disease (AD) and other form of dementia;
- 11 use deep learning classification;
- 41 use a different PET radiotracer or use functional MRI;
- 20 were about the ability to predict the conversion from mild cognitive impairment (MCI) to AD;
- 16 were not open access;
- 1 was a study using mice;
- 12 were not related.

After this selection, 39 papers or books were further added to the pool for better understanding the various topics.

2.1. Alzheimer's disease

The fear of losing oneself and one's own memories is constant these days, mainly for the elderly.

Since the turning of the century, the human lifespan has considerably increased, and people are living longer and with more quality than ever before[1]. And although this longevity allows people to have a happier and more fulfilling life, it also brings cells closer to their biological limit, causing an emergence of new kinds of complications and age-related diseases.

An example of such conditions is dementia, a condition that has been affecting the older population increasingly every year. Dementia is described as *"an acquired global impairment of intellect, memory and personality, but without impairment of consciousness"*[10] and it is classified as an organic mental disorder, in which an underlying causative organic factor is always present[10].

The most common organic origin that results in a state of dementia is Alzheimer's disease (AD)[1–3].

Alzheimer's disease is a neurodegenerative disease characterized by the appearance of two distinct types of lesions in the brain that culminate in a state of dementia, and ultimately lead to the patient's death. These lesions are known as the neurofibrillary tangles (NFTs), composed of tau protein that aggregates in the intracellular space of the neurons, and the neuritic amyloid plaques (or senile plaques), composed of abnormal amyloid beta (Aβ) protein that aggregates on the extracellular space of the neurons[1].

2.1.1. Origin and causes

Genetic evidence indicates that the inheritance of mutations in several genes causes autosomal dominant familial AD (fAD), while the presence of certain alleles of ApoE4 carries a significant risk of putative sporadic disease[1]

Some genetic risk factors for AD include mutations in different chromosomes: on the chromosome 1, a mutation in presenilin (PS)2; on the chromosome 14, a mutation in PS1; on the chromosome 19, different susceptibility alleles of ApoE, and on the chromosome 21, a mutation in the amyloid precursor protein (APP) gene[1,6,11].

The fAD form is highly dependent on genetics and only accounts for 5% of all AD cases[11]. This form is related to the mutation on the PS1, PS2 and APP genes that can in some cases, be inherited as autosomal dominant starting to manifest itself earlier than it would in sporadic cases (<65 years of age)[11].

The PS1 and PS2 genes are highly homologous and conserved multi-pass transmembrane proteins, involved in Notch1 signal pathways, which are critical in the determination of the cell's fate[1]. Also, a mutation in these genes causes the increased production of abnormal Aβ protein[11]. Approximately 50% of early-onset fAD is linked to 90 mutations in PS1 and there are also data suggesting that a small number of PS2 mutations can cause dominant fAD[1].

The APP gene is a type I transmembrane protein, which can be found in different isoforms. It exists abundantly in the nervous system and it is transported anterograde, from the cell's body to the terminals, as stated earlier. Although its specific functions remain to be defined, it is known that some APP mutations are linked to the fact that fAD can change the processing of APP, as well as influence the biology of Aβ, by increasing toxic peptides, which ultimately results in the formation of the senile plaques[1,11].

The sporadic form of the disease occurs in 95% of all the cases of AD. This form is associated with the ApoE gene, though 60% of the sporadic cases are not related to the ApoE gene, leaving a vast number of AD cases without a known cause[11].

ApoE is a lipid transporter present in the blood that binds amyloid proteins. This lipoprotein exists in three possible variations in humans: ApoE2, ApoE3 and ApoE4. The ApoE3 allele is the most common, existing approximately in 78% of the population, but it is the ApoE4 allele who has a major influence in AD[2,11].

This allele exists in 14% of the population, and, according to clinic-based studies, its frequency in patients with late-onset (>65 years of age) is 50%, suggesting that its presence increases the risk of AD[1].

2.1.2. Clinical manifestations and Diagnosis

Clinically, this disease manifests itself most often as a progressive memory loss and cognitive impairment, as a consequence of the dysfunction and death of specific populations of neurons, mainly those involved in processes of memory and cognition, even though a generalized atrophy of the whole brain can also be seen[1]. This loss of neurons results from the slow accumulation of lesions within the brain over a long period of time[5].

Initially, the disease often manifests itself as a clinical entity, known as mild cognitive impairment (MCI), usually characterized by memory complaint and impairments on formal testing, with otherwise intact general cognition, preserved daily activities, as well as an absence of overt dementia[1].

MCI is considered by many as a transitional state between normal aging and early AD. After a period of time, the early stage of MCI either evolves into AD, accompanied by the appearance of the symptoms mentioned earlier, or it can subside and go back to a normal aging state[1].

Currently, there are no specific diagnostic laboratory tests for AD, given that this diagnosis is not an exact method (no perfect standard), except neuropathology after the death of the patient, and it can only be made in a range of probability. Despite this, there are a variety of tests that may enhance diagnostic accuracy[12].

The current diagnostic criteria for AD are positron emission tomography (PET), magnetic resonance imaging (MRI) and the examination of body fluids (eg., CSF) and non-neural tissues[1,12,13].

The use of PET with radiolabelled Pittsburgh Compound B (PIB) often shows AB amyloid burden in vivo. This happens because once the PIB compound binds, with high affinity, to Aβ, it creates patterns that are reflective of the Aβ burden in the brain. Patients with AD show a retention of the radiotracer in areas of the brain in which there is an accumulation of Aβ amyloid (insoluble aggregates)[1].

MRI often discloses regional brain atrophy, particularly in the hippocampus and entorhinal cortex, also providing rates of atrophy which ultimately correlate with the decline of the patient's clinical status. This technique provides consistent data and may have predictive value for the diagnosis[1,12].

Regarding the examination of body fluids as well as of non-neural tissues, patients with AD usually have low levels of Aβ peptides and, sometimes, increased levels

of tau protein in the cerebral spinal fluid (CSF), when compared with cognitively normals. However, these values have high variability between patients, which makes them poor for diagnosis[1,12].

Summarizing, the clinical diagnosis of AD, in its sporadic or familial form, includes the use of biological biomarkers, such as the analysis of CSF, and imaging biomarkers, such as MRI and PET.

2.2. Neuroimaging

Over the past decades, neuroimaging techniques, such as magnetic resonance imaging (MRI), functional MRI (fMRI), positron emission tomography (PET) and diffusion tensor imaging (DTI), have made a profound influence in the advancement of neuroscientific research and clinical application[9].

The fundamental aim of neuroimaging is to discover information about a yet poorly understood disease as well as to provide a set of well-defined biomarkers, capable of making an early diagnosis[14–16].

These imaging biomarkers are being effectively utilized in both diagnosis and prognosis of various neurologic diseases, given their inherent capability of visualization and quantitative measurement of brain structural, functional and metabolic information.

Therefore, it is imperative to develop and improve tools as powerful as these neuroimaging techniques, that may help not only in diagnosing, but also in monitoring, *in vivo*, the disease progression, in the evaluation and the development of novel therapeutic approaches guided by their use as accurate biomarkers[17].

As stated in the previous chapter, the criteria for diagnosing Alzheimer's disease (AD) emphasizes the use of neuroimaging biomarkers[18]. Today, it is possible, using neuroimaging tools, to capture and visualize some of the underlying pathologic traits of AD, *in vivo*. The data extracted from these tools provides a better understanding of the structural, functional and metabolic changes that occur in an AD patient's brain[13,19].

Although neuroimaging is used in the diagnosis of AD, the inherited non-focal disease pathology is able to generate different patterns of structural degeneration of brain tissue, thus making it challenging to identify the most relevant imaging markers[13,19] .

The best and more commonly used neuroimaging method to evaluate structural damage of brain tissue is MRI.

MRI is a non-invasive technique that does not require the use of ionizing radiation, has outstanding contrast between grey and white matter and also an excellent spatial resolution[13], that depending on the acquisition protocol can go as low as 1mm[14,20,21]. Due to these traits and with the use of T1-weighted high-resolution structural images, it is observed that when neuron population loss accures, there is a global atrophy of both grey and white matter tissues as well as an enlargement of the ventricles[13] (Figure 2.1).

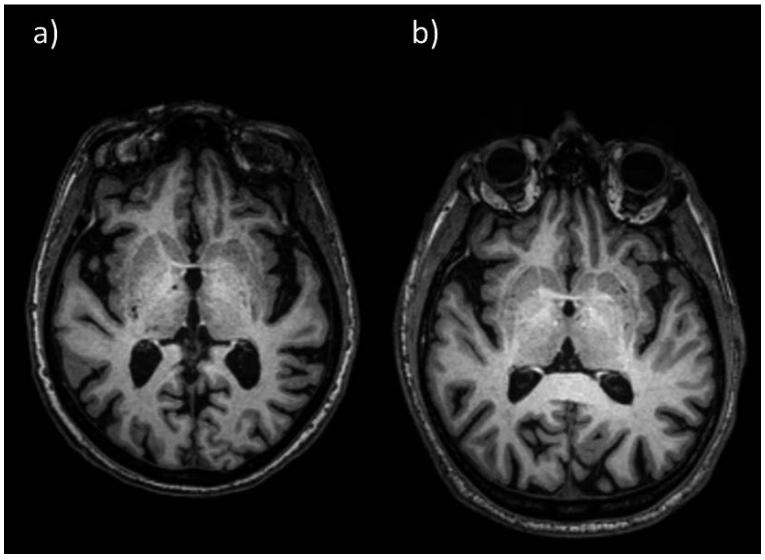

Figure 2.1 Example of the axial view of MRI scan. a) Brain with Alzheimer's. b) Normal aging brain.

MRI images can also provide quantitative features, based on brain morphometry and cortical thickness measures, used in AD/CN classification, thus providing biomarkers capable of assisting in the diagnosis[9,14–16,22,23].

There is a broad spectrum of MRI contrast mechanisms, in addition to T1-weighted imaging, which allows the extraction of different structural or functional biomarkers.

One of the most important contrast mechanisms is the measurement of signal attenuation, resulting from water diffusion. Diffusion tensor imaging (DTI) is able to map and characterize the three-dimensional diffusion of water. The voxels of a DTI image are diffusion tensors and each one of these tensors is a [3x3] matrix (1) that correspond to the diffusion rates in each combination of directions[24], for example, the D_{xx} element of the diffusion tensor represents the diffusion along the x axis.

$$D = \begin{matrix} D_{xx} & D_{xy} & D_{xz} \\ D_{yx} & D_{yy} & D_{yx} \\ D_{zx} & D_{zy} & D_{zz} \end{matrix} \quad (1)$$

The diffusion tensor and, as consequence, the diffusion tensor image is estimated from raw diffusion weighted (DW) images. DW image acquisition is a complex procedure and out of the scope of this thesis. However, the notion that these images are very sensitive to Eddy currents, subject motion and magnetic field inhomogeneities are important to implement DTI studies and to know how to process this type of data[24,25]. The methods used to correct these distortions will be further explored in the section 2.4.

Direct interpretation of DTI at each voxel is a challenging or even and impossible task (Figure 2.2). However, the various characteristics of the diffusion tensor (1) allows the extraction of different values providing an *in vivo* marker of the microstructure and organization of the tissue[24]. Studies using this DTI-derived values reveal a difference between AD and cognitively normal populations. Note that DTI evaluates water diffusion patterns, which are correlated with the structural integrity of the tissues, thus providing structural information[9,13,26].

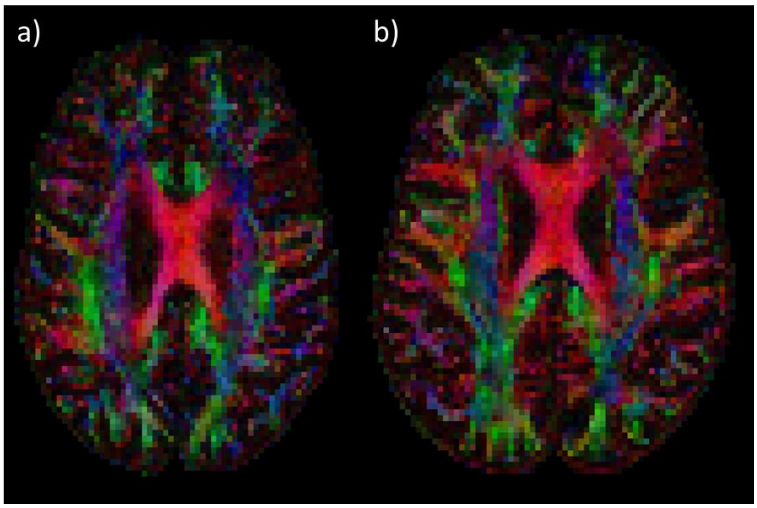

Figure 2.2 Example of the axial view of a DTI. a) Brain with Alzheimer's. b) Normal aging brain.

Structural biomarkers are useful in assisting the diagnosis of AD. However, there are other types of biomarkers, assessing alterations in function and metabolism, that can be extracted from other neuroimaging techniques.

For example, PET has been a powerful imaging method for studying AD, through the use of different radiotracers, namely the 18F-Fluorodeoxyglucose (FDG) tracer,

which provides a better knowledge of the glucose metabolism of the neuronal tissue, and the 11C-Pittsburgh Compound B (PIB) tracer, which enhances the senile plaques, which in turn allow us to identify the abnormal Aβ protein burden (Figure 2.3) [13,15,19].

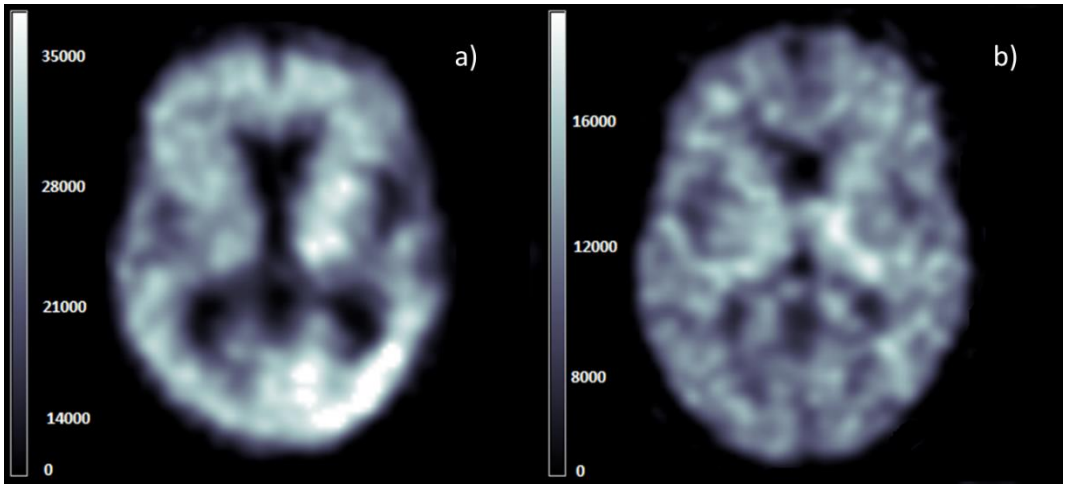

Figure 2.3 Example of the axial projection of a PIB-PET scan. a) Brain with Alzheimer's. b) Normal aging brain.

Through PIB-PET technique, it is possible to acquire quantitative data by calculating the ratio between the tracer accumulation in a specific region of the brain and a non-specific region, that does not accumulate the tracer. This ratio is commonly known as the standardized uptake value ratio (SUVR) and it has also been used as a biomarker capable of assisting in the diagnosis[27,28].

As mentioned before, these imaging biomarkers give, on their own, valuable information for the diagnosis of AD, but a more accurate diagnosis can be provided by combining two or more of these biomarkers. This kind of approach is generally denominated a multimodal approach[9,17].

Despite everything, the analysis of such neuroimaging data is not a straightforward process and may present some challenges.

Neuroimaging provides an enormous dimensional image, containing several measurements on the respective image voxels, and in order to extract the useful biomarkers mentioned earlier, it is necessary to convert this image into something smaller, capable of being analysed without the loss of important information[9,19].

This process was firstly made through manual segmentation of regions of interest, but this approach was slow, prone to human bias and error, and it had a high risk of

losing information. Therefore, a solution that arose naturally was the automation and standardization of the extraction of potential biomarkers useful for diagnosis assistance[9,19]. Although this strategy of automation is indeed better, compared to the one previously used, it also has the disadvantage of creating large amounts of data, which are hard to analyse in its totality.

Hence, there is a need to develop new frameworks capable of analysing and giving meaning to data and machine learning seems to be a great solution, since it offers a systematic and automatic approach to the problem, besides being an objective classification framework able to analyse large amounts of data[7,9].

This kind of tool has the ability to learn from previous data and from that, it can objectively classify/predict two different populations, as well as analyse complex high-dimensional data and detect subtle patterns within these images, which may unfold some knowledge into the unknown physiopathological processes of the disease[7,9].

2.3. Computer-Aided Diagnostics

Developing a systematic and automatic approach, capable of creating objective and accurate classifications or predictions, could have a huge impact in the clinical world.

Also, being able to make an earlier and a definitive diagnostic and therefore starting therapeutic treatments before any irreversible damage has taken place, provides not only a better chance of regaining some function, but it also grants a better life quality of the patient.

Additionally, a faster diagnosis without lengthy specialist investigations and with the removal of human subjectivity, making it only based on quantitative information, could, in theory, be useful to identify individuals that do not show any symptoms and therefore are in a 'preclinical state' of the disease[3,14,15,29,30].

In order to achieve this, computer-aided diagnostics has greatly evolved, and it has shown fantastic results, sometimes being even more accurate than a specialist's opinion[7,8].

Machine learning (ML) has gained a lot of attention recently, because instead of a human, it is a machine who performs the analysis of the data and gives a probability for the final diagnostic conditions.

ML is more than just computer code designed to perform a specific and previously known task[7,8]. In all simplicity, ML is a group of algorithms that can learn from a set of data and posteriorly make predictions or decisions when in contact with new data[7]. Different kinds of learning algorithms can be used in ML, but ultimately their final goal is the same: being able to learn from data in order to predict an outcome[7]. However, ML is supposed to be used as a diagnosis aid tool and never as a replacement for trained professionals.

So how can a machine learn? Just as humans do: teachers in schools use different methods to teach kids new information, and we do the same with machines, also using different methods according to the addressed problem[7].

Relatively to the concept of "learning", ML can be divided into supervised methods or unsupervised methods[7,8].

In supervised methods, both the input data and the expected output set of categories are provided to the computer, and using a suitable algorithm, it constructs a model capable of discriminating the different sets of categories. Therefore, in this learning type, the machine learns from what is already previously known[7,8].

By contrast, in unsupervised methods only the features are provided to the algorithm. In this learning method, the machine does not have previous knowledge of the outcome and must figure out a way to differentiate the input data. This ultimately leads to data mining and knowledge discovery through the identification of patterns invisible to the human eye[7].

One field that is expanding the use of supervised learning methods for computer-aided diagnosis is neurosciences.

Neurodegenerative diseases, such as AD, are being investigated using these techniques which allow the detection of complex and subtle patterns in neuroimaging data[7,8].

Previously it was stated that one of the criteria for AD diagnosis is the use of neuroimaging biomarkers, and that there is high-dimensional data that can be retrieved from them, data that can be posteriorly used in supervised methods of machine learning. This offers a fast, systematic and standardized way of diagnosing AD using only neuroimages, since high-dimensional classifiers can be formed and used as a diagnosis tool[8,9].

A typical AD classifier framework is composed by five building blocks: dataset, feature generation, feature selection, feature-based classification and validation [9,31].

The classifier's aim is to, given a training dataset and the associated output variables, try to solve a model that better adjusts the training data in order to achieve the desirable output variables. So, the ultimate goal is to create a model capable of predicting accurately and automatically the outputs corresponding to new inputs provided by new datasets[20]

Dataset: The dataset must include a large number of individuals, also known as cases, with well-defined clinical properties, that are used for the classification problem, such as a clinical diagnosis and pathologic measures, also known as input data.

This dataset is then divided into a training set and a testing set. The training data set must be large enough to reliably characterize the disease and account for the inter-subject variability[31]. The testing data set is used in order to validate and analyse the performance of the classifier.

Feature Generation: The input data, in this particular case AD-related 'features', such as anatomical or functional characteristics of the disease, is retrieved from the subjects using different neuroimaging modalities: structural MRI (sMRI), functional MRI (fMRI), diffusion tensor imaging (DTI), PIB-PET, and FDG-PET[9,31].

The term "features" refers to the transformation applied to raw medical imaging data and it has the ambition of achieving more informative and quantitative measures. These features must be appropriate and meaningful for the specific classification problem, be comparable across all subjects, and if possible, have a biological value[9,31]. To better understand this, a feature can be as global as the total intracranial volume or as specific as the intensity of a voxel.

Considering that neuroimaging data is high-dimensional, it is predictable that the number of possible features extracted is high-dimensional as well, and for that reason it is often necessary to reduce this high-dimensionality of the input data[9,31]. The feature generation and the dimensional reduction for the different neuroimaging modalities is going to be explained more extensively in section 2.5.

Feature Selection: After the extraction and the dimensional reduction of the features it is possible to select a subset of features in the training set[9,19].

There are two main reasons to justify the use of feature selection. Firstly, using a selected subset tends to improve the global performance of the classification by removing non-informative and confusing features. The other reason is that selecting only the significative features that contribute to the classification problem can help understand the multivariate physiopathology of the underlying disease that is being studied[9,19].

More about feature selection in the Section 2.6.

Classification: This algorithm is responsible for the creation of a model capable of separating the training data, using the output data (also known as labels), into two distinct groups or clinical states: cognitively normal (CN) or diseased subjects, in this case AD patients[9]. The model parameters can also be adjusted in order to achieve a more optimized discrimination between the two labels.

Many algorithms may be used to create this model such as, linear regression (LR), support vector machine (SVM) and neural networks, but their final classification intent ends up always being the same[9,31].

Model Validation: The final phase is the validation of the model created in the previous step, based on new data that were already correctly preprocessed in the same way as the training data. Testing the model through new and different data, not used in the training step, is the best way to evaluate the performance and ability of the developed model in generalizing a systematic rule to divide the data[9,31].

Cross-validation techniques are widely used to unbiasedly estimate the performance values of the classifier, avoiding overfitting[9,31]. An illustration of this is K-fold cross-validation, which consists in dividing the training data into k different subgroups but leaving the one subgroup for testing. This process is then repeated until all subgroups have been tested and the average error is estimated in all the runs[9,31].

More about the algorithms and their evaluation in Section 2.7.

In conclusion, it is worth noting that in recent years a new way of approaching the computer-aided diagnosis has appeared. Instead of taking advantage of only a single

neuroimaging modality, studies are now focusing in the use of a multimodal approach to solve the classification paradigm of AD.

These new classifiers are using the same framework as the previous ones, but now they use features from different combinations of imaging modalities, for example, they join MRI with PIB-PET data. This usually improves the performance of the classifiers[9].

2.4. Preprocessing of Neuroimages

Before generating features from neuroimaging data, standard preprocessing must be executed.

Preprocessing is the term used to describe all of the transformations the original data suffers, in order to prepare and improve for posterior statistical analysis. These applied transformations depend on the type of neuroimaging used.

In this chapter the standard preprocessing of the neuroimaging sources of data used in this particular study (MRI, PET and DTI data) will be characterized.

Magnetic Resonance Imaging (MRI):

MRI brain images allows the detection of brain structural changes in-vivo. To take advantage of this powerful tool various techniques were developed to allow the extraction of relevant data.

The most common way of analysing MRI is through morphometric pattern analysis. However, this analyse techniques require the previous preprocessing of the brain images[30,32].

The standard preprocessing pipeline involves intensity normalization, spatial normalization, tissue segmentation and smoothing[19–22,26,33,34].

Intensity Normalization (Bias Correction): The signal intensity of magnetic resonance acquired from homogeneous tissue is rarely uniform. The same tissue presents a smooth variation in values across an image, which is usually a result of poor radio frequency, gradient-driven Eddy currents and the anatomy of the subject[35].

Although this intensity difference has little or no impact in visual diagnosis, it usually affects the ability to perform automatic segmentation, which requires a certain amount of intensity homogeneity within every tissue[35,36].

The most common way to correct this bias is to use the nonparametric and nonuniform intensity normalization (N3) algorithm[19–22,26,33,34,36].

Tissue Segmentation: Structural data, that is presented in the images, is often unnecessary or not useful for a specific study, since some structures have no biological or statistical value in the classification problem at hand[19,21–23,33,37,38].

A solution to this is segmentation, which appears as a way to divide the brain into distinct tissue types sections, which in turn allows the selection of those mostly involved in the physiopathology of the disease. This segmentation divides the brain into three different tissues: grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF), excluding the skull and other non-brain tissues[19,21,22,33,35,37,38].

Spatial Normalization: Every brain is unique and diverges immensely from all others, namely in terms of size and shape Therefore their differences can only be comparable if the corresponding images are translated onto a common shape and size[35]. This process is commonly called spatial normalization.

Spatial normalization is a procedure that allows the registration of data, from the individual subject-space to a standard-space. In order to normalize the data, a template and a source image are always needed[4,39]. In this case, the template image is the brain on a standard space to which the source image will be registered.

There are numerous standard-spaces such as Tailairach, Tournoux and MNI, and each one uses different coordinate systems.

After selecting the template image, a transformation matrix is constructed from the source image, which allows it to be registered to the template image space.

The normalization step place all images at the same coordinate system permitting to perform statistical analysis and comparison among subjects[39].

Smoothing: Smoothing is typically done by applying a Gaussian linear filter with a full-with half maximum (FWHM) between 4 and 16 mm, and it is used for several reasons[34,40].

On one hand the process of smoothing tends to increase the signal to noise ratio by striping the high frequencies, that represent the small changes at the voxel level. This allows a better visualization of the larger scale changes, which become more apparent if the small changes are removed[40].

Another reason for smoothing is that it allows to mitigate some of the effects of the structural misalignments, resultant from normalization[35,40].

Finally, smoothing provides data that follows a gaussian distribution, making it more suitable for statistical analysis and to be used in classification problems[35,40,41].

Positron Emission Tomography (PET):

PIB-PET data is preprocessed following similar principles as MRI data with some exceptions. Firstly, PIB-PET volumes, that are generally acquired in a long period (~90 minutes)[42], are summed either for the full-time frame or for specific acquisition periods.

The sum volumes are then preprocessed, which involves coregistration, if possible, spatial normalization and finally smoothing[27,28,42–47].

Coregistration: Coregistration allows the combination of functional data from PET with anatomical information provided by structural T1 imaging[48]. The main advantage of coregistration is that it enables the visualization of functional data within an anatomical space, improving the capability of locating the activated zones or zones of accumulation, depending on the utilized radiotracer[48].

This preprocessing step consists in transforming the individual voxels from the PET image (functional images) and align them with the correct voxels in the real brain space (structural image)[48]. Furthermore, this allows transformations applied onto the structural images, such as normalization, to also be applied to the functional ones with better results.

Diffusion Tensor Imaging (DTI):

The most complex and sensitive form of data in this project also involves the most complex form of analysis. Due to the sensitive nature of the diffusion weighted (DW) images previously stated, DTI is sensible to various types of distortions, which must be corrected to allow further analysis of the data[25]. The standard preprocessing pipeline of DTI involves the correction of the distortion caused by Eddy Currents, Head Motion and EPI, Skull stripping and Spatial Normalization[14,18,26,49–52].

Eddy Currents, Head Motion and EPI distortion Correction: Distortions caused by eddy currents, head motion and EPI distortion are the most common artifacts in DTI images. Therefore, a logical preprocessing step is needed for a correction of such artifacts[25].

Eddy currents are loops of electrical current induced by a changing magnetic field. This electric current in turn generates yet another magnetic field that opposes the original one. An ever-changing magnetic field applied in a MRI scan is able to create these currents, that in the case of DTI imaging, since the signal to noise ratio is low, have a huge impact on the image data retrieved[25].

Both eddy currents and head motion can be corrected simultaneously, through registration procedures. For eddy currents corrections they can be achieved through an affine registration to the b_0 image, as for the head motion, those can be corrected using a rigid body registration to b_0[25].

The b-value (b_i) is a parameter of diffusion weighted imaging (DWI) and is defined as the diffusion-weighting. The b-value is measured in s/mm^2 and its optimum value for the brain tissue is situated between 700 and 1300 s/mm^2 with the value of 1000 s/mm^2 being the most commonly used. Therefore, the b_0 image is the non-diffusion image that is acquired in the beginning of each DWI protocol and that is use for the construction of the DTI and for further corrections as seen before[24].

EPI distortions are originated form magnetic field inhomogeneities and causes the anatomy of the brain to appear warped nonlinearly. These distortions are being corrected using field maps and with the correspondence to the anatomical image[25].

Skull Striping: Skull stripping is an optional step also used in the analyse of DTI data. This step is a specific way of segmentation, with the aim of removing the skull, as well as the non-brain areas from the analysis. Although optional, it is widely recommended since it is able to improve coregistration and normalization results, as well as reducing the data size[25].

2.5. Feature Generation

In this chapter the subject of feature generation will be more detailed, containing a detailed analysis of the techniques used in the generation of features from the different medical imaging modalities[9,31].

As stated in the section 2.3, features result from the transformation of raw data and they are used as inputs in classification problems. Features must be a good representation of the data being used, and because of that each medical imaging modality requires a different approach to feature generation[9,31].

The three modalities of imaging that will be used in this study are Magnetic Resonance Imaging (MRI), amyloid-Positron Emission Tomography (PIB-PET) and Diffusion Tensor Imaging (DTI).

2.5.1. Magnetic Resonance Imaging (MRI)

Cerebral neurodegeneration, present in Alzheimer's Disease (AD), is characterized by cellular dysfunction, which leads to a progressively larger scale pathologies, such as damages to synapses, degeneration of axons and ultimately neuronal death (1). This neurodegenerative process, and as a consequence of that, the loss of neuronal mass is a diffuse process and it affects parts of the brain in an unequal severity. Brain structures such as the hippocampus, temporal and cingulate gyri, as well as the precuneus, are commonly the most affected ones[9].

The degradation and atrophy of the brain tissue is generally measured using MRI, more specifically structural MRI, and it serves as a valuable marker to evaluate the progression of the AD pathology. The physiopathology of AD leads to profound structural changes, such as thinning of the cortical surface, variation in the density of the tissue and structural volume alteration.

Knowing that these alterations occur during the onset of the disease, makes the development of different methods for analysing the MRI brain images, more or less obvious. Morphometric pattern analysis, such as voxel-based morphometry (VBM), deformation-based morphometry (DBM), tensor-based morphometry (TBM) and surface-based morphometry (SBM) are the most commonly used methods in AD classification studies[9,22,30,32,53].

However, there are other methods that do not require nonlinear alignments of the voxels, which are used in order to achieve inter-subject correspondence. Instead they use feature-based morphometry[15,54].

2.5.1.1. Morphometric Analysis

The morphometric analysis can be divided in volume-based or surface-based. The volume-based analysis methods are voxel-based morphometry (VBM), deformation-based morphometry (DBM) and tensor-based morphometry (TBM). The surface-based analysis method is SBM.

VBM is based on voxel-wise measures of local brain tissue density after accounting for global differences in the brain's anatomy. From this, brain tissue density changes can be captured and used as biomarkers, by creating density maps that represent volume changes. To assure inter-subject comparison is, in fact, meaningful, each voxel must be aligned to a standard space to guarantee that voxels of different brains are in the same brain region[22,30,32,53,55].

Besides through density, the brain volume changes can be characterized with the use of deformation or tensor measures, such as DBM and TBM.

Both in DBM and TBM, the images are registered into a standard reference space, and the analysis is done by comparing the parameters of the resulting deformation fields, either directly or from derived measures[22,30,32,55].

More specifically, DBM compares directly the deformation fields that describe the moving direction of every voxel in an alignment. It is optimized by an algorithm that aligns the voxels in the brain image to the ones in the template. The moving direction of the voxels can be used as a biomarker since detected differences of values can be used to distinguish between healthy control and diseased brains[22,30,32,55].

In TBM, the most commonly used measure is the determinant of the Jacobian matrix, which is derivative of the deformation field and ultimately it represents the local volume change. Depending on the value of the determinant, it is possible to establish if the brain region is either expanded, if the determinant value is greater than one, or contracted, if it is lesser than one. The expansion and contraction of brain regions can be used as a biomarker of the diseases[16,22,30,32].

Volume-based analysis is the most traditional way of analysing brain structural alterations. However, in recent years, the developing and refinement of other type of tools allowed the development of surface-based analysis methods, that are more sophisticated and capable of studying more specific metrics, such as cortical thickness and surface area.

SBM is a surface-based analysis method and it primarily derives morphometrics measures from geometric models of the cortical surface. This cortical surface is extracted using two lines, that represent the boundaries between the grey/white matter and the grey matter and the meninge dura layer or CSF. These lines are then filled with a mesh of triangles. The point where the triangles meet is called vertex, and its coordinates are stored and posteriorly used for measurements. For example, the area of the sum of this mesh of triangles can be used to calculate the surface area of the cortex[53].

All these analysing tools can be used for feature generation methods in AD related classification studies. The main feature generation methods utilizes density maps or cortical surface maps, which can also be used in combination in order to improve the classification capability[9].

2.5.1.1.1. Density maps-based methods

Density map-based methods are able to quantify the atrophy levels, through the examining of density maps of white matter (WM), grey matter (GM) and cerebrospinal fluid (CSF) (Figure 2.4). The density maps can be constructed using voxel-based morphometry (VBM) or others analyse tools. The creation of feature vectors from density maps can then be divided into two: using density maps as a whole[34] or reduce them in order to construct smaller vectors[8,9,18].

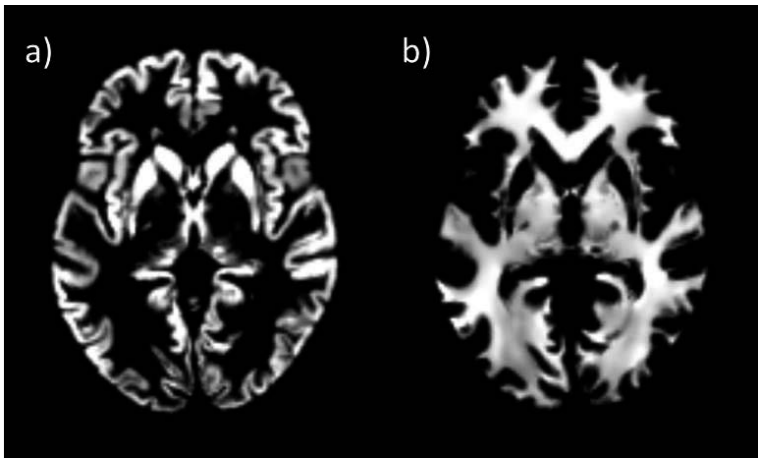

Figure 2.4 Example of tissue Density maps. a) Grey Matter density map. b) White Matter density map.

Using the density maps as a whole seems appealing since all the data can be used as features. However, this approach suffers the drawback of the dimensionality curse. Since each voxel is a feature in this method, the number of features becomes far superior compared with the number of available subjects. Because of this, the classification rule applied to the training dataset is constructed only by chance and thus it prevents the final classification algorithms to select the classification rules capable of generalize to new input data[9].

In order to compensate this potential problem, features have to be reduced, and this is possible using different approaches such as supervised or unsupervised feature-reduction methods, extraction from pre-defined atlases and adaptive ROIs, to reduce dimensionality[9].

Supervised or unsupervised feature-reduction methods are same used for feature selection methods and will be discussed in section 2.6.

Predefined atlases:

Predefined atlases-based methods consist in dividing the brain in several anatomical meaningfully label regions of interest (ROI), built upon a pre-defined atlas template. In these methods the brain images are mapped to the atlas template and the features values generated from each of the defined regions labels[8,9,55].

Various anatomical label atlases can be used, each one constructed on its own predefined atlas template. Examples are automated anatomical labelling (AAL) atlas (Figure

2.5 a), laboratory of neuroimaging (LONI) atlas, Neuromorphometrics (Figure 2.5 b) and many others (1,35, 65).

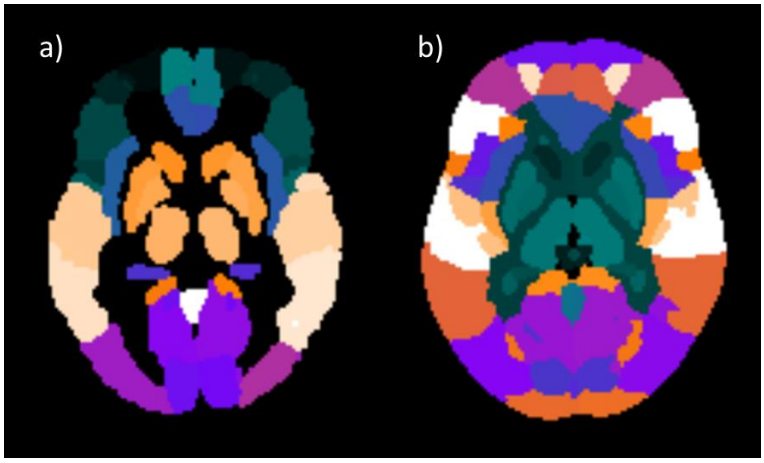

Figure 2.5 Examples of label atlases. a) ALL atlases. b) Neuromorphometrics atlases.

State of the art methods using these methods for feature generation and support vector machine (SVM) have achieved accuracy values as low as 84.00%[33] and high as 94.50%[9].

Furthermore, these methods can be used as single atlas or in multi-atlas approach. The goal of multi-atlas approach is to reduce the bias of mapping the brain images to each atlas template[22,33,55].

Using a multi atlas approach in combination with SVM achieve accuracies between 92.57%[22] and 93.83%[55].

Adaptive regions of interest:

Atlas-based methods apply static label ROI in order to generate meaningful data to study the disease.

However, sometimes knowledge of the abnormal regions is not a possibility and the used of static ROI may not be able to capture the true nature of the progression of the disease that can span multiple predefined ROI.

Contraries to atlas-based methods, adaptive regions of interest methods allow the creation of ROI that are disease specific[9,33].

Adaptive ROIs are generated by performing watershed segmentation on the density maps (Figure 2.6). Watershed segmentation is a traditional image segmentation

approach, vastly used in medical image analysis and it divides images into different regions, according to similarities in the local intensity[22,56].

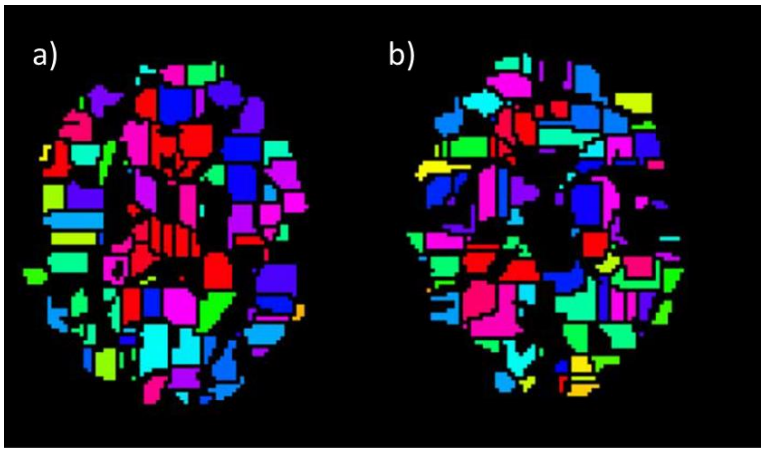

Figure 2.6 Example of adaptive ROIs. a) Adaptive ROIs generated form Grey Matter density maps. b) Adaptive ROIs generated form White Matter density maps. (Adaptation)[56]

Furthermore, adaptive ROIs analyses, as the predefined atlas-based methods, can be done from a single atlas or from multiple atlas template[22,56].

In the single atlas approach the subjects are firstly registered to one particular atlas template, and then followed by the calculation of adaptive ROIs and the corresponding regional volumetric measures. On the other hand, in the multi-atlas approach, each atlas creates a unique set of ROIs, and since one subject is registered to different atlas templates, different tissue maps are generated, according to the atlas used. The different volume measures calculated from the adaptive ROIs can then be concatenated for classification purposes or used in ensemble classification techniques[9,22,56].

Using these methods in combination with SVM it is possible to achieve performance accuracies between 91.64% and 94.30%[9].

2.5.1.1.2. Cortical Surface-based methods

Cortical surface methods use surface-based morphometry (SBM) to create cortical surface maps[9,57] (Figure 2.7), that enable more precise measures of cortical atrophy, since they actually assess the thickness of the cortex in millimetres rather than the difference in volume provided by density maps[57]. These kinds of methods are able to detect subtle changes in the thickness of the cortical surface that are not easily measurable in the early

stages of the disease. Therefore, it can be affirmed that cortical surface data is a powerful tool on its own, as well as a complementary source to regional volumetric maps[9].

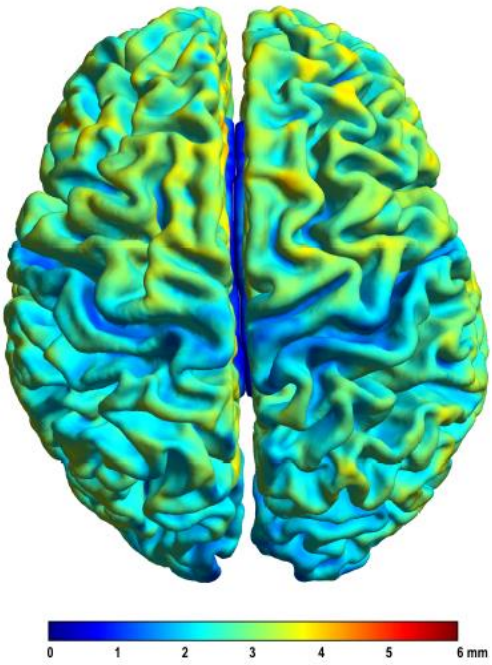

Figure 2.7 Cortical surface map.

From the cortical surface maps measures are extracted from all the vertices of a surface. There are two main sources of features that can extracted from this type of analysis: volumetric and geometric[9].

The features calculated from volumetric analysis include a variety of morphological values, such as cortical thickness, surface area and grey matter volume. From geometric analysis it is possible to extract measures of sulcal depth, metric distortion and mean curvature[9,23].

Similarly to the density maps case, these measures can be used as a whole or they can be reduced applying feature reduction methods, such as supervised/unsupervised feature-reduction-based (section 2.6) and atlas-based methods[9].

Atlas-based methods:

These methods are similar to those described in the previous section "Predefined atlases", however, instead of using structural templates, which are used for volume analysis, they rely on atlases that reflect the surface space[9,57].

In order to extract features, the original brain images are registered onto a standardized stereotaxic space. Posteriorly, cortical maps are constructed using the various regions from existing surface atlas, so that the generated features are representative of different regions, instead of being calculated at each voxel, allowing a reduced number of features[9,57].

Using these methods it is expected to obtain accuracy values between 86.10% and 95.00%[9].

2.5.1.2. Feature-based morphometry

Rather than using traditional methods of analysis for feature generation, the feature-based analysis methods compare brain images based on image features, not requiring the voxel alignment that is applied in the morphometric methods[54].

Images usually have information that can be converted into features, but some of these features may not be suitable for a particular classification problem. Therefore, their extraction must be done in order to obtain only the most informative features of the images for the classification task at hand, which at the same time reduces the complexity of the representation[15].

Studies involving MRI for Alzheimer's classification used two methods in order to generate features: scale-invariant feature transforms (SIFT) and Oriented FAST and Rotated BRIEF (ORB). Other methods can also be used, such as Laguerre Circular Harmonic Functions coefficients[15,54].

These methods are used in computer vision as tools to detect and describe key points which summarise and describe local features in images.

SIFT features are generated using a Gaussian kernel at different scales (variances) convolved with the original image, in order to originate the Difference-Of-Gaussian (DOG), which in turn is used to extract the key points of the images. Features generated by this method are robust to distortion, as well as to noise and to resolution. The main disadvantage of this method is that it imposes a large computational burden[15,54].

The ORB method is computationally more efficient that SIFT and it is based on a FAST key point detector which has similar matching performance, but is less affected by image noise and capable of being used in real-time performance[15].

This kind of approach to feature generation is recent and the features extracted do not have any biological value. However, these techniques can shed some light to new regions of interest that should be further investigated.

2.5.2. PIB-PET

One of the most well-known traits of Alzheimer's Disease (AD) is the accumulation of β-amyloid peptide in the extracellular space of the neurons. This accumulation is considered the core molecular feature of AD[27,28]. Positron emission tomography (PET) imaging with the ^{11}C-Pittsburgh Compound B (PIB) tracer is currently used in nuclear medicine imaging to visualize and evaluate the burden of the β-amyloid in patients with AD[27,28].

The evaluation of the brain amyloid burden is done using a quantitative approach, known as Quantification, that allows the generation of biological relevant metrics that can be used for classification problems.

Quantification is possible given the inherent capability of measuring, at the voxel level, the intensity signal related to the actual quantification of the underlying source distribution. It deduces a direct link between the concentration of activity in organs/tissues across time, and the fundamental physiologic parameters of the biological processes, at a cellular level[58].

The most commonly used quantization method is the Standard Uptake Value (SUV). SUV are metric values calculated form regions of interest (ROI) or from individual voxel values. The SUV value can represent the mean intensity of the voxels in a ROI or can be defined by the maximum intensity value[27,28,42–47].

Using the different SUV values, the standard uptake value ratio (SUVR) can be calculated.

SUVR is a dimensionless value obtained from the ratio between the SUV from a particular tissue or voxel and the SUV from a region where the accumulation of the

radiotracer is not expected, a reference area. The SUVR is posteriorly normalized for body weight and injected dose to allow intrasubject comparison[27,45].

The normalized SUVR from each target region is then used as a feature in the classification problem.

Another option to create a feature is to combine all the regions, forming a composite SUVR. This composite SUVR is used in binary classification, in which subjects above a threshold are considered amyloid positive, and those below are considered amyloid negative[27,28].

Using SUVR as features is expected to achieve accuracy values as low as 85.70%[46] and as high as 96.00%[28,42].

2.5.3. Diffusion Tensor Imaging (DTI)

The physiopathology process of Alzheimer's Disease (AD) is progressive and starts with cellular dysfunction and culminating in cell death. The progression of AD is associated with alterations in the microstructure and organization of the tissues, which normally restrict water motion and thus change the diffusion patterns of the water molecules[24].

The alteration of the water molecules diffusion patterns is detected with DTI and it allows the extraction of biomarkers, useful in the classification of the disease. Since these patterns are sensitive to the alteration of the tissues microstructures, DTI not only allows the characterization of the end stages of the disease, but it also provides a better insight of the small initial structural changes, making DTI a powerful tool in the classification of the disease and also in the prediction of the conversion from MCI to AD[24].

As mentioned in section 2.2, DTI is used to analyse the diffusion patterns of the water assessed by a diffusion tensor. From this diffusion tensor it is possible to extract different diffusion values or to reconstruct the white matter (WM) tracts, showing the connectivity of the different brain regions, known as tractography[9,14]. Both of these properties are being used to extract features for AD/CN classification. And depending on how features are extracted, they can be classified into three different methods: diffusion maps analysis, tractography and connectivity network measures[9].

2.5.3.1. Diffusion maps analysis

Meaningful measurements are impossible to be directly extracted from DTI. The diffusion matrix at each voxel must be simplified creating simple scalar maps similar to those used for MRI analyses[24,25,51,52].

The diffusion tensor can be diagonalized and from this transformation it is possible to extract three eigenvectors (ε_1, ε_2, ε_3), representative of the principal diffusion directions, and the corresponding three eigenvalues (λ_1, λ_2, λ_3), representative of the diffusion magnitudes. Combining the eigenvectors and the eigenvalues it is possible to represent the diffusion as an ellipsoid[24,25,59].

Metrics such as mean diffusion (MD) (2) and fractional anisotropy (FA) (3) can be calculated from the eigenvalues and used to create diffusion maps that reflect the different diffusion patterns of the tissues[14,18,40,49,59]. These values are commonly used for AD/CN classification since empirical data shows that AD patients show lower FA and higher MD values in comparison to CN.

$$MD = \frac{\lambda_1 + \lambda_2 + \lambda_3}{3} \tag{2}$$

$$FA = \sqrt{\frac{(\lambda_1 - MD)^2 + (\lambda_2 - MD)^2 + (\lambda_3 - MD)^2}{2({\lambda_1}^2 + {\lambda_2}^2 + {\lambda_3}^2)}} \tag{3}$$

These diffusion maps (Figure 2.8 a), Figure 2.8 b)) can be used as a whole in voxel based analysis or they can be reduced and analysed according to label atlas based methods[9,14,18,26,40,49,59].

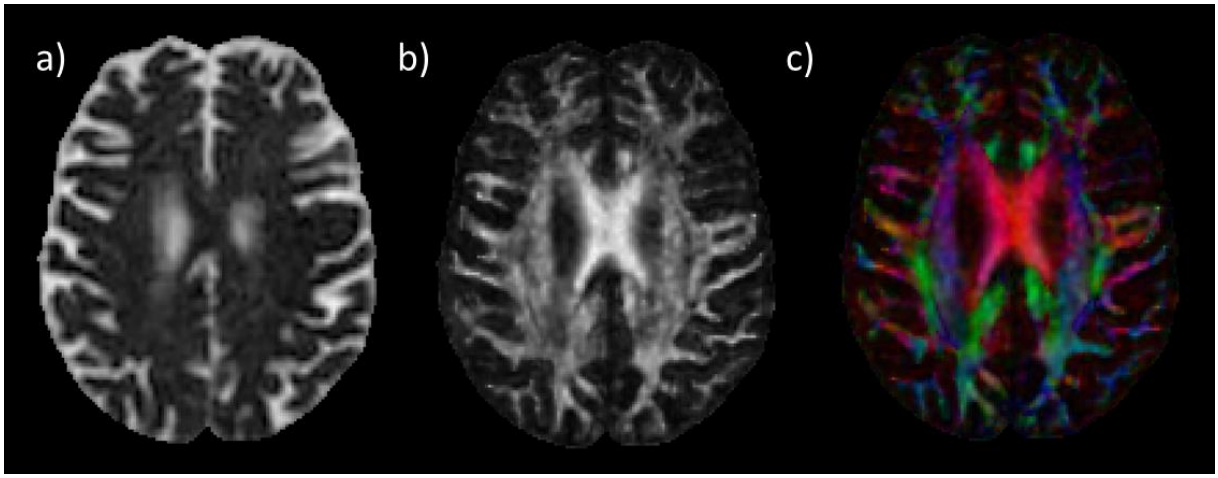

Figure 2.8 Examples of diffusion maps. a) MD diffusion map. b) FA diffusion map. c) FA diffusion map with coloured fiber orientation.

Also, the information from the eigenvectors can be visualized in these diffusion maps by assigning a colour to each diffusion axis[59] (Figure 2.8 c).

The use of FA and MD maps as features for AD/CN classification shows accuracy values between 76.00% and 89.00%[9,14,60].

2.5.3.2. Tractography methods

Tractography can be considered an extension of DTI, through which the directional information, that is stored in the diffusion tensor, is used to construct virtual, three-dimensional, white matter maps. These maps are constructed based on similarities between the diffusion properties of the neighbouring voxels, in terms of the diffusion measurements as well as their orientation[9,59].

Tractography represents the path of least resistance to the water diffusion[59]. White matter tracts are mainly made of myelinated axons and represent all sorts of connections in the central nervous system.

These methods use tractography in order to construct the white matter tracts, instead of using voxels or label atlas as a tool for feature generation. After the construction of the white matter tracts, they are clustered into various tract bundles, based on an anatomical atlas. From these tract bundles, diffusion parameters are calculated and used as features for classification[9,14].

Using tractography for feature generation in AD/CN classification problems lead to accuracy values between 80.60% and 85.00%[9,14].

2.5.3.3. Connectivity Network Measure methods

The obtained DTI images are divided into anatomical regions, which contain fibers, from which several features can be calculated, and in which it is seen a development of connectivity networks that permits a derivation of a collection of network measures, used for classification[9].

The connectivity networks are based on regional features, such as fiber count, averages of on-fiber FA or MD values and the three principal diffusivities. The network measures extracted are raw connectivity matrices, global efficiency, transitivity, path-length, modularity, radius and diameter. These measures may be used as features for classification[9].

These methods are being used in MCI/AD conversion prediction rather than being a simple AD/CN classifier.

2.5.4. Multimodal Studies

The previous sections have given a wider vision of the different modalities of neuroimaging that allow the extraction of biomarkers, which yield complementary information about the disease. Thereby, the use of these modalities improves our understanding of the vast and complex physiopathology processes behind AD[9].

In multimodal studies the goal is to explore and integrate the different biomarkers in a single classification framework. This purpose is achieved with the use of two different strategies: one in which the features from the different neuroimaging modalities are concatenated into one single data set and the other in which specialized fusion methods are used to exploit the multimodal data[9].

2.5.4.1. Straightforward feature concatenation

Straightforward feature concatenation is by far the simplest of the two previously mentioned methods that explore the complementary information between the different data sources. In this method, the features from the neuroimaging modalities are simply concatenated forming a new feature vector that contains all the information. All features from the different modalities are treated as equal in the classification problem[9].

The main component of the multimodal studies is structural MRI. Features from structural MRI have been combined with other data: another imaging modality or some other factors such as genetics, cognitive tests and even demographics data[9].

These methods achieve classification accuracies between 85.00% and 100.00%[9].

2.5.4.2. Specialized Fusion Methods

While the previous method seems appealing, due to its own inherent simplicity, it suffers from a major flaw. By treating all features equally, there is no way to account for their different natures. A major problem arises from this situation when one modality has more features than the rest or when the features from one modality have more variation on a much larger scale. These factors can lead to prediction models, from classification algorithms trained with concatenated features, that ignore some modalities[9].

In order to solve this problem, specialized fusion methods are used to ensure that the complementary information from all modalities is being used, either by combining the results of classification rules from the individual modalities or by using special combination rules to combine features before training[9].

Ensemble classifier techniques offers a variety of methods to study and combine the complementary information from the different sources and will be further explored in more detailed in section 2.8.

When applied to the AD/CN classification problem these techniques shown high performance with accuracy values between 85.00% and 95.95%[9].

2.6. Feature Selection

Neuroimaging data are large in scale, even after the application of the reduction techniques previously mentioned and this high dimensionality presents a serious challenge to the existing learning methods. This is known as the curse of dimensionality[20,61–63].

The large number of features, in relation to the number of cases, causes the learning model to overfit, which results in a degeneration of the performance and in an incapability to adapt to new data. To address this problem, various dimensional reduction techniques have been studied, and they now represent an important branch in machine learning and data mining research areas[61–63].

Feature selection is a widely employed technique that reduces dimensionality. The goal of feature selection is to choose a small subset of relevant features from the original ones, according to an evaluation criterion, leading to a process that provides a better learning performance, lower computational cost and a better model interpretability. Depending on whether the data is labelled or not, the technique can be categorized in supervised, unsupervised and semi-supervised[61–63].

Real world classification problems require supervised learning where the underlying class probabilities and class-conditional probabilities are unknown, and each case is associated with a designed class label[61,62].

The relevance of each variable is unknown in andvance, and due to that, to provide a better representation of the problem, many candidate features are introduced, which in turn results in the existence of irrelevant that do not contribute to the classification but affects the learning process, or redundant features that in nothing contribute to the classification problem. Removing these kinds of features allows a decrease in the learning time and generally produces a better and more robust classifier[61,62].

Regardless of the technique used, a typical feature selection method consists of four steps, subset generation, subset evaluation, stopping criterion and result validation[61,62]. The end goal is to achieve a subset of features that is suited for the problem.

Subset generation produces a set of candidate features using one of the following strategies[62]:

- The specified size of the subset of features that optimizes an evaluation measure;
- The smaller size of subset that satisfies a certain restriction on evaluation measures;
- In general, the subset with the best commitment among size and evaluation measure;

Following the choice of one of the previous strategies, there is an evaluation of the subset, using a specific criterion. These two steps are repeated until the stopping criteria is reached and the best features are selected, taking in account that they are later subjected to validation[63].

Although there are many feature selection methods, this chapter will be mainly focused on the supervised techniques, since they seem to be the best suited to our work. Considering that the features are independent or near-independent, these supervised feature selection methods can be broadly organized into filter, wrapper, embedded and hybrid methods[20]. Additional models have been devised for datasets with structured features, where dependencies exist, and for streaming features[61,63].

The focus will be on methods for independent or near-independent features, mainly those involved in classification tasks.

2.6.1. Filter Methods

Filter methods rely on the data's characteristics and evaluate features without utilizing any classification algorithms[20,61–63]. With filter methods it is possible to rank features individually, univariate scheme or in subsets, multivariate scheme. Therefore, is easy to foresee the capability of the multivariate scheme to handle redundant features.

These methods typically consist of two parts.

Firstly, the features are evaluated and ranked. In this phase, different techniques can be used to develop measures that allow the ranking of each feature. These measures can

be classified into information, distance, consistency, similarity and statistical. In the end, the features with highest ranking are chosen to train the induction algorithm[61–63].

There is a large variety of filtering methods thanks to the combination of the way that features are evaluated, and the measure used to rank them. However, not all filtering methods can be used for the data mining task, and depending on the task, they vary[61–63]. A summary of most commonly used filter methods for classification are listed in Table 2.1.

Table 2.1 Summary of the most common filter methods. (Adaptation)[63]

Name	Ranking Technique
Information gain	information
Gain ration	information
Correlation	statistical
Chi-square	statistical
Inconsistency criterion	consistency
Relief and ReliefF	distance
Laplacian Score	similarity

2.6.2. Embedded Methods

Embedded methods, similarly to wrapper methods, utilize learning algorithms to evaluate the features[20,61–63]. The key difference between these two methods it is that in the embedded methods, feature selection is done during the learning phase, meaning that the learning algorithm is not only ranking the features for posterior selection but is actively selecting the best features as well. These methods are thus embedded in the learning algorithm either as its normal or extended functionality[61–63].

These methods combine the advantages of both the filter and wrapper methods, specifically the speed of filter methods, only running the learning algorithm one time, which in turn leads to a reduction of the computational cost and the feature biases from wrapper methods[61–63].

The most common embedded methods include various types of decision tree algorithms, such as CART and random forest, but also other different algorithms like adaptive lasso and elastic net regularization[61–63].

Referencing section 2.5, these feature selection methods can also be used to reduce the dimensionality of neuroimaging features. By assuming each voxel as a feature, these selection methods allow the reduction of the n voxels to a subset that can be used in classification algorithms. Also, using these methods for reducing the dimensionality of features, it is possible to extract new regions of interest for different diseases[9].

2.7. Classification Algorithms

As seen in section 2.3, Machine Learning (ML) can be used in assisting diagnosis. Supervised machine learning algorithms are applied in classification, being able to categorize new data from prior given information, improving the ability to make a diagnosis.

Choosing the most suitable learning algorithm for the data is an important step and will ultimately affect the performance of the classification model[31,64]. Depending on the problem at hand, the performance of the model can be evaluated based in different classification parameters such as accuracy, sensitivity, specificity and the area under the Receiver Operating Characteristic (ROC) curve (AUC). Prediction accuracy is often the chosen one[21,31,34,64].

The performance parameters can be extracted from the confusion matrix (Figure 2.9).

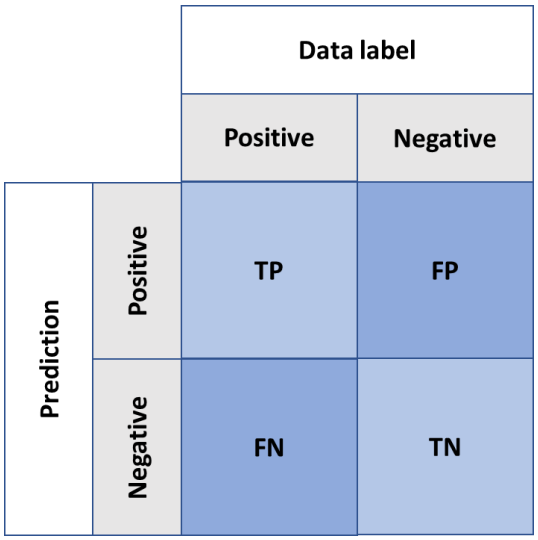

Figure 2.9 Confusion matrix.

In the Figure 2.9 TP stands for true positives and represent subjects correctly identified as positives, TN stands for true negatives and represent subjects correctly identified as negative, FP stands for false positives and represents subjects incorrectly identified as positive and finally FN stands for false negatives and represents subjects incorrectly identified as negative.

Through the confusion matrix, accuracy, sensitivity, specificity and balanced accuracy can mathematically be defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Sensitivity = \frac{TP}{TN + FP} \tag{5}$$

$$Specificity = \frac{TN}{TP + FN} \tag{6}$$

$$Balanced\ Accuracy = (\frac{TP}{TP + FP} + \frac{TN}{TN + FN})/2 \tag{7}$$

These prediction parameters can be calculated through different methods. The simplest approach is to split the data in three parts: two-thirds are for training and the remaining third is used to test the performance[64].

Another approach, known as cross-validation, randomly divides the data in k different and even sized subgroups: k-1 subgroups are used for training and the remaining group is used for testing[31,64]. The process runs iteratively leaving a different group for testing

in each run and computing, at the end, the mean of the evaluation parameters computed at each run.

A special case of cross-validation can also occur, when k is equal to the number of cases in the data. Each subset is composed of only a single instance and the method is known as leave-on-out[31,64].

The results from these methods, not only can evaluate the global performance of the different learning algorithms but also allows to choose the best model that maximizes the prediction parameter[31].

The most common learning algorithms for AD classification are logistic regression (LR) and support vector machine (SVM)[9].

2.7.1. Logistic Regression

Regression analysis is a form of classification technique which investigates the relation between a dependent (label) and an independent (features) variable[65].

Logistic Regression (LR) is one type of regression analysis and it can be used as a classifier. This type of regression is considered a statistical model in which a logit curve is fitted to the data, modelling the probability of occurrence of a class[65–67].

This kind of regression can only be applied to binary problems, where the dependent variable is dichotomous[65,67], but with the use strategies such as 'one against all', it can also be applied to multi class problems, where the dependent variable can assume more than two states.

The first step of a LR classifier consists in building a logit link function[65,66], which is a simple transformation that contains the natural log of the odds of the dependent variable, occurring or not[65,66].

$$\ln\left(\frac{P_{(y|x)}}{1-P_{(y|x)}}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \quad (8)$$

To fit the logistic model, represented in equation (8) towards one set of data, the values of the n unknown parameters β must be estimated[65]. To achieve that purpose different estimation methods are used such as noniterative weighted least squares, discriminant function analysis and maximum likelihood[65].

Maximum likelihood estimation algorithm is the most commonly used method and it is presented as default in the logistic regression routines of the major software's packages[65]. The fitted model can then be used to classify new data.

The main advantages of this learning algorithms are that it can be applied in a wide range of problems, since the output is interpreted as probability and they can handle nonlinearity, interaction effect and power terms. However, it requires a large sample size in order to achieve stable results and it also suffers from multicollinearity problems[67].

As a final note, is important to point out that LR allows the extraction of the different values of the β parameters that reflect how the variable contributes to the final prediction. These values can possibly be used to further study the importance of each feature and to interpret them directly.

2.7.2. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a powerful, state-of-the-art algorithm based on concepts from the Statistical Learning Theory[66,68]. This algorithm was designed as a tool to solve supervised learning problems and has strong regularization properties that theoretically prevent the model from overfitting[67,68].

Geometrically, this method can be interpreted as the search for the optimal separating N-dimensional hyperplane, that separates two data classes[64,68]. The distance between the points of either side of the hyperplane and the hyperplane itself is known as margin. The goal of the algorithm is thus to maximize the margin and create the largest possible distance with a minimum error rate[64,68], which means that the greater the margin, the better the generalization error of the classifier will be[64,68].

SVM can classify linearly separable and non-linearly separable data[64,68].

If the data is linearly separable, the SVM analysis will attempt to find the N-dimensional hyperplane that allows the separation of the classes. Afterwards, the optimal hyperplane, that is the one which maximizes the margin, is chosen from an infinite number of hyperplanes. The data points lying on the margin of the chosen hyperplane are the support vector. The solution for the model is represented as a linear combination of these points only (Figure 2.10)[64,68].

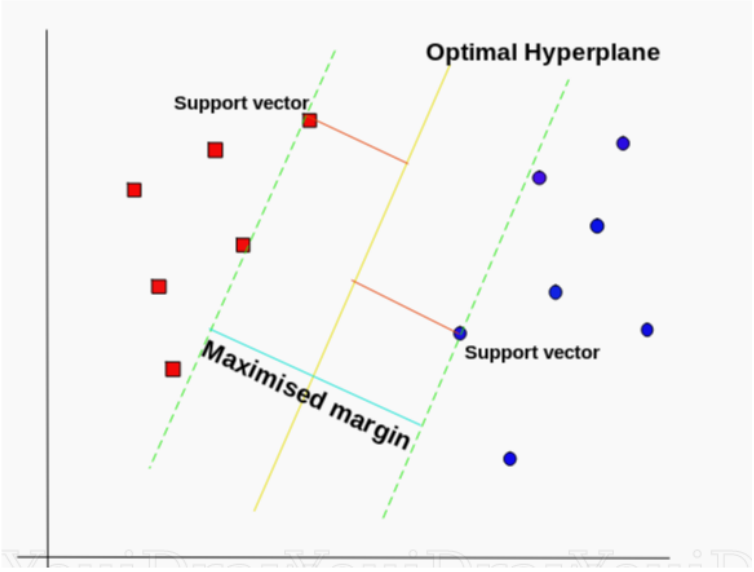

Figure 2.10 Visual representation of the SVM analyses (Adaptation)[69].

However, if the data is non-linearly separable, as in most of real-world problems, the SVM analysis will not be able to find a linear N-dimensional hyperplane capable of dividing the classes. Instead of fitting a nonlinear N-dimensional hyperplane to the data, the SVM algorithm will map the data to a higher-dimensional space and define the hyperplane there[64,68].

This new higher-dimensional space is called feature space. The data is mapped to the feature space through a mapping function **ϕ**, where a linear solution represents a nonlinear solution in the original data space (Figure 2.11)[64,68].

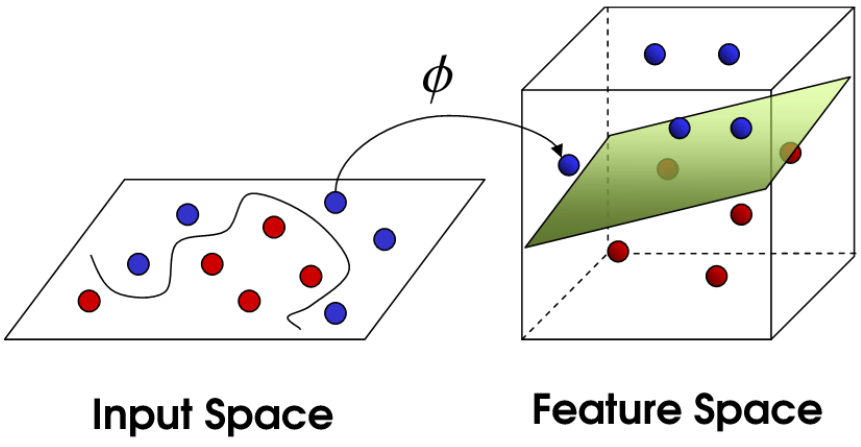

Figure 2.11 Visualization of the mapping function (Adaptation)[70].

In the feature space, the mapping function appears only as the data through dot products in the form[64,68],

$$\phi = \phi(x_1)^T * \phi(x_j) \quad (9)$$

Furthermore, if there is a kernel function K,

$$K(x_i, x_j) \equiv \phi(x_1)^T * \phi(x_j) \quad (10)$$

ϕ does not need to be determined. Instead, K is the only function needed to be used on the training data. Therefore, these kernel functions are a special class of functions that allow the calculation of internal products directly in the feature space[64,66,68].

Many kernel functions may be used, depending on the problem at hand, and choosing the right one will definitely improve the classification performance. The most common kernels used in SVM are the linear kernel, polynomial and the radial basis function (RBF)[68].

$$Linear\ Kernel: K(x_i, x_j) = {x_i}^T * x_j + c \quad (11)$$

$$Polynomial\ Kernel: K(x_i, x_j) = (\gamma {x_i}^T * x_j + r)^d \quad (12)$$

$$RBF\ Kernel: K(x_i, x_j) = \exp\left(-\gamma \left|x_i - x_j\right|^2\right), \gamma > 0 \quad (13)$$

The parameter c in equation (11) is and optimal constant. Furthermore, in the equations (12) and (13), the parameters, γ, r and d are kernel parameters and they can be adjusted to obtain better performances. Also, the parameter c, is the penalty parameter of the error term and it is always greater than 0[68].

Among all of the previous kernels, the most popular is the RBF, because it has lesser hyper parameters in relation to the polynomial kernel, it has less numerical difficulties and it maps samples in a nonlinear way[68].

The major flaws of these methods are their complexity, in the choice of the kernel and respective parameters and their limitation in speed and size[67,68].

However, they present high performance, they have a good generalization ability and they prevent overfitting[67].

2.8. Classifier Ensemble

Inspired by human social learning behaviours, ensemble methods are statistical and computational learning methods, that seek several opinions before making a decision[71].

Ensembles represent the combination of individual classifier characteristics, such as their prediction or learning algorithm, and have the goal of improving upon the individual classifier's performances[64,71–73].

Empirical data has shown that the combination of different classifiers can indeed improve the final prediction and it be applied in different fields[71,73].

One field in particular is when different data sources are available[71]. As discussed in the section 2.5.4.2, the specialized fusion techniques applied to multimodal approaches are none other than ensemble methods.

The reasons used in order to explain the better performance of these methods over a single learner can be classified roughly into three categories: overfitting avoidance, computational advantage and representation[71,72].

The first reason is referent to the cases where small amounts of data are available. In these types of cases, the individual learning may find different hypotheses predicting the training data, with lower training error - optimal solution - but are incapable of generalizing this to future given data[71,72]. Therefore, the combination of different classifiers reduces the risk of choosing only one of these hypotheses, and as consequence improves the general performance[71,72].

The second reason is concerning the sense that individual classifiers may get stuck at a local optimal solution. During the training phase, an error function is minimized in order to find the best solution. However, the error function can present different local minima that unable the learning algorithm to reach the global optimal solution. By combining different models, ensemble methods can avoid this suboptimal solutions and find a better solution[71,72].

The last reason arises from the possibility that the optimal hypothesis could be outside the space of any single model. Therefore, the combination of different models

expands the space of representable functions that now can embrace the true optimal hypothesis[71,72].

Ensemble techniques can be divided into two phases[71–73]. The first phase involves the process of generation of a set of different base classifiers[72,73]. This is followed by the second phase, the combination phase, where the output of the different base classifiers is combined into a single output[72,73].

There is a great variety of methods for both the generation and the combination of the base classifiers. This makes the classification and organization of the ensemble methods difficult[71].

Regarding the generation phase, ensemble methods can be classified as either non generative or generative[71].

Generative methods focus more on the generation phase. These methods actively generate sets of base classifiers and their main focus is to intervene on the learning of these classifiers or on the structure of the data, in order to improve diversity as well as the accuracy of the base classifiers[71,72]. This ought to be achieved using several approaches such as input manipulation, manipulated learning algorithm, partitioning and hybridization[72,73].

On the other hand, non-generative methods do not actively generate new base classifiers. Instead, these methods restrict themselves to combine, in the best suitable way, a set of previously well-designed base classifiers[71]. The focus is not on the generation phase, but only on the combination phase.

In regard to the combination phase, ensemble methods can be divided into either fusion or selection methods[71,73].

Fusion methods refer to the process of integration of the base classifier's outputs into a single output. Methods such as weighting and meta-learning are commonly used[71–73].

In weighting methods, the ensemble output is given through assigning weights to each base classifier[72]. These methods are best suited when the base classifiers have comparable performances[72].

The most common and simplest weighting method is majority voting[71,72]. The idea behind this method is to define the final output as the class which collects the majority of votes from the base classifiers[71,72].

Meta-learning methods have more than one learning stage[71,72], which means that the output of the base classifiers serves as input to the meta-learner, which in turn generates

the final output[71,72]. These methods are suitable for cases where the base classifiers have different performances and are trained in different subspaces[72].

Selection methods are another approach to combine the decision of base classifiers. These methods try to find the optimal subset of base classifiers generated and presuppose that not all base classifiers contribute to a final decision and that only a restrict group can achieve the best possible performance[73]. After the selection of a final set of classifiers, the output of the ensemble model can be divided in picking the base classifier with the best performance or in combining, in some way, the set of classifiers[71]. Static and dynamic selection are the selection methods used in this combination technique and further information can be found in[71,73].

Combining different techniques from both generation and combination, a vast number of ensemble methods can be constructed.

3. METHODS

Chapter 3 will focus in the representation and exploration of the general workflow of this study. The chapter is divided in 6 sections, describing the different methods used to analyse and evaluate each neuroimaging modalities.

3.1. Participants

To achieve the goal of the thesis the original data were divided into four groups.

Each group was constructed in the most balanced way possible and used for different objectives.

In Table 3.1 we summarize the information of the four groups.

Table 3.1 Demographics for the study population.

	CN (n=21)	*AD (n=20)*
Age (years)	65.9±6.8	66.3±6.9
Gender (male/female)	10/11	10/10
a) MRI		
	CN (n=21)	AD (n=17)
Age (years)	65.9±6.8	66.4±7.3
Gender (male/female)	10/11	8/9
b) PIB		
	CN (n=20)	AD (n=17)
Age (years)	66.4±6.5	65.8±7.3
Gender (male/female)	10/10	9/8
c) DTI		
	CN (n=20)	AD (n=16)
Age (years)	66.4±6.5	65.6±7.4
Gender (male/female)	10/10	9/7
d) Ensemble		

Age value is defined as mean ± standard deviation

3.2. Preprocessing

The first common step between all the image modalities was converting all the images from DICOM format to NII format. Afterwards, standard preprocessing was applied, before further data processing.

The preprocessing of the T1-weight MR brain imagens started by manually aligning each image, so that origin of the image was the anterior commissure (AC). This step is important since the preprocessing tools used assume that the origin of the image is there.

After the alignment, T1-weighted MR brain images were preprocessed using the Computational Anatomy Toolbox version 12 (CAT12), (www.neuro.uni-jena.de/cat/index.html), for Statistical Parametric Mapping 12 (SPM12), ((www.fil.ion.ucl.ac.uk/spm/software/spm12/), in the MatLab environment, (www.mathworks.com). The images were preprocessed using the segment data option of the toolbox. This preprocessing can be roughly summed to the following: tissue inhomogeneity correction; tissue segmentation into grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF); spatial normalization into the ICBM space using the Dartel template.

PET-PIB images were preprocessed using SPM12. Firstly, the sum image that reflects the total accumulation was calculated and then coregisted to the corresponded T1-weighted image. This allows the application of the same spatial normalization transformation from the T1-weighted image to the PET-PIB image, providing a more accurate spatial normalization.

The sum image was spatially normalized to the T1 MRI template ICBM152, given by SPM12. The normalized images were then visually inspected in order to verify the existence of flagrante imperfections.

Lastly, the normalized imagens were smoothed, using SPM12 smoothing and a gaussian smoothing kernel, with full width at half maximum (FWHM) of 8mm.

From the Diffusion Weighted images (DWIs), the Diffusion Tensor Images (DTI) were constructed using ExploreDTI, (www.exploredti.com). DTI was then corrected for head motion, eddy currents and EPI distortions, with deformation axes set to [1 0 0] and image type set to FA. After the correction, the DTI was spatially normalized to FA atlas template SRI24, (www.nitrc.org/frs/?group_id=214) or to FA atlas template IIT Human

Brain Atlas, (www.nitrc.org/frs/?group_id=432) or to FA atlas template ICBM, (identifiers.org/neurovault.collection:264).

3.3. Feature Generation

After preprocessing the neuroimages, features were generated in order to be posteriorly be used in classification problems.

Tissue density maps, from the preprocessed T1-weighted MR images, were then analysed using region or label-based morphometry (RBM). RBM, is a type of predefined atlas analysis that allows the estimation of regional tissue volumes and cortical thickness values from different volumes or surface based atlas maps.

This method was implemented using the CAT12 toolbox for SPM 12, in MatLab environment, and both volumetric features and surface features were generated.

The Cobra, Hammers, Neuromorphometrics and lpba40 atlas were used to generate volumetric features from the different brain tissues: White Matter (WM), Grey Matter (GM) and Cerebrospinal Fluid (CSF). The resulting absolute values were then used as features for the classification.

Surface features were generated from a2009, DK40 and HCP brain surface atlas. The feature values extracted from the different atlas regions of interest were Thickness and Gyrification values.

In Table 3.2 we summarize the results of MRI feature generation and how we define each model.

Table 3.2 MRI-based models

Atlas	*Modality*	*Model Name*
Cobra	GM	$Cobra_{GM}$
	WM	$Cobra_{WM}$
Hammers	GM	$Hammers_{GM}$
	WM	$Hammers_{WM}$
	CSF	$Hammers_{CSF}$
Lpba40	GM	$Lpba40_{GM}$
Neuromorphometrics	GM	$Neuro_{GM}$
	CSF	$Neuro_{CSF}$

A2009	Gyrification	$a2009_{Gyri}$
	Thickness	$a2009_{Thick}$
DK40	Gyrification	$DK40_{Gyri}$
	Thickness	$DK40_{Thick}$
HCP	Gyrification	HCP_{Gyri}
	Thickness	HCP_{Thick}

For the preprocessed PET-PIB images, 19 regions of interest masks were applied, and the mean value of intensity extracted. These regions values were then normalized to dose injected and the body mass index (BMI), defined as (14):

$$BMI = \frac{weight\ (kg)}{height\ (m)^2} \tag{14}$$

Furthermore, the final set of features were generated using Standard Uptake Value Ratio (SUVR). The SUVR was calculated for all the regions of interest using different reference regions: Cerebellum, WM and the GM, and the final models are defined as $SUVR_{Cere}$, $SUVR_{WM}$ and $SUVR_{GM}$ respectively.

All three different regions were defined on the T1 MRI template ICBM152.

DTI features were generated following a similar structure to T1-weighted MR images, but instead of generating volumetric or surface values, in this case diffusion metrics were extracted from different label atlas: lpba40, (www.nitrc.org/frs/?group_id=214), Desikan, Destrieux, (www.nitrc.org/frs/?group_id=432), Hammers and JHU, (identifiers.org/neurovault.collection:264). Using the ExploreDTI software, (www.exploredti.com/), the diffusion metrics were generated from the different atlas.

Only the mean fractional anisotropy (FA) and mean diffusivity (MD) of the various atlas regions were considered as features.

Table 3.3 summarizes the results of DTI feature generation and how we define each model.

Table 3.3 DTI-based models.

Atlas	*Modality*	*Model Name*
Lpba40	FA	$Lpba40_{FA}$
	MD	$Lpba40_{MD}$
Desikan	FA	$Desikan_{FA}$
	MD	$Desikan_{MD}$
Destrieux	FA	$Destrieux_{FA}$
	MD	$Destrieux_{MD}$
Hammers	FA	$Hammers_{FA}$
	MD	$Hammers_{MD}$
JHU	FA	JHU_{FA}
	MD	JHU_{MD}

3.4. Feature Selection

After feature generation, feature selection methods were applied using R Study environment, https://www.r-project.org/, in order to select the best feature subset of each model. Due to the limited number of cases a limit of 8 features was set, 1 feature for each 5 cases to be selected.

In order to validate the selection, two different methods of feature selection were applied: a filter based method (FBM) and an embedded based method (EBM).

Both feature selection methods are composed of two phases, the first is a ranking phase here the features are ranked using a defined criterion and the second is a refined selecting phase that intends to remove irrelevant and redundant features.

The filter based method consists of using Student's t-test to rank each feature individually, accordingly to their t-value. To achieve this the data is divided into two groups, one with the CN cases and other with the AD cases, and the t-student test is applied for all the variables individually extracting the t and p values for each variable.

After sorting the features, from the highest to the lowest t-values, they were filtered, using the criteria of $p<0.05$. Only features with p values smaller than 0.05 are selected and used in the next phase.

Afterwards, a refined filter was applied, using the Pearson correlation (r). The Pearson correlation was calculated for all variables and a nxn matrix created. In this matrix each row represents a feature and the colons the Pearson correlation between the variable and all the others.

Using this correlation matrix, we removed all the features that had an absolute correlation value greater than 0.55 ($|r| > 0.55$) in relation to the most important feature, which is the one with highest t-value and placed at the first row of the correlation matrix.

Furthermore, the resultant features were evaluated, from the second row until the n-1 row. The cut-off criteria in this case was an absolute correlation value greater than 0.70 ($|r| > 0.70$), those that exceed that value were excluded (Figure 3.1 a).

Theoretically, the main objective of this refined filter is to try to achieve orthogonality between the main feature and all the other features as well as to remove the redundant information from features.

The embedded based method is similar to the filter based method, applying the same refined filter technique. The main difference is on how the features are ranked.

In this method the features are ranked according to the importance of each variable, which is given by a classification algorithm, in this case, we used random forest.

All variables were divided in sets of 8 features constructed randomly and the data was divided in two groups to simulate the final environment that will be used to train the final classifiers. The learning algorithm was used on the different sets of features and the importance of each feature stored. This process was repeated 2000 times and the final value of importance was established as the mean importance value of each feature.

After extracting the final importance value, the features were sorted form the biggest to the smallest importance values and a filter was applied to remove all the features that exhibited an importance value lesser than 0.55 times of the most important feature. This was done to prevent the use of possible irrelevant features (Figure 3.1 b).

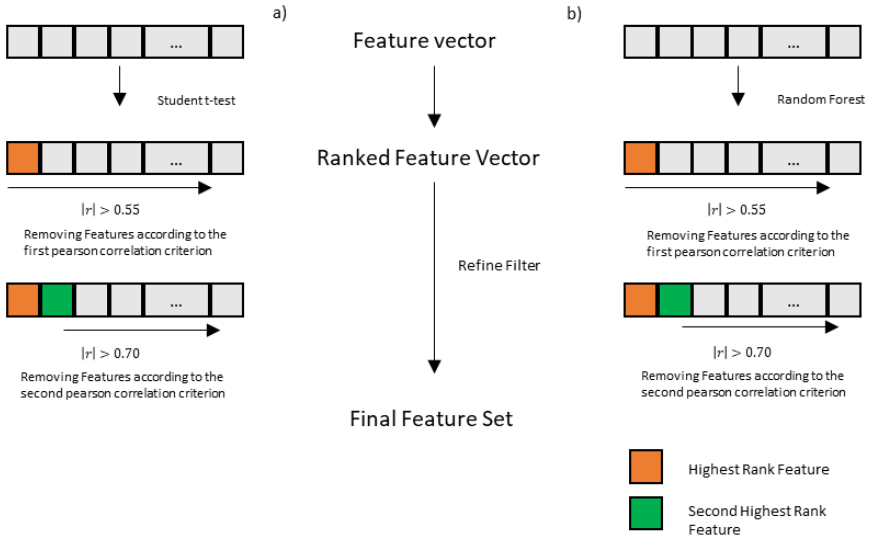

Figure 3.1 Illustration of the feature selection methods used. a) Filter Based Method (FBM). b) Embedded Based Method (EBM).

3.5. Classification

All the classifiers were constructed using Python 3.7, https://www.python.org/, and resorting to the scikit-learn package, https://scikit-learn.org/stable/.

Before any learning algorithm has been applied, all feature vectors were standardized to zero mean and variance one. This was achieved by subtracting the mean to each feature vector, and then dividing it by the standard deviation (15). This was meant to improve the performance of the learning algorithm.

$$x' = \frac{x - \bar{x}}{\sigma} \tag{15}$$

With all the data standardized, the data was split into two groups, 80% used for training and 20% for testing. The splitting was performed such that both groups were as balanced as possible.

The training group was evaluated, using 8-fold cross validation, and assured that every fold was balanced. The best performance model from the training set was then selected using the criteria of best accuracy and validated using the test data (Figure 3.2). The values

of accuracy, sensitivity, specificity and the ROC curve and AUC, from this last evaluation where stored and used to evaluate the final performance of the model.

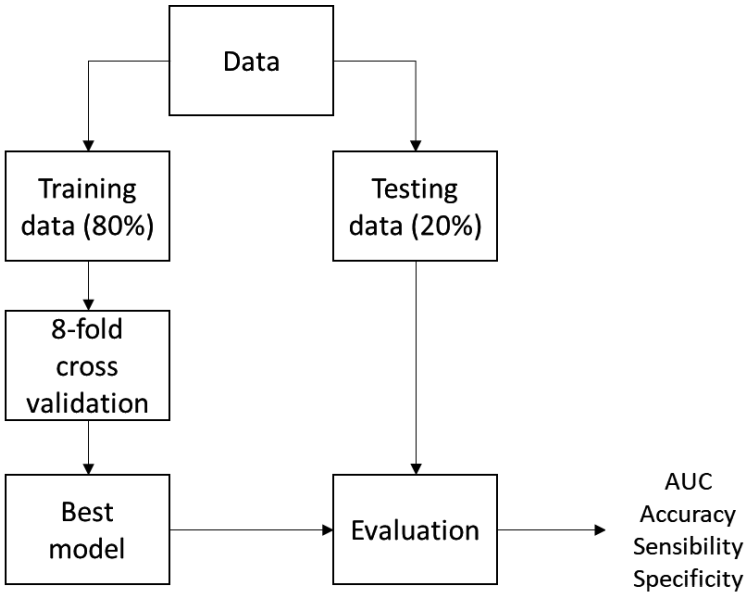

Figure 3.2 Scheme of the process used for validation of all models.

This process was repeated 2000 times, and the classifiers overall performance was primarily evaluated using the mean ROC curve and the mean AUC.

Furthermore, parameters such as mean accuracy, sensitivity and specificity were used in order to distinguish the classifiers that have similar performance measurements, as well as to select those that will be used in the ensemble phase.

The learning algorithms applied were the support vector machine (SVM) and the logistic regression.

SVM classifier was constructed using the radial basis function (RBF) kernel, the gamma value set to "scale" and the value of C set to the default value of 0.1.

The logistic regression was constructed with the random state set to 0 and the multi class set to "auto", and was solved using the "lbfgs" solver, taking into account that all of the other parameters were left as default.

The goal here was to construct three robust base classifiers, one for each image modality, that can be used in the ensemble phase.

3.6. External Validation

In order to perform external validation, we gather another dataset, obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu).

The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

The data downloaded from ADNI database, were organized into three groups of data that were constructed aiming to validate. The demographics information of each group is summarized in Table 3.4.

Table 3.4 Demographics for the ADNI external data.

	CN (n=164)	*AD (n=166)*
Age (years)	76.5±5.9	75.1±7.9
Gender (male/female)	77/87	83/83
a) External MRI		
	CN (n=31)	AD (n=24)
Age (years)	79.5±5.8	74.4±8.0
Gender (male/female)	19/12	18/6
b) External PIB		
	CN (n=71)	AD (n=77)
Age (years)	72.7±7.2	73.7±8.4
Gender (male/female)	28/43	40/37
c) External DTI		

Age value is defined as mean ± standard deviation

To evaluate the performance of our classifiers, the most promising classifiers were test with these external data. For each modality, the external data were processed in the

same way as before. Following this, the classifiers were evaluated from the most promising to the least until one, from each image modality, could perform in a similar pattern as in the internal data.

For each classifier, the performance of each classifier was evaluated using a similar strategy as before (Figure 3.2). Here 100 models were create using the previous method and the one with the best performance selected to be validated with the external data. This process was repeated 1000 times and for each run the values of accuracy, sensitivity, specificity and balanced accuracy stored. As in the previous method the final performance of the classifier was defined as the mean values of the performance parameters extracted.

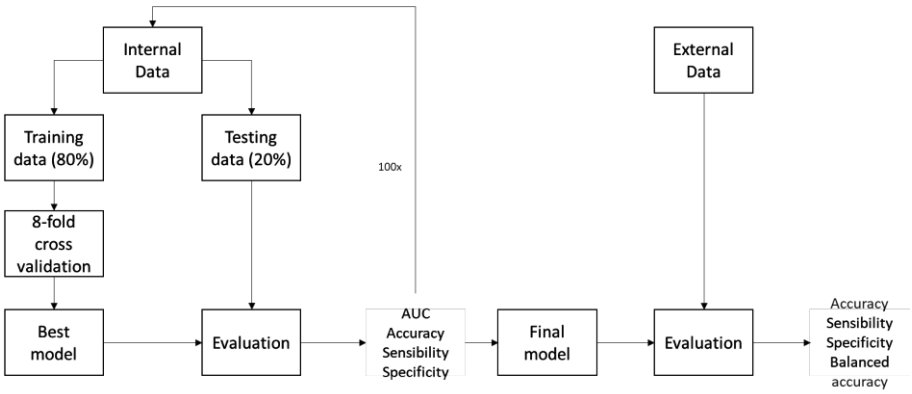

Figure 3.3 Scheme of the process used for external validation of the models.

3.7. Ensemble Classifiers

The performance of the final classifiers was tested with all the internal data and the potential of combining all the image modalities was assessed.

To ensemble the classifiers a non-generative weighted fusion technique was used, and the previously constructed and validated classifiers were combined in order to generate a final decision.

The base classifiers were combined using a weight value between 0 and 1, that was applied to the predicted probability, given from each of the base learning algorithm.

$$Y_i = c_1 y_{1_i} + c_2 y_{2_i} + c_3 y_{3_i} \quad (16)$$

In equation (16), Y_i stands for the final ensemble prediction probability, y_{1_i} for the predicted probability, given by the classifier trained using a MRI-based model, y_{2_i} for the predicted probability, given by the classifier trained using a PIB-based model and y_{3_i} for the predicted probability, given by the classifier trained using a DTI-based model. Lastly, c_1, c_2 and c_3 are weights applied to the different modalities, such as $c_1+c_2+c_3 = 1$.

In order to evaluate the overall performance of the ensemble method, the individual classifiers, one for each modality, were constructed as in the same way as before (Figure 3.2). As in the external validation, also here, 100 models were constructed for each classifier and the one with the best performance selected to predict the ensemble data. The prediction of each classifier were then combined using the equation (16) and the performance parameters of accuracy, sensibility and specificity stored. The final performance of the ensemble method was defined as the mean value of the performance parameters. Furthermore, the mean ROC curves from the individual classifiers and the ensemble classifier where also calculated.

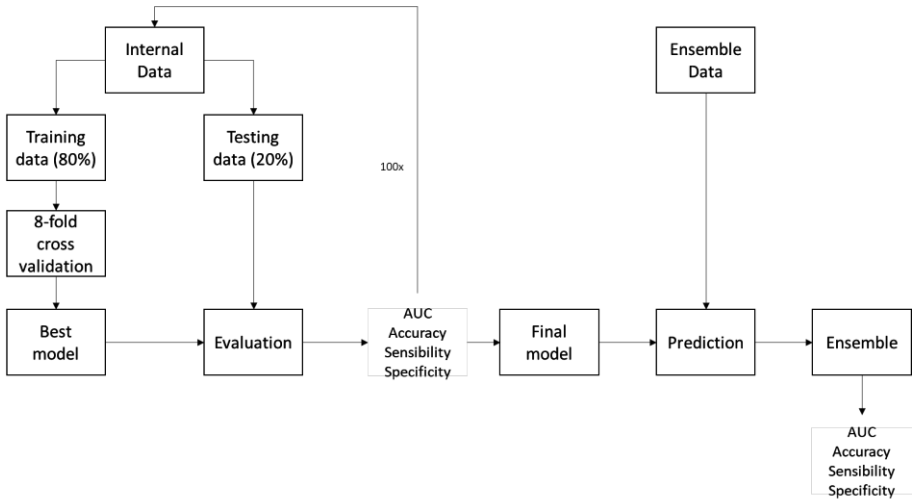

Figure 3.4 Scheme of the process used for validation of the ensemble technique.

The validation of the ensemble technique was also tested using the external data, but unfortunately, only the validation through the combination of MRI and DTI models was possible from the external data collected.

4. RESULTS AND DISCUSSION

4.1. Feature Selection

In this section we present the results of the feature selection methods.

Due to the high number of models tested, Table 4.1 shows a summary of the results. The full list can be consulted in Table S.1.

Table 4.1 Summary of the feature selection methods

Model	Feature Selection Method	Number of total Features	Final Number of Features
$Neuro_{GM}$	EBM	142	41
	FBM		10
HCP_{Thick}	EBM	362	202
	FBM		54
$Desikan_{FA}$	EBM	84	58
	FBM		6
$SUVR_{WM}$	EBM	19	2
	FBM		3

EBM: Embedded Based Method; FBM: Filter Based Method

From Table 4.1 it is possible to conclude that the general tendency is that the embed based method (EBM) allows the selection of more features in relation to the filter based method (FBM). The exception is the $SUVR_{WM}$ model. These tendencies remain true for all the other models, not shown in this table, but presented in Table S.1.

Another important subject to analyse is about the features that are selected by each method.

In Figure 4.1, we present the results of the ranking phase of the two feature selection methods for the $Neuro_{GM}$model. Due to the number of features we only choose to present the top 15 features.

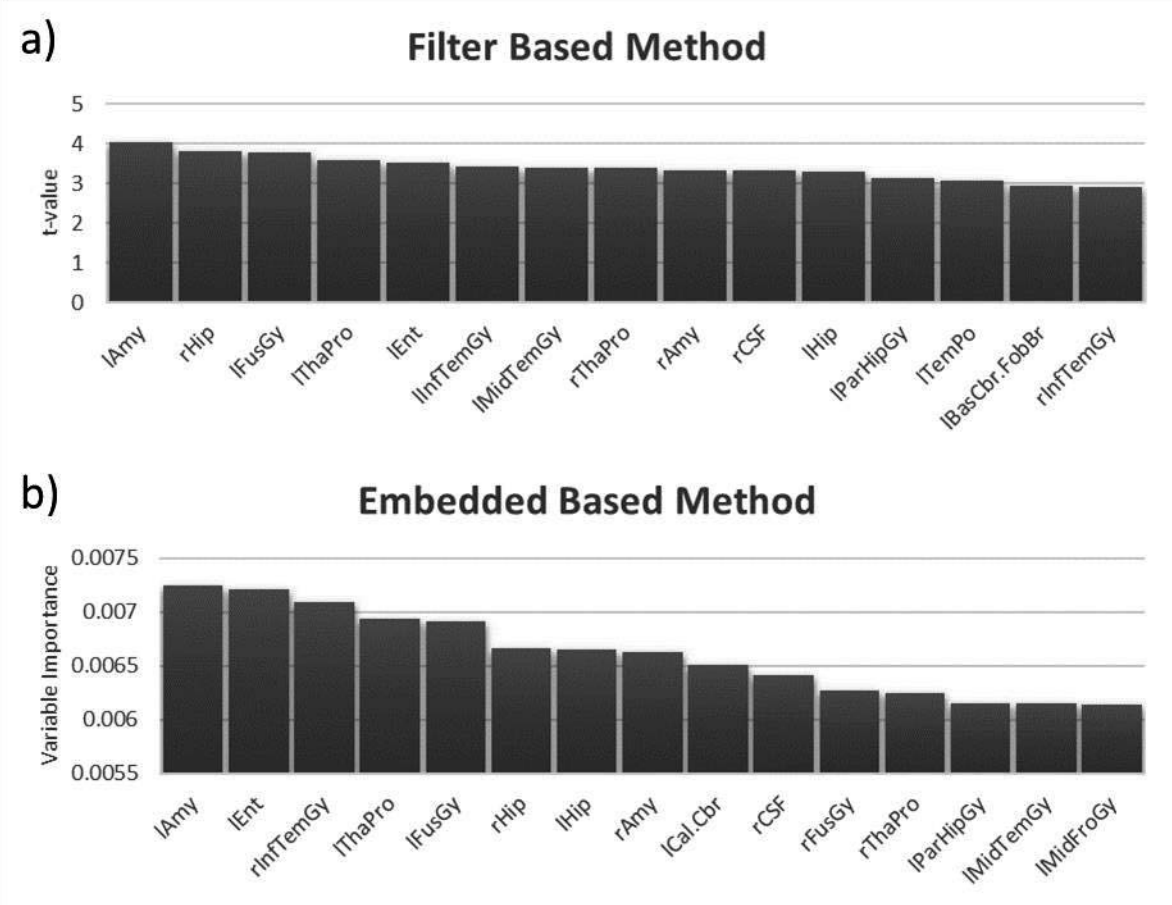

Figure 4.1 Results of feature ranking for $\boldsymbol{Neuro_{GM}}$ model, a) FBM features ranking. b) EBM features ranking. A glossary for the acronym's names can be found in Table S.5.

Although in different order, is possible to perceive from Figure 4.1 that both methods considered the same bag of features as the most relevant. The fact that two different methods consider the same features as the best, allows us to the be more confident that, for each model, the final set of features will be composed of the most relevant for the problem in question.

Furthermore, due to ours feature selection method refined filter phase, the final sets of features depend on the most important feature and might differ slightly or in some cases differ a lot. Figure 4.2 shows the final feature set of the previous model shown. It is possible to notice the difference between the two final sets even though they rank the same feature as the most important.

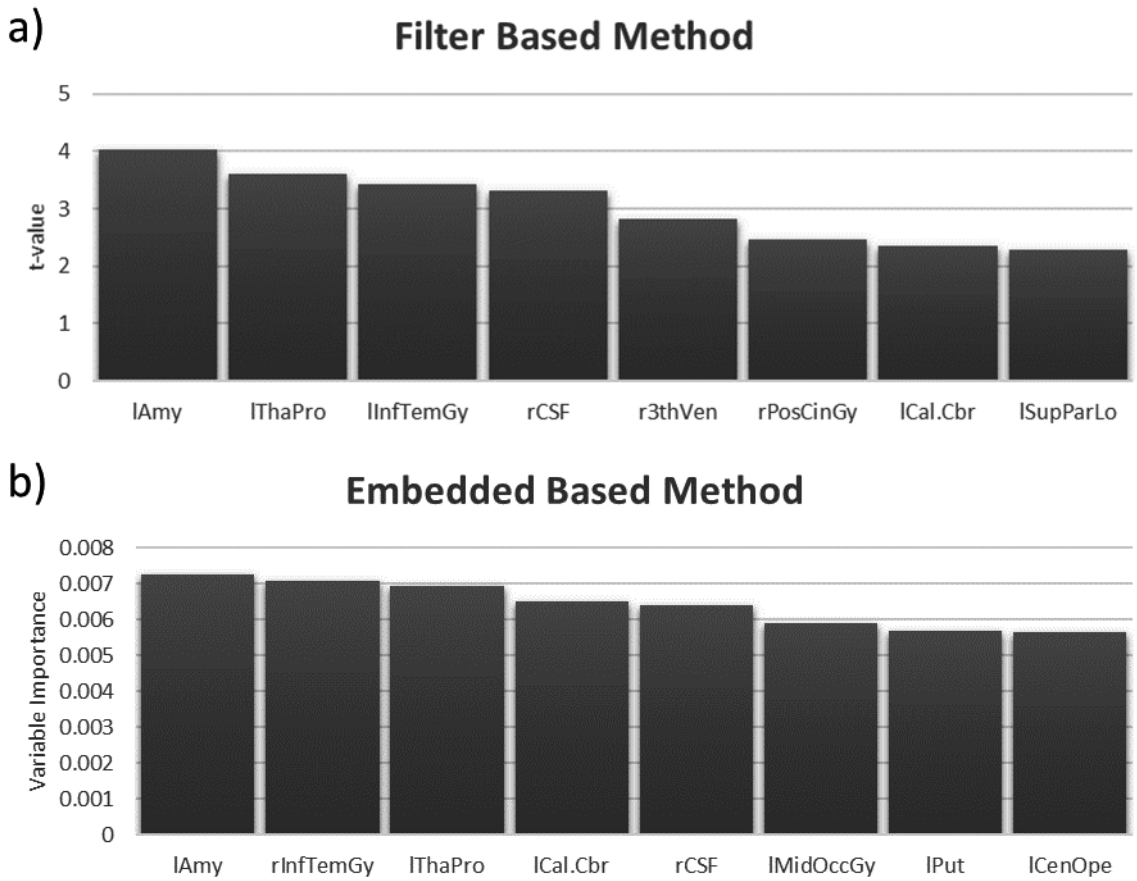

Figure 4.2 Final feature set for the $\boldsymbol{Neuro_{GM}}$ model. a) FBM final set of features. b) EBM final set of features. A glossary for the acronym's names can be found in Table S.5.

Although only the results from one model are being presented, all of the other models follow the same trend, here the feature ranking method selects the same bag of features as the most important and the final set differ from each other due to the refined filter phase. The remaining results can be consulted in Supplement A.

4.2. Base Classifiers Evaluation

In this section we present the performance of the base classifier's models for AD/CN classification, using SVM and Logistic Regression (LR).

To facilitate the analysis this section is divided into three subsections, each one corresponding to modality of image. The full list of the classifier's performance can be found in Table S.6.

4.2.1. MRI-Based Classifiers

Table 4.2 and Table 4.3 present a summary of the performance of the volume-based and surface-based classifiers, respectively. For each model, the area under the Receiving Operating Characteristic (ROC) curve (AUC), mean accuracy, mean sensitivity and mean specificity are shown. The full data can be found in Table S.6.

Table 4.2 Summary of the classification performance of volume-based classifiers.

Model	Feature Selection Method	SVM				Logistic Regression			
		AUC	Mean Accuracy	Mean Sensitivity	Mean Specificity	AUC	Mean Accuracy	Mean Sensitivity	Mean Specificity
$Cobra_{GM}$	**EBM**	0.87±0.13	76.13%	76.06%	76.19%	0.84±0.14	76.14%	73.28%	79.00%
	FBM	0.87±0.13	82.08%	73.63%	90.54%	0.83±0.14	72.76%	74.09%	71.44%
$Hammers_{GM}$	**EBM**	0.85±0.15	74.79%	66.14%	83.44%	0.86±0.12	78.68%	74.60%	82.75%
	FBM	0.76±0.21	72.92%	67.26%	78.57%	0.83±0.14	73.09%	71.51%	74.68%
$Lpba40_{GM}$	**EBM**	0.76±0.21	70.28%	61.16%	79.40%	0.78±0.16	70.21%	66.65%	73.78%
	FBM	0.77±0.17	73.93%	67.23%	80.64%	0.85±0.13	73.99%	76.66%	71.32%
$Neuro_{GM}$	**EBM**	***0.96±0.07**	***92.05%**	86.78%	97.32%	0.93±0.09	85.13%	76.98%	93.27%
	FBM	***0.93±0.09**	***86.02%**	82.70%	89.34%	0.95±0.08	86.98%	84.98%	89.53%
$**Cobra_{WM}$	**EBM**	0.87±0.13	76.13%	76.06%	76.19%	0.92±0.10	80.64%	78.41%	79.00%
	FBM	0.86±0.13	77.06%	73.63%	90.54%	0.91±0.10	79.34%	77.54%	81.54%

AUC value is defined as (mean ± standard deviation)

SVM performance was evaluated primarily by the AUC values and was divided in two (Table 4.2). SVM EBM shows a similar median performance for $Cobra_{GM}$ (0.87±0.13) and the $Hammers_{GM}$ (0.85±0.15) models, and ranges from $Lpba40_{GM}$ (0.76±0.21) to $Neuro_{GM}$ (0.96±0.07) model. Regarding SVM FBM it can be notice that $Hammers_{GM}$ and $Lpba40_{GM}$ presented the smallest AUC (0.76±0.21) and (0.77±0.17), respectively. The $Neuro_{GM}$ (0.93±0.09) model performs the best, and $Cobra_{GM}$ (0.87±0.13) has a median performance.

LR performance was evaluated in the same way as SVM (Table 4.2). LR EBM shows a similar pattern to SVM EBM with the smallest performance for $Lpba40_{GM}$ (0.78±0.16) model, two median performance models $Cobra_{GM}$ (0.84±0.14) and

$Hammers_{GM}$ (0.86±0.12) and a higher performance $Neuro_{GM}$ (0.93±0.09). LR FBM shows three similar performance $Cobra_{GM}$ (0.83±0.14), $Hammers_{GM}$ (0.83±0.14) and $Lpba40_{GM}$ (0.85±0.13) models and a higher performance $Neuro_{GM}$ (0.95±0.08).

Figure 4.3 depicts the mean ROC curves for each analysis presented in Table 4.2. Graphs a) and c) were derived from SVM using the EBM and FBM feature selection methods respectively. Graphs b) and d) were derived from LR using EBM and FBM feature selection methods respectively.

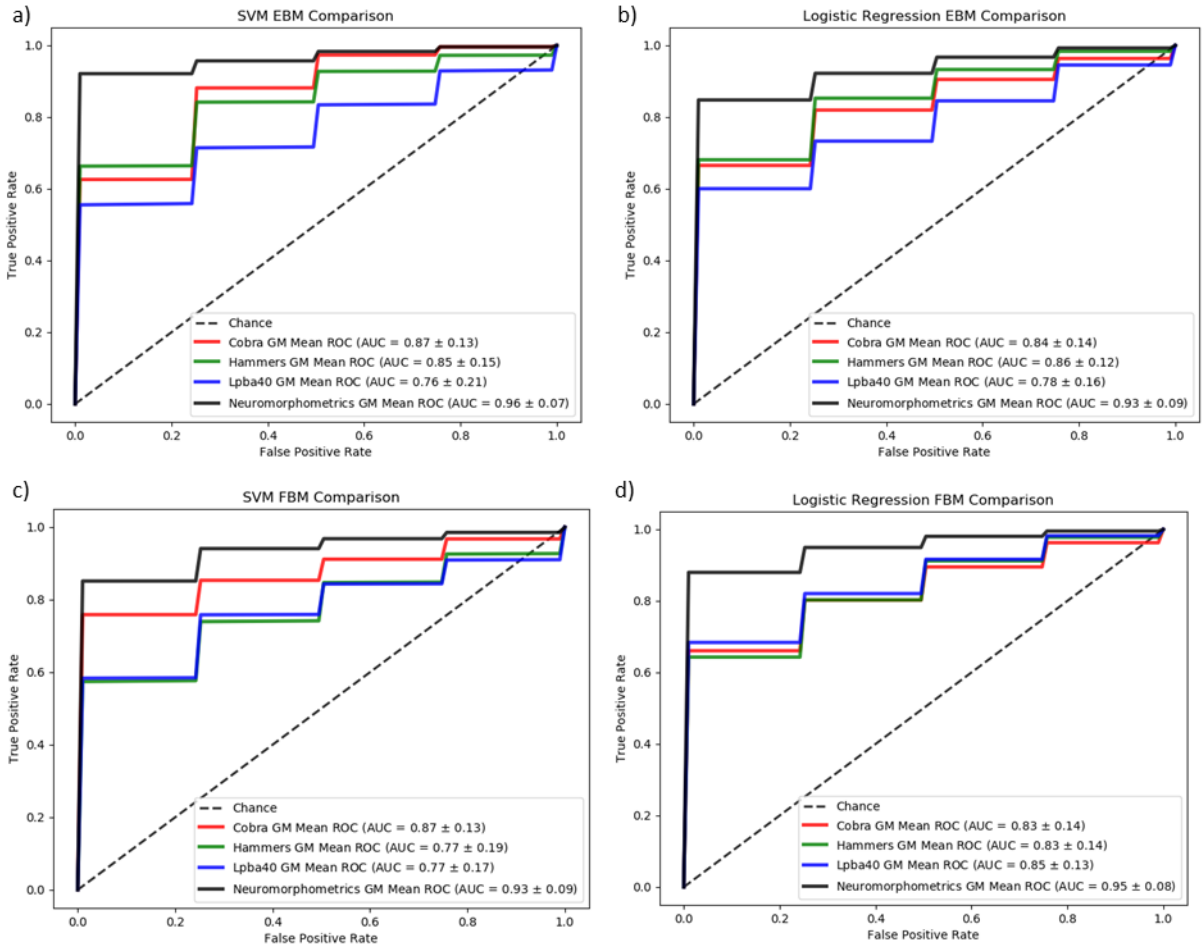

Figure 4.3 Mean ROC curve of Volume-based models. a) SVM classifiers using EBM feature set. b) Logistic Regression classifiers using EBM feature set. c) SVM classifiers using FBM feature set. d) Logistic Regression classifiers using FBM feature set.

Table 4.3 Summary of the classification performance of surface-based classifiers.

Model	Feature Selection Method	SVM				Logistic Regression			
		AUC	Mean Accuracy	Mean Sensitivity	Mean Specificity	AUC	Mean Accuracy	Mean Sensitivity	Mean Specificity
$a2009_{Gyri}$	**EBM**	0.91±0.10	83.01%	86.98%	79.04%	0.89±0.11	81.91%	81.84%	81.99%
	FBM	0.85±0.13	76.45%	77.70%	75.20%	0.89±0.11	81.37%	83.37%	79.46%
$a2009_{Thick}$	**EBM**	0.90±0.11	81.89%	81.93%	81.87%	0.92±0.10	84.78%	84.01%	85.54%
	FBM	0.84±0.14	81.47%	82.76%	80.18%	0.85±0.13	80.13%	79.53%	80.74%
$DK40_{Gyri}$	**EBM**	0.76±0.18	69.13%	73.14%	65.11%	0.80±0.15	67.88%	66.79%	68.98%
	FBM	0.78±0.17	69.57%	73.99%	65.15%	0.82±0.14	68.84%	69.65%	68.03%
HCP_{Thick}	**EBM**	***0.93±0.09**	***84.28%**	83.055	85.495	0.89±0.11	83.16%	80.71%	85.60%
	FBM	***0.93±0.09**	***88.07%**	84.59%	91.55%	0.93±0.08	87.02%	85.15%	88.89%

AUC value is defined as (mean ± standard deviation)

The surface-based classifiers (Table 4.3) were evaluated in the same way as the volume-based classifiers (Table 4.2).

SBM EBM shows a $DK40_{Gyri}$ model with the smallest AUC (0.76±0.18) and three similar high performances for the $a2009_{Gyri}$ (0.91±0.10), $a2009_{Thick}$ (0.90±0.11) and HCP_{Thick} (0.93±0.09) models. Regarding SVM FBM it is possible to notice a similar performance for $a2009_{Gyri}$ (0.85±0.13) and $a2009_{Thick}$ (0.84±0.14) models. The $DK40_{Gyri}$ model performs the worst (0.78±0.17) and the HCP_{Thick} (0.93±0.09) model performs the best.

LR EBM shows a similar pattern to SVM EBM with three similar high performances from the $a2009_{Gyri}$ (0.89±0.11), $a2009_{Thick}$ (0.92±0.10) and HCP_{Thick} (0.89±0.11) models and a $DK40_{Gyri}$ model with the smallest AUC value (0.80±0.15). Lastly, LR FBM as SVM FBM shows a similar median performance for $a2009_{Gyri}$ (0.89±0.11) and $a2009_{Thick}$ (0.85±0.13) models, and ranges from $DK40_{Gyri}$ (0.82±0.14) to HCP_{Thick} (0.93±0.08).

Figure 4.4 illustrates the mean ROC curves for each analysis presented in Table 4.3. Graphs a) and c) were derived from SVM using the EBM and FBM feature selection methods respectively. Graphs b) and d) were derived from LR using EBM and FBM feature selection methods respectively.

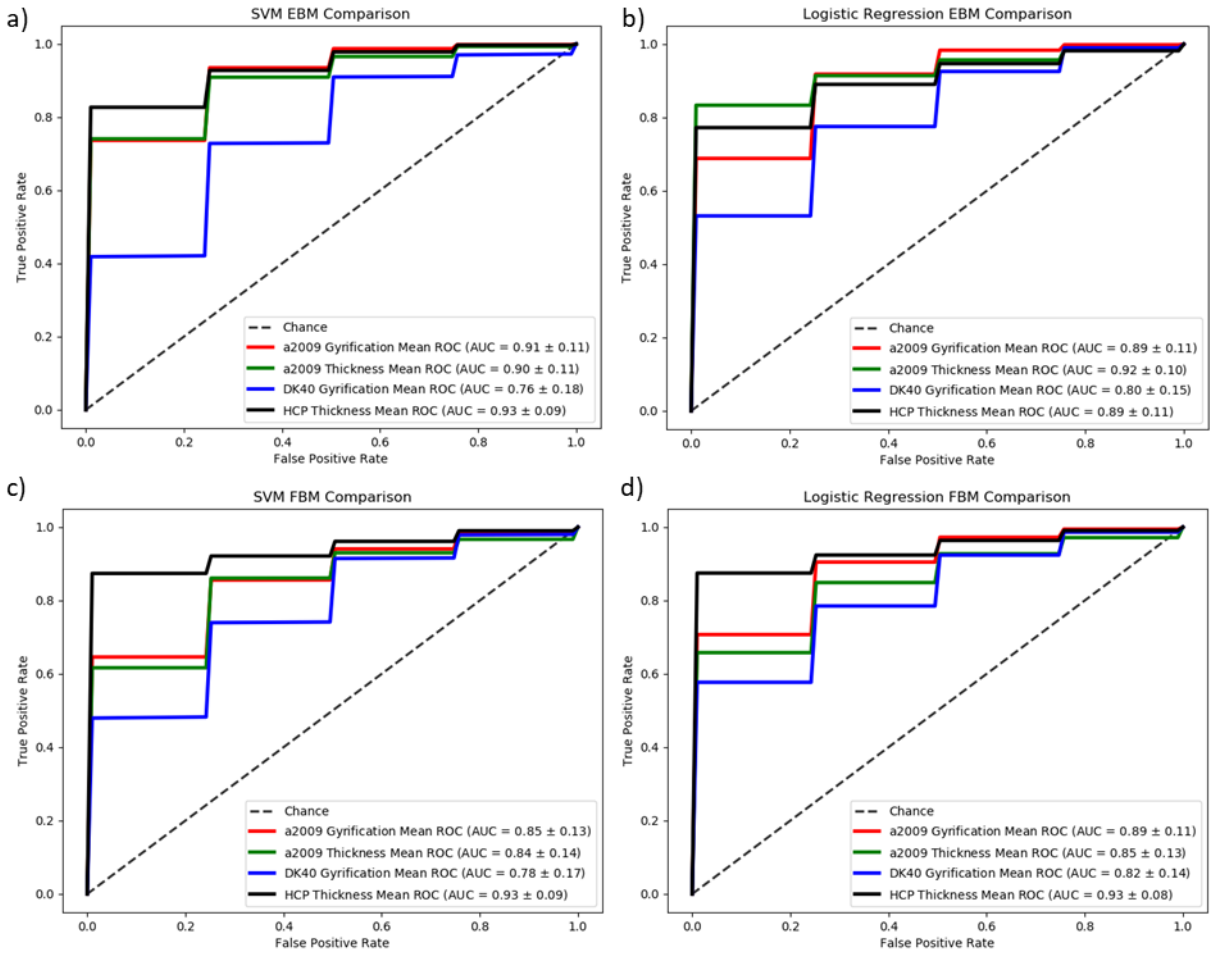

Figure 4.4 Mean ROC curve of Surface-based classifiers. a) SVM classifiers using EBM feature set. b) Logistic Regression classifiers using EBM feature set. c) SVM classifiers using FBM feature set. d) Logistic Regression classifiers using FBM feature set.

The results shown that our MRI-based classifiers have an overall great potential in the AD/CN classification. Also, when comparing the performance of our models with those used in the state of the art, our best performing model for both volume (92.05%) and surface-based (88.07%) methods has shown a similar performances to those used in the state of the art with performance intervals of (84.00% and 94.50%)[9,22,33,55] and (86.70 and 95.00%)[9] respectively.

Our models show in general high accuracy and a better specificity than sensitivity. Furthermore, it is possible to see that EBM models have the tendency of performing in a similar or better way than the FBM models which was also expected.

Regarding the learning algorithms we did not see a great disparity between SVM and LR models. Both have similar performances, though, we can make the case that overall

the SVM models can have a slight advantage in their performance when considering the mean accuracy, sensitivity and specificity.

Even with no optimization of our classifier's parameters, they have achieved an excellent ability for AD/CN classification and can only be improved.

One model present in Table 4.2 and which is not mentioned in the analysis is the $Cobra_{WM}$ model. WM models are not usually used in the study of AD/CN classification. However, the $Cobra_{WM}$ model has shown a performance similar to the GM models and the study of the WM regions selected for this classifier (Supplement A) could be an interesting study topic.

4.2.2. PIB-Based Classifiers

Similarly to the previous analyses we start by presenting in Table 4.4 the results of the classification performance of PIB-based classifiers. For each model we present the area under the Receiving Operating Characteristic (ROC) curve (AUC), mean accuracy, mean sensitivity and mean specificity. The full data can be found in Table S.6.

Table 4.4 Classification performance of PIB based models.

Model	Feature Selection Method	SVM				Logistic Regression			
		AUC	Mean Accuracy	Mean Sensitivity	Mean Specificity	AUC	Mean Accuracy	Mean Sensitivity	Mean Specificity
$SUVR_{Cere}$	**EBM**	0.94±0.10	86.96%	96.03%	80.16%	0.96±0.07	91.34%	97.13%	87.0%
	FBM	0.95±0.08	87.55%	95.77%	81.59%	0.97±0.05	87.99%	93.10%	84.16%
$SUVR_{GM}$	**EBM**	***0.97±0.06**	***92.34%**	98.63%	87.63%	0.95±0.07	88.12%	93.05%	84.43%
	FBM	0.97±0.06	87.66%	96.27%	81.20%	0.96±0.07	87.65%	91.60%	84.69%
$SUVR_{WM}$	**EBM**	***0.93±0.10**	***90.53%**	92.00%	89.43%	0.95±0.07	89.81%	88.93%	90.48%
	FBM	***0.93±0.11**	***93.80%**	99.63%	89.43%	0.96±0.07	90.81%	90.93%	90.71%

AUC value is defined as (mean ± standard deviation)

Here both SVM and LR analyses, as well as EBM and FBM analyses, shows similar performing models. The values of AUC for all models are always above 0.90 and very close to each other.

Like in the previous analysis, Figure 4.5 shows the mean ROC curves for each analysis presented in Table 4.4. Graphs a) and c) were derived from SVM using the EBM and FBM feature selection methods respectively. Graphs b) and d) were derived from LR using EBM and FBM feature selection methods respectively.

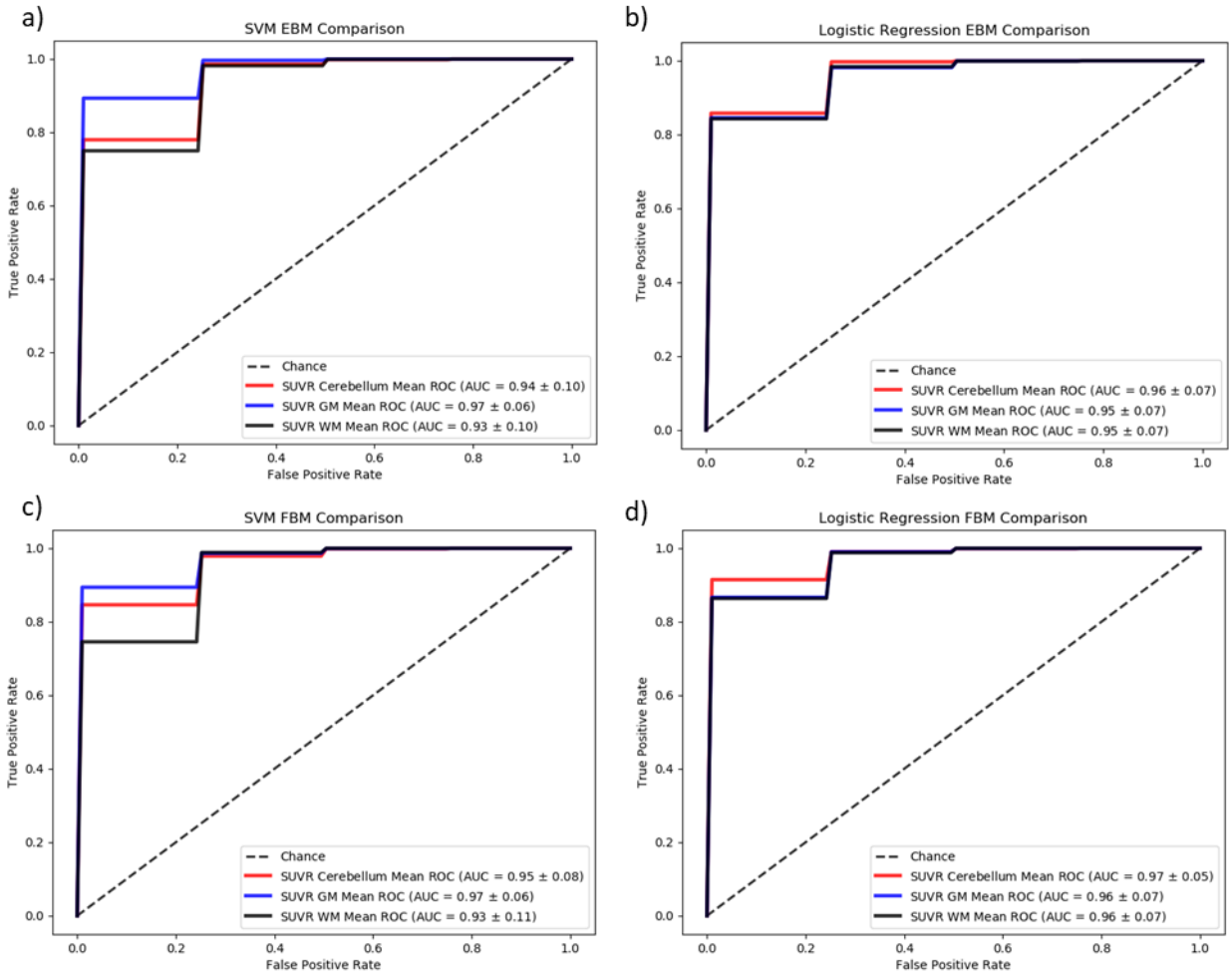

Figure 4.5 Mean ROC curve of PBI-based classifiers. a) SVM classifiers using EBM feature set. b) Logistic Regression classifiers using EBM feature set. c) SVM classifiers using FBM feature set. d) Logistic Regression classifiers using FBM feature set.

Even though our classifiers were not optimized, PIB-based classifiers demonstrate a great ability in the AD/CN classification with excellent values of AUC for all the models.

These classifiers show an overall high value of accuracy and contrarily to MRI-based classifiers a higher sensitivity compare to specificity.

SVM and LR have similar performances and as in the MRI-based methods, and as in the previous results, we can see a general better performance with SVM than with LR.

Furthermore, all the regions used as reference show an equal great potential to be used in classification problems

In addition, all our classifiers shown performances that are in accordance with those find in the in the state of the art with performances between 85.70% and 96.00%[28,42,46].

4.2.3. DTI-Based Classifiers

In Table 4.5 we present a summary of the performance of DTI-based classifiers. For each model we present the area under the Receiving Operating Characteristic (ROC) curve (AUC), mean accuracy, mean sensitivity and mean specificity. The full data can be found in Table S.6.

Table 4.5 Summary of the classification performance of DTI based classifiers.

Model	Feature Selection Method	SVM				Logistic Regression			
		AUC	Mean Accuracy	Mean Sensitivity	Mean Specificity	AUC	Mean Accuracy	Mean Sensitivity	Mean Specificity
$Desikan_{FA}$	**EBM**	***0.86±0.14**	***79.84%**	76.17%	82.09%	0.81±0.17	70.77%	78.16%	73.54%
	FBM	0.86±0.14	79.06%	71.43%	84.78%	0.83±0.16	77.41%	73.65%	80.23%
$Destrieux_{MD}$	**EBM**	0.84±0.15	77.29%	72.12%	81.18%	0.82±0.15	73.54%	65.37%	79.68%
	FBM	***0.92±0.10**	***88.12%**	81.67%	92.96%	0.92±0.11	87.92%	83.13%	91.51%
JHU_{FA}	**EBM**	0.80±0.19	76.69%	66.15%	84.59%	0.78±0.17	74.26%	67.72%	79.18%
	FBM	***0.83±0.17**	***79.55%**	71.55%	86.16%	0.81±0.17	76.35%	71.78%	79.77%
$Lpba40_{MD}$	**EBM**	0.85±0.15	75.02%	67.57%	80.61%	0.83±0.16	73.71%	63.70%	81.21%
	FBM	0.75±0.21	73.36%	67.45%	77.80%	0.83±0.15	75.06%	69.65%	79.11%

AUC value is defined as (mean ± standard deviation)

As similarly to the previous analyses both SVM and LR assessment were performed in both feature selection methods and the AUC was the criteria chosen to primarily evaluate their performance (Table 4.5).

SVM EBM shows an overall similar performance between all models with AUC values ranging from 0.80 to 0.90. The JHU_{FA} (0.80±0.19) model seems to have a slight lower performance than the $Desikan_{FA}$ (0.86±0.14), $Destrieux_{MD}$ (0.84±0.15) and $Lpba40_{MD}$ (0.85±0.15) models. SVM FBM analyses reveal a lower performing $Lpba40_{MD}$ (0.75±0.21) model, one median performing JHU_{FA} (0.83±0.17) model and two higher performing $Desikan_{FA}$ (0.86±0.14) and $Destrieux_{MD}$ (0.92±0.10) models.

LR EBM analyses shows that all the models have similar performance. The four models have AUC values around 0.80. The LR FBM analyses show three equally performing for the $Desikan_{FA}$ (0.83±0.16), JHU_{FA} (0.81±0.17) and $Lpba40_{MD}$ (0.83±0.15) models and a slightly higher performance for the $Destrieux_{MD}$ (0.92±0.11) model.

Figure 4.6 illustrates the mean ROC curves for each analysis presented in Table 4.5. Graphs a) and c) were derived from SVM using the EBM and FBM feature selection methods respectively. Graphs b) and d) were derived from LR using EBM and FBM feature selection methods respectively.

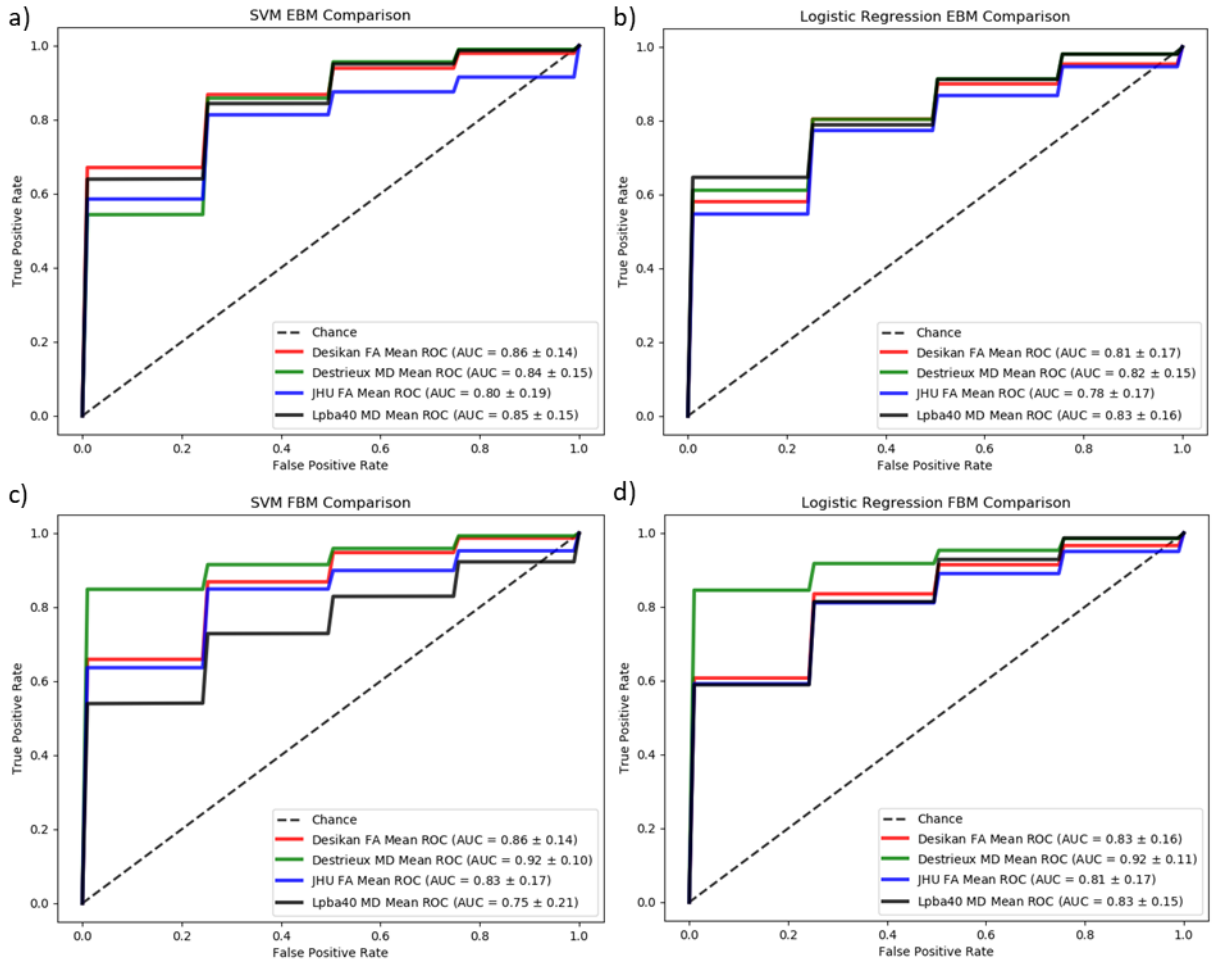

Figure 4.6 Mean ROC curve of DTI-based classifiers. a) SVM classifiers using EBM feature set. b) Logistic Regression classifiers using EBM feature set. c) SVM classifiers using FBM feature set. d) Logistic Regression classifiers using FBM feature set.

The DTI-based classifiers show a lower overall performance than MRI-based and PIB-based classifiers in AD/CN classification but follow the same pattern as the MRI-based classifiers with higher specificity and lower sensitivity.

SVM classifiers show a better overall performance than the LR classifiers.

Contrarily to the MRI and PIB-based models, the FBM methods show a slightly better performance than the EBM, opposing the general notion that EBM models perform better than FBM models.

Alone, these models show some potential for AD/CN classification, though, better feature generation methods and classifier optimization is needed to possibly achieve the performance of both the MRI and PIB-based classifiers. However, it is noteworthy to

point out that we managed to achieve a comparable and equal performance to those used in state DTI-based classifiers with performance between 76.00% and 89.00%[9,14,60].

The global analyse of all the base classifiers seems to favour the use of SVM rather than LR. With this in mind, in the ensemble method we will use SVM based classifiers to stud the complementarity between the data. Furthermore, the most promising SVM models were tested, by order of performance, until one achieves a reasonable performance with the external data.

4.3. External Validation

This section is dedicated to present the performance results from the most promising based classifiers using the ADNI external data. Table 4.6 presents the most promising classifiers from each modality. All the classifiers in Table 4.6 were SVM based.

Table 4.6 Best performing classifiers for each modality.

Image Modality	*Models*	*AUC*	*Mean Accuracy*	*Mean Sensitivity*	*Mean Specificity*
MRI	$Neuro_{GM}$ EBM	0.96±0.07	92.05%	86.78%	97.32%
	HCP_{Thick} FBM	0.93±0.09	88.07%	84.59%	91.55%
	$Neuro_{GM}$ FBM	0.93±0.09	86.02%	82.70%	89.34%
PIB-PET	$SUVR_{GM}$ EBM	0.97±0.06	92.34%	98.63%	87.63%
	$SUVR_{WM}$ FBM	0.93±0.11	93.80%	99.63%	89.43%
	$SUVR_{WM}$ EBM	0.93±0.10	90.53%	92.00%	89.43%
DTI	$Destrieux_{MD}$ FBM	0.92±0.10	88.12%	81.67%	92.96%
	$Desikan_{FA}$ EBM	0.86±0.14	79.84%	76.17%	82.09%
	JHU_{FA} FBM	0.83±0.17	79.55%	71.55%	86.16%

AUC value is defined as (mean ± standard deviation)

Starting with the MRI-based classifiers (Table 4.6 a)) we tested the $Neuro_{GM}$ EBM classifier. Using this classifier in the external data we obtain a 78.19% accuracy, 74.12% sensitivity, 82.29% specificity and a 78.20% balanced accuracy. Furthermore, the mean ROC curve was also calculated showing an AUC value of 0.85±0.02.

These results show that the $Neuro_{GM}$ EBM classifier has a lower performance with the external data than with the internal data as expected. However, it is still a good result specially since the classifier was trained with a sample that only have 40 cases and tested with a sample that have 330 cases. Furthermore, the classifier presents the same performance pattern in both internal and external data in what concerns specificity and sensitivity values. Figure 4.7 shows the comparison of the performance of the $Neuro_{GM}$ EBM classifier between internal and external data.

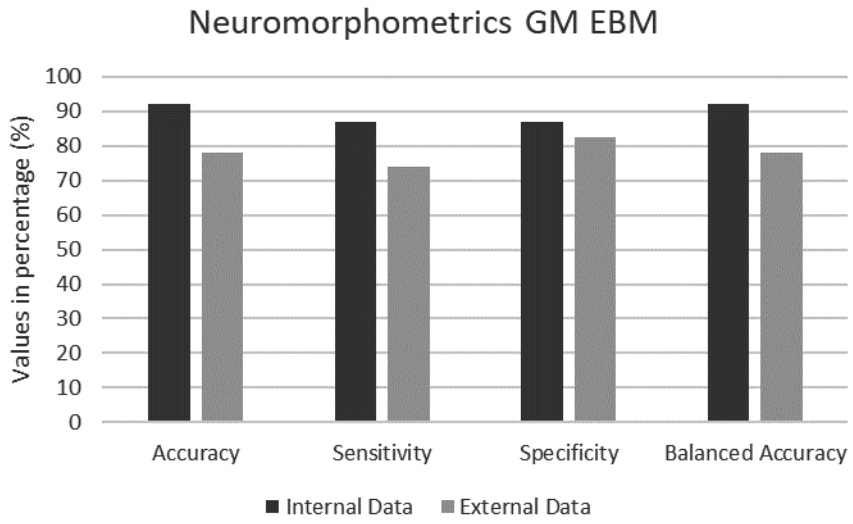

Figure 4.7 Comparison of the performance of the $\boldsymbol{Neuro_{GM}}$ EBM classifier. Internal data balanced accuracy is considered as the same as the accuracy since the internal data is almost perfectly balanced.

With these results, and for time reasons, we opt to not test more MRI-based classifiers besides the $Neuro_{GM}$ EBM classifier which will be used to represent the MRI modality in the ensemble phase.

Following the analysis of the MRI-based classifiers, we analysed the PIB-based classifiers. From Table 4.6 b), we start by evaluating the $SUVR_{GM}$ EBM. Using this classifier in external data we obtained an accuracy of 59.73%, a sensitivity of 63.93%, a specificity of 56.48%, a balanced accuracy of 60.21% and an AUC of 0.71±0.03. This classifier had a disappointing performance in the external data, far from what we originally expected. However, it has shown the pattern of better sensitivity in relation to specificity previously

seen in the in internal data. In the Figure 4.8 we show a comparison between the performance of the $SUVR_{GM}$ EBM classifier both in the internal and external data.

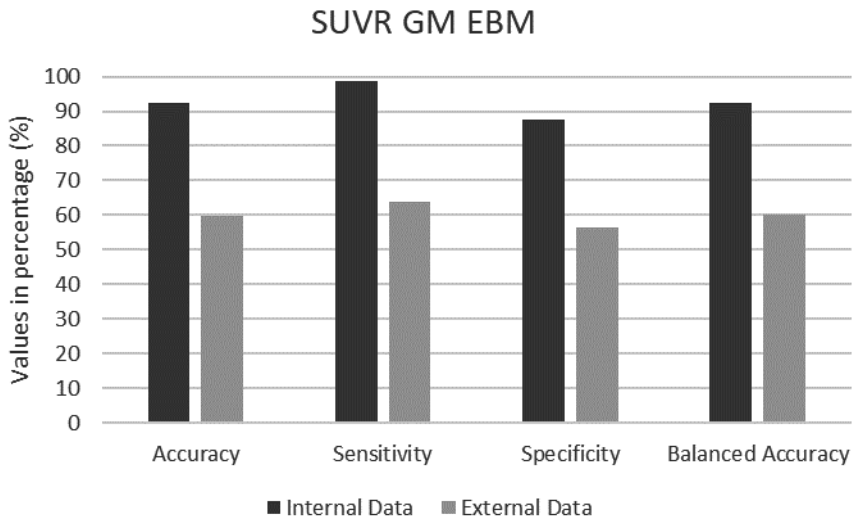

Figure 4.8 Comparison of the performance of the $\boldsymbol{SUVR_{GM}}$ EBM classifier. Internal data balanced accuracy is considered as the same as the accuracy since the internal data is almost perfectly balanced.

Due to the overall close performance of the PIB-based methods, we decided that we were not satisfied with this external performance and decided to further evaluate the other classifiers.

The next in line was the $SUVR_{WM}$ FBM. When tested in the external data this classifier obtained an accuracy of 69.55%, a sensitivity of 81.69%, a specificity of 60.16%, a balance accuracy of 70.92% and an AUC of 0.81±0.02. This classifier shown a better performance than previous and the same performance pattern of the PIB-based classifiers. Figure 4.9 portrays a comparison between the performance of the $SUVR_{WM}$ FBM classifier in both the internal and external data.

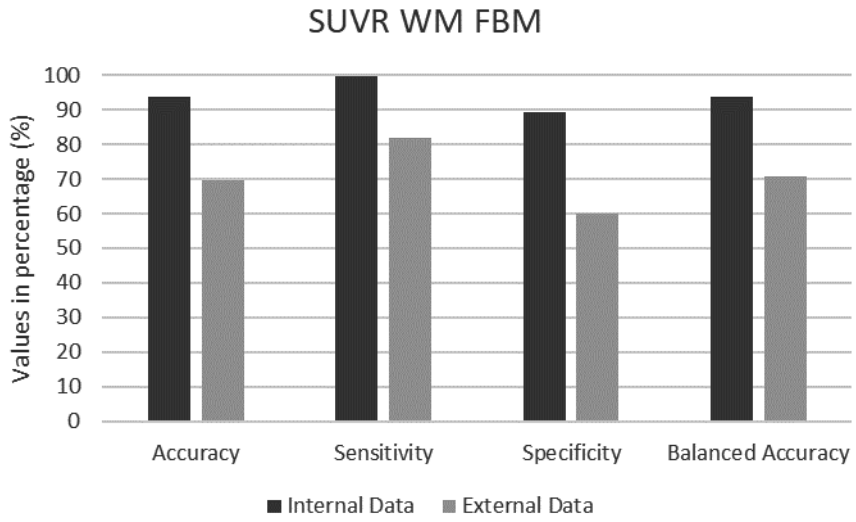

Figure 4.9 Comparison of the performance of the $\boldsymbol{SUVR_{WM}}$ FBM classifier. Internal data balanced accuracy is considered as the same as the accuracy since the internal data is almost perfectly balanced.

Due to this improvement in performance, and the fact that the next classifier has a similar performance as this one, we decided to also evaluate it in the external data.

The evaluation of the $SUVR_{WM}$ EBM classifier with the external reveals an accuracy of 76.87%, sensitivity of 87.90%, specificity of 68.33%, a balance accuracy of 78.11% and an AUC of 0.81±0.04.

Figure 4.10 shows a comparison between the performance of the $SUVR_{WM}$ FBM classifier in both the internal and external data.

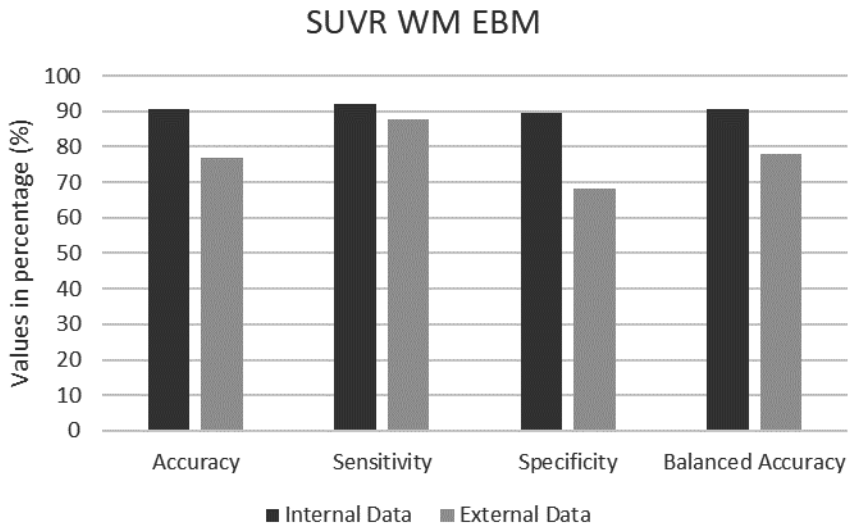

Figure 4.10 Comparison of the performance of the $\boldsymbol{SUVR_{WM}}$ EBM classifier. Internal data balanced accuracy is considered as the same as the accuracy since the internal data is almost perfectly balanced.

Comparing all the PIB-based classifiers we can conclude that the $SUVR_{WM}$ EBM classifier shows the better performance in the external validation. Furthermore, this classifier has a comparable behaviour in the internal and external data. With this in mind, we decided that the $SUVR_{WM}$ EBM classifier will be the one that represents the PIB modality in the ensemble.

Finally, we evaluated the DTI-based classifiers. Similarly, to the previous two analyses we started by evaluating the most promising classifier. The $Destrieux_{MD}$ FBM, was tested with the external data and it was obtained an accuracy of 47.97%, sensitivity of 1.29%, specificity of 98.59%, balance accuracy of 49.96 and an AUC of 0.83±0.03. In the Figure 4.11 we show a comparison between the performance of the $Destrieux_{MD}$ FBM classifier both in the internal and external data.

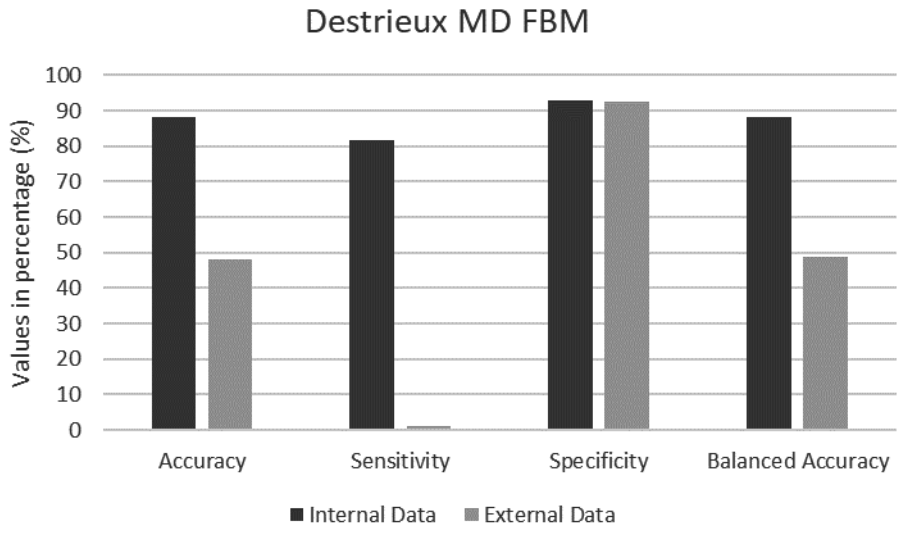

Figure 4.11 Comparison of the performance of the $\boldsymbol{Destrieux_{MD}}$ FBM classifier. Internal data balanced accuracy is considered the same as the accuracy since the internal data is almost perfectly balanced.

Although it is a promising classifier in the internal data, the external validation shows that the $Destrieux_{MD}$ FBM classifier has an extremely poor performance. This classifier shows a different performance pattern in the external data comparing with the internal data. Furthermore, this classifier is practically uncapable of identifying AD cases and it identifies all the cases as CN.

The next logical classifier, the $Desikan_{FA}$ EBM, was tested with the external data and it was obtained an accuracy of 62.79%, sensitivity of 54.31%, specificity of 71.98%, balance accuracy of 63.15 and an AUC of 0.69±0.05. Figure 4.12 describes a comparison between the performance of the $Desikan_{FA}$ EBM classifier both in the internal and external data.

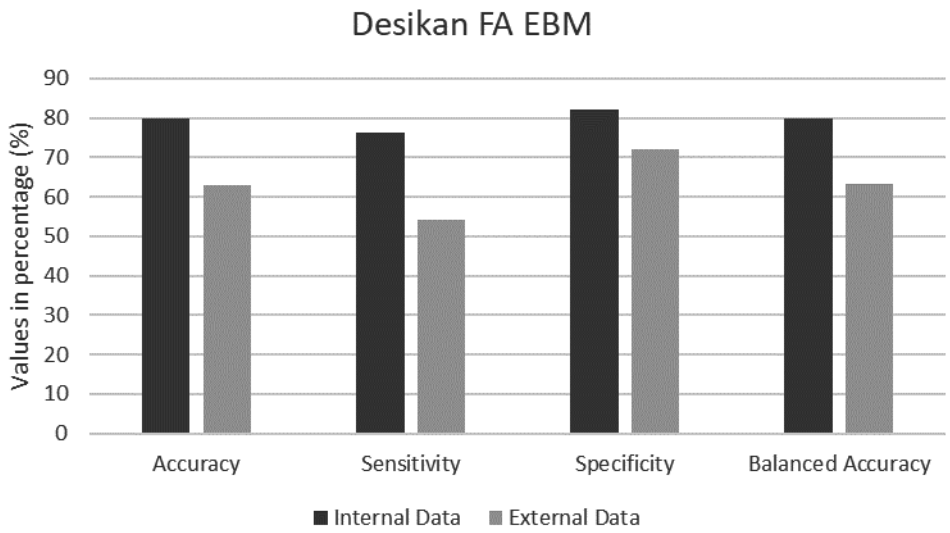

Figure 4.12 Comparison of the performance of the $\boldsymbol{Desikan_{FA}}$ EBM classifier. Internal data balanced accuracy is considered as the same as the accuracy since the internal data is almost perfectly balanced.

The results from this classifier shown a better correspondence between the internal and external data results. Also, this classifier shows a similar performance pattern regarding sensitivity and specificity in the external and internal data.

The performance of this classifier was considered satisfactory and the $Desikan_{FA}$ EBM was chosen to represent the DTI modality in the ensemble.

In conclusion, we were able to construct at least three robust classifiers, one for each modality, that not only have a good performance in the internal data but also perform reasonably well in the external data. Special attention should be given to the MRI and PIB-based classifiers that show an incredibly capacity for generalization. The DTI-based classifier shows some ability for generalization, but for alone usage it is not recommendable.

4.4. Ensemble Classification

In this section we present the evaluation of the ability of our ensemble method as well as the complementarity between all the neuroimaging modalities in the internal data and the combination effects of MRI and DTI-based classifiers in the external data.

Table 4.7, summarizes the results of each classifier selected as well as the ensemble result of combining all three modalities. The weights applied were 1/3, 1/3 and 1/3 for MRI, PIB and DTI-based classifiers respectively.

Table 4.7 Results from the ensemble of all the neuroimaging modalities on the internal data.

Weights (1/3, 1/3, 1/3)	*Mean Accuracy*	*Mean Sensitivity*	*Mean Specificity*
MRI-based classifier	95.09%	90.17%	99.02%
PIB-based classifier	88.76%	88.13%	89.26%
DTI-based classifier	92.59%	90.76%	94.06%
Ensemble classifier	98.11%	99.16%	97.27%

Analysing the Table 4.7, it is possible to notice that the ensemble result is superior than individual base classifiers except for the specificity. This suggests that there is complementary information between the neuroimaging modalities that make possible to improve the global final performance of the classification.

Figure 4.13 shows the mean ROC curves for each classifier presented in Table 4.7

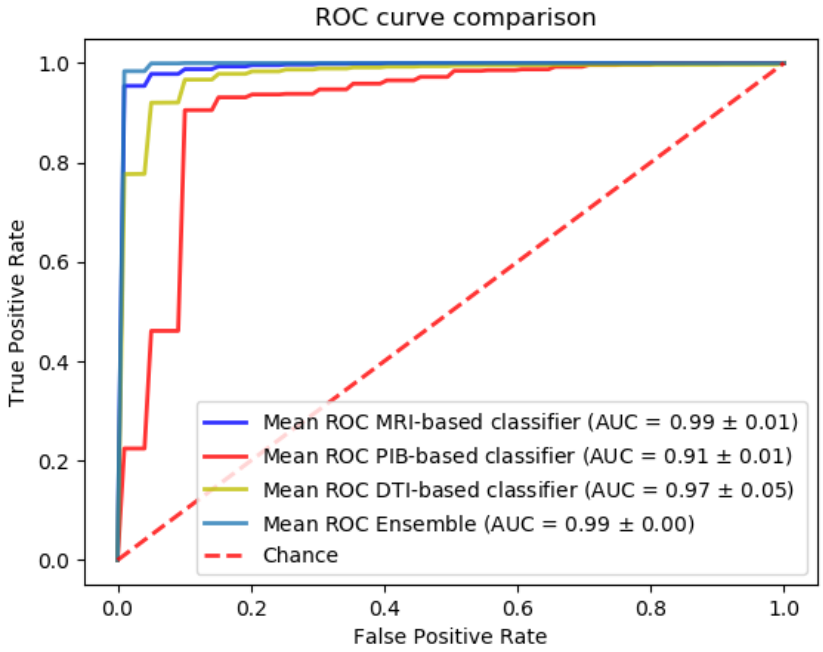

Figure 4.13 Ensemble and all base classifier's mean ROC curve comparison.

Furthermore, we evaluated the effects of only combining the modalities in pairs.

In the Table 4.8, we show the results of only combining MRI and PIB-based classifiers. The classifiers were combined using an equal weight of 1/2 for both modalities.

Table 4.8 Results from the ensemble of MRI and PIB-based classifiers on the internal data.

Weights (1/2, 1/2)	*Mean Accuracy*	*Mean Sensitivity*	*Mean Specificity*
MRI-based classifier	95.13%	90.24%	99.04%
PIB-based classifier	88.75%	88.23%	89.17%
Ensemble classifier	98.05%	98.59%	97.62%

Comparing the ensemble results from Table 4.7 and Table 4.8, it is possible to observe that the addition of the DTI-based classifier only had a small impact. Moreover, the combination of only MRI and PIB-based classifiers shows an improvement in performance except for specificity.

Figure 4.14 shows the mean ROC curves for each classifier presented in Table 4.8.

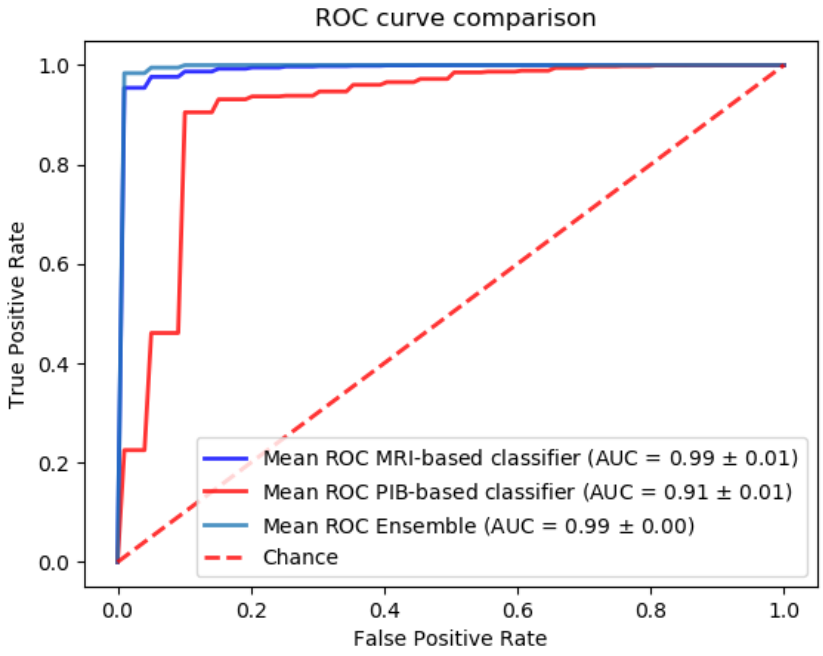

Figure 4.14 Ensemble and (MRI+PIB) base classifier's mean ROC curve comparison.

We also evaluated the performance of combining the MRI and DTI-based classifiers. Table 4.9, presents the results of this combination. The classifiers were combined using an equal weight of 1/2 for both modalities.

Table 4.9 Results from the ensemble of MRI and DTI-based classifiers on the internal data.

Weights (1/2, 1/2)	*Mean Accuracy*	*Mean Sensitivity*	*Mean Specificity*
MRI-based classifier	95.04%	90.04%	99.04%
DTI-based classifier	92.55%	90.58%	94.13%
Ensemble classifier	97.30%	94.62%	99.43%

From the results of the Table 4.9 it is possible to see that MRI and DTI-based classifiers have a similar performance as both the combination of all modalities and the combination of only MRI and PIB. Furthermore, in this combination all the performance parameters were superior in the ensemble than in the individual base classifiers. However, it is noteworthy to see that the individual DTI-based classifiers have a slightly better performance than the individual PIB-based classifier.

Figure 4.15 presents the mean ROC curves for each classifier presented in Table 4.9.

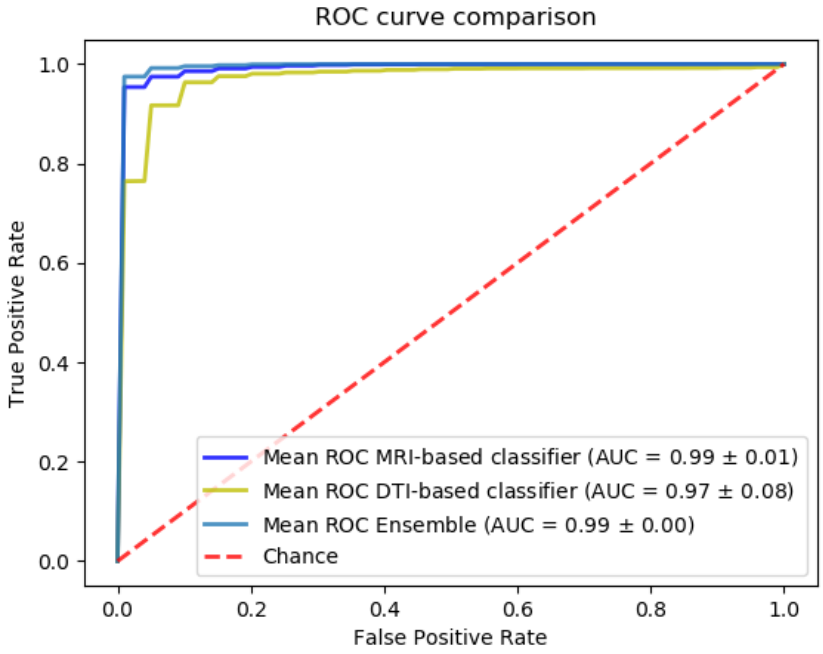

Figure 4.15 Ensemble and (MRI+DTI) base classifier's mean ROC curve comparison.

The last analysis from the internal data was to evaluate the effects of combining PIB and DTI-based classifiers.

Table 4.10 Results from the ensemble of PIB and DTI-based classifiers on the internal data.

Weights (1/2, 1/2)	*Mean Accuracy*	*Mean Sensitivity*	*Mean Specificity*
PIB-based classifier	88.75%	88.21%	89.18%
DTI-based classifier	92.34%	90.27%	94.10%
Ensemble classifier	91.28%	92.61%	90.22%

From the Table 4.10 it is possible to perceive that the combination of PIB and DTI-based classifiers had an overall negative impact only improving the sensitivity of the ensemble classifier.

In the Figure 4.16 we show the mean ROC curves for each classifier presented in Table 4.10.

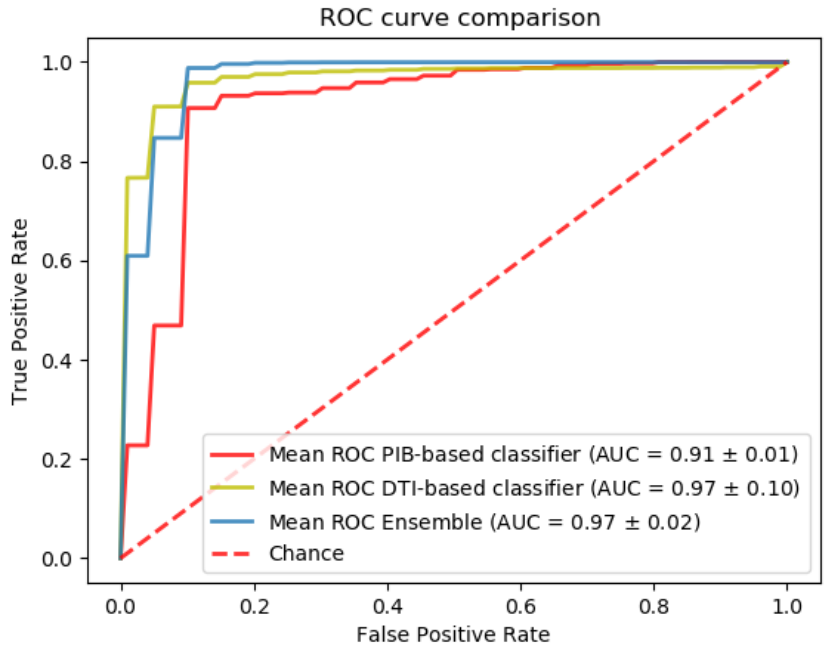

Figure 4.16 Ensemble and (PIB+DTI) base classifier's mean ROC curve comparison.

The results presented above allows the conclusion that using MRI as a base and combining it with one other modality produces the best results. Furthermore, it is possible

to see that there is complementarity between MRI and PIB (Table 4.8) as well as between MRI and DTI (Table 4.9). In contrast, there is no evidence of complementarity between PIB and DTI (Table 4.10).

Also, all the ensemble results shown have similar performance than those find in the state of the art that had shown performance accuracy values between 85.00% and 95.95%[9].

In the Figure 4.17 we present the comparison of performance between all the ensemble classifiers.

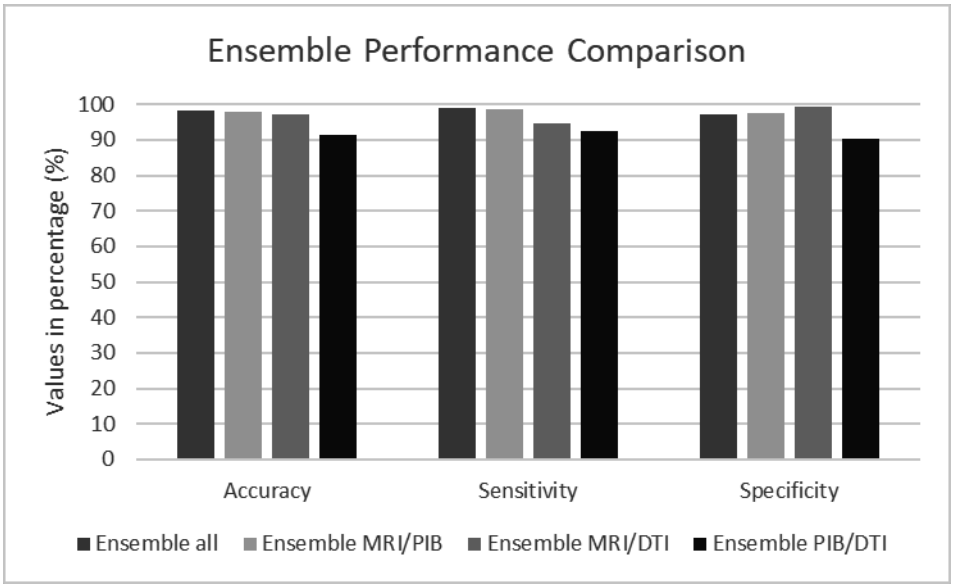

Figure 4.17 Comparison of the ensemble performances.

Analysing Figure 4.17 we see that the combination of all the neuroimaging modalities has the same performing impact as combining only two except for when using PIB and DTI. It is then arguable that the use of the all three neuroimaging modalities for AD classification does not improve the final classification ability.

Lastly, we evaluated the ensemble performance in the external data. In the Table 4.11 we show the results of only combining MRI and DTI-based classifiers on the external data. The classifiers were combined using an equal weight of 1/2 for both modalities.

Table 4.11 Results from the ensemble of MRI and DTI-based classifiers on the external data.

Weights (1/2, 1/2)	*Mean Accuracy*	*Mean Sensitivity*	*Mean Specificity*	*Mean Balanced Accuracy*
MRI-based classifier	78.02%	73.92%	82.47%	78.19%
DTI-based classifier	62.73%	54.46%	71.70%	63.08%
Ensemble classifier	78.59%	77.15%	80.16%	78.66%

Form the analysis of the Table 4.11 it is possible to establish that the ensemble of the classifiers had some effect improving the overall performance except for the specificity. This behaviour is similar to the ensemble of MRI and PIB based when using the internal data.

Moreover, it is important to point out that the DTI-based classifier had a much lower performance with the external data than the MRI-based classifier and despite this fact it contributes in a relatively significant way to the improvement of the global ensemble performance specially improving the sensitivity.

In the Figure 4.18 we show the mean ROC curves for each classifier presented in Table 4.11.

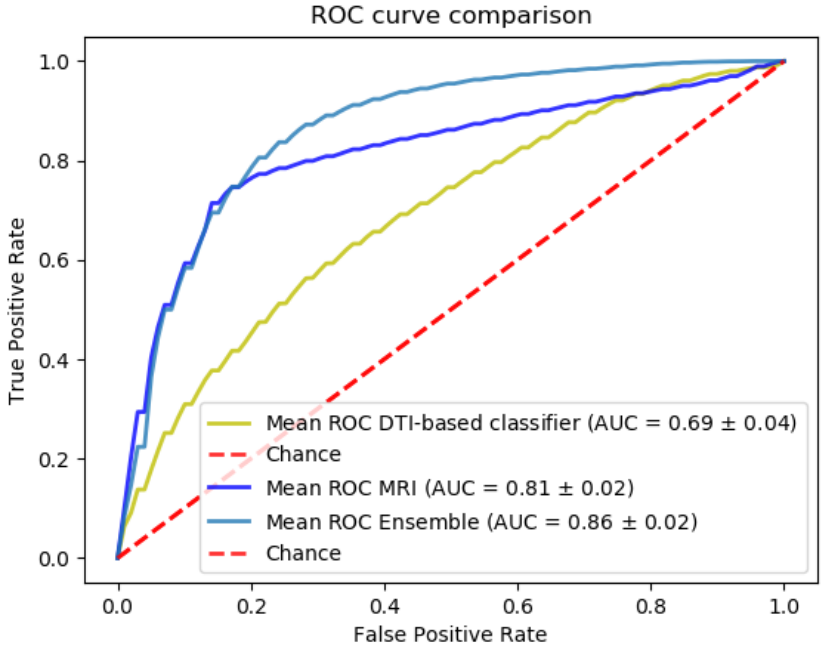

Figure 4.18 Ensemble and (MRI+DTI) base classifiers ROC curve comparison using external data.

5. CONCLUSION AND FUTURE WORK

The major goal of this thesis was to evaluate the effects of a multimodal approach that combined three different imaging modalities (MRI, PIB and DTI). From the results of our ensemble technique we can conclude that the combination of all three imaging modalities does not represent a significant improvement in relation to the traditional use of two imaging modalities.

Furthermore, we can see that the combination of just MRI+PIB or just MRI+DTI had comparable performances, which showed that both PIB and DTI provide complementary information towards the AD/CN classification problem. Also, the combination of PIB+DTI did not provide any improvement, which may be because they provide similar information regarding the AD/CN classification.

Due to this, it could be possible in the future to base the diagnosis of AD in only MRI and DTI imaging data without the need of using PIB-PET images. This could bring some benefits for the patient, that will not be required to be submitted to ionization radiation from the radioactive tracers as well as a lower discomfort caused from the lengthy acquisition protocols of PIB-PET (~90 minutes)[42] in comparison to those of MRI and DTI (between 5 and 15 minutes)[24].

Moreover, the use of only MRI and DTI images only requires the use of one machine. However, the use of PIB-PET images is still a powerful tool for the study of the disease's pathology.

Other conclusions that we can also extract from this work can be summarized into three topics: feature extraction, feature selection and classification.

Our way of extracting features from of the all images provided some good results but it can be further optimized or replaced with more efficient methods that may allow better results in the classification process.

Relatively to feature selection both of our methods seem to achieve good and similar results. However, when putting one against the other in the external data analysis, the embedded based methods show a better ability to generalization and a better performance.

Classification performances were generally high in the internal data despite the lower number of cases. The evaluation of the methods show that DTI-based classifiers have lower performance values than MRI and PIB-based classifiers. This tendency was also verified in the external data were the MRI and PIB-based classifiers showed an excellent performance and the DTI-based classifier struggled to perform better than chance. However, when using all the data in the ensemble, the DTI-based classifier had a similar performance as the others and even outperformed the PIB-based classifier.

One classification finding that was relatively expected was the decrease in performance seen when validating using the external data. It is possible that this decrease is a consequence of the use of different acquisition protocols when obtaining the images due to the type of scanner that was used. Other possible reason for this difference could be the in the morphology of the data itself, however, since both the internal and external data have similar demographic profiles is unlikely that the difference observed between the two analysis can be attributed to this.

It is also noteworthy to refer that all the classifiers used in this thesis were not optimized and performing grid search for some parameters the classification may be improved. Furthermore, the relatively small number of cases that we initial had to perform this work was a limitation, reducing the number of possible features that we can used.

Since our classifier's already have a considerable high performance, feature work should consider the use of more features. This could improve significantly the ability of generalization of the classifier, that is more important for diagnosis purposes.

Future work, build on the findings present here, should be focused in constructing a larger data sample and on, if possible, improving the base classifiers, specially the DTI-based. It could be interesting to study the effects of using the external data to construct the initial classifiers and then validate in the internal data or even to combine the two data sets into an even larger one. However, this kind of analyse is very time intensive and due to the nature of this work it would be impossible to develop all these ideas in the short time given.

ACKNOWLEDGMENTS

Internal data collection was supported by BIGDATIMAGE, CENTRO-01-0145-FEDER-000016 financed by Centro 2020 FEDER, COMPETE; PAC – MEDPERSYST POCI-01-0145- FEDER-016428, and Prémio Santa Casa Mantero Belard.

External Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

REFERENCES

1. Wong, P. C., Li, T. & Price, D. L. Neurobiology of Alzheimer's Disease. in *Basic Neurochemistry MOLECULAR, CELLULAR AND MEDICAL ASPECTS* (eds. Siegel, G. J., Albers, R. W., Brady, S. T. & Price, D. L.) 781–790 (Elsevier, 2006).
2. Scheltens, P. *et al.* Alzheimer's disease. *Lancet* **388**, 505–17 (2016).
3. Wachinger, C. & Reuter, M. Domain adaptation for Alzheimer's disease diagnostics. *Neuroimage* **139**, 470–479 (2016).
4. Duraisamy, B., Shanmugam, J. V. & Annamalai, J. Alzheimer disease detection from structural MR images using FCM based weighted probabilistic neural network. *Brain Imaging Behav.* (2019). doi:10.1007/s11682-018-9831-2
5. Perl, D. P. Neuropathology of Alzheimer ' s Disease Address. *Mt. Sinai J. Med.* **77**, 32–42 (2010).
6. Vommaro, P. A. Correlation of Alzheimer Disease Neuropathologic Changes With Cognitive Status: A Review of the Literature. **71**, 11–36 (2010).
7. Bryan, R. N. Machine Learning Applied to Alzheimer Disease. *Radiology* (2016). doi:10.1148/radiol.2016162151
8. Rondina, J. M. *et al.* Selecting the most relevant brain regions to discriminate Alzheimer's disease patients from healthy controls using multiple kernel learning: A comparison across functional and structural imaging modalities and atlases. *NeuroImage Clin.* **17**, 628–641 (2018).
9. Rathore, S., Habes, M., Iftikhar, M. A., Shacklett, A. & Davatzikos, C. A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *NeuroImage* **155**, 530–548 (2017).
10. Bay, H. & Valley, M. What is dementia? *Acta Neurol. Scand.* **94**, 22–24 (1996).
11. Piaceri, I., Nacmias, B. & Sorbi, S. Genetics of familial and sporadic Alzheimer's disease. *Front. Biosci.* 167–177 (2013).
12. McKhann, G. *et al.* Clinical diagnosis of Alzheimer ' s disease : *Neurology* **34**, 939 (1984).
13. Tsao, S. *et al.* Feature selective temporal prediction of Alzheimer's disease progression using hippocampus surface morphometry. *Brain Behav.* **7**, 1–11 (2017).

14. Schouten, T. M. *et al.* Individual classification of Alzheimer's disease with diffusion magnetic resonance imaging. *Neuroimage* **152**, 476–481 (2017).
15. Previtali, F., Bertolazzi, P., Felici, G. & Weitschek, E. A novel method and software for automatically classifying Alzheimer's disease patients by magnetic resonance imaging analysis. *Comput. Methods Programs Biomed.* **143**, 89–95 (2017).
16. Archana, M. & Ramakrishnan, S. Detection of Alzheimer disease in MR images using structure tensor. *2014 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBC 2014* 1043–1046 (2014). doi:10.1109/EMBC.2014.6943772
17. Farzan, A., Mashohor, S., Ramli, A. R. & Mahmud, R. Boosting diagnosis accuracy of Alzheimer's disease using high dimensional recognition of longitudinal brain atrophy patterns. *Behav. Brain Res.* **290**, 124–130 (2015).
18. Dyrba, M. *et al.* Predicting Prodromal Alzheimer's Disease in Subjects with Mild Cognitive Impairment Using Machine Learning Classification of Multimodal Multicenter Diffusion-Tensor and Magnetic Resonance Imaging Data. *J. Neuroimaging* **25**, 738–747 (2015).
19. M., L., D., Z. & D., S. Identifying informative imaging biomarkers via tree structured sparse learning for AD diagnosis. *Neuroinformatics* (2014). doi:10.1007/s12021-013-9218-x LK - http://limo.libis.be/resolver?&sid=EMBASE&issn=15590089&id=doi:10.1007%2Fs12021-013-9218-x&atitle=Identifying+informative+imaging+biomarkers+via+tree+structured+sparse+learning+for+AD+diagnosis&stitle=Neuroinformatics&title=Neuroinformatics&volume=12&issue=3&spage=381&epage=394&aulast=Liu&aufirst=Manhua&auinit=M.&aufull=Liu+M.&coden=&isbn=&pages=381-394&date=2014&auinit1=M&auinitm=
20. Tohka, J., Moradi, E. & Huttunen, H. Comparison of Feature Selection Techniques in Machine Learning for Anatomical Brain MRI in Dementia. *Neuroinformatics* **14**, 279–296 (2016).
21. Kim, J. & Lee, B. Automated discrimination of dementia spectrum disorders using extreme learning machine and structural T1 MRI features. *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS* 1990–1993 (2017). doi:10.1109/EMBC.2017.8037241

22. Liu, M., Zhang, D. & Shen, D. View-centralized multi-atlas classification for Alzheimer's disease diagnosis. *Hum. Brain Mapp.* **36**, 1847–1865 (2015).
23. Cai, K. *et al.* Identification of early-stage Alzheimer's disease using sulcal morphology and other common neuroimaging indices. *PLoS One* **12**, 1–15 (2017).
24. Alexander, A. L., Lee, J. E., Lazar, M. & Field, A. S. Diffusion Tensor Imaging of the Brain. *Neurotherapeutics* **4**, 316–329 (2007).
25. Soares, J. M., Marques, P., Alves, V. & Sousa, N. A hitchhiker's guide to diffusion tensor imaging. *Front. Neurosci.* **7**, 31 (2013).
26. Dyrba, M., Grothe, M., Kirste, T. & Teipel, S. J. Multimodal analysis of functional and structural disconnection in Alzheimer's disease using multiple kernel SVM. *Hum. Brain Mapp.* **36**, 2118–2131 (2015).
27. Cattell, L. *et al.* Classification of amyloid status using machine learning with histograms of oriented 3D gradients. *NeuroImage Clin.* **12**, 990–1003 (2015).
28. Oliveira, F. *et al.* Data driven diagnostic classification in Alzheimer's disease based on different reference regions for normalization of PiB-PET images and correlation with CSF concentrations of Aβ species. *NeuroImage Clin.* **20**, 603–610 (2018).
29. Cao, P. *et al.* Nonlinearity-aware based dimensionality reduction and over-sampling for AD/MCI classification from MRI measures. *Comput. Biol. Med.* **91**, 21–37 (2017).
30. Chen, Y. *et al.* Detecting brain structural changes as biomarker from magnetic resonance images using a local feature based SVM approach. *J. Neurosci. Methods* **221**, 22–31 (2014).
31. Klöppel, S. *et al.* Diagnostic neuroimaging across diseases. *Neuroimage* **61**, 457–463 (2012).
32. Koikkalainen, J. *et al.* Multi-template tensor-based morphometry: Application to analysis of Alzheimer's disease. *Neuroimage* **56**, 1134–1144 (2011).
33. Min, R., Wu, G., Cheng, J., Wang, Q. & Shen, D. Multi-atlas based representations for Alzheimer's disease diagnosis. *Hum. Brain Mapp.* (2014). doi:10.1002/hbm.22531
34. Beheshti, I. & Demirel, H. Feature-ranking-based Alzheimer's disease classification from structural MRI. *Magn. Reson. Imaging* **34**, 252–263 (2016).
35. Kurth, F., Luders, E. & Gaser, C. Voxel-Based Morphometry. *Elsevier* **1**, 345–349

(2015).

36. Sled, J. G., Zijdenbos, A. P. & Evans, A. C. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Trans. Med. Imaging* **17**, 87–97 (1998).
37. Casanova, R., Hsu, F. C. & Espeland, M. A. Classification of Structural MRI Images in Alzheimer's Disease from the Perspective of Ill-Posed Problems. *PLoS One* **7**, (2012).
38. Doan, N. T. *et al.* Distinguishing early and late brain aging from the Alzheimer's disease spectrum: consistent morphological patterns across independent samples. *Neuroimage* **158**, 282–295 (2017).
39. Yang, S.-T. *et al.* Discrimination between Alzheimer's Disease and Mild Cognitive Impairment Using SOM and PSO-SVM. *Comput. Math. Methods Med.* **2013**, 1–10 (2013).
40. Smith, S. M., Kindlmann, G. & Jbabdi, S. Cross-Subject Comparison of Local Diffusion MRI Parameters. in *Diffusion MRI: From Quantitative Measurement to In vivo Neuroanatomy: Second Edition* 209–239 (Elsevier, 2013). doi:10.1016/B978-0-12-396460-1.00010-X
41. Linn, K. A. *et al.* Control-group feature normalization for multivariate pattern analysis of structural MRI data using the support vector machine. *Neuroimage* **132**, 157–166 (2016).
42. Li, Y. *et al.* Regional analysis of FDG and PIB-PET images in normal aging, mild cognitive impairment, and Alzheimer's disease. *Eur. J. Nucl. Med. Mol. Imaging* **35**, 2169–2181 (2008).
43. Lei, B., Yang, P., Wang, T., Chen, S. & Ni, D. Relational-Regularized Discriminative Sparse Learning for Alzheimer's Disease Diagnosis. *IEEE Trans. Cybern.* **47**, 1102–1113 (2017).
44. Vandenberghe, R. *et al.* Binary classification of 18F-flutemetamol PET using machine learning: Comparison with visual reads and structural MRI. *Neuroimage* **64**, 517–525 (2013).
45. Camus, V. *et al.* Using PET with 18F-AV-45 (florbetapir) to quantify brain amyloid load in a clinical environment. *Eur. J. Nucl. Med. Mol. Imaging* **39**, 621–631 (2012).
46. Fu, L. *et al.* Comparison of dual-biomarker PIB-PET and dual-tracer PET in AD

diagnosis. *Eur. Radiol.* **24**, 2800–2809 (2014).

47. El-Gamal, F. E. A. *et al.* A novel early diagnosis system for mild cognitive impairment based on local region analysis: A pilot study. *Front. Hum. Neurosci.* **11**, (2018).
48. Kiebel, S. J., Ashburner, J., Poline, J. B. & Friston, K. J. MRI and PET coregistration - A cross validation of statistical parametric mapping and automated image registration. *Neuroimage* **5**, 271–279 (1997).
49. Li, M., Qin, Y., Gao, F., Zhu, W. & He, X. Discriminative analysis of multivariate features from structural MRI and diffusion tensor images. *Magn. Reson. Imaging* **32**, 1043–1051 (2014).
50. Lee, W., Park, B. & Han, K. Classification of diffusion tensor images for the early detection of Alzheimer's disease. *Comput. Biol. Med.* **43**, 1313–1320 (2013).
51. Mesrob, L. *et al.* DTI and Structural MRI Classification in Alzheimer's Disease. *Adv. Mol. Imaging* **02**, 12–20 (2012).
52. Wen, J. *et al.* Using Diffusion Mri for Classification and Prediction of Alzheimer'S Disease: a Reproducible Study. *Alzheimer's Dement.* **14**, P97–P98 (2018).
53. Greve, D. N. An Absolute Beginner's Guide to Surface- and Voxel-based Morphometric Analy. in *International Society for Magnetic Resonance Medicine* (2011).
54. Li, T. & Zhang, W. Classification of brain disease from magnetic resonance images based on multi-level brain partitions. *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS* **2016-Octob**, 5933–5936 (2016).
55. Liu, M., Zhang, D., Adeli, E. & Shen, D. Inherent Structure-Based Multiview Learning with Multitemplate Feature Representation for Alzheimer's Disease Diagnosis. *IEEE Trans. Biomed. Eng.* (2016). doi:10.1109/TBME.2015.2496233
56. Fan, Y., Shen, D., Gur, R. C., Gur, R. E. & Davatzikos, C. COMPARE: Classification of morphological patterns using adaptive regional elements. *IEEE Trans. Med. Imaging* **26**, 93–105 (2007).
57. Ahn, H. J. *et al.* The cortical neuroanatomy of neuropsychological deficits in mild cognitive impairment and Alzheimer's disease: A surface-based morphometric analysis. *Neuropsychologia* **49**, 3931–3945 (2011).
58. Zaidi, H. & Karakatsanis, N. Towards enhanced PET quantification in clinical

oncology. *Br. J. Radiol.* **91**, 20170508 (2017).

59. Mukherjee, P., Berman, J. I., Chung, S. W., Hess, C. P. & Henry, R. G. Diffusion tensor MR imaging and fiber tractography: Theoretic underpinnings. *Am. J. Neuroradiol.* **29**, 632–641 (2008).
60. Wen, J. *et al.* Reproducible evaluation of diffusion MRI features for automatic classification of patients with Alzheimer's disease. 1–51 (2018).
61. Tang, J., Alelyani, S. & Liu, H. Feature selection for classification: A review. *Data Classif. Algorithms Appl.* 37–64 (2014).
62. Kumar, V. Feature Selection: A literature Review. *Smart Comput. Rev.* **4**, (2014).
63. Jović, A., Brkić, K. & Bogunović, N. A review of feature selection methods with applications. *2015 38th Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO 2015 - Proc.* 1200–1205 (2015). doi:10.1109/MIPRO.2015.7160458
64. Kotsiantis, S. B., Zaharakis, I. D. & Pintelas, P. E. Machine learning: A review of classification and combining techniques. *Artif. Intell. Rev.* **26**, 159–190 (2006).
65. Hosmer, D. W. & Lemeshow, S. *Applied Logistic Regression Second Edition*. *Applied Logistic Regression* (2000). doi:10.1002/0471722146
66. Lorena, A. C. *et al.* Comparing machine learning classifiers in potential distribution modelling. *Expert Syst. Appl.* **38**, 5268–5275 (2011).
67. Amanpreet Singh; Narina Thakur ; Aakanksha Sharma. A review of supervised machine learning algorithms. *Ieee* 1310–1315 (2016).
68. Bhavsar, H. & Panchal, M. H. A Review on Support Vector Machine for Data Classification. *Int. J. Adv. Res. Comput. Eng. Technol.* **1**, 185–189 (2012).
69. Drakos, G. Support Vector Machine vs Logistic Regression. (2018). Available at: https://towardsdatascience.com/support-vector-machine-vs-logistic-regression-94cc2975433f. (Accessed: 12th August 2019)
70. Jordam, J. Support vector Machines. (2017). Available at: https://www.jeremyjordan.me/support-vector-machines/. (Accessed: 12th August 2019)
71. Matteo, R. & Giorgio, V. *Ensemble methods: a review*. *Data Mining and Machine Learning for Astronomical Applications* (2001).
72. Sagi, O. & Rokach, L. Ensemble learning: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **8**, 1–18 (2018).

73. Jurek, A., Bi, Y., Wu, S. & Nugent, C. A survey of commonly used ensemble-based classification techniques. *Knowl. Eng. Rev.* **29**, 551–581 (2013).

SUPPLEMENT A

Supplementary data can be found here, and the codes used to obtain all the results can be found in https://github.com/Daniel-Agostinho/Thesis.git.

Feature Selection all data

Table S.1 Feature Selection results.

Model	Feature selection method	Total Number of features	Final Number of features
$Cobra_{GM}$	EBM	52	16
	FBM		2
$Cobra_{WM}$	EBM	52	20
	FBM		6
$Hammers_{GM}$	EBM	68	14
	FBM		7
$Hammers_{WM}$	EBM	68	13
	FBM		1
$Hammers_{CSF}$	EBM	68	3
	FBM		4
$Lpba40_{GM}$	EBM	56	13
	FBM		2
$Neuro_{GM}$	EBM	142	41
	FBM		10
$Neuro_{CSF}$	EBM	142	11
	FBM		5
$a2009_{Gyri}$	EBM	152	144
	FBM		25
$a2009_{Thick}$	EBM	152	57
	FBM		16

$Dk40_{Gyri}$	EBM	72	67
	FBM		13
$Dk40_{Thick}$	EBM	72	20
	FBM		9
HCP_{Gyri}	EBM	362	358
	FBM		41
HCP_{Thick}	EBM	362	202
	FBM		54
$SUVR_{Cere}$	EBM	19	1
	FBM		1
$SUVR_{GM}$	EBM	19	2
	FBM		5
$SUVR_{WM}$	EBM	19	2
	FBM		3
$Lpba40_{FA}$	EBM	56	21
	FBM		1
$Lpba40_{MD}$	EBM	56	12
	FBM		6
$Desikan_{FA}$	EBM	84	58
	FBM		6
$Desikan_{MD}$	EBM	84	26
	FBM		9
$Destrieux_{FA}$	EBM	164	130
	FBM		9
$Destrieux_{MD}$	EBM	164	61
	FBM		19
$Hammers_{FA}$	EBM	68	37
	FBM		5
$Hammers_{MD}$	EBM	68	24
	FBM		4
JHU_{FA}	EBM	48	24
	FBM		7
JHU_{MD}	EBM	48	18
	FBM		5

Atlases Region of Interest (ROI) Glossary

Table S.2 Cobra volume atlas acronym's names.

ROI acronym	***ROI name***
Striatum	Striatum
GloPal	Globus Pallidus
Tha	Thalamus
AntCerebLI_II	Anterior Cerebellar Lobule I-II
AntCerebLIII	Anterior Cerebellar Lobule III
AntCerebLIV	Anterior Cerebellar Lobule IV
AntCerebLV	Anterior Cerebellar Lobule V
SupPostCerebLVI	Superior Posterior Cerebellar Lobule VI
SupPostCerebCI	Superior Posterior Cerebellar Lobule Crus I
SupPostCerebCII	Superior Posterior Cerebellar Lobule Crus II
SupPostCerebLVIIB	Superior Posterior Cerebellar Lobule VIIB
InfPostCerebLVIIIA	Inferior Posterior Cerebellar Lobule VIIIA
InfPostCerebLVIIIB	Inferior Posterior Cerebellar Lobule VIIIB
InfPostCerebLIX	Inferior Posterior Cerebellar Lobule IX-
InfPostCerebLX	Inferior Posterior Cerebellar Lobule X
AntCerebWM	Cerebellar White Matter
Amy	Amygdala
HCA1	Hippocampus CA1
Sub	Subiculum
For	Fornix
CA4	CA4/Dentate Gyrus
CA2_3	CA2/CA3
Stratum.L…	Stratum, Stratum Radiatum/Lacunosum/Moleculare
Fimbra	Fimbria
MamBody	Mammillary body
Alveus	Alveus

Table S.3 Hammers volume atlas acronym's names.

ROI acronym	***ROI name***
Hip	Hippocampus
Amy	Amygdala
AntMedTeLo	Anterior Medial Temporal Lobe
AntLatTeLo	Anterior Lateral Temporal Lobe
Amb+ParHipGy	Ambient and Parahippocampus Gyri
SupTemGy	Superior Temporal Gyrus
InfMidTemGy	Inferior Middle Temporal Gyri
FusGy	Fusiform Gyrus
Cbe	Cerebellum
Bst	Brainstem
Ins	Insula
LatOcLo	Lateral Occipital Lobe
AntCinGy	Anterior Cinguli Gyrus
PosCinGy	Posterior Cinguli Gyrus
MidFroGy	Middle Frontal Gyrus
PosTeLo	Posterior Temporal Lobe
InfLatPaLo	Inferior Lateral Pariatal Lobe
CauNuc	Caudate Nucleus
AccNuc	Accumbens Nucleus
Put	Putamen
Tha	Thalamus
Pa	Pallidum
CC	Corpus Callosum
LatTemVen	Lateral Temporal Ventricle
3thVen	Third Ventricle
PrcGy	Precentral Gyrus
RecGy	Gyrus Rectus
OrbFroGy	Orbito-Frontal Gyri
InfFroGy	Inferior Frontal Gyrus
SupFroGy	Superior Frontal Gyrus
PoCGy	Postcentral Gyrus
SupParGy	Superior Parietal Gyrus
LinGy	Lingual Gyrus
Cun	Cuneus

Table S.4 Lpba40 volume atlas acronym's names.

ROI acronym	***ROI name***
SupFoG	Superior Frontal Gyrus
MidFoG	Middle Frontal Gyrus
InfFroG	Inferior Frontal Gyrus
PrcG	Precentral Gyrus
MidOrbG	Middle Orbitofrontal Gyrus
LatOrbG	Lateral Orbitofrontal Gyrus
rRecG	Gyrus Rectus
rPoCG	Postcentral Gyrus
SupParG	Superior Parietal Gyrus
SupMarG	Supramarginal Gyrus
AngG	Angular Gyrus
PCu	Precuneus
SupOccG	Superior Occipital Gyrus
MidOccG	Middle Occipital Gyrus
InfOccG	Inferior Occipital Gyrus
Cun	Cuneus
SupTemG	Superior Temporal Gyrus
MidTemG	Middle Temporal Gyrus
InfTemG	Inferior Temporal Gyrus
ParHipG	Parahippocampal Gyrus
LinG	Lingual Gyrus
FusG	Fusiform Gyrus
Ins	Insula
CinG	Cingulate Gyrus
Cau	Caudate
Put	Putamen
Hip	Hippocampus
CBeL	Cerebellar Lobe
Bst	Brainstem

Table S.5 Neuromorphometrics volume atlas acronym's names.

ROI acronym l	***ROI name***
3thVen	Third Ventricle
4thVen	Fourth Ventricle
Acc	Accumbens
Amy	Amygdala
Bst	Brainstem
Cau	Caudate
ExtCbe	Exterior Cerebellum
CbeWM	Cerebellum White Matter
CbrWM	Cerebral White Matter
CSF	CSF
Hip	Hippocampus
InfLatVen	Inferior Lateral Ventricle
LatVen	Lateral Ventricle
Pa	Pallidum
Put	Putamen
ThaPro	Thalamus Proper
VenVen	Ventral Ventricle
OC	Optic Chiasm
CbeLoCbe1-5	Cerebellar Lobule Cerebellar Vermal Lobules I-V
CbeLoCbe6-7	Cerebellar Lobule Cerebellar Vermal Lobules VI-VII
CbeLoCbe8-10	Cerebellar Lobule Cerebellar Vermal Lobules VIII-X
BasCbr+FobBr	Basal Cerebrum and Forebrain Brain
AntCinGy	Anterior Cingulate Gyrus
AntIns	Anterior Insula
AntOrbGy	Anterior Orbital Gyrus
AngGy	Angular Gyrus
Ca+Cbr	Calcarine and Cerebrum
CenOpe	Central Operculum
Cun	Cuneus
Ent	Entorhinal Area
FroOpe	Frontal Operculum
FroPo	Frontal Pole
FusGy	Fusiform Gyrus
RecGy	Gyrus Rectus
InfOccGy	Inferior Occipital Gyrus
InfTemGy	Inferior Temporal Gyrus
LinGy	Lingual Gyrus
LatOrbGy	Lateral Orbital Gyrus
MidCinGy	Middle Cingulate Gyrus
MedFroCbr	Medial Frontal Cerebrum
MidFroGy	Middle Frontal Gyrus

MidOccGy	Middle Occipital Gyrus
MedOrbGy	Medial Orbital Gyrus
MedPoCGy	Medial Postcentral Gyrus
MedPrcGy	Medial Precentral Gyrus
SupMedFroGy	Superior Medial Frontal Gyrus
MidTemGy	Middle Temporal Gyrus
OccPo	Occipital Pole
OccFusGy	Occipital Fusiform Gyrus
InfFroGy	Inferior Frontal Gyrus
InfFroOrbGy	Inferior Frontal Orbital Gyrus
PosCinGy	Posterior Cingulate Gyrus
PCu	Precuneus
ParHipGy	Parahippocampus Gyrus
PosIns	Posterior Insula
ParOpe	Parietal Operculum
PoCGy	Postcentral Gyrus
PosOrbGy	Posterior Orbital Gyrus
Pa	Planum Polare
PrcGy	Precentral Gyrus
Tem	Temporal
SCA	Subcallosal Area
SupFroGy	Superior Frontal Gyrus
Cbr+Mot	Cerebrum and Motor
SupMarGy	Supramarginal Gyrus
SupOccGy	Superior Occipital Gyrus
SupParLo	Superior Parietal Lobule
SupTemGy	Superior Temporal Gyrus
TemPo	Temporal Pole
InfFroAngGy	Inferior Frontal Angular Gyrus
TemTraGy	Temporal Transverse Gyrus

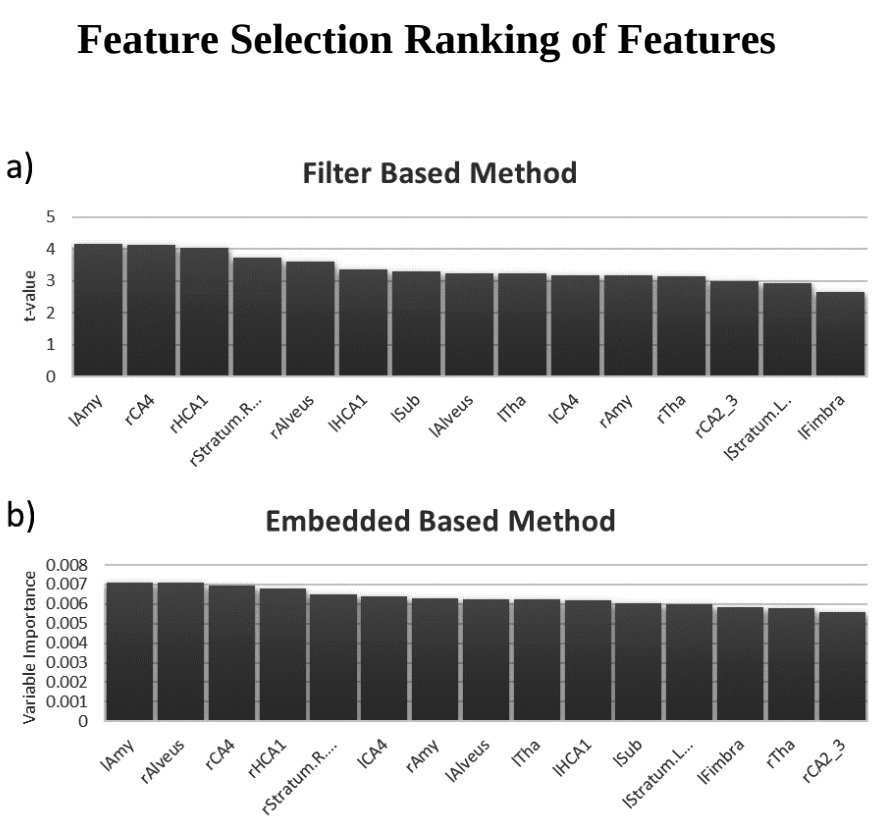

Figure S.1 Top ranking features for the Cobra GM model. a) FBM method. b) EBM method.

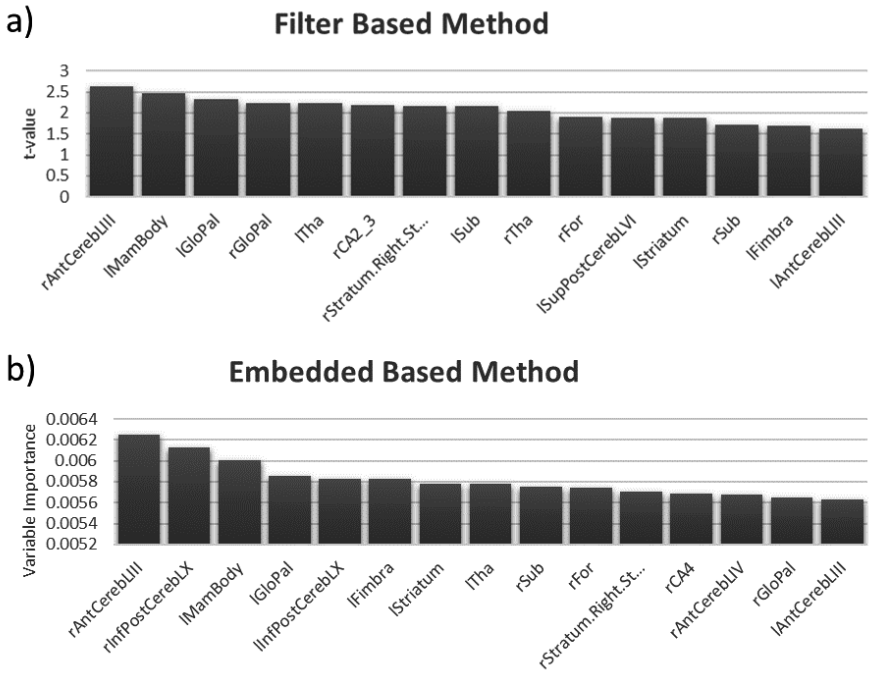

Figure S.2 Top ranking features for the Cobra WM model. a) FBM method. b) EBM method.

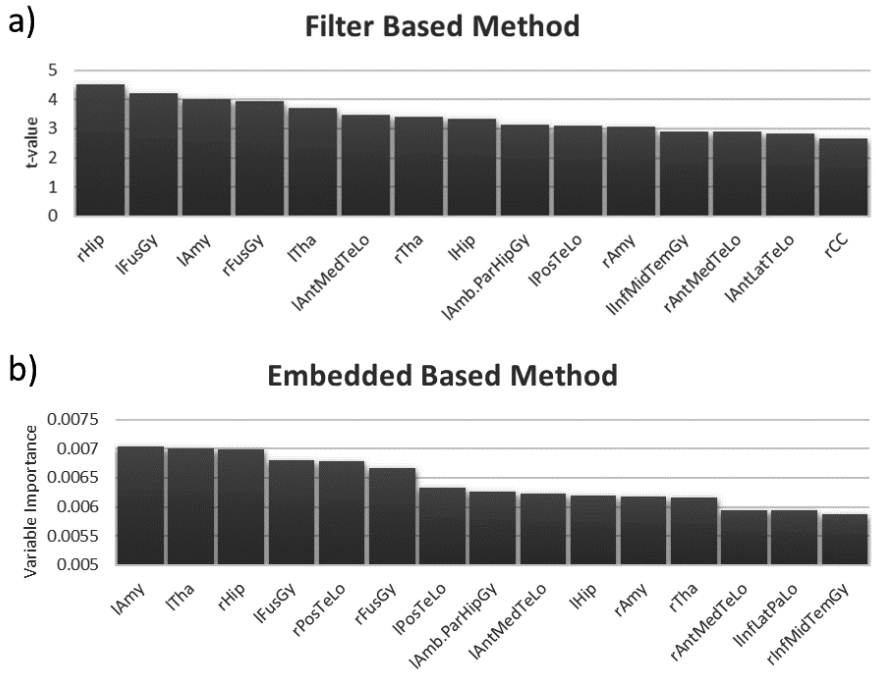

Figure S.3 Top ranking features for the Hammers GM model. a) FBM method. b) EBM method.

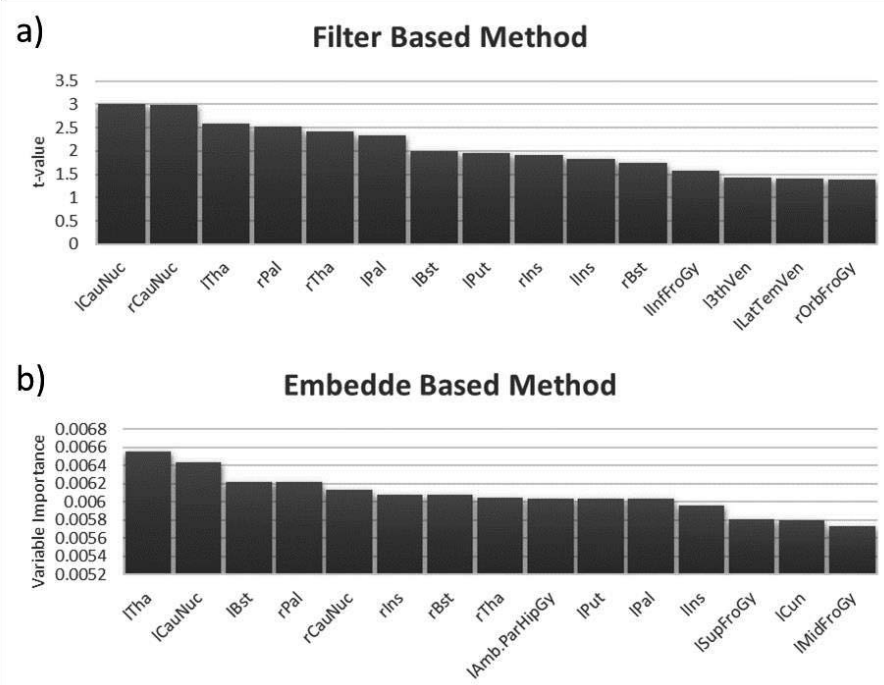

Figure S.4 Top ranking features for the Hammers WM model. a) FBM method. b) EBM method.

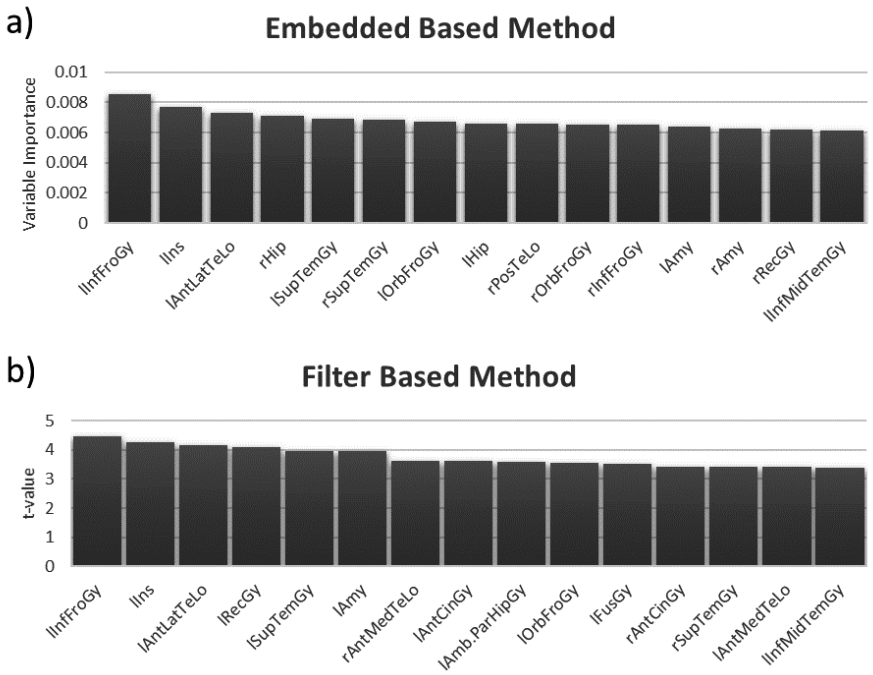

Figure S.5 Top ranking features for the Hammers CSF model. a) FBM method. b) EBM method.

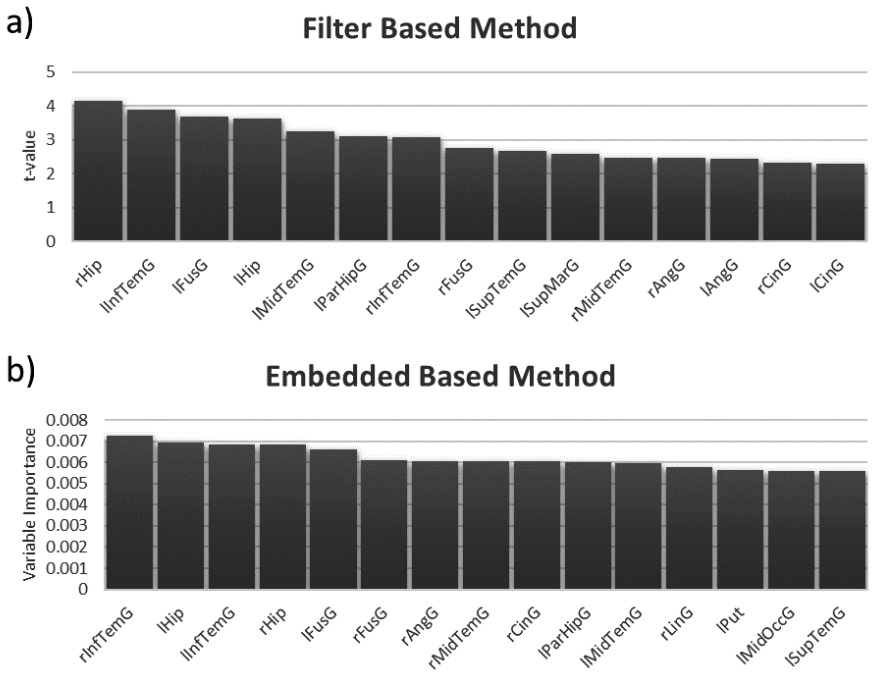

Figure S.6 Top ranking features for the Lpba40 GM model. a) FBM method. b) EBM method.

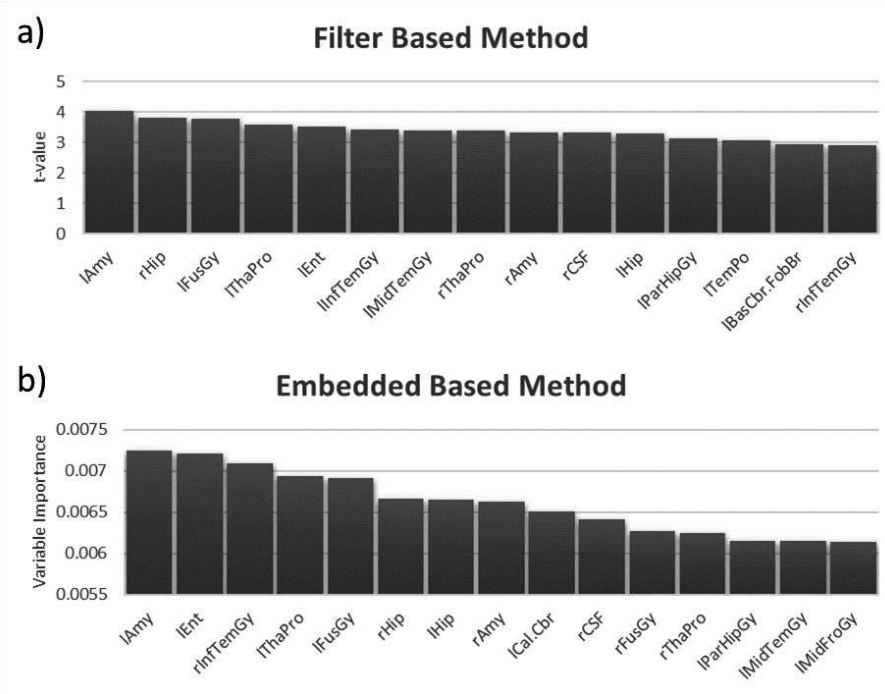

Figure S.7 Top ranking features for the Neuromorphometrics GM model. a) FBM method. b) EBM method.

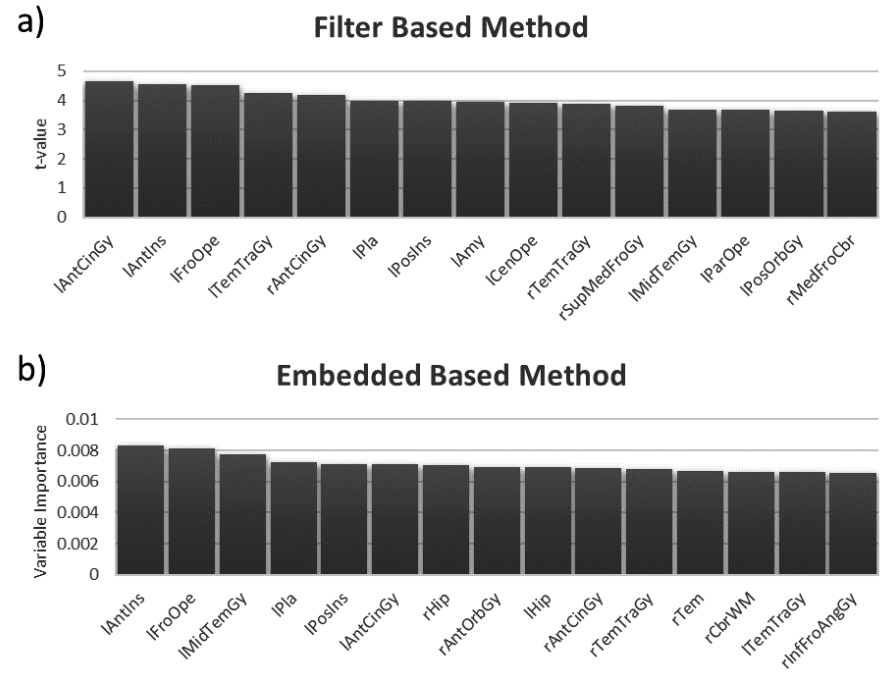

Figure S.8 Top ranking features for the Neuromorphometrics CSF model. a) FBM method. b) EBM method.

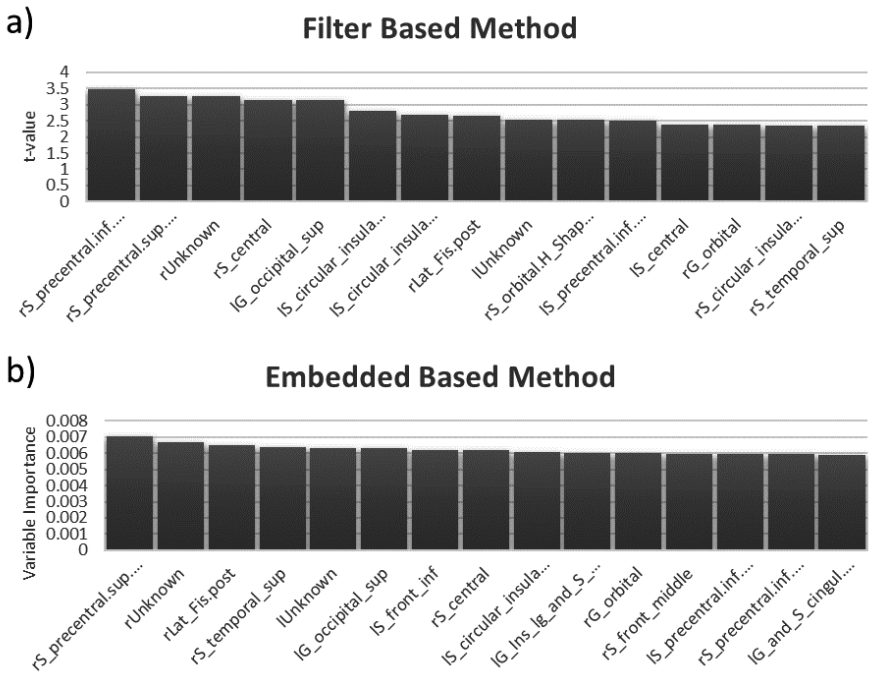

Figure S.9 Top ranking features for the a2009 Gyrification model. a) FBM method. b) EBM method.

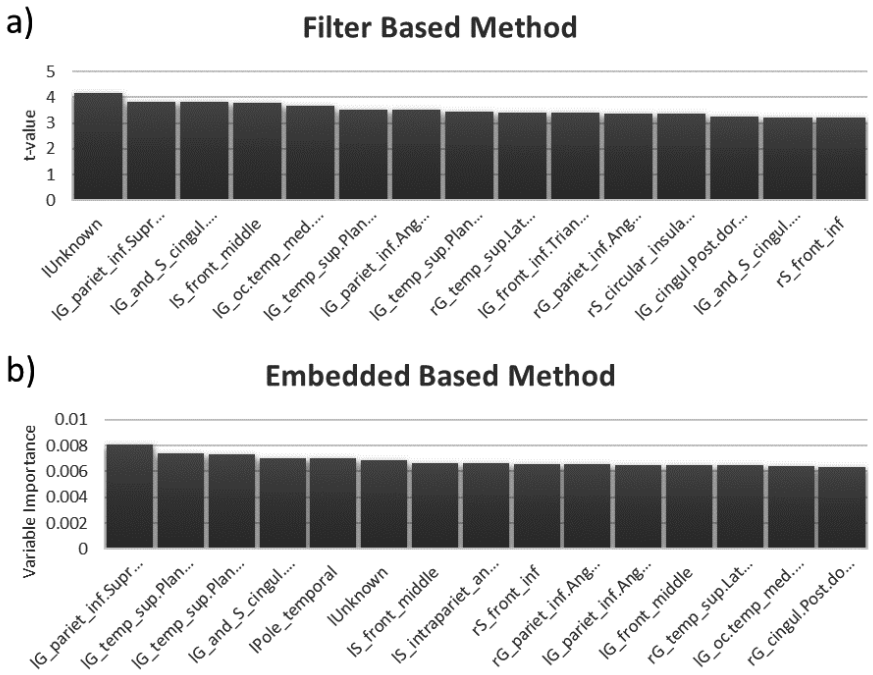

Figure S.10 Top ranking features for the a2009 Thickness model. a) FBM method. b) EBM method.

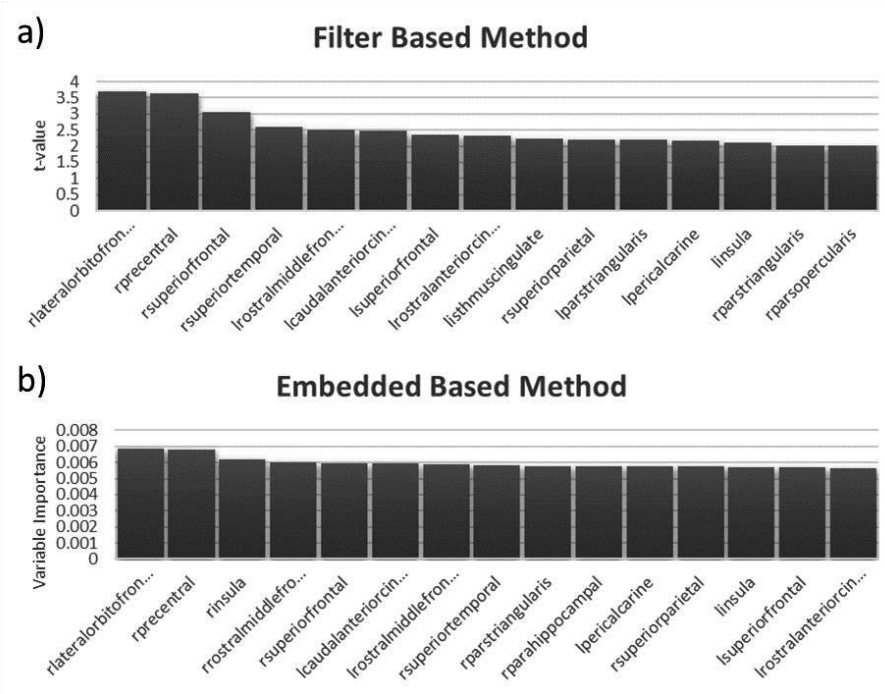

Figure S.11 Top ranking features for the DK40 Gyrification model. a) FBM method. b) EBM method.

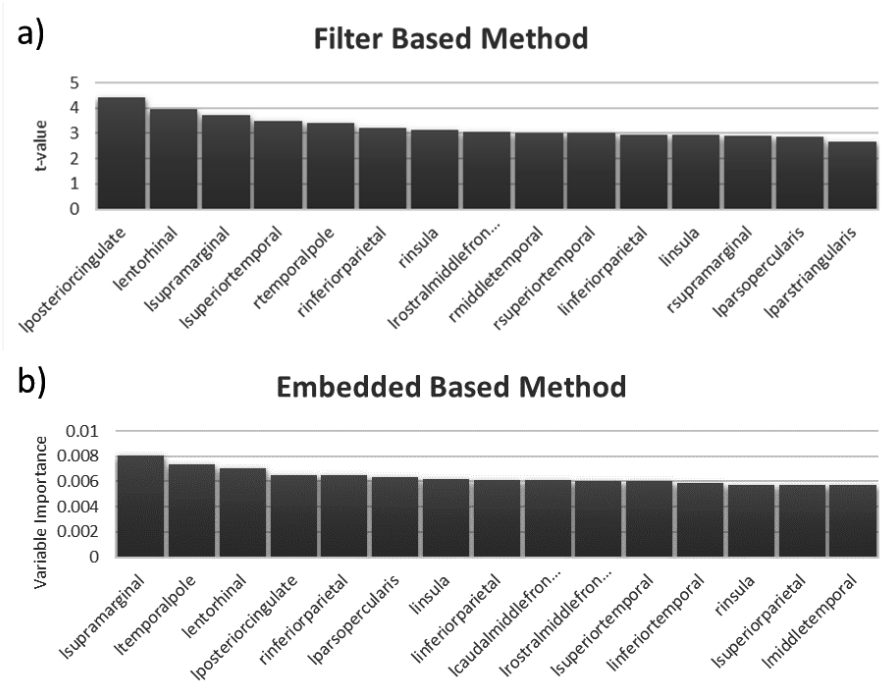

Figure S.12 Top ranking features for the DK40 Thickness model. a) FBM method. b) EBM method.

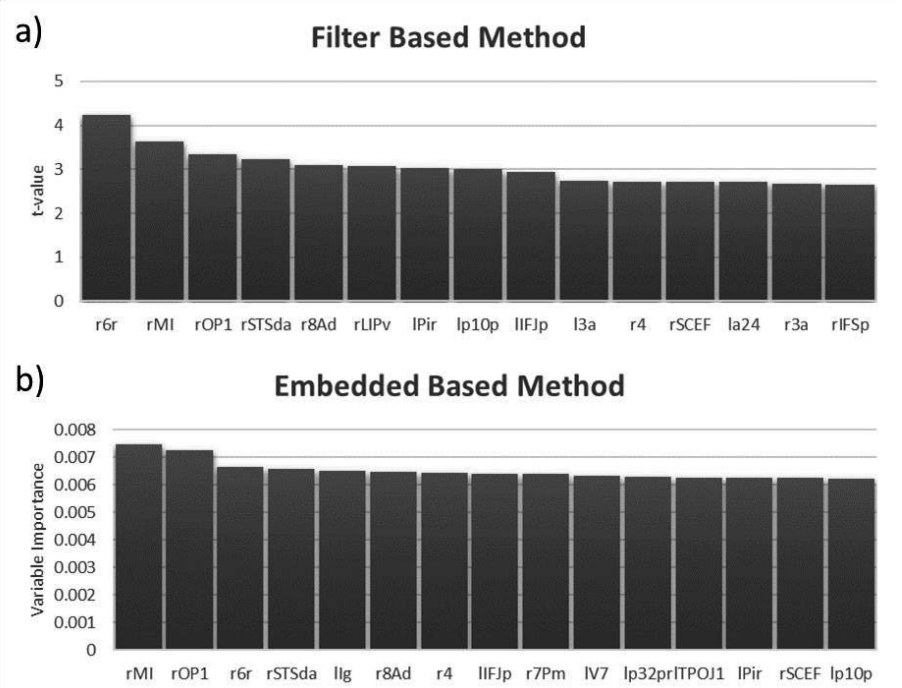

Figure S.13 Top ranking features for the HCP Gyrification model. a) FBM method. b) EBM method.

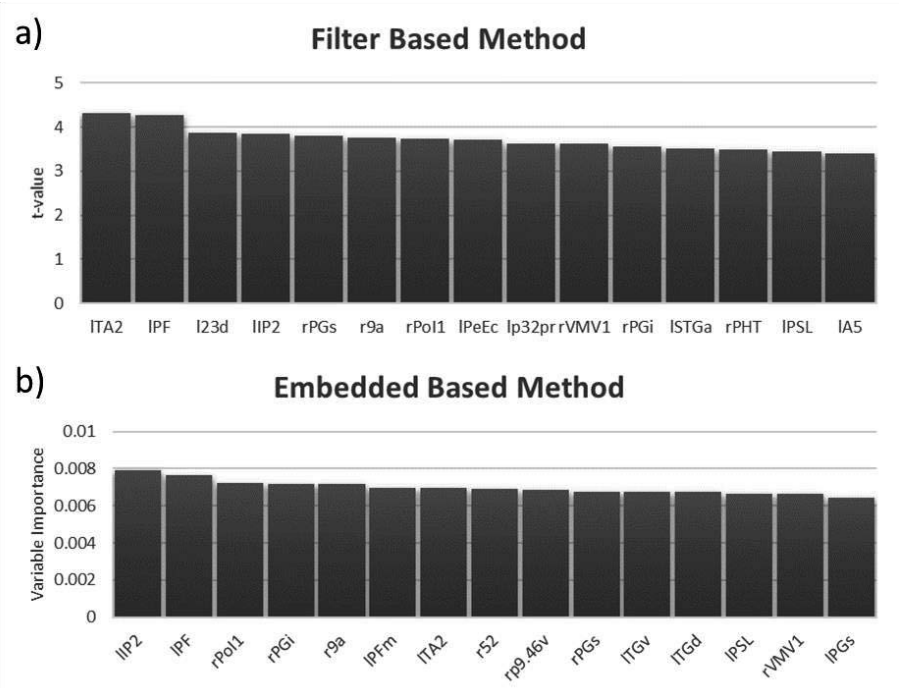

Figure S.14 Top ranking features for the HCP Thickness model. a) FBM method. b) EBM method.

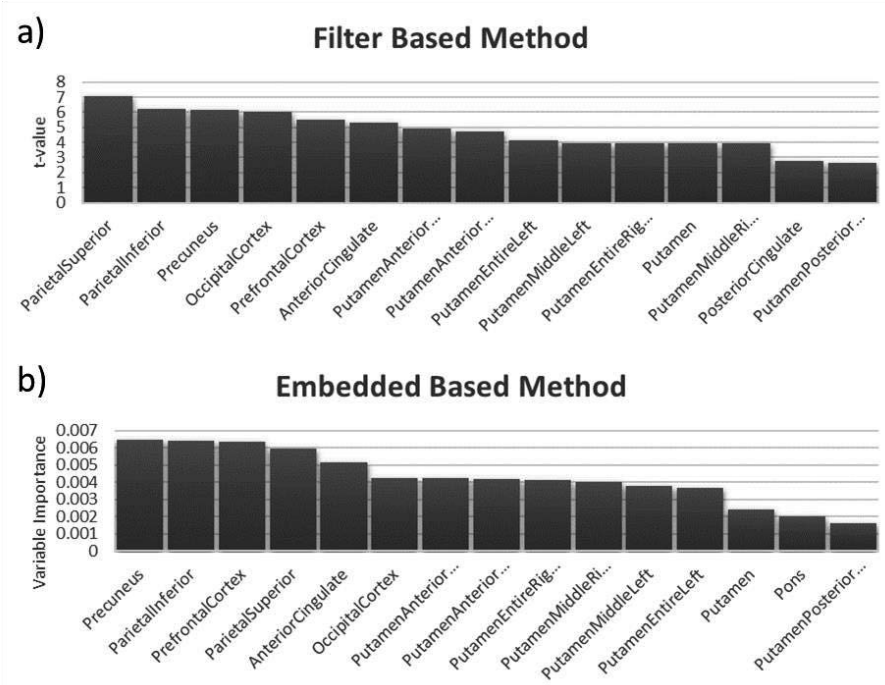

Figure S.15 Top ranking features for the SUVR Cerebellum model. a) FBM method. b) EBM method.

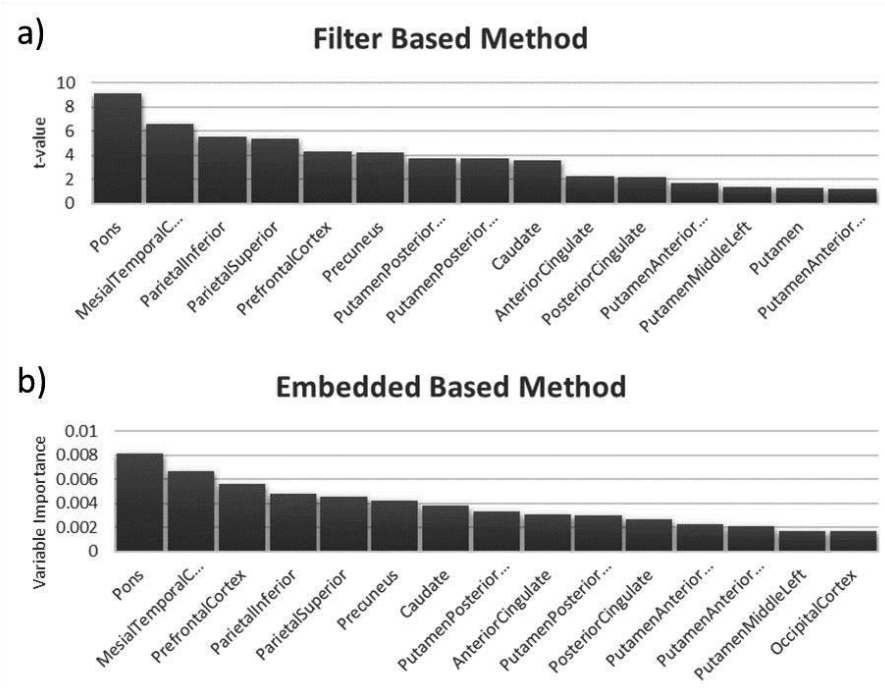

Figure S.16 Top ranking features for the SUVR GM model. a) FBM method. b) EBM method.

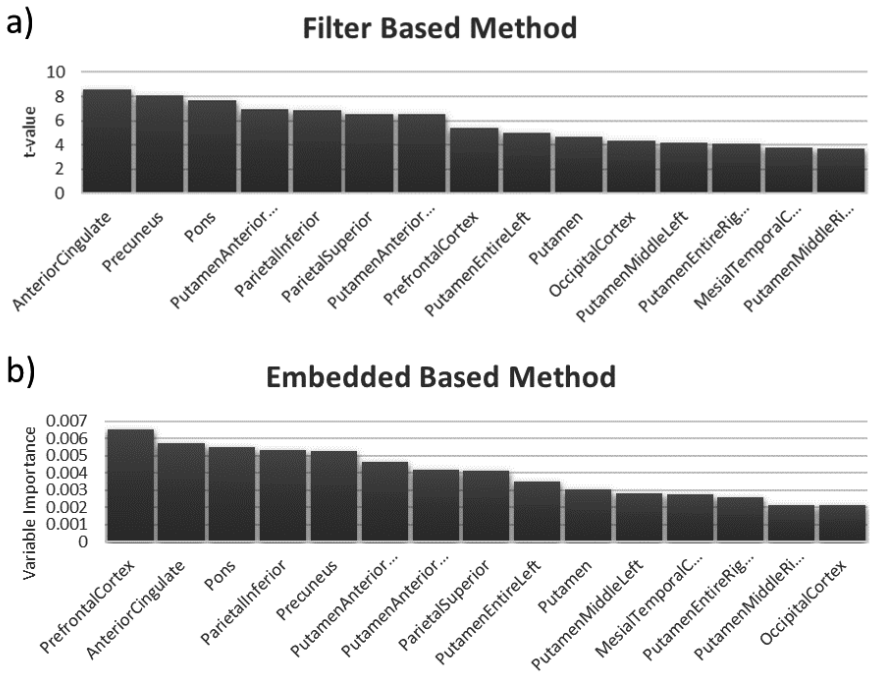

Figure S.17 Top ranking features for the SUVR WM model. a) FBM method. b) EBM method.

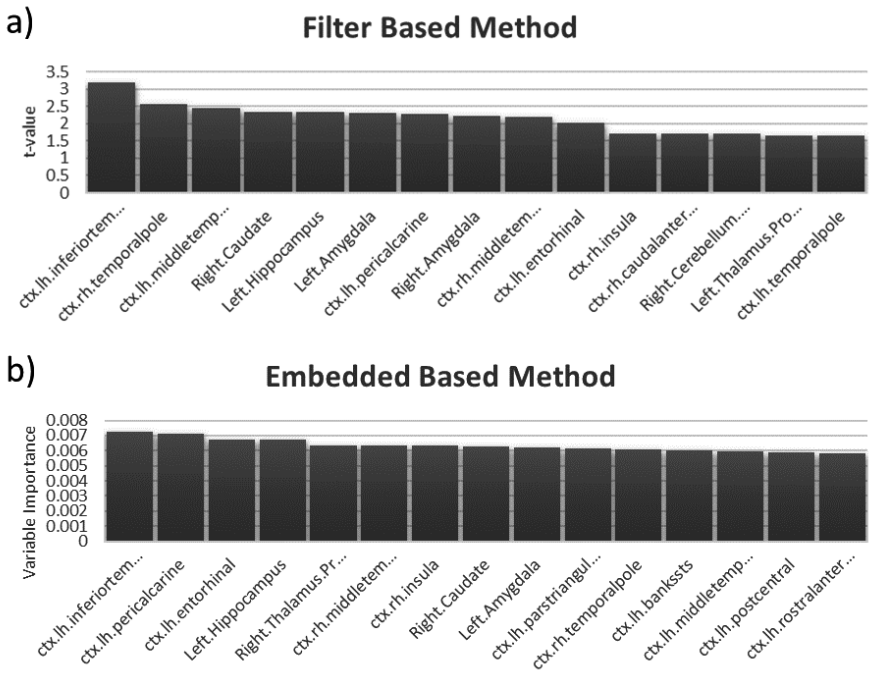

Figure S.18 Top ranking features for the Desikan FA model. a) FBM method. b) EBM method.

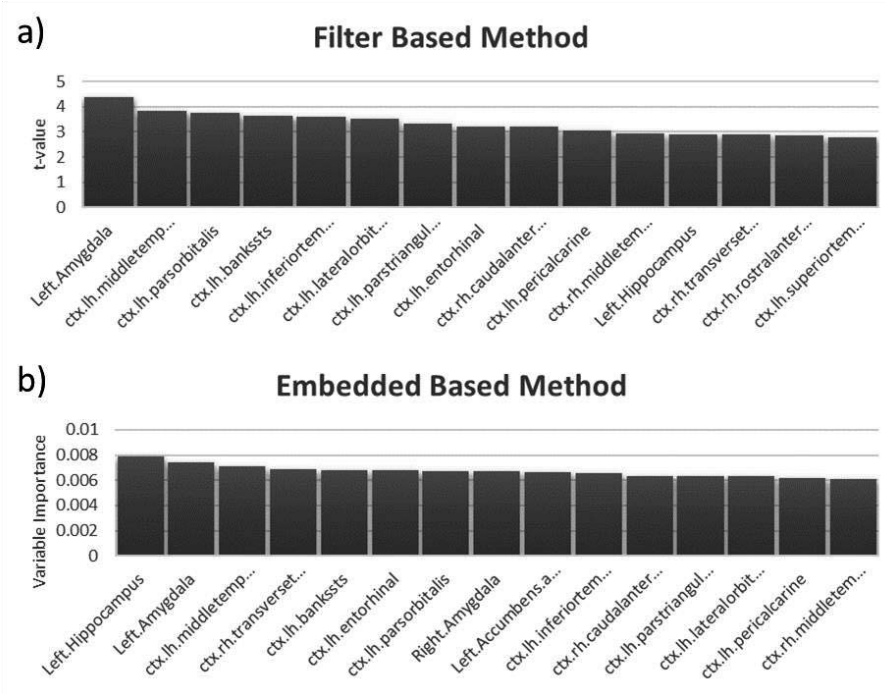

Figure S.19 Top ranking features for the Desikan MD model. a) FBM method. b) EBM method.

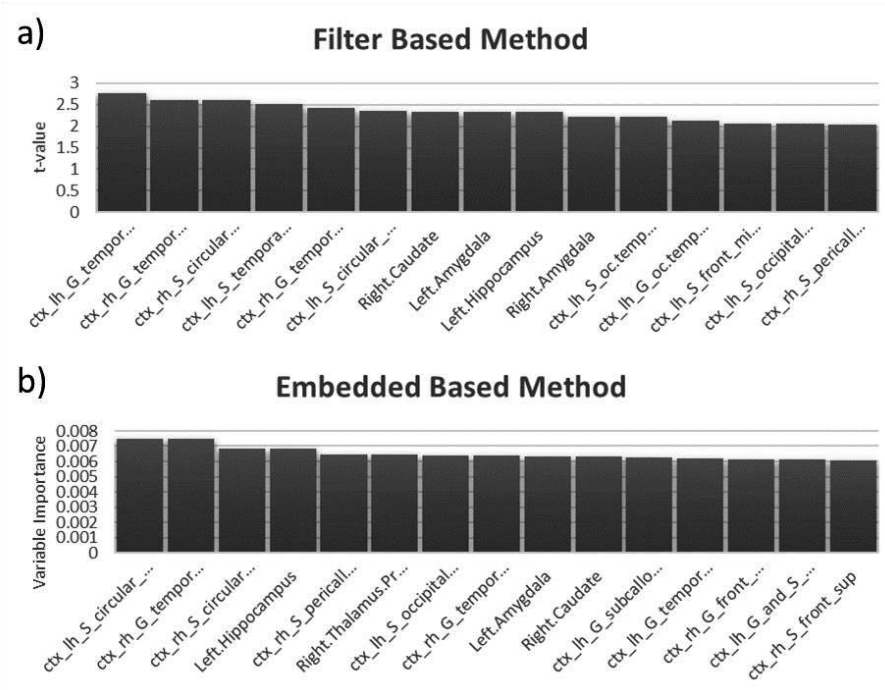

Figure S.20 Top ranking features for the Destrieux FA model. a) FBM method. b) EBM method.

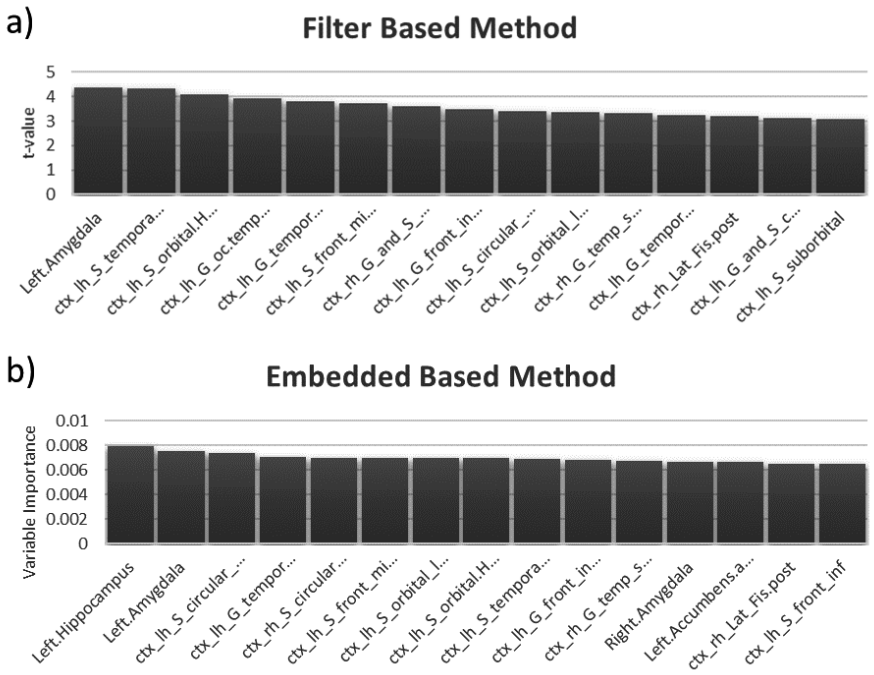

Figure S.21 Top ranking features for the Destrieux MD model. a) FBM method. b) EBM method.

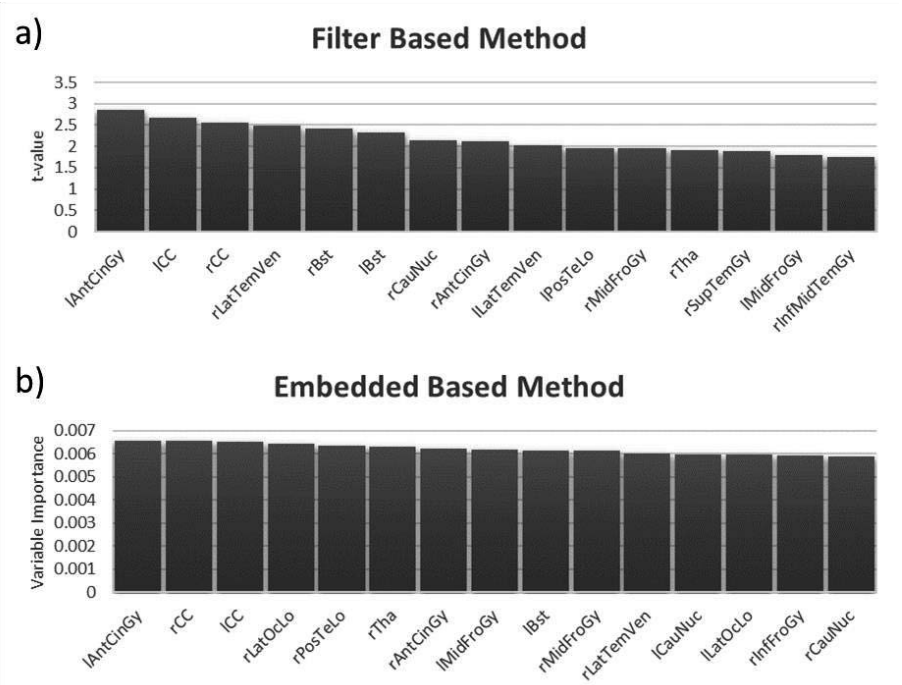

Figure S.22 Top ranking features for the Hammers FA model. a) FBM method. b) EBM method.

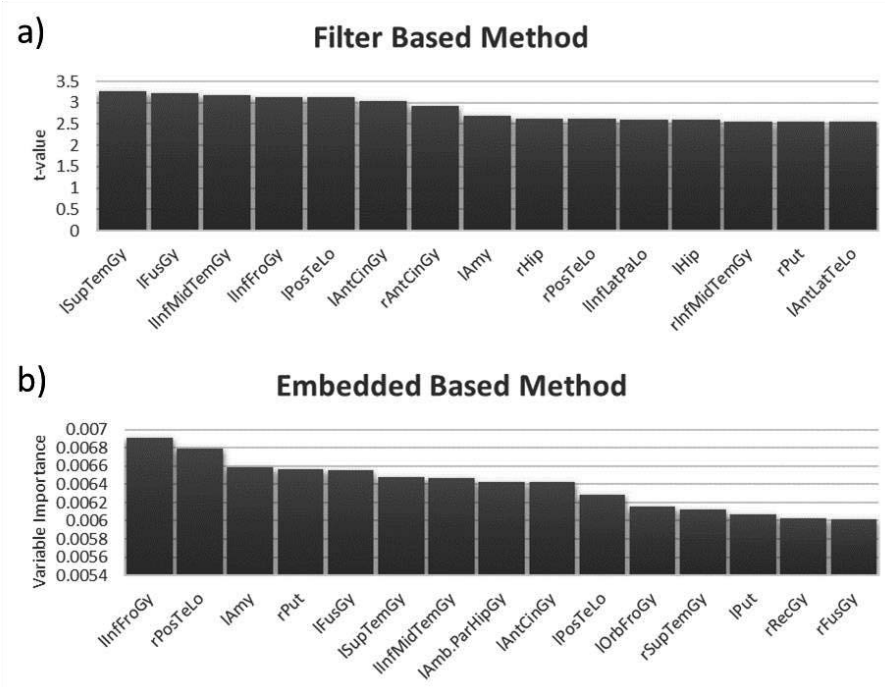

Figure S.23 Top ranking features for the Hammers MD model. a) FBM method. b) EBM method.

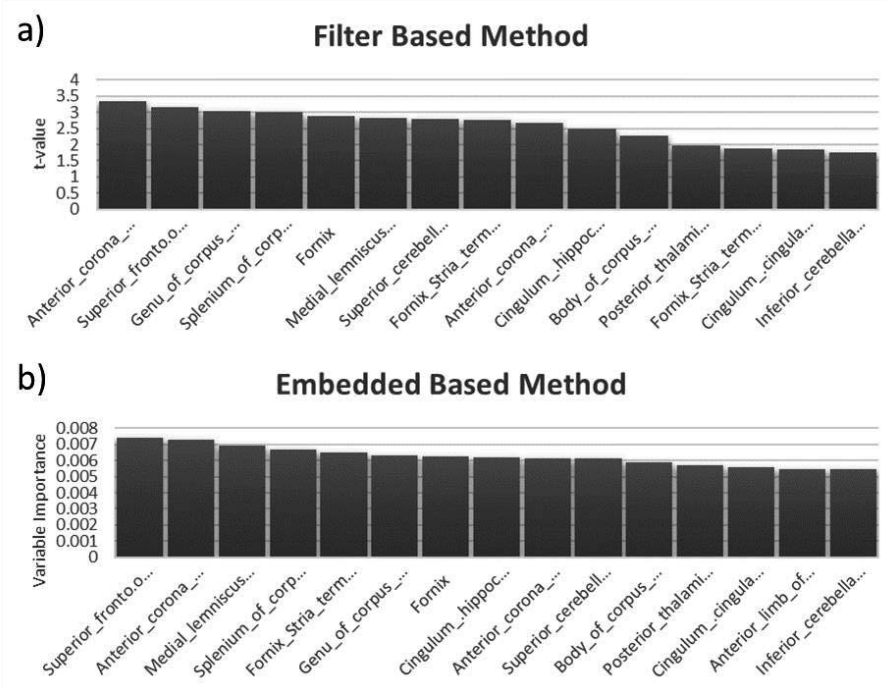

Figure S.24 Top ranking features for the JHU FA model. a) FBM method. b) EBM method.

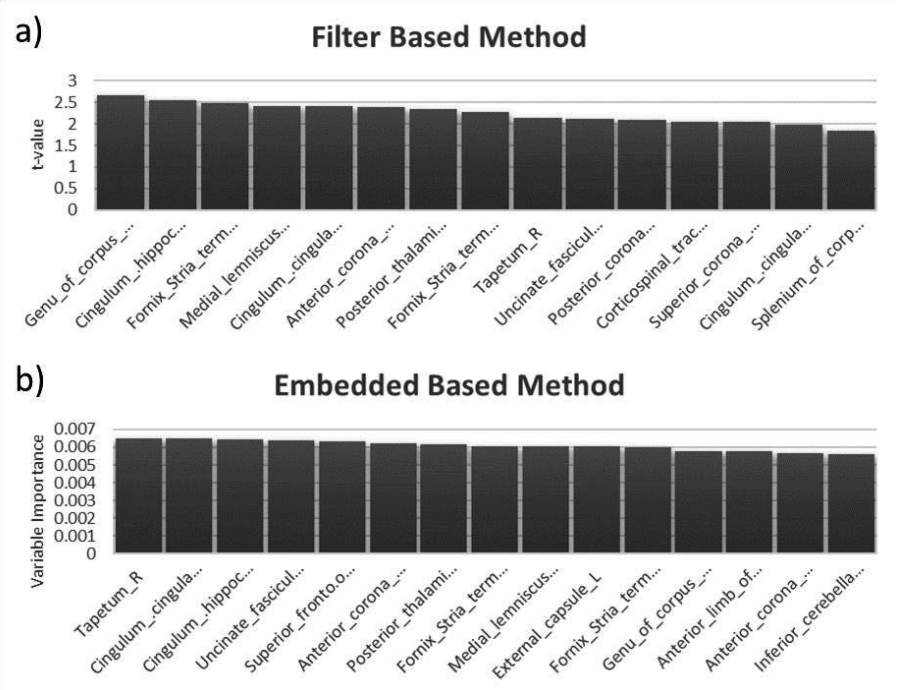

Figure S.25 Top ranking features for the JHU MD model. a) FBM method. b) EBM method.

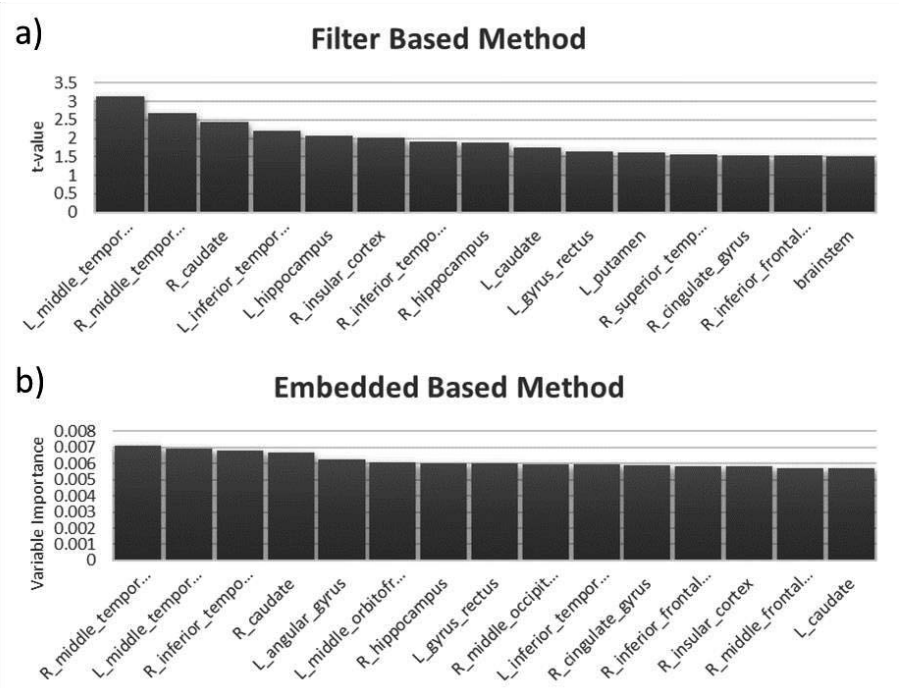

Figure S.26 Top ranking features for the Lpba40 FA model. a) FBM method. b) EBM method.

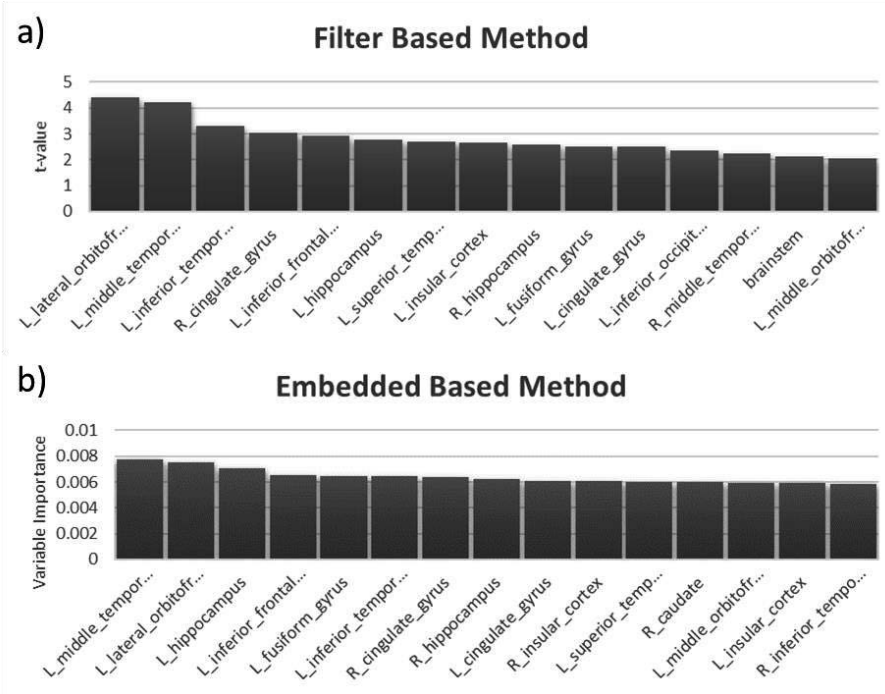

Figure S.27 Top ranking features for the Lpba40 MD model. a) FBM method. b) EBM method.

Feature Selection Final Feature Sets

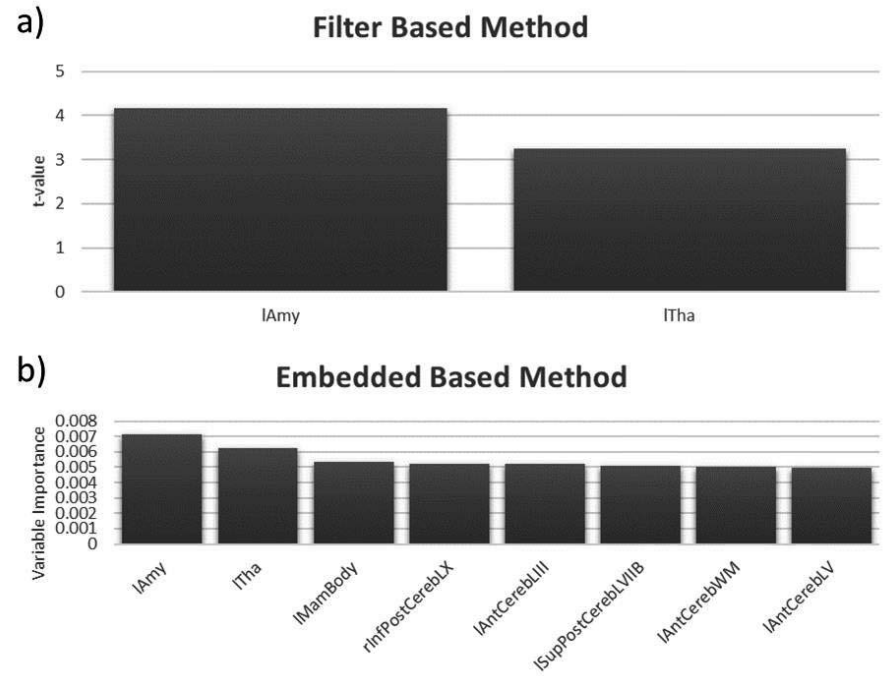

Figure S.28 Final feature set for the Cobra GM model. a) FBM method. b) EBM method.

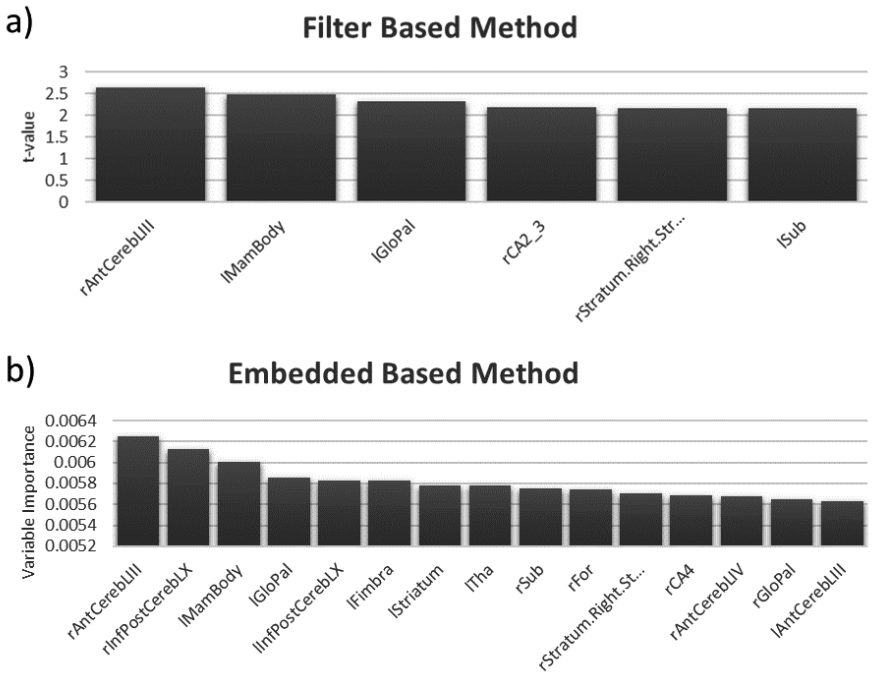

Figure S.29 Final feature set for the Cobra WM model. a) FBM method. b) EBM method.

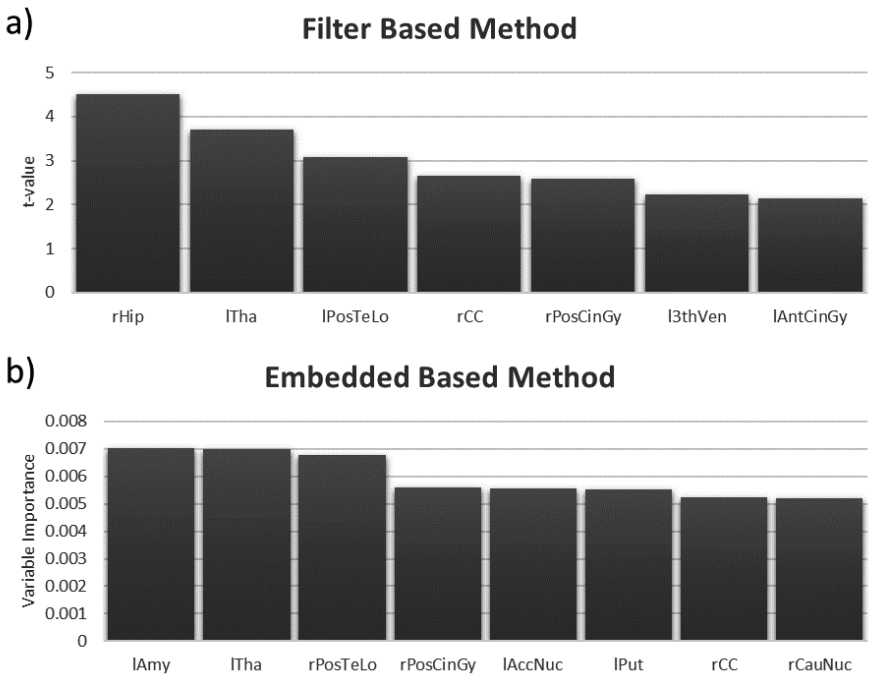

Figure S.30 Final feature set for the Hammers GM model. a) FBM method. b) EBM method.

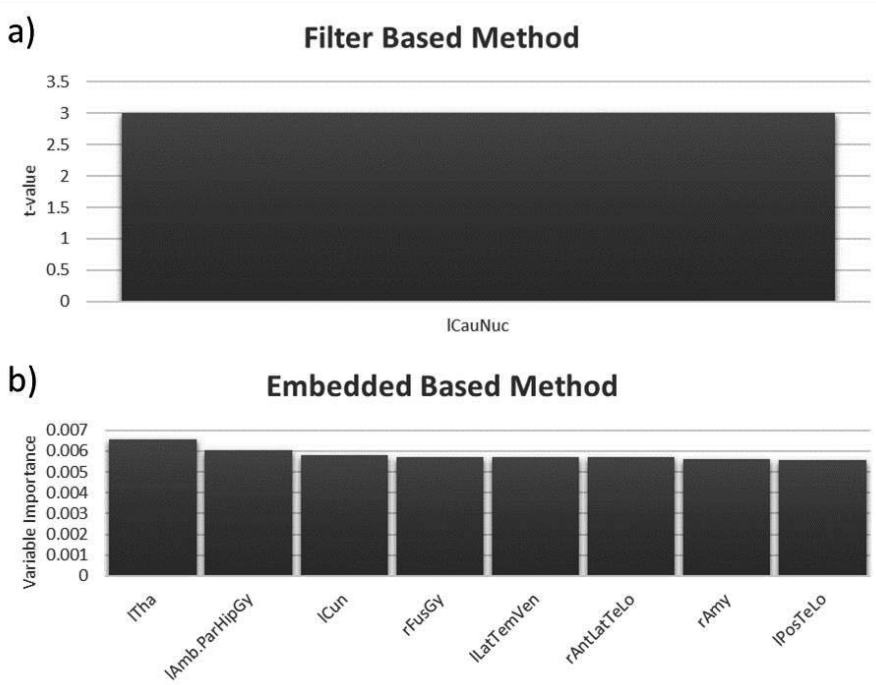

Figure S.31 Final feature set for the Hammers WM model. a) FBM method. b) EBM method.

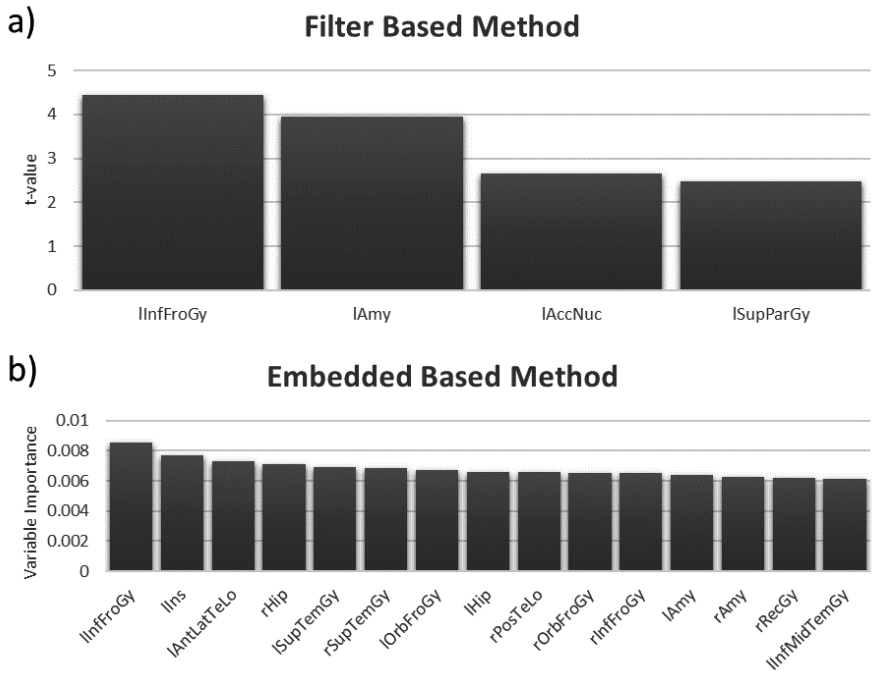

Figure S.32 Final feature set for the Hammers CSF model. a) FBM method. b) EBM method.

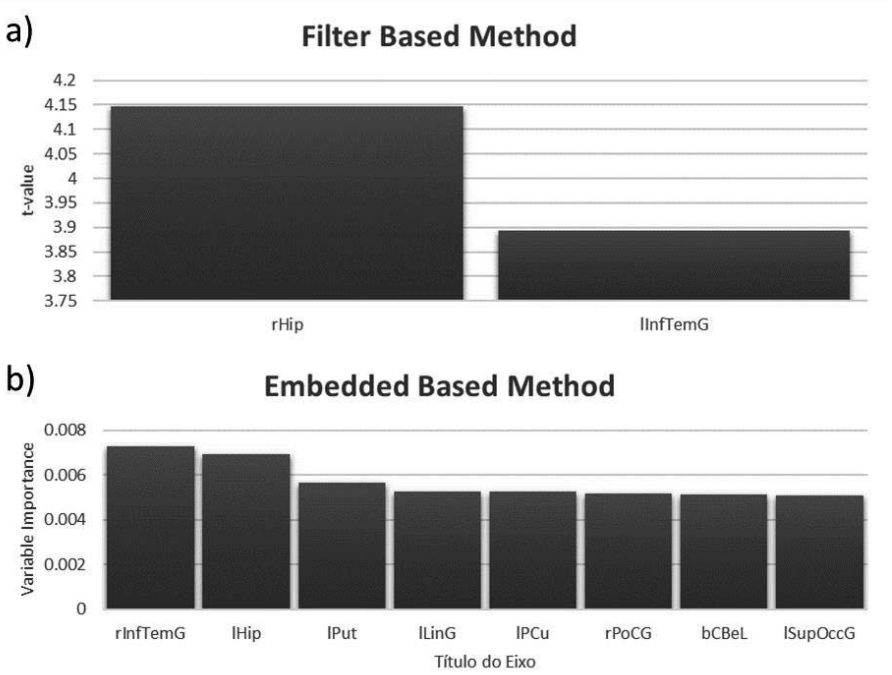

Figure S.33 Final feature set for the Lpba40 GM model. a) FBM method. b) EBM method.

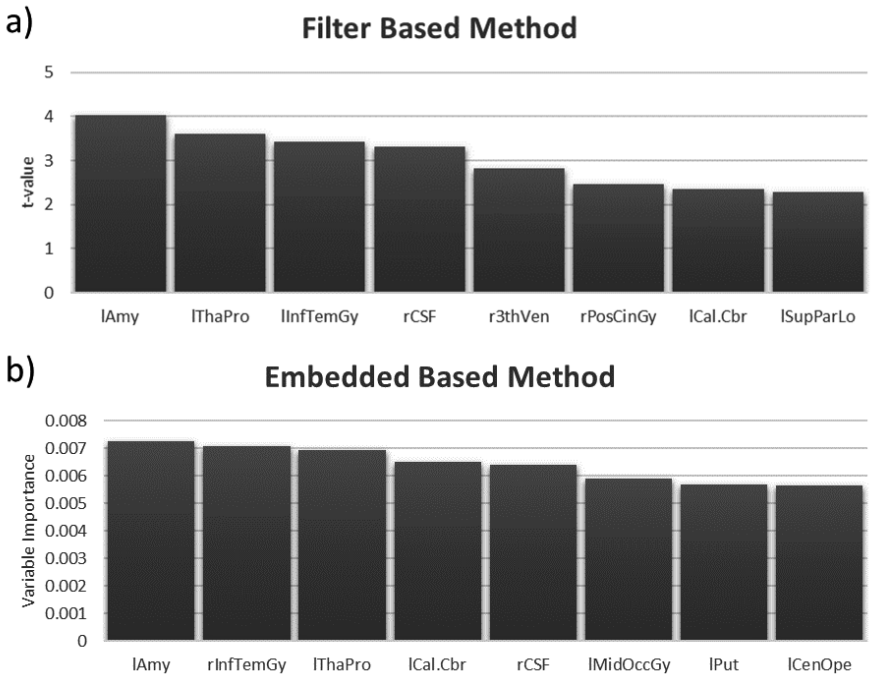

Figure S.34 Final feature set for the Neuromorphometrics GM model. a) FBM method. b) EBM method.

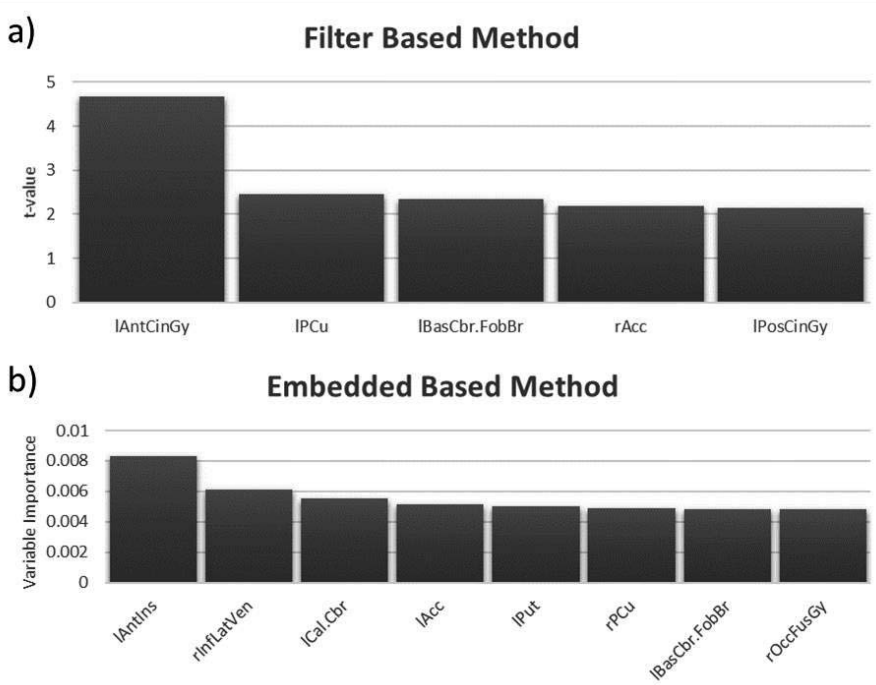

Figure S.35 Final feature set for the Neuromorphometrics CSF model. a) FBM method. b) EBM method.

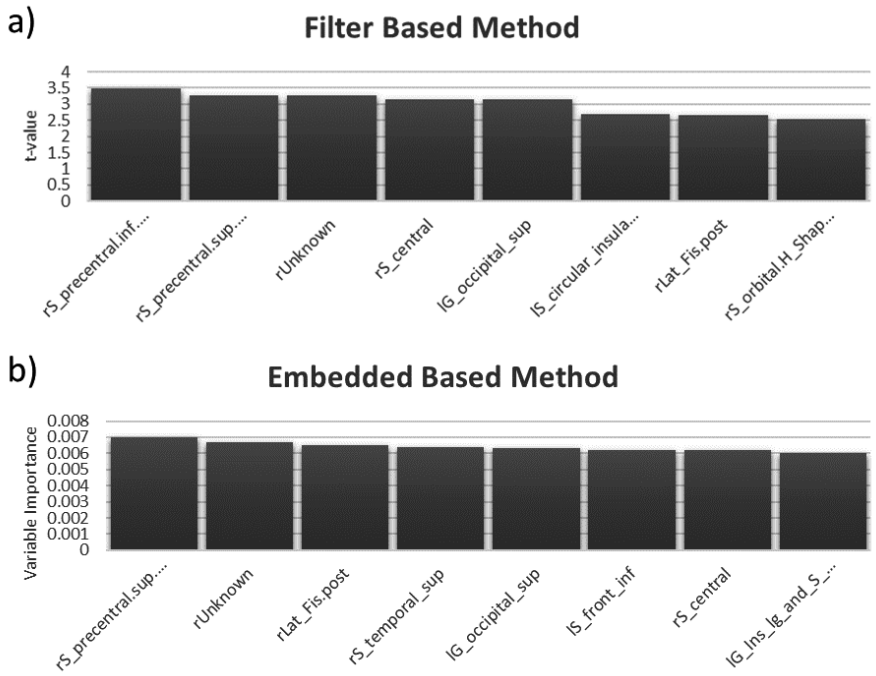

Figure S.36 Final feature set for the a2009 Gyrification model. a) FBM method. b) EBM method.

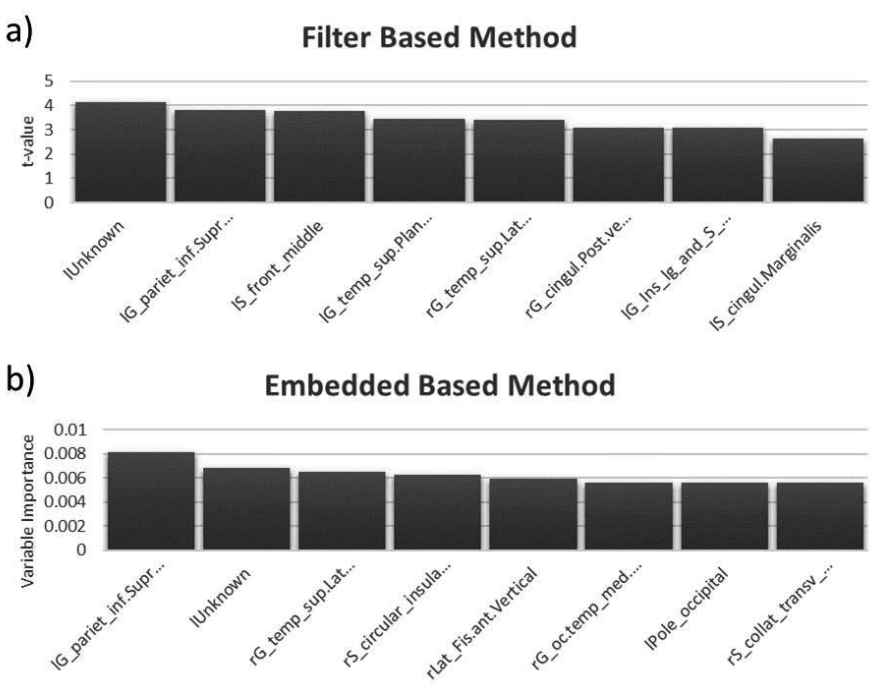

Figure S.37 Final feature set for the a2009 Thickness model. a) FBM method. b) EBM method.

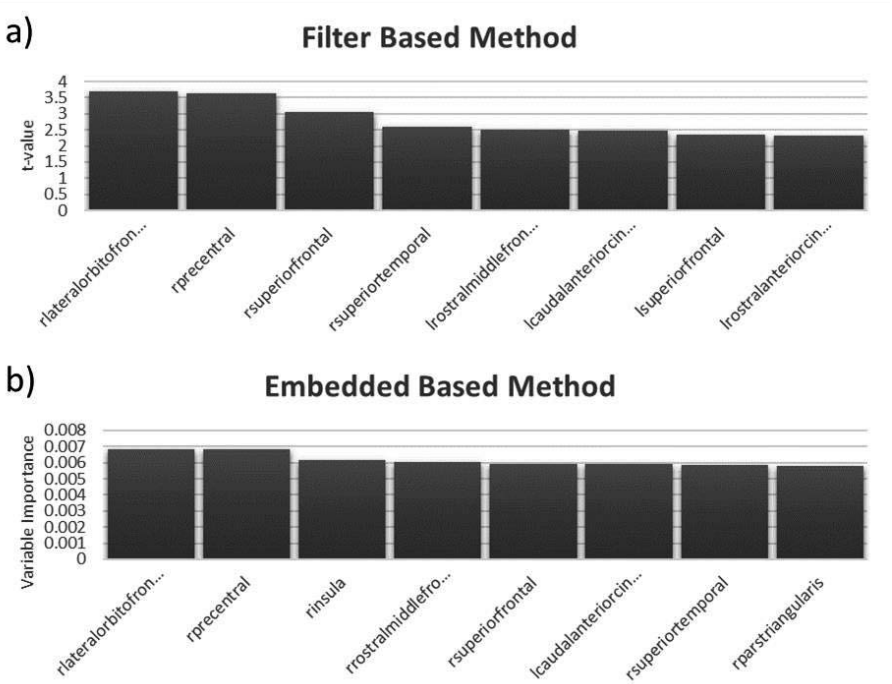

Figure S.38 Final feature set for the DK40 Gyrification model. a) FBM method. b) EBM method.

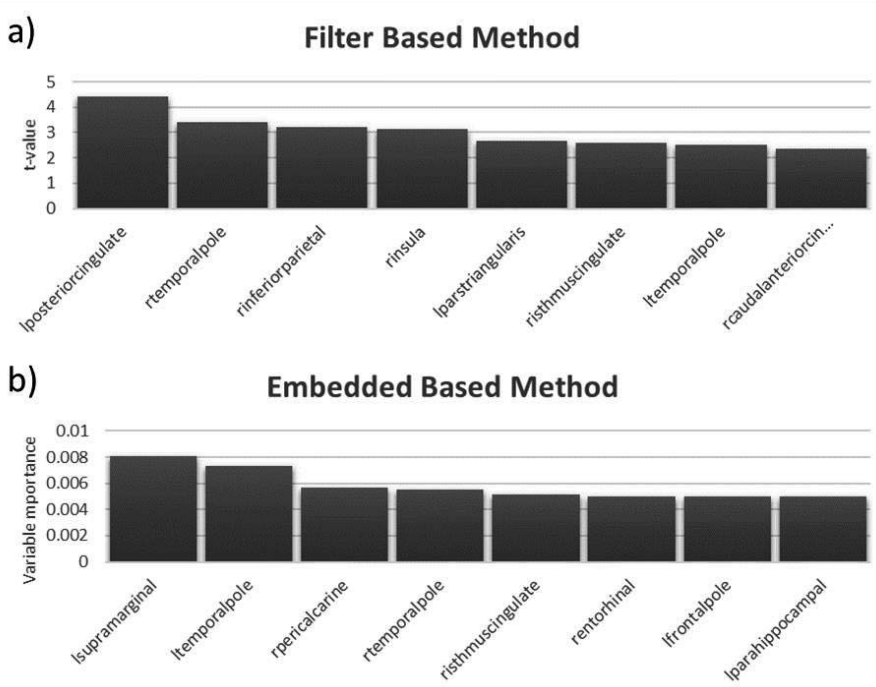

Figure S.39 Final feature set for the DK40 Thickness model. a) FBM method. b) EBM method.

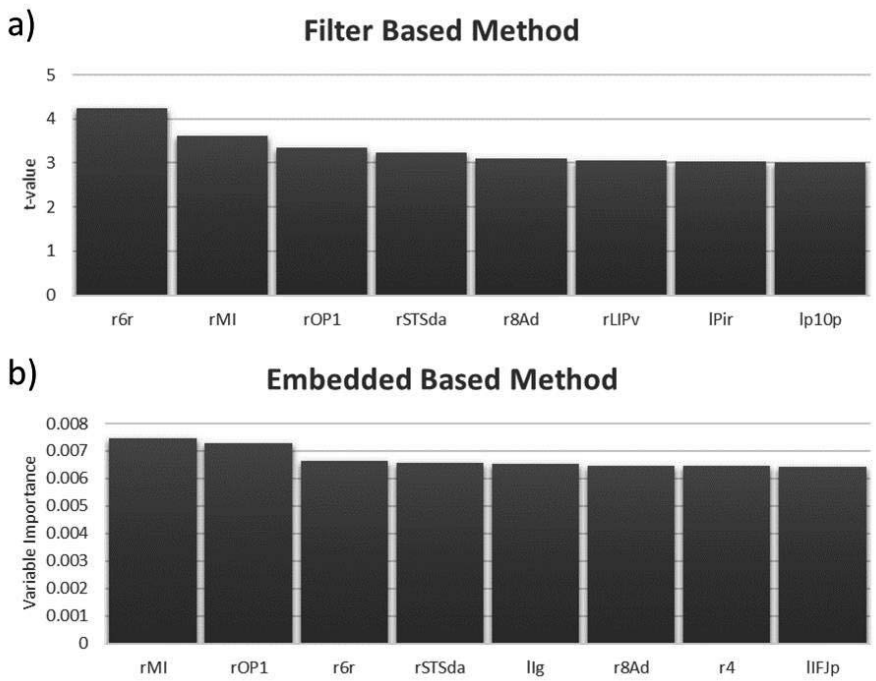

Figure S.40 Final feature set for the HCP Gyrification model. a) FBM method. b) EBM method.

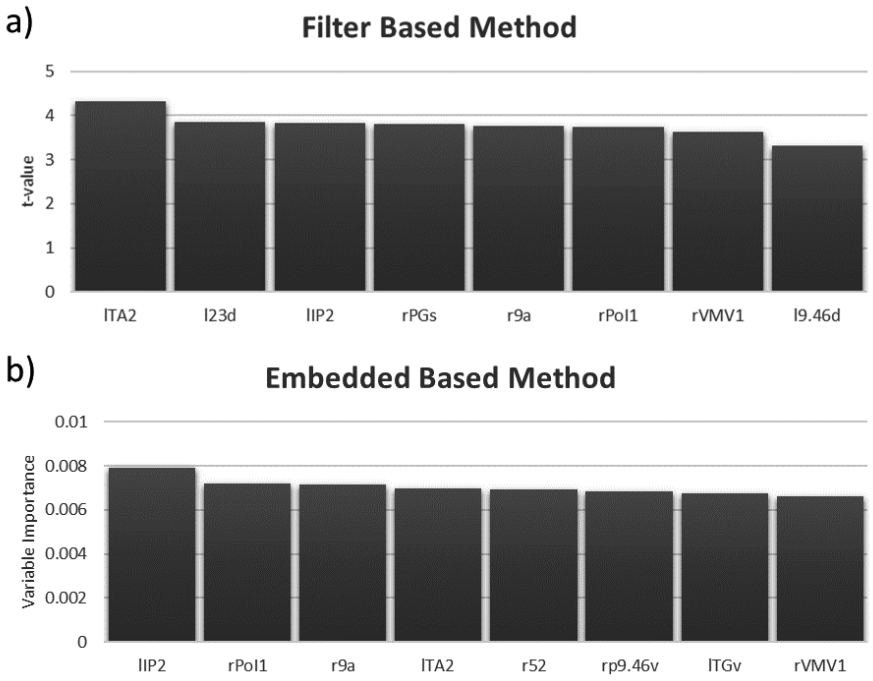

Figure S.41 Final feature set for the HCP Thickness model. a) FBM method. b) EBM method.

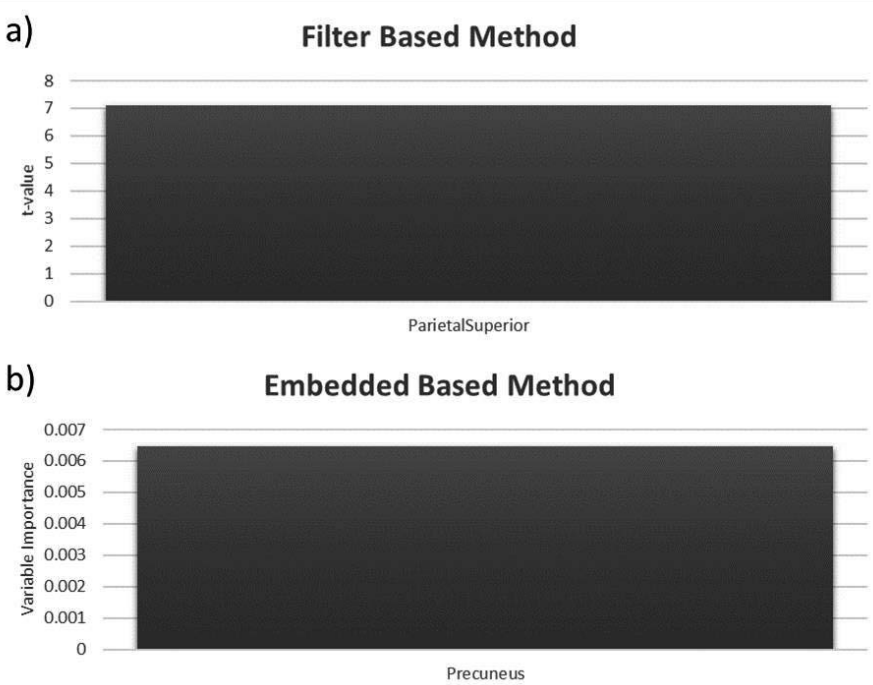

Figure S.42 Final feature set for the SUVR Cerebellum model. a) FBM method. b) EBM method.

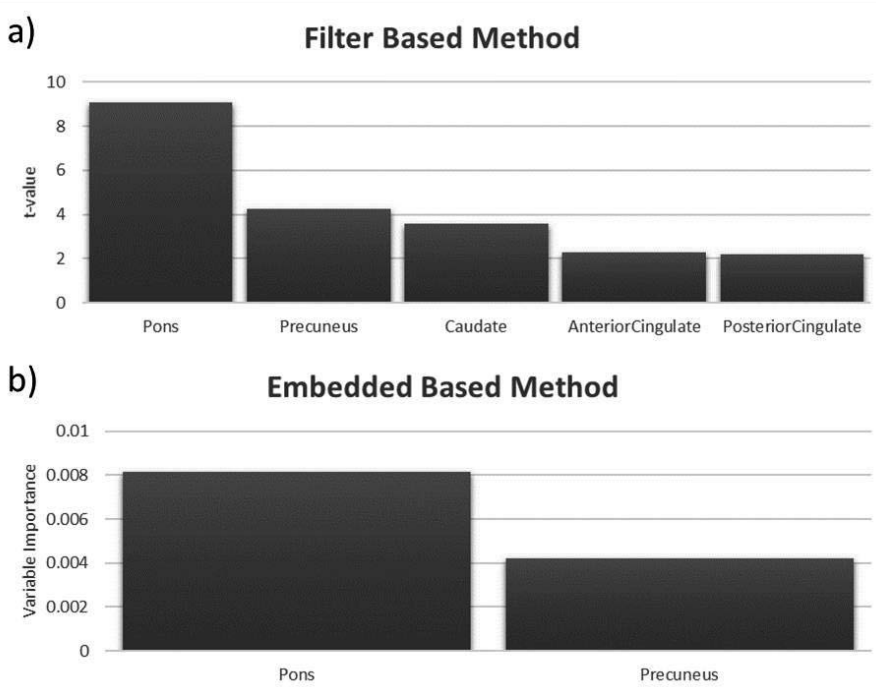

Figure S.43 Final feature set for the SUVR GM model. a) FBM method. b) EBM method.

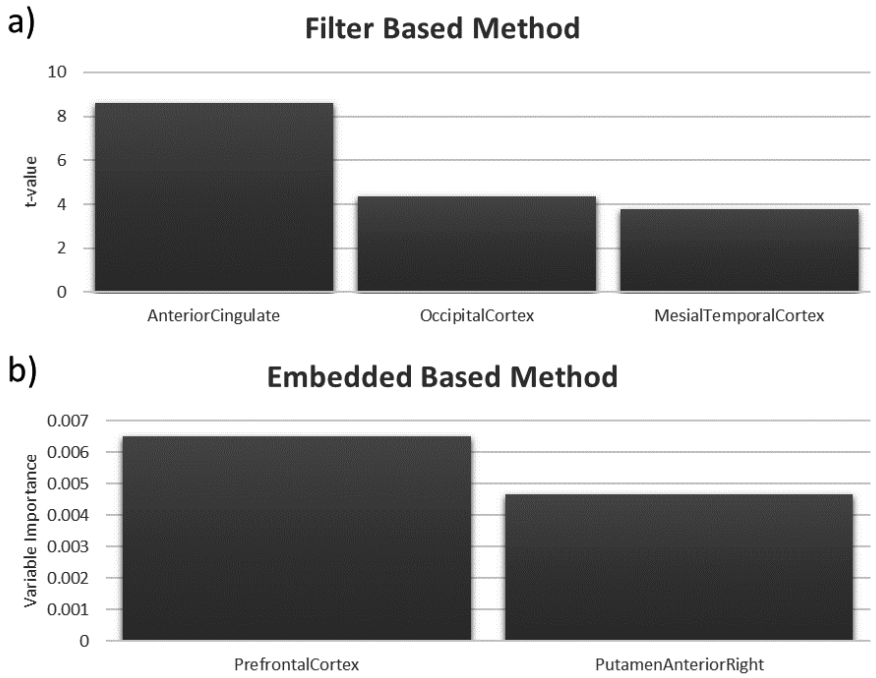

Figure S.44 Final feature set for the SUVR WM model. a) FBM method. b) EBM method.

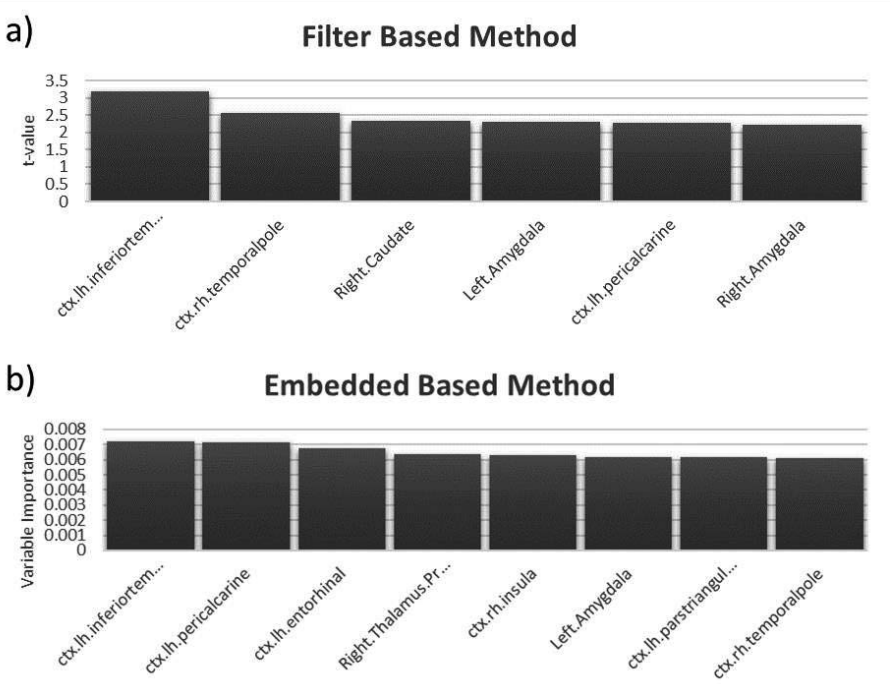

Figure S.45 Final feature set for the Desikan FA model. a) FBM method. b) EBM method.

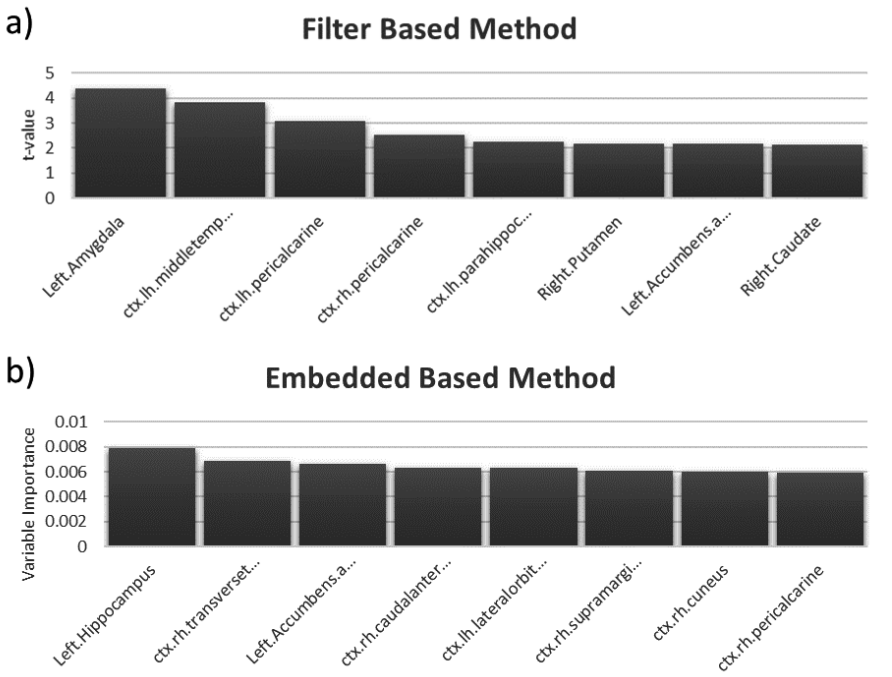

Figure S.46 Final feature set for the Desikan MD model. a) FBM method. b) EBM method.

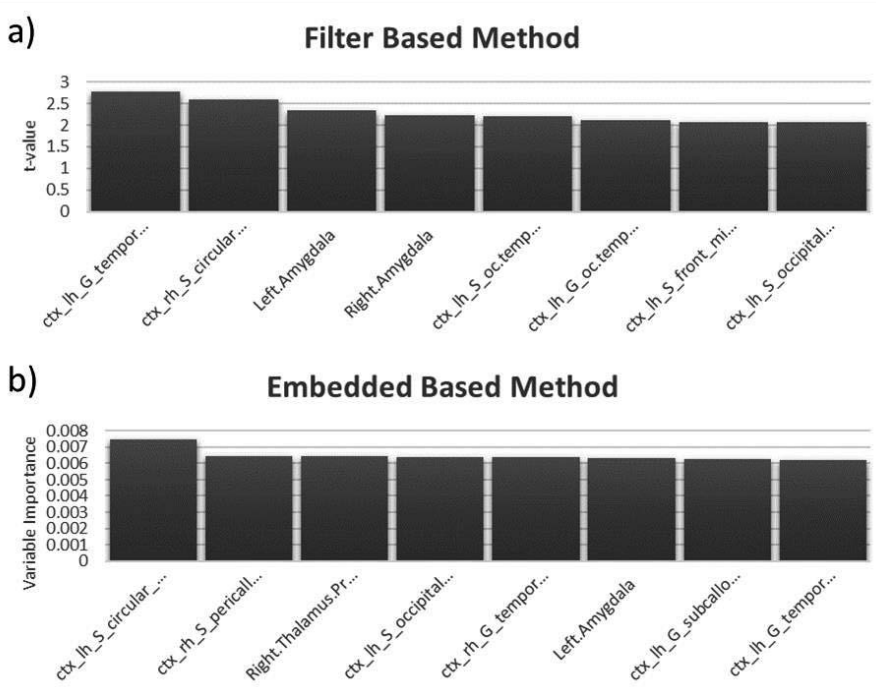

Figure S.47 Final feature set for the Destrieux FA model. a) FBM method. b) EBM method.

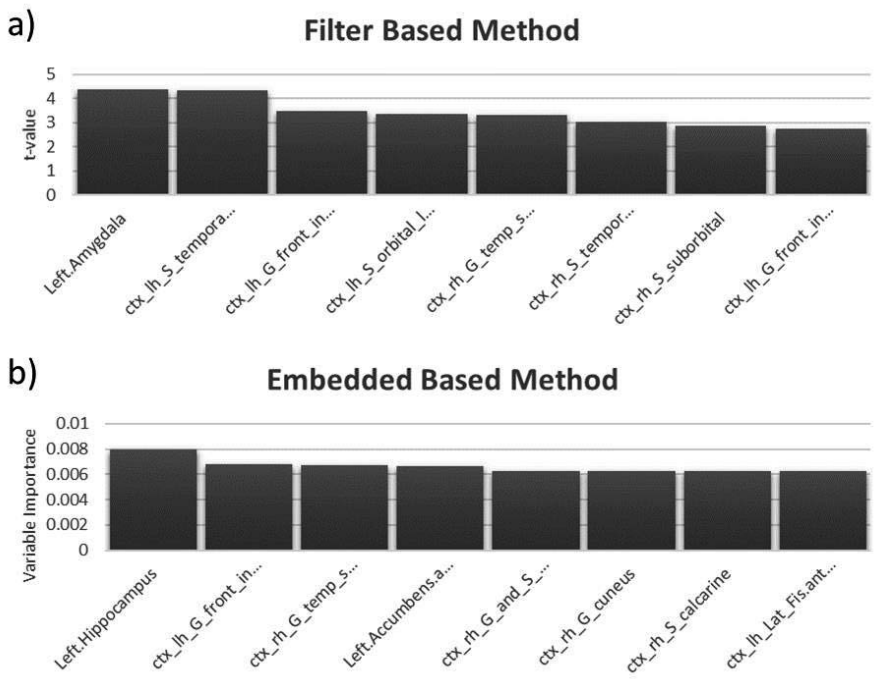

Figure S.48 Final feature set for the Destrieux MD model. a) FBM method. b) EBM method.

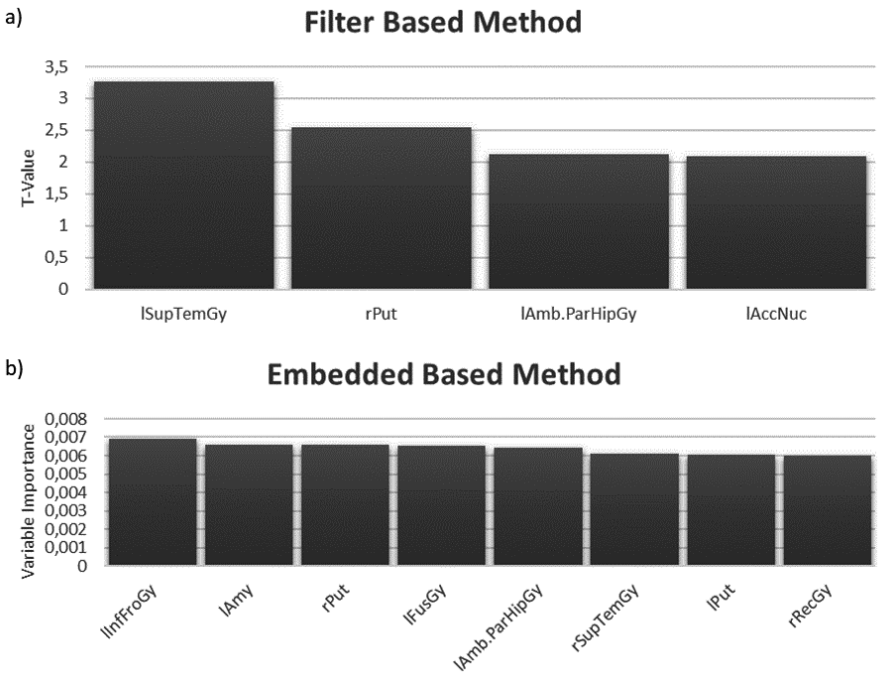

Figure S.49 Final feature set for the Hammers FA model. a) FBM method. b) EBM method.

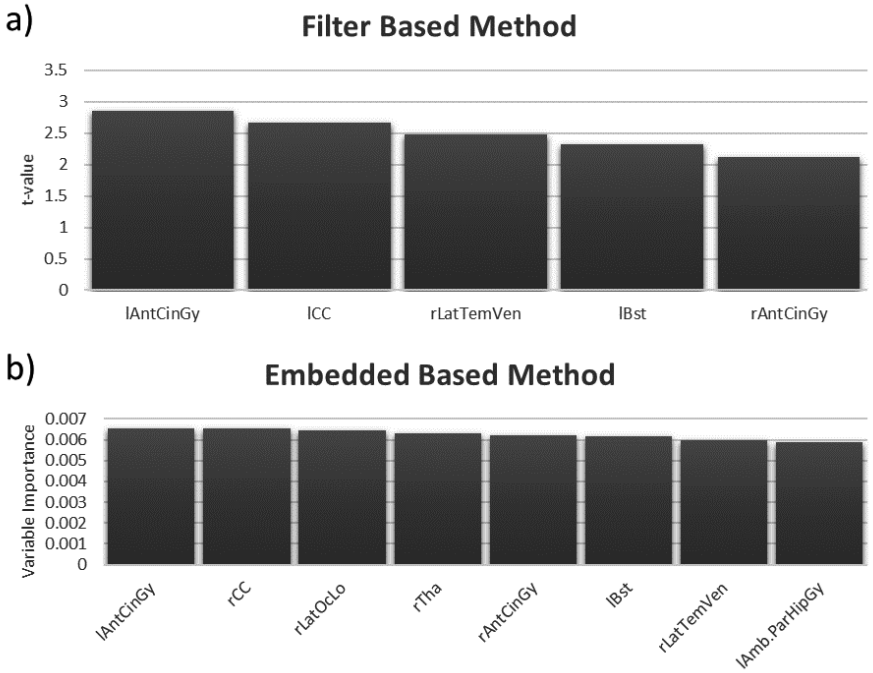

Figure S.50 Final feature set for the Hammers MD model. a) FBM method. b) EBM method.

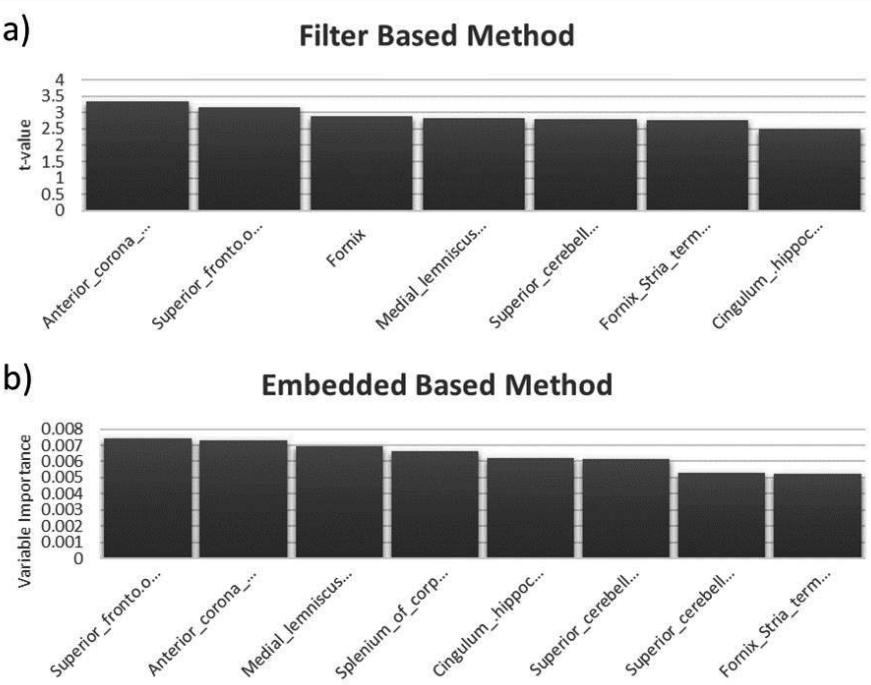

Figure S.51 Final feature set for the JHU FA model. a) FBM method. b) EBM method.

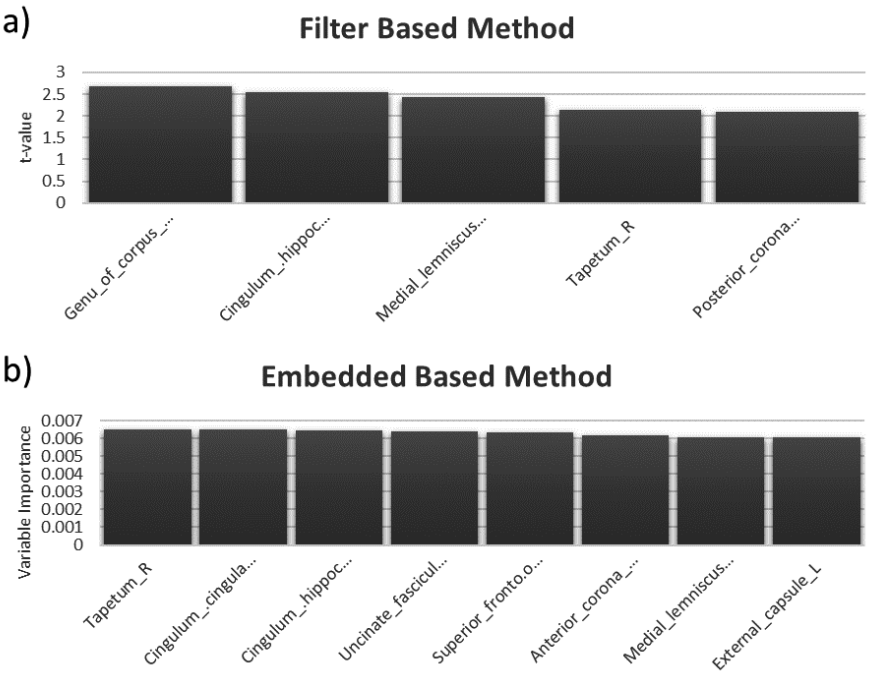

Figure S.52 Final feature set for the JHU MD model. a) FBM method. b) EBM method.

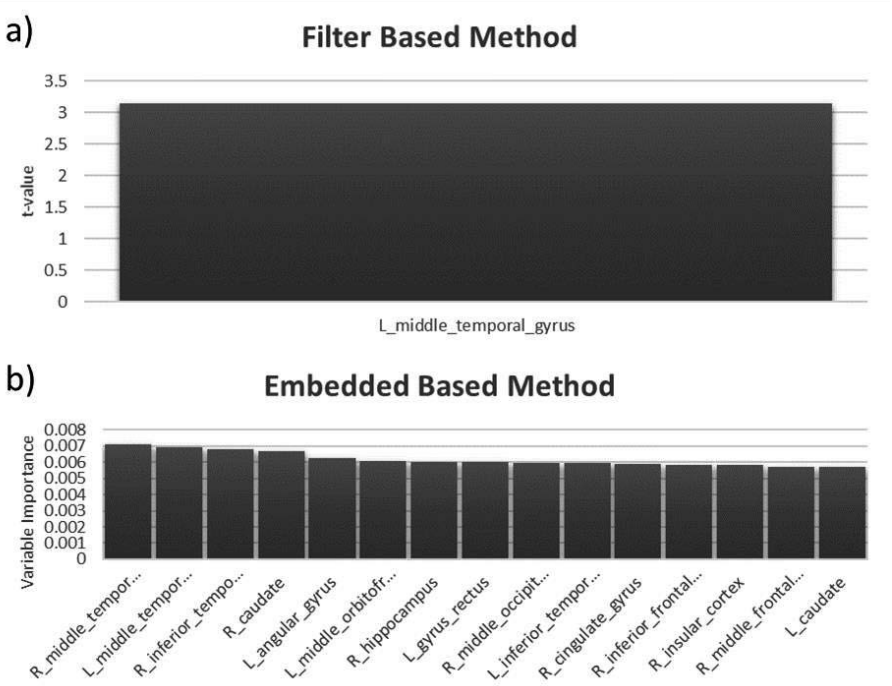

Figure S.53 Final feature set for the Lpba40 FA model. a) FBM method. b) EBM method.

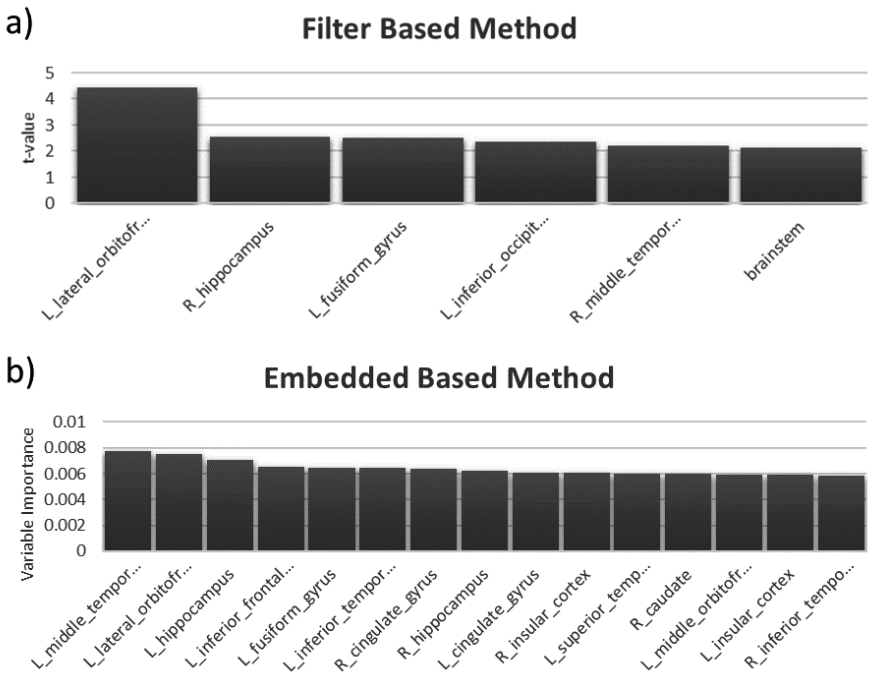

Figure S.54 Final feature set for the Lpba40 MD model. a) FBM method. b) EBM method.

Individual classifier performance

Table S.6 Performance results for all models.

Model	Feature Selection Method	SVM				Logistic Regression			
		AUC	Mean Accuracy	Mean Sensitivity	Mean Specificity	AUC	Mean Accuracy	Mean Sensitivity	Mean Specificity
$Cobra_{GM}$	**EBM**	0.87±0.13	76.13%	76.06%	76.19%	0.84±0.14	76.14%	73.28%	79.00%
	FBM	0.87±0.13	82.08%	73.63%	90.54%	0.83±0.14	72.76%	74.09%	71.44%
$Cobra_{WM}$	**EBM**	0.88±0.12	78.43%	79.04%	77.04%	0.92±0.10	80.64%	78.41%	82.88%
	FBM	0.86±0.13	77.06%	77.69%	76.44%	0.91±0.10	79.34%	77.54%	81.54%
$Hammers_{GM}$	**EBM**	0.85±0.15	74.79%	66.14%	83.44%	0.86±0.13	78.68%	74.60%	82.75%
	FBM	0.77±0.20	72.92%	67.26%	78.57%	0.83±0.14	73.09%	71.51%	74.68%
$Hammers_{WM}$	**EBM**	0.51±0.24	58.36%	58.24%	58.49%	0.71±0.18	63.22%	60.34%	66.10%
	FBM	0.50±0.23	58.27%	56.57%	59.96%	0.74±0.17	64.29%	62.30%	66.27%
$Hammers_{CSF}$	**EBM**	0.85±0.14	75.37%	77.01%	73.73%	0.83±0.15	73.51%	65.70%	81.33%
	FBM	0.79±0.17	71.65%	72.24%	71.06%	0.86±0.12	74.94%	70.09%	79.79%
$Lpba40_{GM}$	**EBM**	0.76±0.20	70.28%	61.16%	79.40%	0.77±0.17	70.21%	66.65%	93.27%
	FBM	0.78±0.17	73.93%	72.24%	71.06%	0.85±0.13	73.99%	76.66%	71.32%
$Neuro_{GM}$	**EBM**	0.96±0.07	92.05%	86.78%	97.32%	0.93±0.09	85.13%	76.8%	93.27%
	FBM	0.93±0.09	86.02%	82.70%	80.64%	0.95±0.08	86.98%	84.098%	89.53%
$Neuro_{CSF}$	**EBM**	0.84±0.15	76.13%	80.34%	71.91%	0.84±0.15	75.96%	69.91	82.00%
	FBM	0.61±0.23	60.94%	64.78%	57.10%	0.79±0.16	68.48%	66.69%	70.28%
$a2009_{Gyri}$	**EBM**	0.91±0.10	83.01%	86.98%	79.04%	0.89±0.12	81.91%	81.84%	81.99%
	FBM	0.86±0.13	76.45%	77.70%	75.20%	0.89±0.11	81.37%	83.37%	79.46%
$a2009_{Thick}$	**EBM**	0.90±0.11	81.89%	81.93%	81.87%	0.92±0.1	84.78%	84.01%	85.54%
	FBM	0.84±0.15	81.47%	82.76%	80.18%	0.84±0.13	80.13%	79.53%	80.74%
$DK40_{Gyri}$	**EBM**	0.75±0.19	69.13%	73.14%	65.11%	0.80±0.15	67.88%	66.79%	68.98%
	FBM	0.78±0.18	69.57%	73.99%	65.15%	0.81±0.14	68.84%	69.65%	68.03%
$DK40_{Thick}$	**EBM**	0.79±0.17	72.19%	71.01%	73.36%	0.83±0.14	72.95%	71.94%	73.96%
	FBM	0.82±0.15	75.07%	74.86%	75.28%	0.82±0.14	75.19%	75.78%	74.60%

HCP_{Gyri}	**EBM**	0.94±0.08	83.11%	73.65%	92.56%	0.95±0.07	88.25%	83.85%	92.66%
	FBM	0.96±0.07	85.59%	79.51%	91.98%	0.96±0.06	89.49%	86.05%	92.93%
HCP_{Thick}	**EBM**	0.93±0.09	84.27%	83.05%	85.49%	0.90±0.11	83.16%	80.71%	85.60%
	FBM	0.93±0.09	88.07%	84.59%	91.55%	0.93±0.09	87.02%	85.15%	88.89%
$SUVR_{Cere}$	**EBM**	0.94±0.09	87.71%	95.92%	81.26%	0.96±0.07	91.01%	97.47%	86.18%
	FBM	0.95±0.07	88.06%	96.67%	81.60%	0.97±0.05	87.69%	93.02%	83.70%
$SUVR_{GM}$	**EBM**	0.97±0.06	91.98%	98.66%	87.01%	0.96±0.07	88.44%	93.23%	84.85%
	FBM	0.97±0.06	87.15%	95.73%	80.71%	0.96±0.06	87.56%	91.02%	84.98%
$SUVR_{WM}$	**EBM**	0.93±0.10	90.68%	92.78%	89.10%	0.95±0.07	89.99%	89.20%	90.58%
	FBM	0.94±0.10	93.52%	99.72%	88.90%	0.96±0.07	90.64%	90.73%	90.56%
$Desikan_{FA}$	**EBM**	0.86±0.15	79.84%	76.67%	82.23%	0.81±0.16	75.13%	70.88%	78.31%
	FBM	0.86±0.14	79.68%	72.48%	85.08%	0.83±0.16	77.63%	73.73%	80.55%
$Desikan_{MD}$	**EBM**	0.66±0.24	62.00%	52.70%	68.98%	0.74±0.18	64.90%	57.81%	70.21%
	FBM	0.85±0.15	76.66%	63.72%	86.38%	0.81±0.16	70.64%	58.27%	79.91%
$Destrieux_{FA}$	**EBM**	0.66±0.25	65.66%	63.23%	67.49%	0.68±0.19	62.29%	57.81%	70.21%
	FBM	0.66±0.24	66.31%	55.63%	74.31%	0.76±0.18	71.51%	69.02%	73.38%
$Destrieux_{MD}$	**EBM**	0.84±0.15	77.60%	72.15%	81.69%	0.83±0.15	74.21%	66.25%	80.19%
	FBM	0.93±0.10	88.14%	81.68%	92.68%	0.92±0.11	87.45%	82.90%	90.88%
$Hammers_{FA}$	**EBM**	0.79±0.19	76.77%	63.80%	86.50%	0.70±0.21	69.06%	65.72%	71.58%
	FBM	0.72±0.22	74.12%	57.68%	86.45%	0.76±0.20	75.53%	72.95%	77.46%
$Hammers_{MD}$	**EBM**	0.72±0.22	69.52%	55.63%	79.94%	0.70±0.20	67.11%	54.17%	76.81%
	FBM	0.73±0.22	70.83%	53.67%	83.70%	0.71±0.20	68.23%	53.43%	79.33%
JHU_{FA}	**EBM**	0.80±0.19	77.24%	67.15%	84.81%	0.79±0.17	74.71%	68.15%	79.64%
	FBM	0.83±0.17	79.35%	71.05%	85.58%	0.80±0.17	76.16%	71.20%	79.89%
JHU_{MD}	**EBM**	0.61±0.26	63.86%	47.47%	76.16%	0.74±0.19	68.29%	62.08%	72.94%
	FBM	0.68±0.25	69.22%	53.53%	80.99%	0.75±0.19	71.86%	66.28%	76.05%
$Lpba40_{FA}$	**EBM**	0.72±0.24	71.34%	65.37%	75.81%	0.78±0.18	72.96%	63.43%	80.10%
	FBM	0.67±0.26	69.26%	53.77%	80.88%	0.77±0.02	73.58%	68.72%	77.23%
$Lpba40_{MD}$	**EBM**	0.86±0.15	75.29%	6.32%	81.04%	0.82±0.16	76.18%	65.78%	83.98%
	FBM	0.75±0.20	73.76%	67.78%	78.25%	0.83±0.16	74.91%	69.98%	78.60%

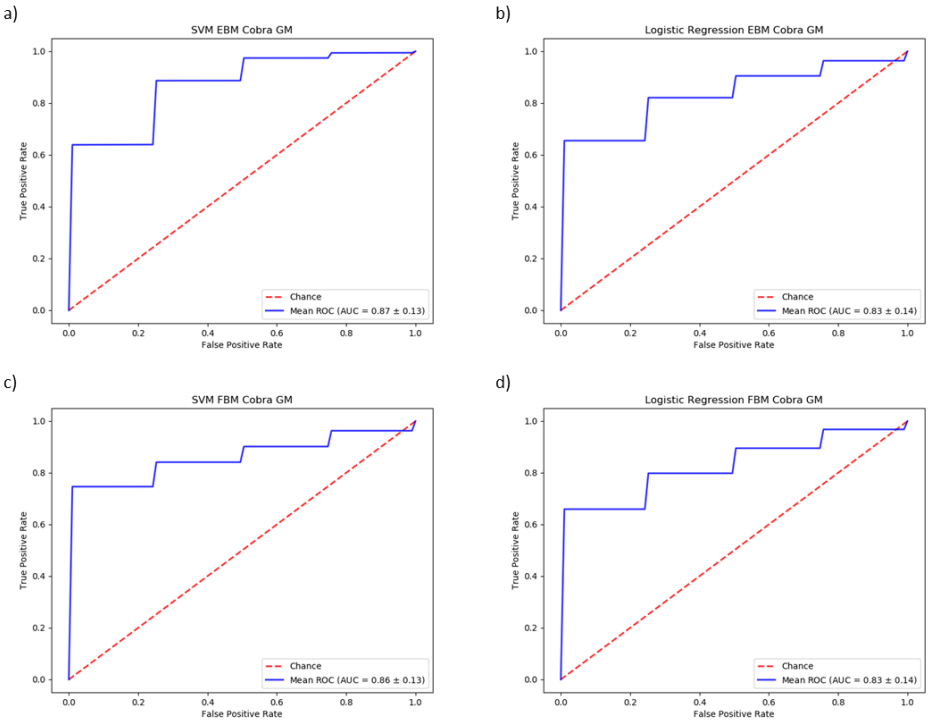

Figure S.55 Mean ROC curves of the Cobra GM model. a) SVM EBM. b) LR EBM. c) SVM FBM. d) LR FBM.

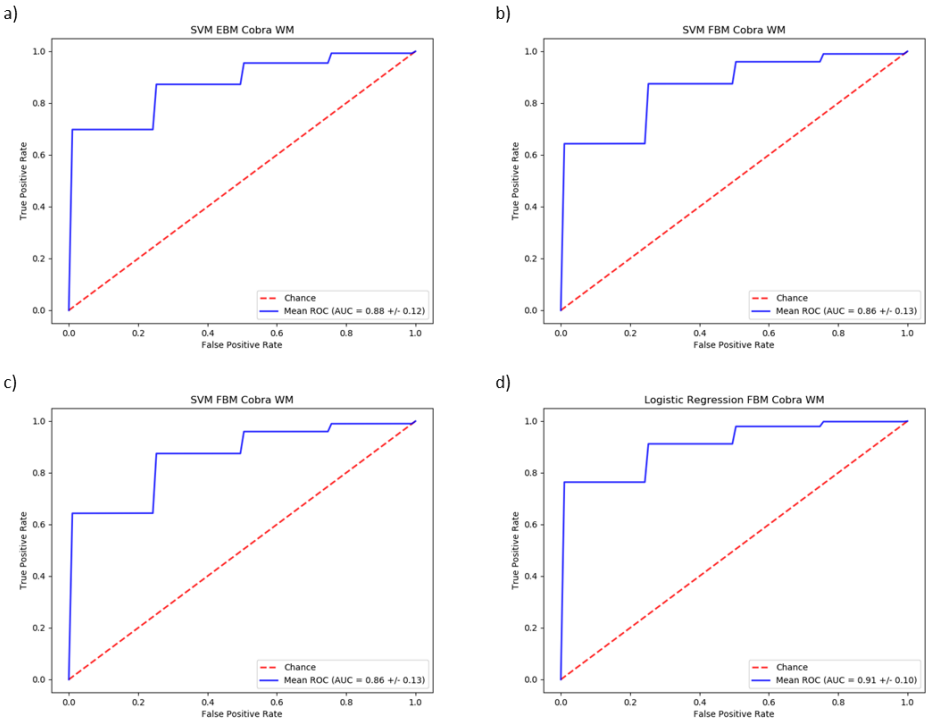

Figure S.56 Mean ROC curves of the Cobra WM model. a) SVM EBM. b) LR EBM. c) SVM FBM. d) LR FBM.

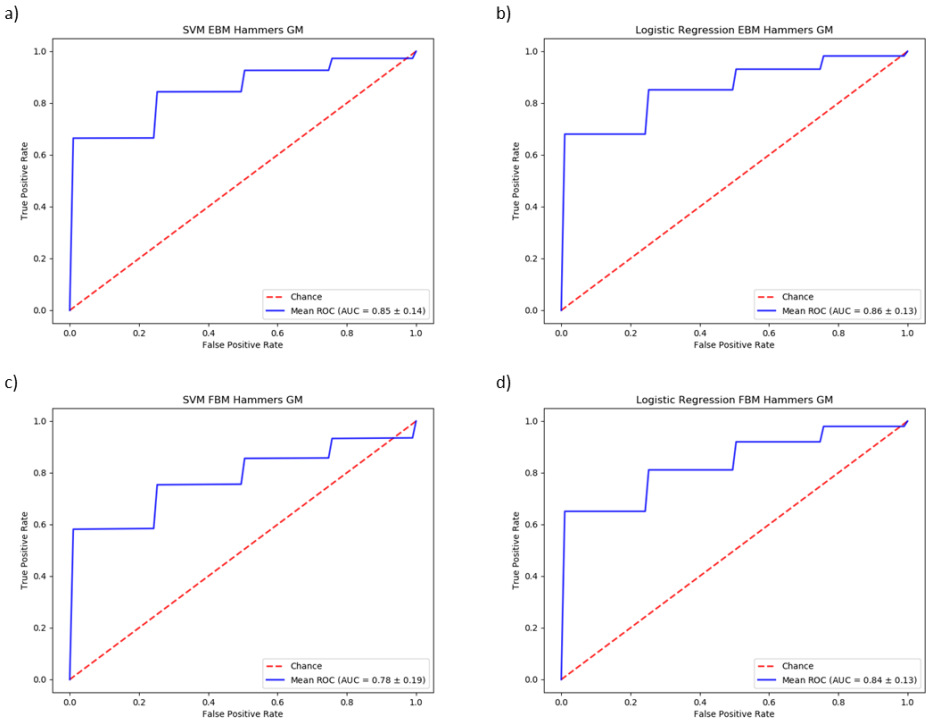

Figure S.57 Mean ROC curves of the Hammers GM model. a) SVM EBM. b) LR EBM. c) SVM FBM. d) LR FBM.

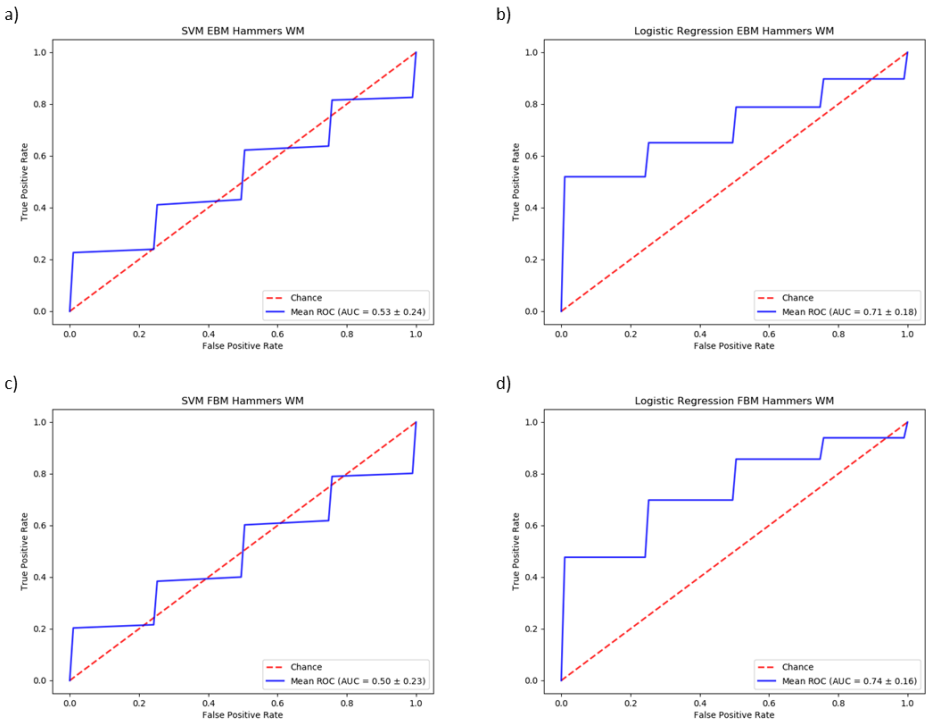

Figure S.58 Mean ROC curves of the Hammers WM model. a) SVM EBM. b) LR EBM. c) SVM FBM. d) LR FBM.

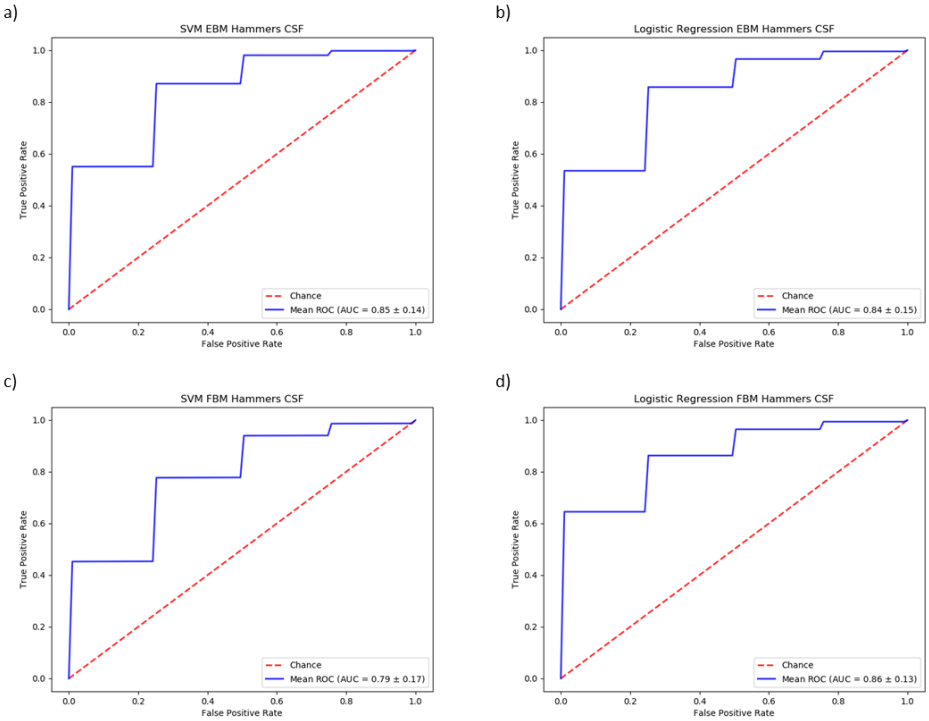

Figure S.59 Mean ROC curves of the Hammers CSF model. a) SVM EBM. b) LR EBM. c) SVM FBM. d) LR FBM.

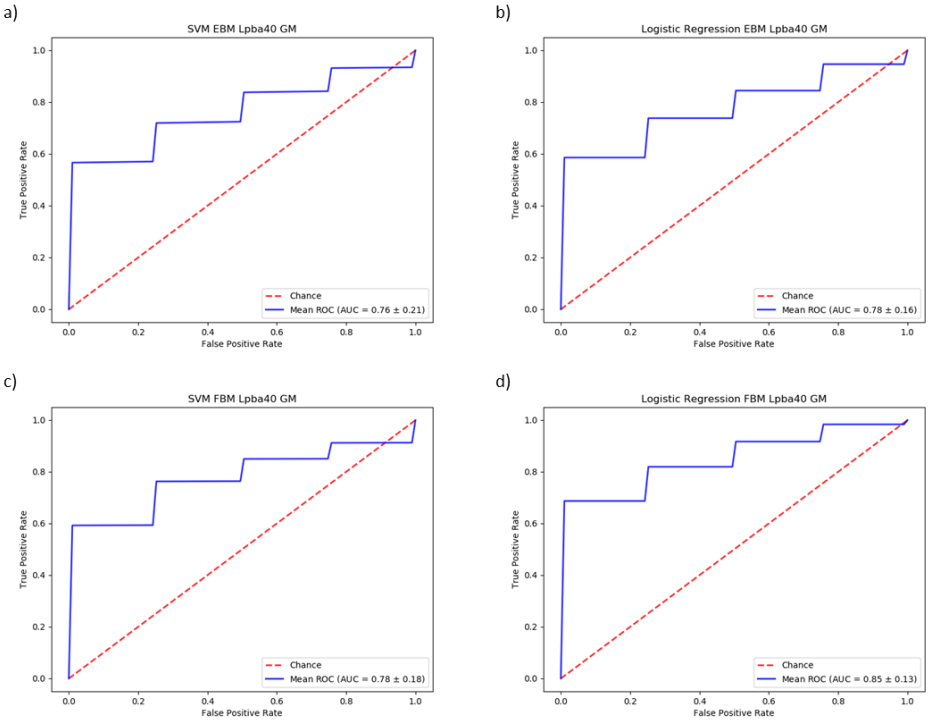

Figure S.60 Mean ROC curves of the Lpba40 GM model. a) SVM EBM. b) LR EBM. c) SVM FBM. d) LR FBM.

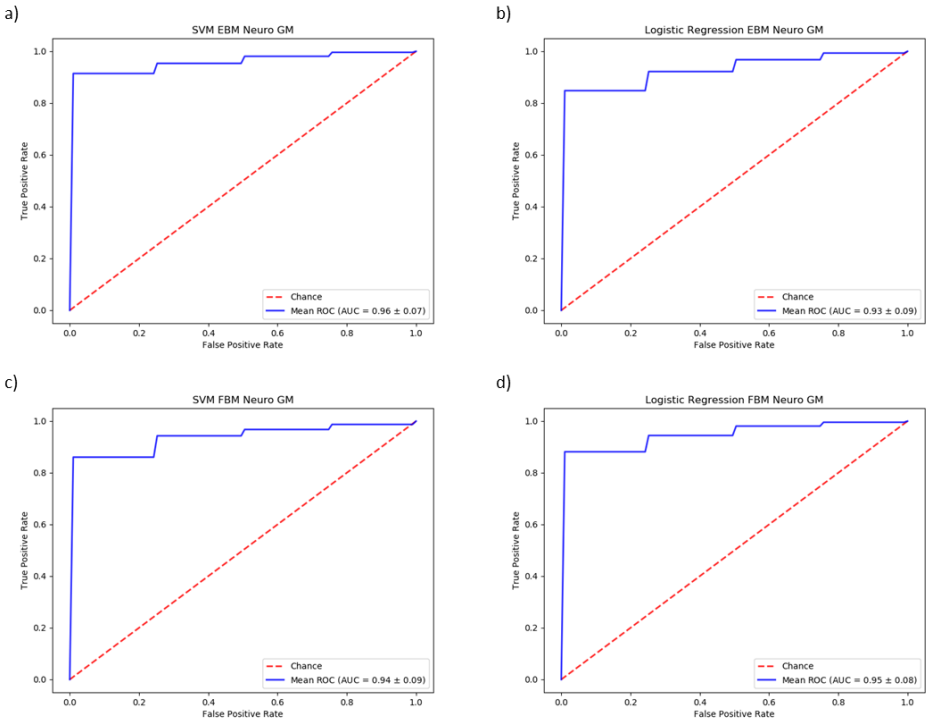

Figure S.61 Mean ROC curves of the Neuro GM model. a) SVM EBM. b) LR EBM. c) SVM FBM. d) LR FBM.

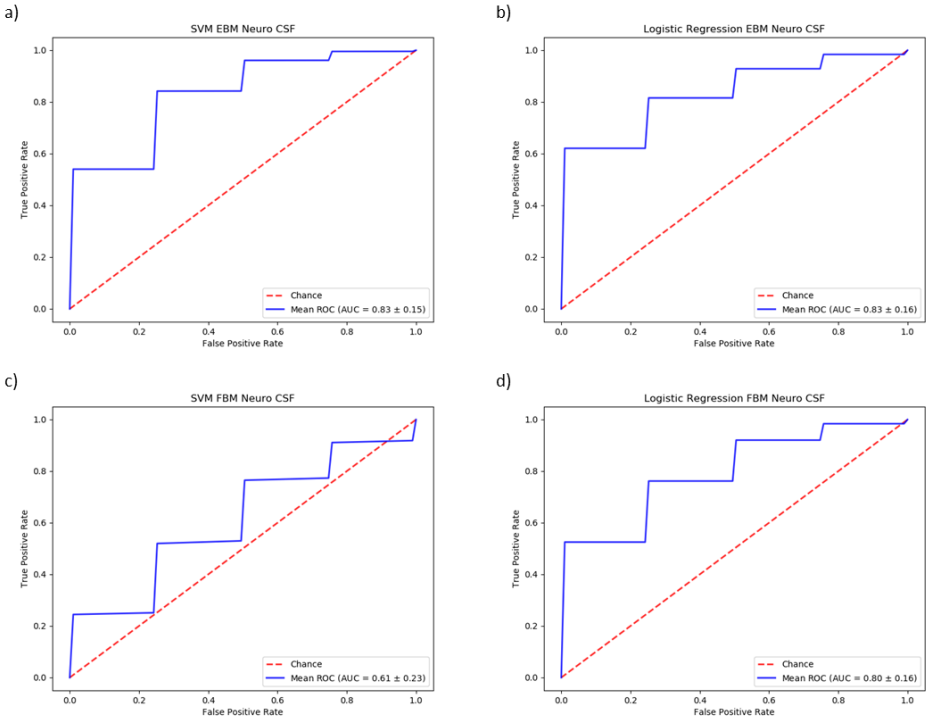

Figure S.62 Mean ROC curves of the Neuro CSF model. a) SVM EBM. b) LR EBM. c) SVM FBM. d) LR FBM.

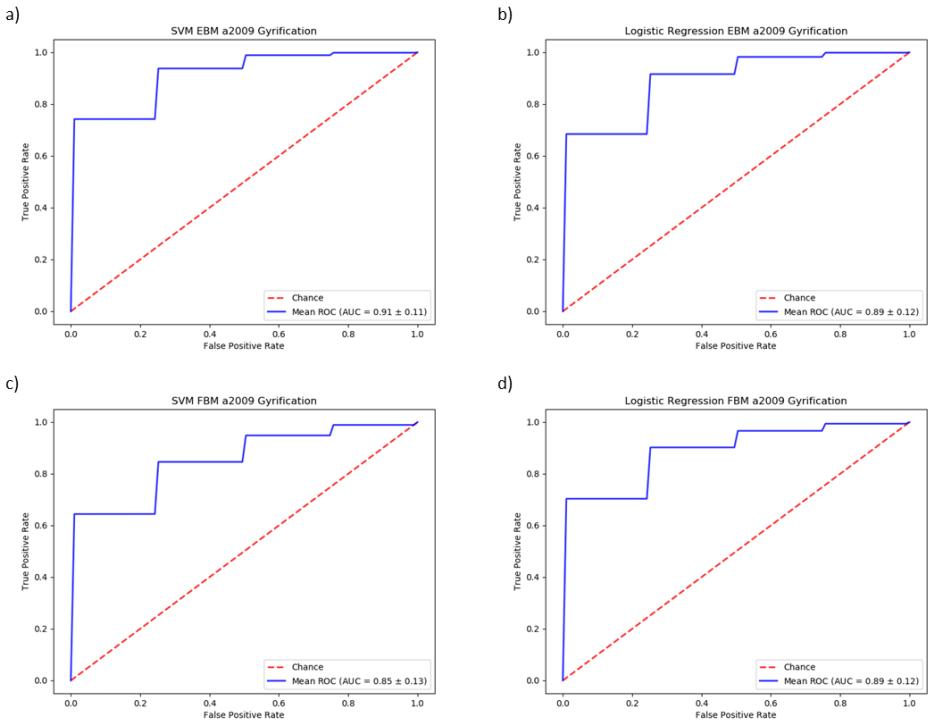

Figure S.63 Mean ROC curves of the a2009 Gyrification model. a) SVM EBM. b) LR EBM. c) SVM FBM. d) LR FBM.

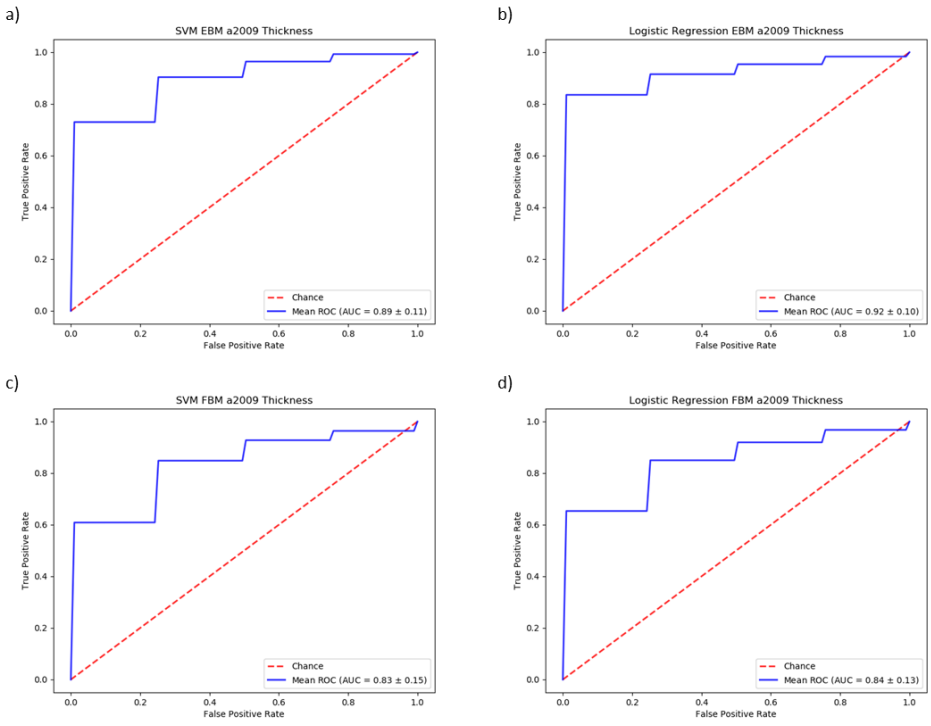

Figure S.64 Mean ROC curves of the a2009 Thickness model. a) SVM EBM. b) LR EBM. c) SVM FBM. d) LR FBM.

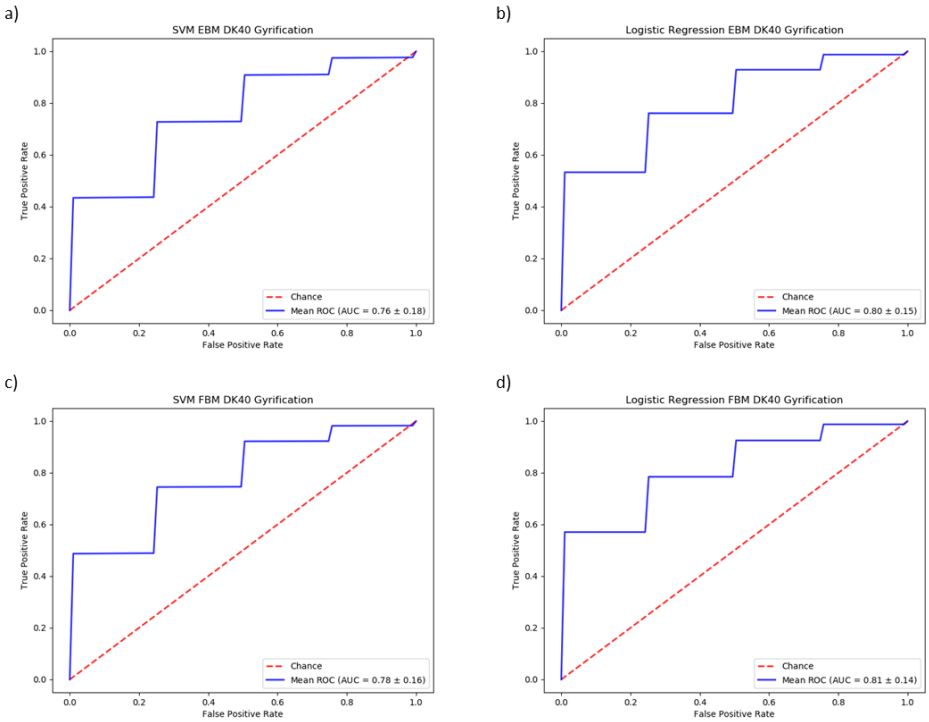

Figure S.65 Mean ROC curves of the DK40 Gyrification model. a) SVM EBM. b) LR EBM. c) SVM FBM. d) LR FBM.

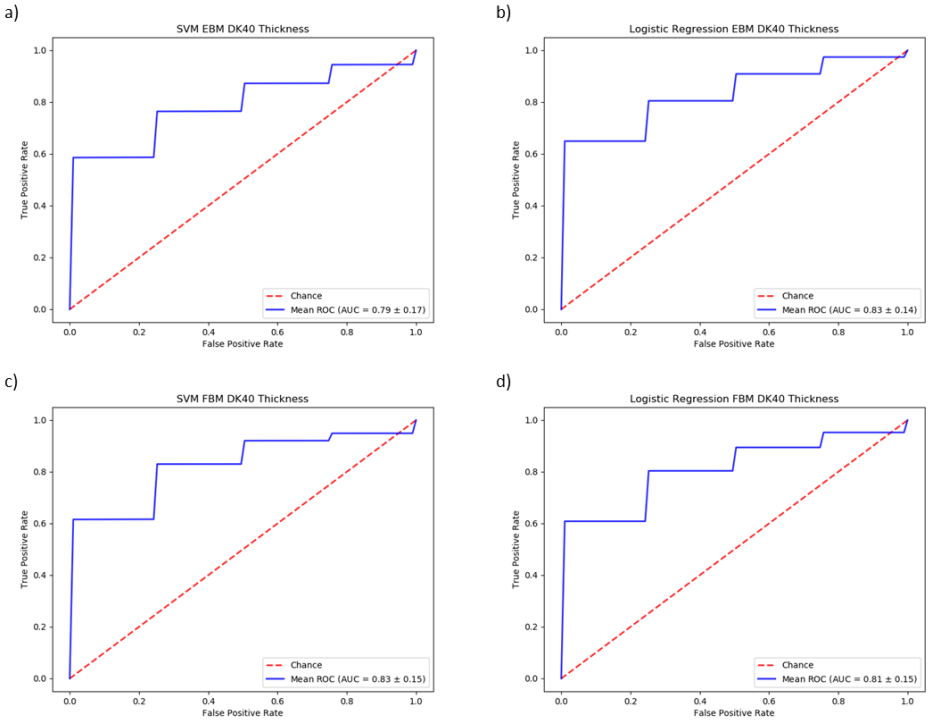

Figure S.66 Mean ROC curves of the DK40 Thickness model. a) SVM EBM. b) LR EBM. c) SVM FBM. d) LR FBM.

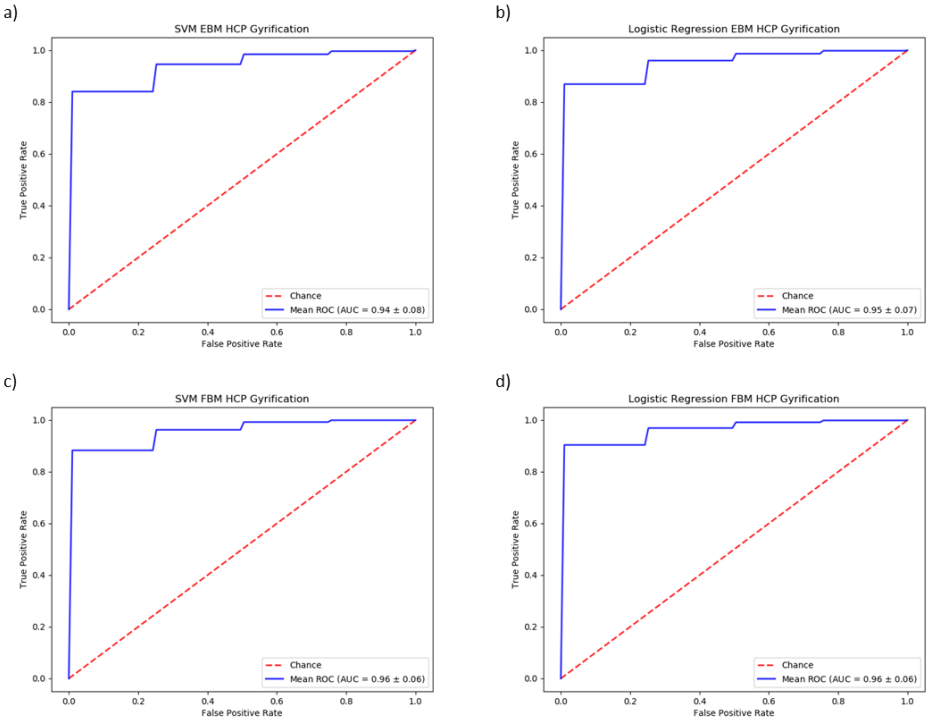

Figure S.67 Mean ROC curves of the HCP Gyrification model. a) SVM EBM. b) LR EBM. c) SVM FBM. d) LR FBM.

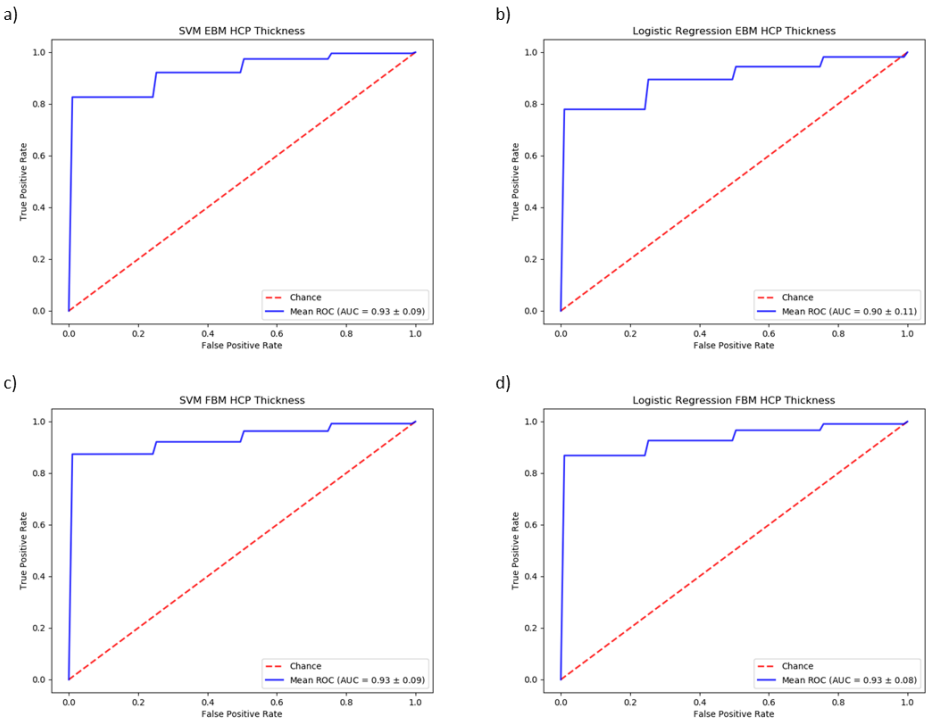

Figure S.68 Mean ROC curves of the HCP Thickness model. a) SVM EBM. b) LR EBM. c) SVM FBM. d) LR FBM.

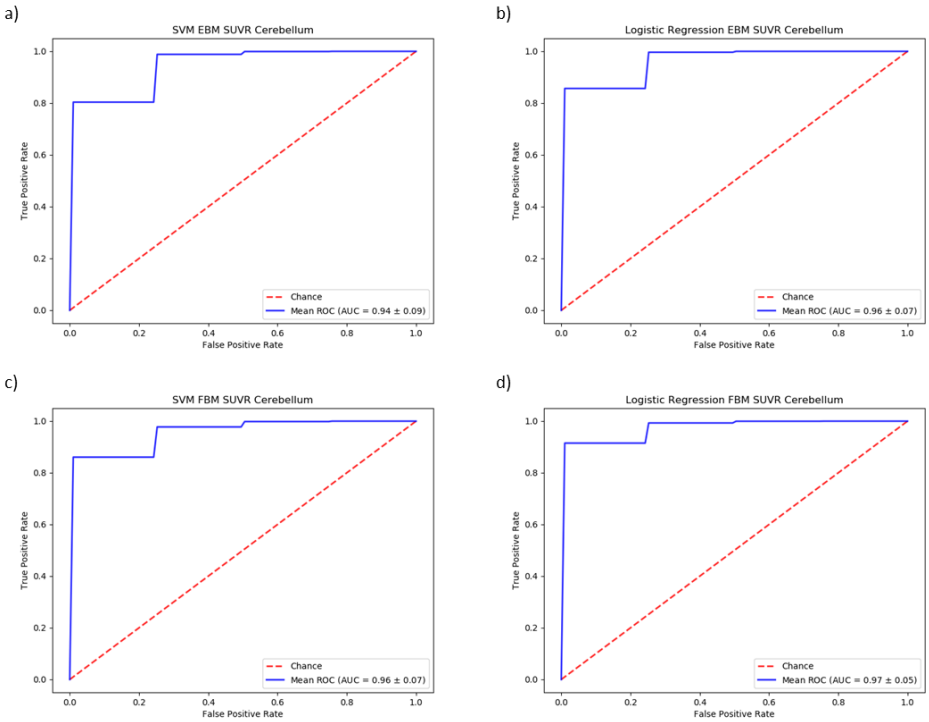

Figure S.69 Mean ROC curves of the SUVR Cerebellum model. a) SVM EBM. b) LR EBM. c) SVM FBM. d) LR FBM.

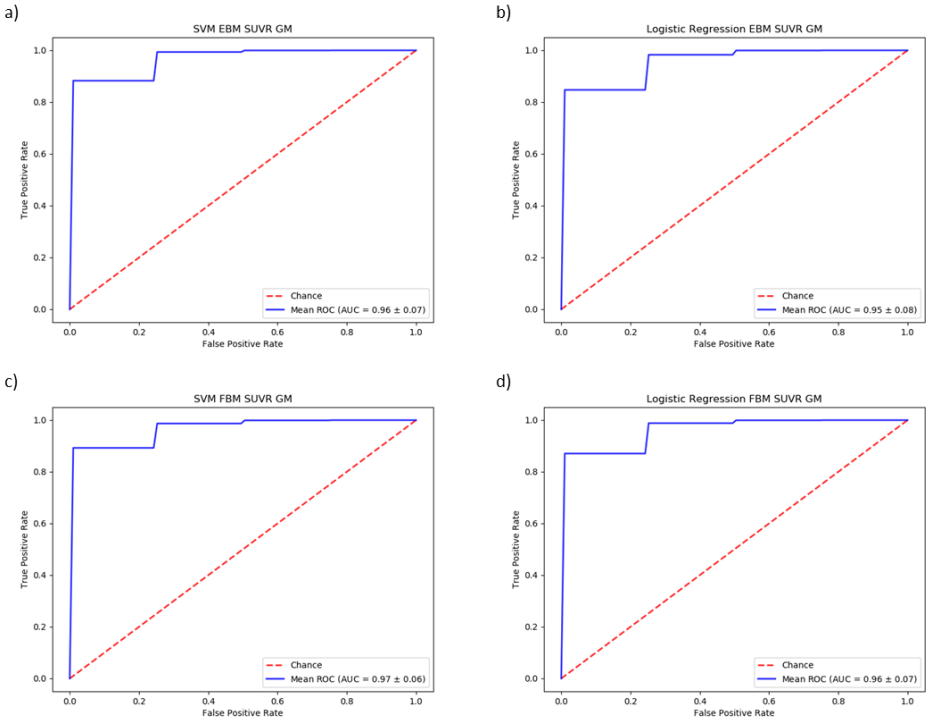

Figure S.70 Mean ROC curves of the SUVR GM model. a) SVM EBM. b) LR EBM. c) SVM FBM. d) LR FBM.

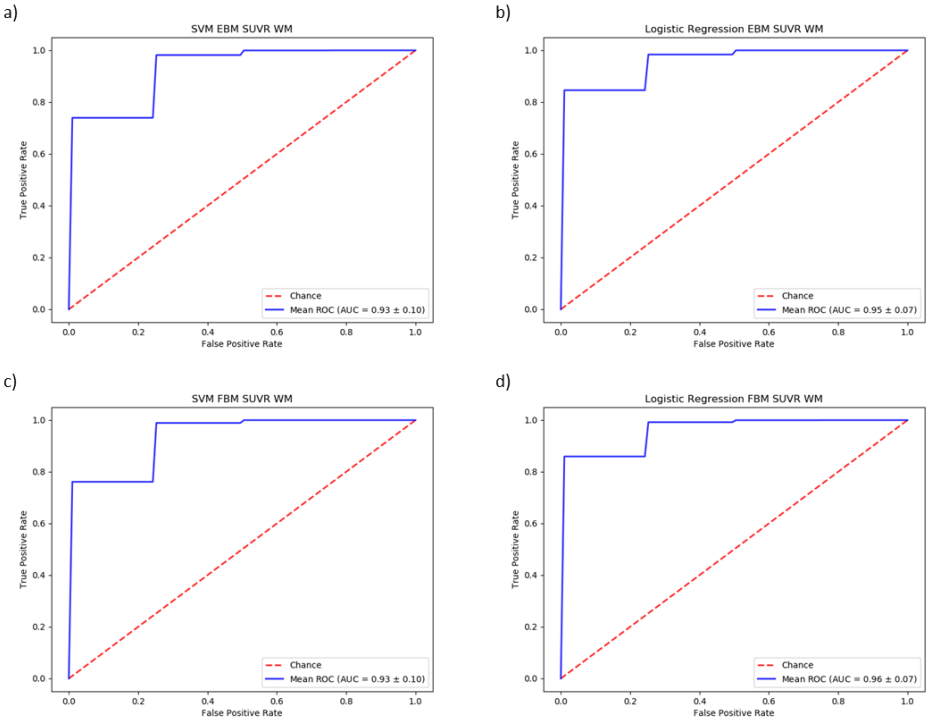

Figure S.71 Mean ROC curves of the SUVR WM model. a) SVM EBM. b) LR EBM. c) SVM FBM. d) LR FBM.

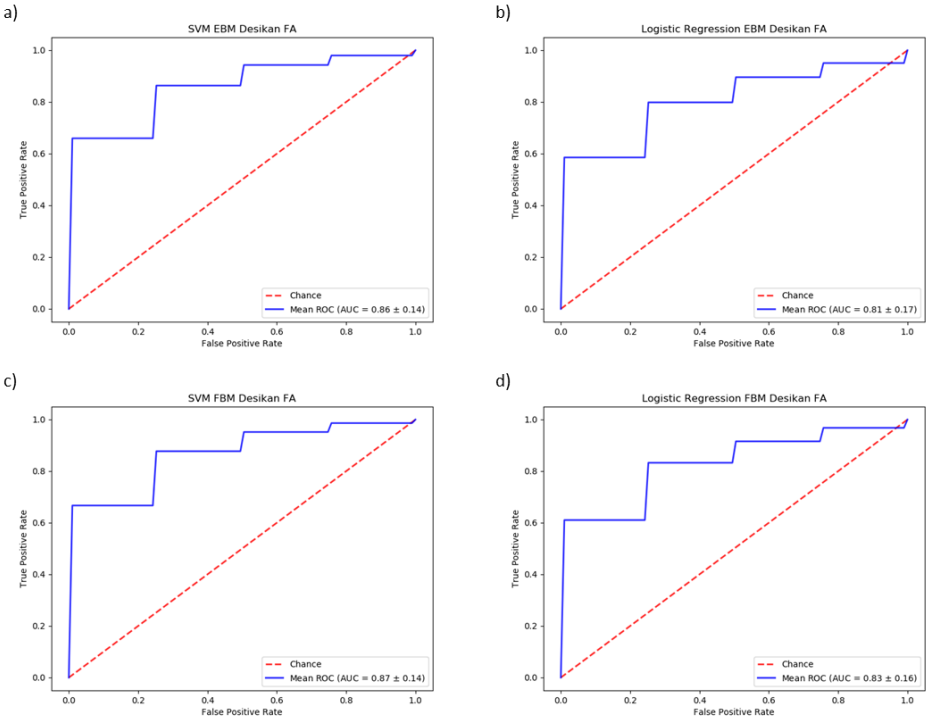

Figure S.72 Mean ROC curves of the Desikan FA model. a) SVM EBM. b) LR EBM. c) SVM FBM. d) LR FBM.

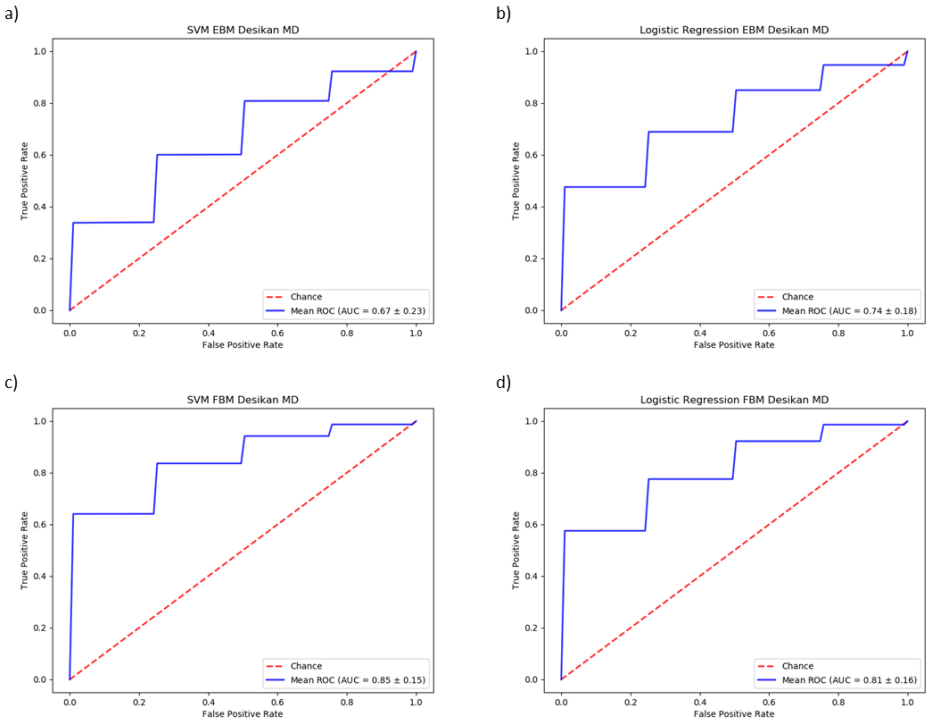

Figure S.73 Mean ROC curves of the Desikan MD model. a) SVM EBM. b) LR EBM. c) SVM FBM. d) LR FBM.

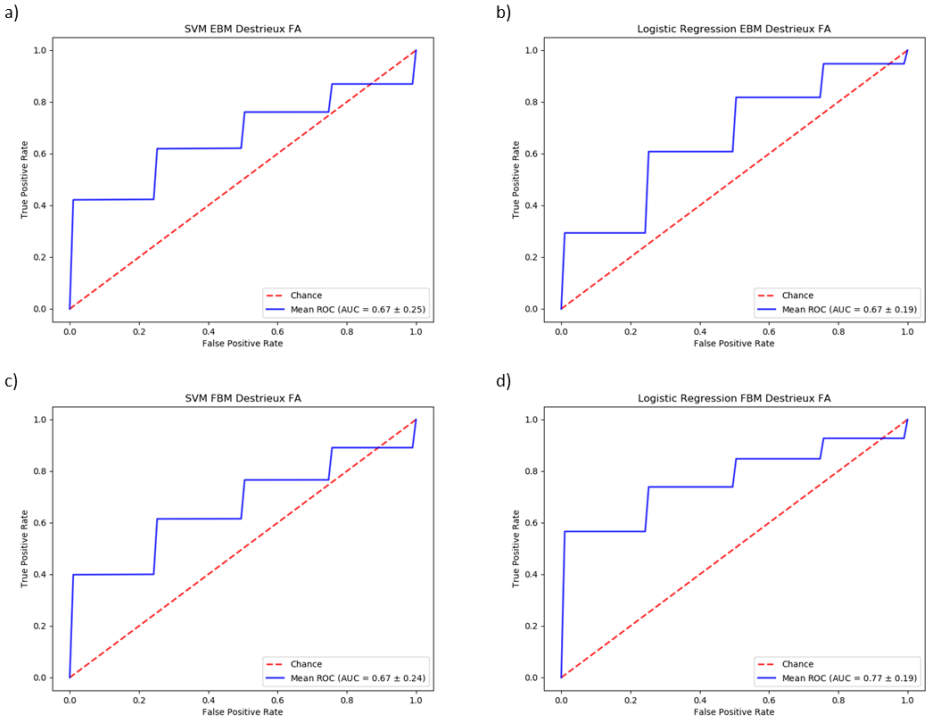

Figure S.74 Mean ROC curves of the Destrieux FA model. a) SVM EBM. b) LR EBM. c) SVM FBM. d) LR FBM.

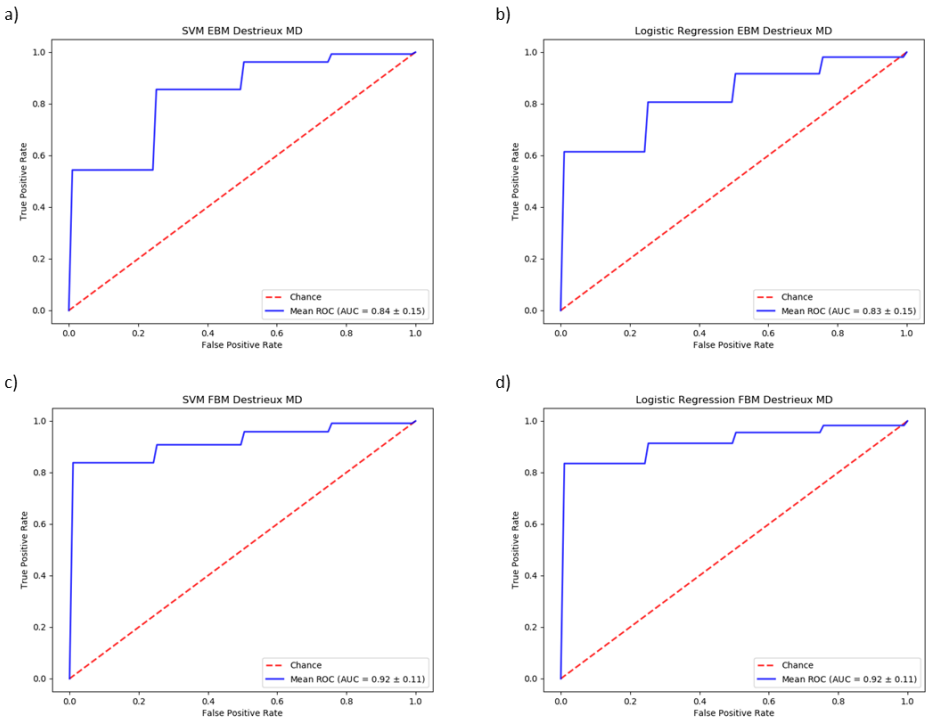

Figure S.75 Mean ROC curves of the Destrieux MD model. a) SVM EBM. b) LR EBM. c) SVM FBM. d) LR FBM.

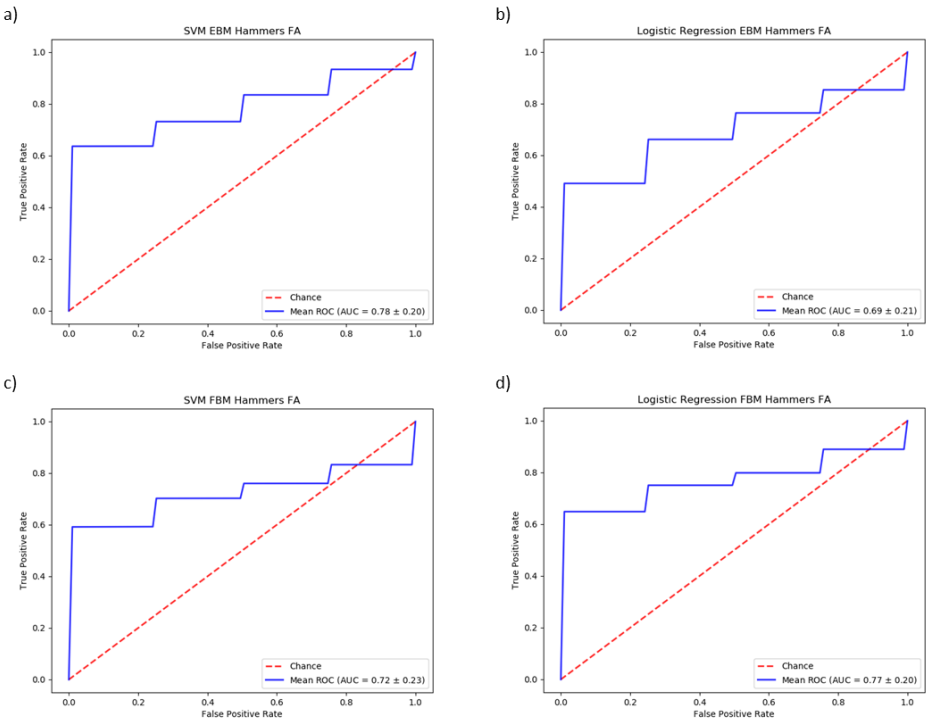

Figure S.76 Mean ROC curves of the Hammers FA model. a) SVM EBM. b) LR EBM. c) SVM FBM. d) LR FBM.

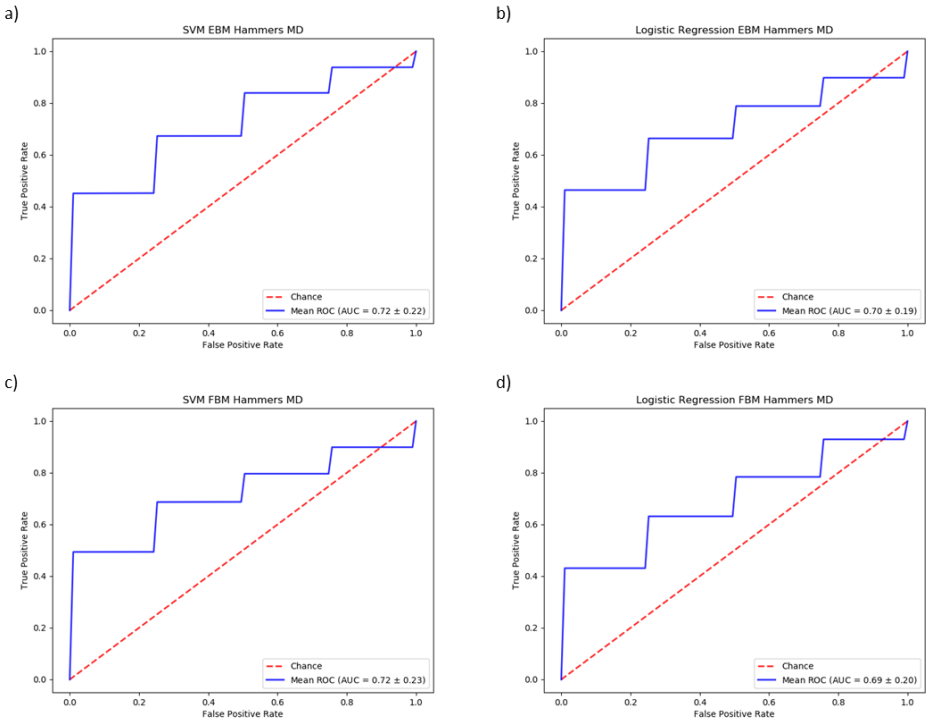

Figure S.77 Mean ROC curves of the Hammers MD model. a) SVM EBM. b) LR EBM. c) SVM FBM. d) LR FBM.

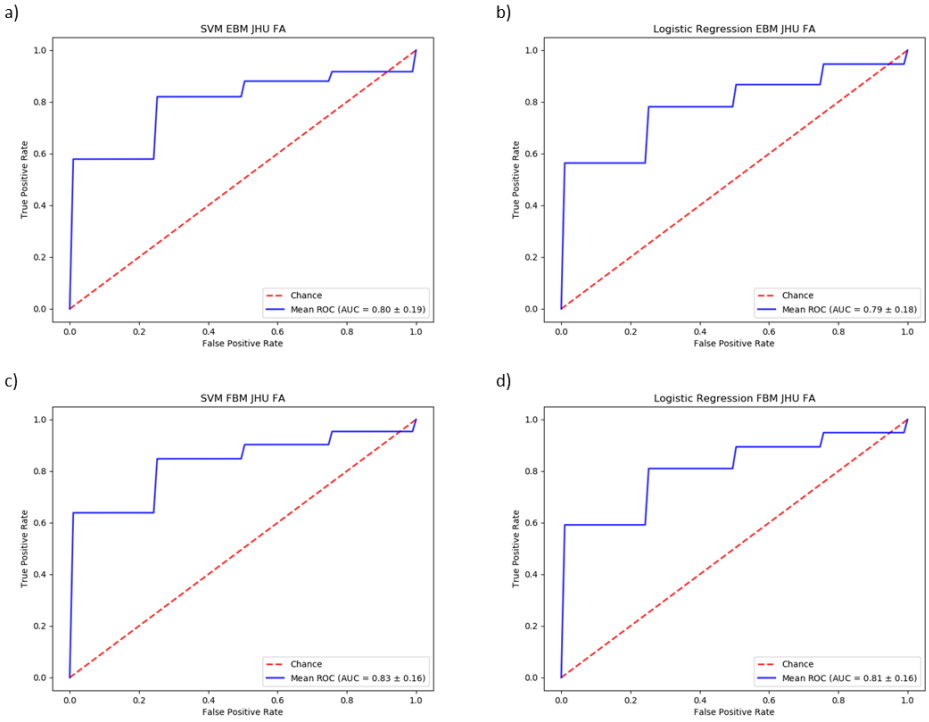

Figure S.78 Mean ROC curves of the JHU FA model. a) SVM EBM. b) LR EBM. c) SVM FBM. d) LR FBM.

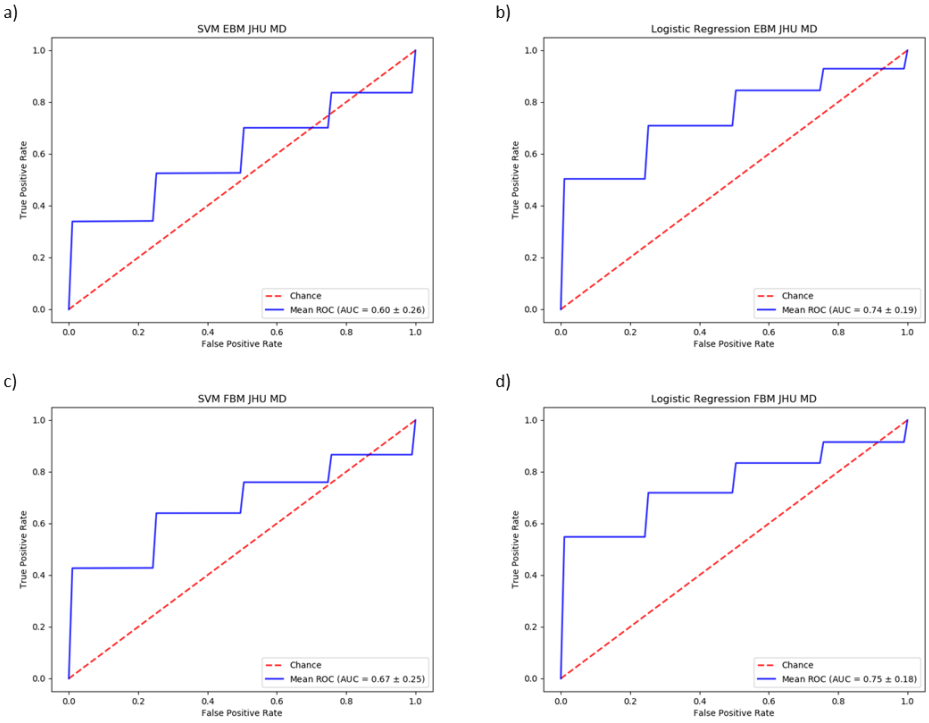

Figure S.79 Mean ROC curves of the JHU MD model. a) SVM EBM. b) LR EBM. c) SVM FBM. d) LR FBM.

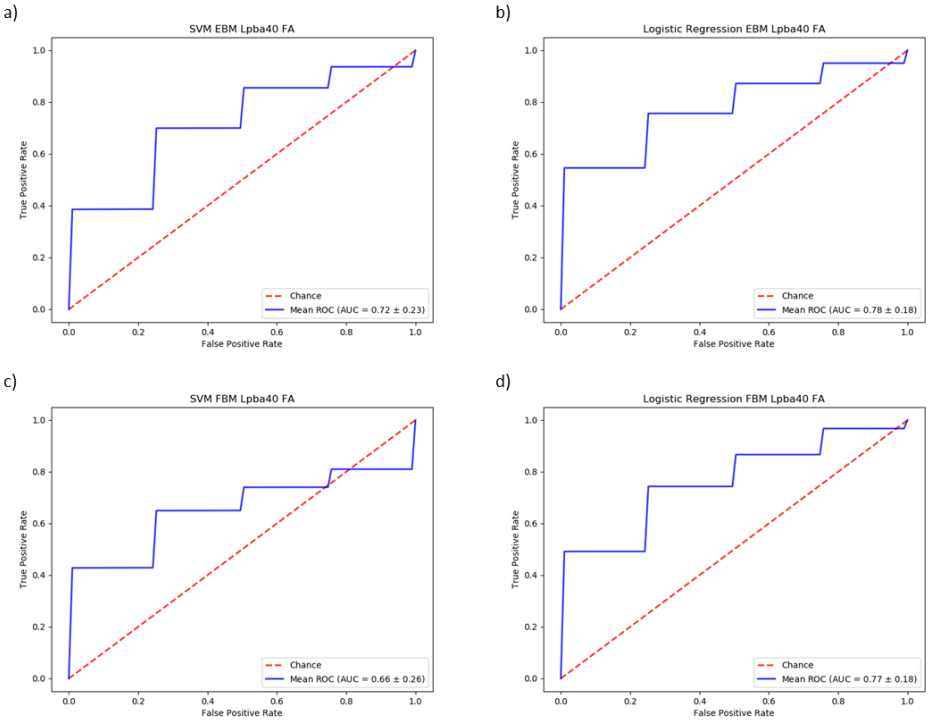

Figure S.80 Mean ROC curves of the Lpba40 FA model. a) SVM EBM. b) LR EBM. c) SVM FBM. d) LR FBM.

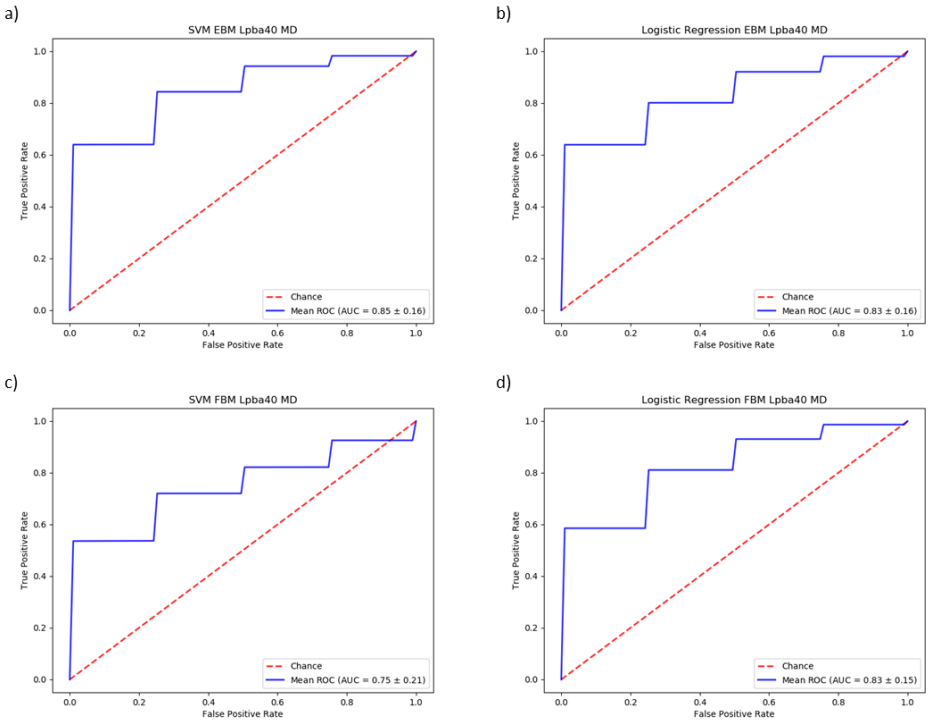

Figure S.81 Mean ROC curves of the Lpba40 MD model. a) SVM EBM. b) LR EBM. c) SVM FBM. d) LR FBM.