UNIVERSIDAD
DE GRANADA

TESIS DOCTORAL

ENHANCING DIAGNOSTIC ACCURACY IN NEUROIMAGING THROUGH MACHINE LEARNING: ADVANCEMENTS IN STATISTICAL CLASSIFICATION AND MAPPING

Autora:

Carmen Jiménez Mesa

Directores:

Juan Manuel Górriz Sáez
Javier Ramírez Pérez de Inestrosa
John Suckling

Programa de Doctorado en Tecnologías de la Información y la Comunicación

Octubre 2023

Editor: Universidad de Granada. Tesis Doctorales
Autor: Carmen Jiménez Mesa
ISBN: 978-84-1195-093-0
URI: https://hdl.handle.net/10481/85701

Enhancing Diagnostic Accuracy in Neuroimaging through Machine Learning: Advancements in Statistical Classification and Mapping

Carmen Jiménez Mesa

carmenj@ugr.es

THESIS SUPERVISORS

Dr. Dr. Juan Manuel Górriz Sáez

Signal Processing and Biomedical Applications (Universidad de Granada)

Dr. Javier Ramírez Pérez de Inestrosa

Signal Processing and Biomedical Applications (Universidad de Granada)

Dr. John Suckling

Department of Psychiatry (University of Cambridge)

Tesis para la obtención del título de doctora por la Universidad de Granada
Programa de Doctorado en Tecnologías de la Información y la Comunicación

Financiación

Esta tesis doctoral ha sido financiada por el Programa Predoctoral Formación de Profesorado Universitario (ref. FPU18/04902) concedido a Carmen Jiménez Mesa por el Ministerio de Universidades (España). Otra fuente de financiación ha sido el Programa de Proyectos de Investigación Precompetitivos para Jóvenes Investigadores del Plan Propio de la UGR (ref. PPJIB2021-14), donde Carmen Jiménez Mesa y José María González Peñalver han sido IPs del proyecto.

Siempre debes tener fe en las personas. Y, lo más importante,
siempre debes tener fe en ti misma.

— Elle Woods (Una rubia muy legal, 2001)

A Pilar Taboada. Te quiero hasta el cielo.

AGRADECIMIENTOS

En primer lugar, quiero darle las gracias a mis directores de tesis por brindarme la oportunidad de investigar en este campo tan interesante. Juanma, gracias por un día en clase plantearme la opción de trabajar con vosotros; Javier, gracias por haber estado pendiente de mi evolución estos años; y John, gracias por acogerme en Cambridge, esa ciudad tan bonita donde se respira tanta ciencia. Estoy especialmente agradecida de formar parte del grupo SiPBA, el cual está formado por investigadores magníficos y excelentes personas como Juan Eloy, Paco, Diego C., Diego S., David, Fermín e Ignacio.

Esta tesis me ha permitido conocer a personas maravillosas por el camino: mis compañeros de despacho(s), las personas que conocí en Cambridge, mis alumnos, los compañeros del 3MT, o a una gran parte del grupo de Psicología Experimental de la UGR. Chema, muchas gracias por pasar de ser un desconocido a ser un amigo y un apoyo fundamental estos años.

Me gustaría también agradecer a todos mis amigos que se han preocupado estos años por mí, interesados por lo que hago, y que incluso se han activado alertas para saber cuándo publico artículos. Qué suerte tengo de teneros. A mis profesores de distintas etapas que me encuentro y me animan a continuar en este camino, gracias. En especial, quiero mencionar a Juan Antonio Jiménez Tejada y Pilar López Varo, por ser los primeros que vieron en mí la capacidad de ser investigadora. Pilar, gracias por ser una mentora tan excepcional.

Alex, gracias por ser el mejor compañero de aventuras que he podido tener estos años, y que espero tener muchos más. Gracias por los abrazos y las risas que alejan los problemas. Gracias porque, cuando me dices que soy una campeona, me lo creo. Tú también eres un campeón para mí.

Por último, esta tesis no hubiera sido posible sin mi familia: mis padres, mi hermana y mi abuelo. Ellos son los que me soportan y alientan día sí, y día también. No han sido unos años fáciles en casa, pero en ningún momento me habéis dejado tirar la toalla. Mi abuela Pili no ha podido verme acabar esta etapa, pero sí como he ido avanzando en ella. Gracias por estar siempre tan orgullosa de mi, y quererme tanto.

Gracias a todos.

ABSTRACT

In recent years, the application of artificial intelligence (AI) techniques in health and medicine, including neuroimaging, has grown exponentially. Neuroimaging plays a crucial role in studying the central nervous system and supporting clinical diagnosis through non-invasive examination of the human brain. Computer-aided diagnosis (CAD) systems have been developed to assist clinicians in this process by using pattern recognition and prediction capabilities. These systems, created through multidisciplinary collaboration, incorporate AI algorithms to improve diagnostic accuracy and reduce clinician workload. However, these systems face certain challenges, such as the lack of interpretability of AI models and the small sample sizes inherent in neuroimaging studies, which complicate system learning and performance. Addressing these challenges is crucial for establishing CAD systems as a standard for clinical diagnostic support.

This thesis focuses on exploring various machine learning (ML) approaches to enhance the accuracy and utility of CAD systems while ensuring optimal interpretability for clinical analysis.

To enhance reliability, one approach involves tackling the *curse of dimensionality* by reducing the number of features. The significance of feature selection and extraction stages is emphasised in the development of an optimised multiclass classification system. Additionally, validation methods implemented so far in neuroimaging studies are questioned. The viability of an upper-bounded resubstitution as a validation method is demonstrated through a non-parametric statistical inference framework, particularly suitable in studies with small sample sizes. Moreover, the use of classical statistics in neuroimaging, which usually rely on assumptions that are frequently violated, is also questioned. To address this, a proposed data-driven approach for generating statistical inference maps is tested in brain disorders such as Alzheimer's Disease and Parkinson's Disease, and compared with a parametric approach.

Regarding interpretability, two systems are designed to provide easily interpretable results for clinical experts. These systems place emphasis on the use of explainable AI techniques. One system focuses on analysing sulcal morphology in individuals with schizophrenia, while the other system proposes a method for detecting patterns in the Clock-Drawing Test to assess cognitive impairment.

In summary, this thesis demonstrates the utility of ML techniques in neuroimaging for brain mapping, feature detection, and classification. It explores the reliability and interpretability of CAD systems, identifies potential improvements, and emphasises

the need for further research to develop techniques tailored to neuroimaging's unique conditions. This advancements will enable CAD systems to become a standard for clinical diagnostic support, ultimately improving the quality of life for patients through earlier and more accurate diagnoses.

RESUMEN AMPLIO EN CASTELLANO

Motivación

En los últimos años, ha habido un crecimiento exponencial en la aplicación de técnicas de inteligencia artificial (IA) en el campo de la salud y la medicina, y la neuroimagen no ha sido una excepción. La neuroimagen es un campo especializado en el estudio estructural, funcional y de conectividad del cerebro. Se ha convertido en una herramienta invaluable para el diagnóstico clínico debido a su naturaleza no invasiva, lo que proporciona una forma conveniente de investigar el cerebro humano. En este contexto, se están desarrollando sistemas de ayuda al diagnóstico asistido por computadora (CAD, por sus siglas en inglés) que brindan asistencia a los médicos en el proceso de diagnóstico a través de un conjunto de herramientas computarizadas con capacidades de reconocimiento de patrones y predicción. Estos sistemas se diseñan a partir de la colaboración multidisciplinaria en áreas como matemáticas, ciencia de datos, IA o la estadística, entre otras. La inclusión de algoritmos de IA en los sistemas CAD es especialmente importante, ya que ha llevado a mejoras significativas en el proceso de diagnóstico, con sistemas más precisos y una reducción de la carga de trabajo para los médicos. Además de su aplicación en la medicina clínica, los sistemas CAD también se utilizan en áreas como la psicología, las ciencias del comportamiento, la neurociencia cognitiva y la psiquiatría para comprender mejor el funcionamiento cerebral, así como el origen, etapas y consecuencias de los trastornos cerebrales. Esto demuestra la versatilidad y el impacto potencialmente amplio de los sistemas CAD en diferentes disciplinas relacionadas con el estudio del cerebro.

La estructura de los sistemas CAD generalmente consta de los mismos pasos. En primer lugar, los datos que se introducen en el sistema, generalmente imágenes cerebrales, se procesan para estandarizarlos y garantizar un rendimiento óptimo del sistema. La siguiente etapa es dotar de inteligencia al sistema para el aprendizaje deseado, ya sea la clasificación de diferentes condiciones o la detección de regiones de interés (ROIs, por sus siglas en inglés). En ambos casos, diversos algoritmos son aplicados para operar con las características de entrada y para el proceso de aprendizaje en sí del sistema. Finalmente, se evalúa el rendimiento del sistema en un último paso.

Las imágenes cerebrales proporcionan una amplia gama de información que debe ser procesada para su análisis. En este contexto, la IA y específicamente el aprendizaje automático o *machine learning* (ML) desempeñan un papel fundamental. Aprovechando la capacidad computacional disponible en la actualidad, se están proponiendo enfoques

basados en ML para el análisis de neuroimagen. En este campo de estudio se han logrado avances significativos que van desde esquemas lineales y de baja dimensionalidad, hasta arquitecturas de redes neuronales profundas para la selección, extracción y clasificación de características. Estas técnicas contribuyen de manera sustancial a mejorar la precisión y la capacidad de reconocimiento al analizar patrones complejos presentes en las imágenes cerebrales. Por tanto, estos avances tienen un gran potencial en el campo del análisis de neuroimagen y ofrecen nuevas perspectivas para la comprensión y el diagnóstico de trastornos cerebrales. Sin embargo, estos sistemas se enfrentan a ciertos desafíos significativos. Uno de ellos es la falta de interpretabilidad de los modelos de IA, que se vuelve más indescifrable a medida que aumenta la complejidad de los sistemas, especialmente aquellos basados en aprendizaje profundo o *deep learning* (DL). La opacidad de estos modelos dificulta la comprensión de los procesos internos y las razones detrás de sus decisiones, lo que limita su confiabilidad y aceptación en el ámbito clínico. Otro desafío importante es el pequeño tamaño de las muestras inherente a los estudios de neuroimagen. Estos estudios se basan en conjuntos de datos obtenidos de un número limitado de sujetos, lo cual dificulta el aprendizaje adecuado del sistema. El tamaño reducido de la muestra puede resultar en modelos con baja generalización y susceptibles a errores, lo que afecta negativamente su rendimiento y precisión. Abordar estos desafíos es crucial para poder establecer los sistemas CAD como un estándar fiable y efectivo de apoyo al diagnóstico clínico.

Objetivos

La finalidad de esta tesis es explorar diferentes enfoques de ML para conseguir métodos fiables e interpretables en neuroimagen. Obtener métodos con estas propiedades conduciría a sistemas CAD de mayor precisión y utilidad en la práctica clínica, lo cual es el fin último de cualquier estudio de neuroimagen. Con este propósito, se pueden definir los siguientes objetivos:

- Desarrollar y evaluar diferentes métodos para aumentar la fiabilidad de los sistemas CAD, abordando los problemas asociados al tamaño de la muestra y al número de características disponibles.

- Desarrollar y evaluar algoritmos con una interpretabilidad óptima para el análisis clínico, lo cual mejora la confianza de los profesionales de la salud y facilita la toma de decisiones basadas en evidencia.

Se han realizado diversos estudios para lograr los objetivos propuestos. Para el primer objetivo, centrado en la fiabilidad de los sistemas CAD, se han llevado a cabo investigaciones que cuestionan las técnicas utilizadas actualmente en neuroimagen, como los métodos de validación de resultados y las técnicas de generación de mapas estadísticos cerebrales. Todo ello con el objetivo último de obtener sistemas CAD robustos. A continuación, se enumeran los estudios realizados en este sentido:

1. Un sistema CAD optimizado y centrado en la capacidad predictiva de un clasificador multiclase basado en técnicas de ML. Este estudio realiza un análisis exhaustivo de la relevancia de las fases de extracción y selección de características.

2. Un marco de inferencia estadística no paramétrica para evaluar la significancia estadística de la tasa de acierto en modelos de ML y DL. Esto también permite comparar el rendimiento de los métodos de validación tradicionales con respecto a un enfoque novedoso basado en la Teoría del Aprendizaje Estadístico (SLT, por sus siglas en inglés).

3. Una metodología para la generación de mapas estadísticos en imágenes cerebrales basada en técnicas de ML (aprendizaje agnóstico), el cual es adaptable a diferentes técnicas de imagen médica. Este método es comparado al modelo tradicional basado en el modelo linear general (aprendizaje paramétrico), implementado en la herramienta SPM (*Statistical Parametric Mapping*).

Enfocándose en el segundo objetivo, centrado en el diseño de sistemas CAD con una interpretabilidad óptima, se han realizado los siguientes estudios que integran técnicas de IA explicables, conocidas comúnmente como XAI (*Explainable AI*), junto con los algoritmos propios de un sistema CAD:

4. Un sistema CAD para la detección de patrones de interés en la morfología de los surcos cerebrales, en el cual se compara el rendimiento obtenido usando técnicas paramétricas y no paramétricas de significancia estadística. Además, los resultados de clasificación se analizan utilizando técnicas XAI para mejorar la interpretabilidad.

5. Un sistema CAD basado en técnicas de DL y XAI para la detección de patrones en imágenes, con el propósito de ser una herramienta útil en el análisis clínico.

Contribuciones

En la Parte II de esta tesis, se presentan los capítulos que contienen las descripciones y análisis de los estudios mencionados anteriormente. A continuación, se ofrece un breve resumen de cada uno de ellos:

Aplicación de Machine Learning en Clasificación Multiclase

En este estudio, descrito en el capítulo 6, se propone un sistema CAD para detección multiclase de Alzheimer. En este caso, se consideraron cuatro condiciones: sujetos controles, sujetos con deterioro cognitivo leve, pacientes de Alzheimer y personas que durante el seguimiento del estudio habían pasado de estar diagnósticos con deterioro cognitivo a Alzheimer. La base de datos empleada procede de un concurso internacional

sobre predicción automatizada del deterioro cognitivo leve, y contiene datos extraídos de imágenes de resonancias magnéticas (MRI, por sus siglas en inglés) cerebrales.

El método implementado en el sistema se basa en un enfoque uno vs. uno tanto para selección y extracción de características como para clasificación. Los datos también son inicialmente preprocesados para detección de valores atípicos, los cuales son corregidos previamente a las etapas principales del sistema. Son varios los parámetros y clasificadores evaluados, con especial atención a la optimización de las etapas de selección y extracción de características. El proceso finalmente seleccionado se compone de varias etapas. Para la selección de las características más relevantes, se utiliza un método de ordenación y filtrado basado en el t-test. Las características seleccionadas se transforman mediante PLS (*Partial Least Squares*) para extraer características de menor dimensionalidad, las cuales se introducen en el clasificador. Este clasificador consiste en una Máquina de Vectores de Soporte (SVM, por sus siglas en inglés). Los clasificadores binarios obtenidos se combinan utilizando la codificación ECOC (*Error Correcting Output Codes*), y la condición asociada a cada muestra es aquella con mayor peso en el código obtenido.

El método propuesto, el cual es posterior a la realización del concurso, alcanza una tasa de acierto del 67 % en la clasificación multiclase, superando todas las propuestas que se presentaron al concurso. Además, el método también es coherente con hallazgos recientes aplicando sistemas CAD para la enfermedad de Alzheimer y, la metodología seguida podría aplicarse a otros problemas de clasificación multiclase.

Un Método de Inferencia Estadística No Paramétrica

En el capítulo 7 se detalla una metodología a seguir para estimar la significancia estadística de los resultados obtenidos con sistemas CAD. Esto es de especial interés en campos como la neuroimagen, donde el tamaño muestral (normalmente limitado) limita la capacidad de generalización del clasificador, al no haber suficientes muestras de las que aprender y validar los resultados.

Esta metodología propuesta consiste en una inferencia de efectos aleatorios basada en una prueba de permutación de etiquetas. La significancia estadística de los resultados se evalúa a partir del poder estadístico y la capacidad de control del error Tipo I. Son varios los escenarios analizados en este estudio, conjuntos balanceados, desbalanceados, binarios y multiclase, todos ellos con el denominador común de que son conjuntos muestrales de tamaño limitado y que están relacionados con el Alzheimer. Los sistemas CAD aplicados en éstos se basan en Autoenconders para la fase de extracción de características y en SVM para la etapa de clasificación. Para el análisis del rendimiento de estos sistemas, se aplican dos métodos diferentes de validación, el más conocido y empleado, la validación cruzada basada en K-fold, y la resubstitución acotada con límite superior, la cual se ha denominado RUB (*resubstitution with upper bound correction*).

Los resultados indican que RUB rinde ligeramente mejor que K-fold como método de validación en escenarios de pequeño tamaño muestral, especialmente en aquellos donde las muestras son más heterogéneas. Por tanto, se podría considerar RUB como un método válido y eficaz en neuroimagen en términos de coste computacional, varianza

y sesgo.

Statistical Agnostic Mapping (SAM)

Se propone en el capítulo 8 una metodología para generación de mapas estadísticos en imágenes cerebrales. Dicha metodología se ha nombrado como *Statistical Agnostic Mapping* o SAM. Como su nombre indica, este método evita las técnicas paramétricas en su proceso, a diferencia del marco ampliamente aceptado en la comunidad de neuroimagen (SPM), siguiendo un enfoque basado en datos. Se trata de un método de análisis por regiones, donde las imágenes cerebrales son parceladas por regiones, por ejemplo, procedentes de un atlas, y éstas son procesadas y clasificadas una a una. Se reduce la dimensionalidad de dicha región (vóxeles) mediante PLS y las características son introducidas en un clasificador SVM lineal, donde datos procedentes de sujetos controles y aquellos de la condición bajo análisis son contrastados aplicando RUB como método de validación. Finalmente, para detectar las regiones más significativas, se aplica un test de proporciones a partir de las tasas de acierto obtenidas de cada región y se detectan aquellas más significativas en función de un nivel de significancia prefijado.

Esta metodología se ha aplicado satisfactoriamente en escenarios con diferentes técnicas de imagen, como MRI o la tomografía computarizada por emisión de fotón único (SPECT), diversas condiciones (Alzheimer o Parkinson) y tamaños de efecto que van desde grandes hasta triviales. Se ha comprobado que genera un control efectivo sobre falsos positivos, así como resultados consistentes en diferentes tamaños de muestra.

Los mapas obtenidos han sido comparados con los que se obtendrían aplicando SPM, concluyendo con su utilidad como una alternativa altamente competitiva y complementaria a SPM, cuyas suposiciones paramétricas en las que se basa son a menudo son difíciles de cumplir. Además, se aborda el potencial de la herramienta para ser aplicada en otros tipos de contrastes o funcionalidades. Concretamente, se explora la posibilidad de aplicarlo en estudios de electroencefalograma (EEG), es decir, pasar de estudios espaciales a temporales. Sin embargo, el rendimiento no ha sido óptimo, siendo necesario refinar el método para que sea de utilidad en este campo. Por ejemplo, esto implicaría modificar el proceso de selección de instantes temporales de interés y adoptar un límite superior para aplicar RUB que se ajuste mejor a este tipo de datos.

Explorando características relevantes: Interpretabilidad de los Modelos de Machine Learning

En este estudio se propone un sistema CAD para explorar la utilidad de características procedentes de los surcos cerebrales para discriminar entre pacientes con esquizofrenia y controles. Dicho estudio se describe en el capítulo 9, el cual muestra el potencial de combinar diversas técnicas estadísticas, de ML y de DL con características procedentes de los surcos cerebrales para abordar la tarea de clasificación.

En primer lugar, las características de los surcos cerebrales empleadas son analizadas en una etapa de selección de características relevantes, donde se comparan técnicas paramétricas y no paramétricas para realizar dicha selección, con el objetivo comprobar cómo se comportan dichas técnicas en escenarios de pequeño tamaño muestral. Las características seleccionadas también son comparadas en rendimiento de clasificación con aquellas extraídas mediante PLS, usando el algoritmo SVM lineal como clasificador. Además, se incluye la comparativa de rendimiento entre aplicar RUB o K-fold como métodos de validación. Por último, se plantea un escenario de DL en el cual se aplican técnicas XAI para comprobar la relevancia de las características durante el proceso de aprendizaje de la red.

En los resultados obtenidos se observa un rendimiento similar entre aplicar técnicas paramétricas y no paramétricas, así como en el uso de RUB y K-fold. Por tanto, podría evitarse el uso de técnicas basadas en asunciones que no se cumplen fácilmente, especialmente en relación con el típico tamaño muestral limitado. Las técnicas no paramétricas permitirían abordar de manera efectiva estas dificultades inherentes de datos limitados.

Desde una perspectiva clínica, los hallazgos obtenidos son altamente significativos por dos razones principales. En primer lugar, ninguna investigación previa ha automatizado el análisis de características procedentes de los surcos de todo el córtex cerebral, lo que representa un avance notable. En segundo lugar, la forma en que se procede en el estudio ha permitido un estudio exhaustivo de las áreas cerebrales relevantes en las diferentes etapas del sistema CAD. Por ejemplo, se han reproducido hallazgos de estudios anteriores en áreas como la temporal y la precentral, y se ha observado la importancia del hemisferio izquierdo. Estos resultados proporcionan información valiosa que contribuye a una comprensión más profunda de la condición en cuestión.

Inteligencia Artificial Explicable para Imagen

Finalmente, en el capítulo 10 se detalla el sistema CAD propuesto para automatizar el diagnóstico del deterioro cognitivo a partir de la versión clásica del Test del Reloj, un test cognitivo altamente empleado en el entorno clínico para evaluar el estado cognitivo de los sujetos. La base de datos empleada en este caso se compone de más de 7000 muestras de dibujos realizados a manos de dicho test por personas sanas y con deterioro cognitivo. A diferencia de en los estudios anteriores, tal cantidad de muestras evita el problema del pequeño tamaño muestral, permitiendo una buena capacidad de generalización del algoritmo de clasificación.

El sistema CAD está compuesto por varias etapas. La primera de ella es el preprocesado de las imágenes. Puesto que se tratan de dibujos escaneados en distintos centros médicos españoles, ni la forma ni el tamaño de los dibujos estaban normalizados, incluso contenían comentarios clínicos del paciente. Es por ello que esta primera etapa de preprocesado es de vital importancia para estandarizar las imágenes al mismo tamaño (224x224) y eliminar elementos irrelevantes del dibujo. Estas imágenes son introducidas en una red neuronal convolucional, donde las primeras capas, convolucionales, tienen el objetivo de extraer patrones relevantes para la clasificación, y las

capas finales, lineales, están orientadas a la clasificación.

La tasa de precisión de este modelo fue del 75.65 %, con un área bajo la curva ROC de 0.83, al aplicar un subconjunto equilibrado de 3282 muestras. Con un conjunto de datos más grande de 7009 muestras, la tasa de precisión fue del 70.04 %. Estos valores son consistentes según la literatura previa en este tipo de prueba. Se aplicaron varias técnicas XAI para mejorar la interpretación gráfica del modelo, como los mapas de saliencia y el algoritmo Grad-CAM (*Gradient-weighted Class Activation Mapping*). Estas técnicas destacaron la importancia de la posición de las manecillas del reloj en el dibujo, ya que las personas sanas tienden a colocarlas a la hora establecida en la prueba (a las once y diez), mientras que las personas con deterioro cognitivo muestran mayor variabilidad en su colocación. La capacidad del sistema para obtener resultados fiables e interpretables demuestra su utilidad en el ámbito clínico, incluso en regiones con recursos limitados donde se utiliza la versión analógica de la prueba.

Conclusiones

Esta tesis ha tenido como objetivo demostrar la utilidad de las técnicas de ML no solo en escenarios de clasificación, sino también en el mapeo cerebral y la detección de características relevantes. Además, se han examinado la fiabilidad e interpretabilidad de los sistemas CAD y se han explorado formas de mejorar estos aspectos. En resumen, las conclusiones extraídas de las contribuciones realizadas en esta tesis son las siguientes:

- La aplicación de una resubstitución acotada en los diversos capítulos ha demostrado su viabilidad en neuroimagen como método de validación. Esta estimación teórica del error o riesgo asociado a un algoritmo de clasificación mejora su capacidad de generalización al aprovechar la información de todas las muestras disponibles. Esto es especialmente valioso en escenarios de pequeño tamaño muestral, comunes en los estudios de neuroimagen.

- Otro enfoque para mejorar la fiabilidad del sistema es considerar las etapas de selección y extracción de características. Estas etapas desempeñan un papel vital en la identificación de cualquier inconsistencia en los datos y en la optimización del rendimiento de los clasificadores. Además, estos pasos tienen como objetivo reducir la carga computacional, que puede ser significativamente altas al analizar datos de alta dimensionalidad en neuroimagen. De hecho, se propone un sistema CAD en el capítulo 6, en el cual la etapa de extracción y selección de características fue de suma relevancia. El estudio detallado de esta etapa permitió una mejora considerable en el rendimiento del sistema. Además, el método también fue coherente con hallazgos recientes en la enfermedad de Alzheimer.

- La comparación más completa entre RUB y K-fold CV se realizó en el capítulo 7 utilizando una metodología no paramétrica para analizar la certeza de predicción en sistemas CAD. Para ello, se exploró el equilibrio entre la potencia estadística y el error de Tipo I. Ambos enfoques de validación, CV y RUB, obtuvieron una

tasa de FP muy cercana al nivel de significación para cualquier dimensión de entrada. Además, ambos ofrecieron una potencia estadística aceptable, aunque ligeramente inferior utilizando CV. Todo esto con la ventaja de utilizar todo el conjunto de muestras. Por lo tanto, considerando también que el costo computacional por iteración es menor utilizando RUB que CV, se recomienda su uso para estudios estadísticos.

- La metodología SAM, propuesta en el capítulo 8, es un enfoque basado en datos. Ofrece un enfoque factible para obtener mapas estadísticos agnósticos. A través de experimentos realizados en varios marcos experimentales y conjuntos de datos, este enfoque multivariado ha demostrado su capacidad para evaluar cambios significativos en los volúmenes cerebrales. Exhibió un control efectivo sobre los falsos positivos y arrojó resultados consistentes en diferentes tamaños muestrales. En consecuencia, SAM puede verse como una alternativa altamente competitiva y complementaria a SPM, ampliamente aceptado en la comunidad de neuroimagen.

- Se ha observado potencial en SAM para adaptar el método a otros tipos de contrastes o análisis factoriales, aprovechando los numerosos avances en ML en los últimos años. De hecho, se ha explorado la posibilidad de extrapolar la funcionalidad espacial a estudios de EEG temporal. Sin embargo, para mejorar su rendimiento, es necesario refinar el método. Esto implicaría modificar el proceso de selección de los instantes temporales de interés y adoptar un límite superior menos conservador que se ajuste mejor a este tipo de datos.

- El estudio en el capítulo 9 reveló el potencial de combinar diversas técnicas estadísticas, de ML y de DL con características procedentes de surcos cerebrales para abordar la clasificación de sujetos con esquizofrenia y controles. Los métodos utilizados en el estudio mostraron la eficacia de las técnicas de extracción y selección de características, así como los métodos de validación como RUB, para abordar de manera efectiva las dificultades inherentes asociadas con pequeños tamaños muestrales. Además, el estudio confirmó el valor de las técnicas XAI para mejorar la interpretabilidad y obtener información sobre la importancia de cada característica en el proceso de clasificación.

- Desde una perspectiva clínica, los hallazgos obtenidos en este estudio son muy intrigantes, ya que ninguna investigación anterior ha automatizado el análisis de características de los surcos cerebrales en todo el córtex cerebral. Este proceso ha permitido replicar los hallazgos de estudios previos en las áreas temporal y precentral, al tiempo que proporciona nuevos conocimientos sobre los mecanismos subyacentes en la esquizofrenia. Estos resultados ofrecen información valiosa que contribuye a una comprensión más profunda de la afección.

- Por último, el sistema CAD propuesto en el capítulo 10 obtuvo un rendimiento acorde a las expectativas al emplear una versión análoga del Test del Reloj. El uso de un gran número de muestras reales aseguró la fiabilidad de los resultados.

Además, se aplicaron métodos teóricos de grafos para mejorar la interpretabilidad e identificar patrones relevantes de información. Específicamente, se detectó la posición de las agujas del reloj como ROI. Estos hallazgos demostraron la idoneidad del sistema para su implementación en centros médicos de todo el mundo, especialmente en regiones con recursos limitados donde se utiliza la versión análoga del test.

Considerando las fortalezas y limitaciones encontradas durante el desarrollo de esta tesis, existen varias direcciones prometedoras para investigaciones futuras. Por ejemplo, el trabajo presentado está limitado a la aplicación de la resubstitución acotada para clasificadores lineales. Trabajos futuros podrían centrarse en implementar y probar otros límites en clasificadores más complejos, por ejemplo, en redes neuronales, y comparar su rendimiento con otros métodos de validación. Otra línea de investigación interesante sería ampliar el análisis de las características de los surcos cerebrales para investigar sus interrelaciones y establecer comparaciones entre las características morfológicas de los surcos y las circunvoluciones. De esta manera, se podría obtener una comprensión más amplia del papel de la morfología cortical, no solo en la esquizofrenia, si no en diversas afecciones neurológicas.

Contents

List of Figures

List of Tables

ACRONYMS

AAL	Automated Anatomical Labeling
AD	Alzheimer's Disease
ADNI	Alzheimer Disease Neuroimaging Initiative
AE	Autoencoder
AI	Artificial Intelligence
ANOVA	Analysis of Variance
APOE	apolipoprotein E
AUC	area under the ROC curve
CAD	Computer Aided Diagnosis
CDR	Clinical Dementia Rating
CDT	Clock Drawing Test
CIMCYC	Centro de Investigación Mente, Cerebro y Comportamiento – Centre for Mind, Brain and Behaviour Research
CI	Cognitive Impairment
CIBERNED	Centro de Investigación Biomédica en Red de Enfermedades Neurodegenerativas – Biomedical Research Networking Centre for Neurodegenerative Diseases
CIEN	Centro de Investigación de Enfermedades Neurológicas – Centre for Research in Neurological Diseases
CNN	Convolutional Neural Network
cMCI	converter MCI
CSF	cerebro-spinal fluid
CT	Computed Tomography

CV cross-validation

dCDT digital CDT

DIAD Dominantly Inherited Alzheimer's Disease

DIAN Dominantly Inherited Alzheimer Network

DL Deep Learning

DSM-IV-TR Diagnostic and Statistical Manual of Mental Disorders

DWI Diffusion-Weighted Imaging

ECOC Error Output Correcting codes

EC Euler Characteristic

EEG electroencephalography

FDR False Discovery Rate

fMRI functional MRI

FDG Fluorodeoxyglucose

FN False Negative

FP False Positive

FWE Family Wise Error

GLM General Linear Model

GM grey matter

Grad-CAM Guided Gradient Class Activation Map

HC Healthy Controls

HCP Human Connectome Project

KNN K-Nearest Neighbours

KS Kolmogorov-Smirnov

LIME Local Interpretable Model-agnostic Explanations

LOAD Late Onset AD

LOO Leave-One-Out

MC mutation carriers

MCI	Mild Cognitive Impairment
MEG	magnetoencephalography
ML	Machine Learning
MLP	Multilayer Perceptron
MMSE	Mini-Mental State Examination
MNI	Montreal Neurological Institute
MRA	Magnetic Resonance Angiography
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
MVPA	Multivariate Pattern Analysis
N	Negative
NC	non-carriers
NIA-AA	National Institute on Aging-Alzheimers Association
NN	neural network
NPV	Negative Predictive Value
P	Positive
PCA	Principal Component Analysis
PD	Parkinson's Disease
PAC	Probably Approximately Correct
PET	Positron Emission Tomography
PLS	Partial Least Squares
PPV	Positive Predictive Value
RFT	Random Field Theory
ReLU	Rectified Lineal Unit
ROC	Receiver Operating Characteristics
ROI	Region of Interest
RUB	resubstitution with upper bound correction

PPMI	Parkinson's Progression Markers Initiative
SAM	Statistical Agnostic Mapping
SCZ	Schizophrenia
SHAP	SHapley Additive exPlanations
SLT	Statistical Learning Theory
SMHC	Shanghai Mental Health Centre
sMRI	structural MRI
SNR	Signal-to-Noise Ratio
std	Standard Deviation
SPECT	Single Photon Emission Computed Tomography
SPM	Statistical Parametric Mapping
SPM12	Statistical Parametric Mapping v.12
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
TPOI	Time Point of Interest
VBM	Voxel Based Morphometry
VC	Vapnik-Chervonenkis
WM	white matter
XAI	Explainable Artificial Intelligence

Part I

FUNDAMENTALS

1 | INTRODUCTION

1.1 Motivation

In recent years, the application of Artificial Intelligence (AI) techniques in health and medicine has increased exponentially [1], and neuroimaging is no exception [2]. Neuroimaging is a specialised field that focuses on studying the structure, function, and connectivity of the brain. It has emerged as a valuable tool for clinical diagnosis due to its non-invasive nature, providing a convenient means to investigate the human brain. To this end, Computer Aided Diagnosis (CAD) systems are developed, which offer such assistance to clinicians in the diagnostic procedure through a set of computerised tools with pattern recognition or prediction capabilities. These tools are implemented based on multidisciplinary collaboration in areas such as mathematics, data science, AI or statistics, among others. Particularly important was the inclusion of AI algorithms in CAD systems, which led to significant improvements in the diagnostic procedure, with more accurate systems and reductions in clinician workload. Furthermore, CAD systems are not only used in clinical medicine, but are also applied in other areas such as psychology [3], behavioural science [4], cognitive neuroscience [5] or psychiatry [6] to better understand the way the brain behaves or the origin, stages and consequences of brain disorders.

The structure of CAD systems usually consists of the same steps. First of all, data to be fed into the system, usually brain images, are processed to standardise them and ensure optimal system performance. The next stage is to provide intelligence to the system to learn the desired outcome, whether it is classification of different conditions or detection of Region of Interests (ROIs). In either case, algorithms are applied to handling the input features and for the actual learning of the system. Finally, the performance of the system is evaluated in a last step.

Brain imaging scans as input features provide a high volume of information that needs to be processed. This is where AI comes into play, or more specifically, Machine Learning (ML), one of its branches. Given such large volume of information, versatile approaches based on ML are currently suggested for neuroimaging analysis, which is possible thanks to the computational capacity available today. Advances in this field of study range from linear and low-dimensional schemes [7] to deep neural network (NN) architectures for selection, extraction and classification [8]. All of these techniques contribute to enhancing accuracy and recognition rates when analysing complex patterns in high-dimensional contexts, even with the presence of subgroups in the different conditions [9]. It is precisely the latter (NNs) that are currently most widely applied, increasing the complexity of the learning techniques in the systems implemented, and which are known as Deep Learning (DL). Nevertheless, this increase in complexity does not always translate into a noticeable improvement in the systems' performance, although it does increase the computational cost and decreases the interpretability of the model. This generates several problems, such as the reluctance of clinicians to use these methods because of their lack of physical interpretation, which reduces their capacity for interpretation, or concerns about the learning capacity of the algorithms due to the opacity that shrouds them [10, 11]. Therefore, in order to establish CAD systems as a standard for clinical diagnostic support, it is necessary to address these issues, which are not the only ones.

One of the fundamental neuroimaging challenges, the sample size, persists over time [12, 13]. Neuroimaging studies are based on data, e.g. biomarkers, obtained from subjects. The number of these subjects hardly reaches a thousand, and is usually less than a hundred. Such modest sample size leads to statistical uncertainty in the studies: imprecise measurements, variable effect sizes or occurrence of False Positives (FPs) and False Negatives (FNs), i.e. the incorrect detection by the system of the presence of a condition and the failure to detect the presence of such a condition when it does exist, respectively [14, 15]. All this negatively affects the statistical power of the studies conducted and the issue is commonly known as the *Small Sample Size Problem* [16, 17].

A directly-related common issue in neuroimaging is the *Curse of Dimensionality* [18]. This is related to the disproportion that often exists in the ratio between the number of samples (sample size) and the number of features fed into the computational system. In contrast to the small sample size, the number of features normally available in neuroimaging is considerable. Brain imaging scans contain a vast number of voxels, even a million. This number is increasing due to technological advances that improve spatial and temporal resolutions [19]. In addition, features related to measurements of anatomy and physiology are increasingly used as biomarkers in conjunction with imaging scans [20].

These issues limit the learning capacity of the implemented systems, since it is extremely difficult to exploit all the information offered by the biomarkers with the typical small sample size. This is especially true when applying DL algorithms in neuroimaging. The complexity that characterises them allows for more abstract

calculations. Therefore, it might be assumed that a more accurate performance will be obtained, but these algorithms require a large number of samples to tune them properly. Not forgetting their lack of interpretability. The ideal solution to most of the issues described above would be to conduct studies with larger sample sizes. However, this is not easy to implement in practice. Limitations include the difficulty of recruiting participants in the studies (lack of volunteers, limited budget, etc.) or conducting multi-centre studies (different acquisition protocols or processing methods, diverse clinical conditions, etc.) [21]. Fortunately, in recent years, some initiatives are appearing to overcome these limitations such as the Human Connectome Project (HCP) [22] or UK Biobank [23] with thousands of samples currently collected. Another approach that is becoming popular is to use data augmentation techniques [24]. However, the samples generated are synthetic, so clinically this alternative is of less interest.

An alternative approach to mitigate the *Curse of Dimensionality* is to reduce the number of features. For this purpose, feature selection or feature extraction techniques are normally applied, which have been widely and successfully implemented in a large number of CAD systems [7, 25, 26, 27]. Nevertheless, this approach is oriented towards traditional ML algorithms, as its application in DL is meaningless. This is because DL architectures themselves serve as a method of feature extraction and subsequent classification [28, 29].

Another proposal to improve the reliability of CAD systems, which encompasses all types of algorithms, is to question the validation methods implemented so far and to propose alternatives that are better adapted to the conditions of neuroimaging [15, 30, 31]. The same applies to classical statistics used in neuroimaging, especially in those studies involving voxel-wise analyses, and that relies on assumptions that are frequently violated [32, 33, 34].

Further research is needed in this line to cover these crucial questions in neuroscience. Methods and techniques should be developed adapted to the particular conditions of the field, such as sample size or be valid for any type of data, in a reliable and interpretable ways. This will allow CAD systems moving a step forward, enabling them to become a standard for clinical diagnostic support and thus improving the quality of life of patients with earlier and more accurate diagnoses.

1.2 Aims and objectives

This thesis aims to explore different ML approaches to achieved reliable and interpretable methods in neuroimaging. Obtaining methods with these properties would lead to CAD systems of enhanced accuracy and utility in clinical practice or other areas, which is the ultimate aim of any neuroimaging study. For this purpose, the following objectives can be defined:

- Develop and evaluate different methods to increase the reliability of CAD systems,

addressing the issues associated with sample size and the number of features available.

- Develop and evaluate algorithms with optimal interpretability for clinical analysis, which enhances healthcare professionals' confidence and facilitates evidence-based decision-making.

Several studies have been conducted to achieve the proposed objectives. For the first one, which focuses on system reliability, research has been carried out questioning the techniques currently used in neuroimaging, such as validation methods and statistical brain mapping generation techniques. Additionally, robust CAD systems have been implemented. The following studies have been conducted in this regard:

1. An optimised CAD system focusing on the predictive capability of a multiclass classifier based on ML techniques. This study includes an exhaustive examination of the relevance of the feature extraction and selection steps.
2. A non-parametric statistical inference framework to assess the statistical significance of accuracies in ML and DL models. This framework also enables the comparison of the performance between traditional validation methods and a novel approach based on Statistical Learning Theory (SLT).
3. A methodology for generating statistical maps in brain images based on ML techniques (agnostic learning), which is adaptable to different medical imaging techniques. This method is compared to the traditional model based on the General Linear Model (GLM) (parametric learning), implemented in the framework Statistical Parametric Mapping (SPM).

Focusing on the second objective, which is centered on the design of CAD systems with optimal interpretability, the following studies have been conducted, where Explainable Artificial Intelligence (XAI) techniques have been incorporated alongside the algorithms inherent to a CAD system:

4. A CAD system for detecting patterns of interest in the morphology of brain sulci, in which the performance obtained using parametric and non-parametric techniques of statistical significance is compared. In addition, the classification outcomes are analysed using XAI techniques to enhance the interpretability.
5. A CAD system based on DL and XAI techniques for pattern detection in images, aiming to be a valuable tool in clinical analysis.

1.3 Organisation of this thesis

This thesis is organised in three main parts, each consisting of several chapters. Part I begins with an introduction (chapter 1), where the motivation and primary

objectives of this research are outlined. Following that, a comprehensive review of the state of the art is presented, covering topics such as the morphological features of the brain and neuroimaging techniques (chapter 2), as well as the statistical techniques (chapter 3) and AI techniques (chapter 4) commonly applied in neuroimaging. Finally, chapter 5 provides a detailed description of the datasets used in this thesis.

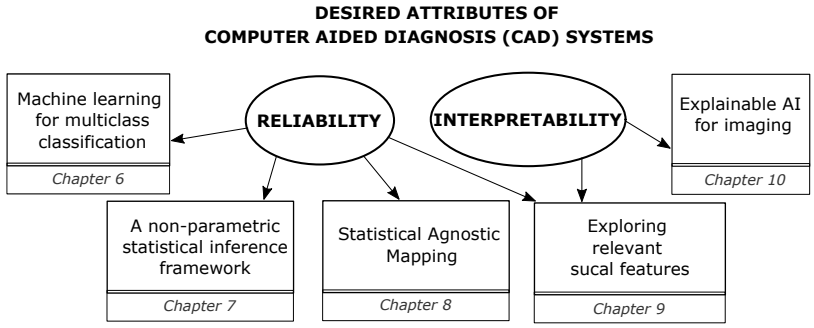


Figure 1.1: Structured scheme of the content of the contributions of this thesis.

Part II encompasses the studies conducted to support this thesis, which are divided into various topics and chapters as illustrated in Figure 1.1. The initial chapters primarily focus on analysing the reliability of CAD systems. In chapter 6, the relevance of the feature extraction and feature selection phase in implementing a multiclass CAD system is assessed. Chapter 7 examines the statistical significance of ML models based on the chosen validation method. Additionally, chapter 8 demonstrates the potential of ML techniques to generate statistical brain maps with performance comparable to standard methods, eliminating the need for statistical assumptions. Following these chapters, the analysis shifts towards the interpretability of the results, where XAI techniques are applied in CAD systems. Chapter 9 described a CAD system to assess the relevance of sulcal features in Schizophrenia (SCZ), as well as comparing the performance of the system for parametric and non-parametric techniques. Furthermore, chapter 10 introduces a CAD system that detects patterns of interest in drawings from the Clock Drawing Test (CDT), which is usually employed for assessing Cognitive Impairment (CI).

Finally, Part III, which only consists of chapter 11, presents a general discussion of the achieved results as well as the conclusions about the implications that these results may have in neuroimaging.

1.4 Contributions

Part of the content of this thesis, including figures and tables, has been published in several international journal articles and conference presentations. These contributions are detailed below.

Articles

C. Jimenez-Mesa, I. A. Illan, A. Martin-Martin, D. Castillo-Barnes, F. J. Martinez-Murcia, J. Ramirez, and J. M. Gorriz, "Optimized one vs one approach in multiclass classification for early Alzheimer's disease and Mild Cognitive Impairment diagnosis," *IEEE Access*, vol. 8, pp. 96981–96993, 2020 (**chapter 6**)

C. Jimenez-Mesa, J. Ramirez, J. Suckling, J. Vöglein, J. Levin, and J. M. Gorriz, "A non-parametric statistical inference framework for deep learning in current neuroimaging," *Information Fusion*, vol. 91, pp. 598–611, mar 2023 (**chapter 7**)

J. Gorriz, C. Jimenez-Mesa, R. Romero-Garcia, F. Segovia, J. Ramirez, D. Castillo-Barnes, F. Martinez-Murcia, A. Ortiz, D. Salas-Gonzalez, I. Illan, C. Puntonet, D. Lopez-Garcia, M. Gomez-Rio, and J. Suckling, "Statistical agnostic mapping: A framework in neuroimaging based on concentration inequalities," *Information Fusion*, vol. 66, pp. 198–212, feb 2021 (**chapter 8**)

Invited paper - C. Jiménez-Mesa, J. E. Arco, M. Valentí-Soler, B. Frades-Payo, M. A. Zea-Sevilla, A. Ortiz, M. Ávila-Villanueva, D. Castillo-Barnes, J. Ramírez, T. del Ser-Quijano, C. Carnero-Pardo, and J. M. Górriz, "Using explainable artificial intelligence in the clock drawing test to reveal the cognitive impairment pattern," *International Journal of Neural Systems*, jan 2023 (**chapter 10**)

Conferences

C. Jimenez-Mesa, J. M. Peñalver, D. Lopez-Garcia, J. Ramirez, C. Gonzalez-Garcia, F. Segovia, J. Suckling, and J. M. Gorriz, "Standarization of agnostic learning techniques in Neuroimaging: a case study in EEG," in *2022 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, pp. 1–4, IEEE, (Conference proceedings not yet published), 2022 (**chapter 8**)

C. Jimenez-Mesa, D. Castillo-Barnes, J. E. Arco, F. Segovia, J. Ramirez, and J. M. Górriz, "Analyzing statistical inference maps using MRI images for Parkinson's disease," in *Artificial Intelligence in Neuroscience: Affective Analysis and Health Applications*, pp. 166–175, Springer International Publishing, 2022 (**chapter 8**)

C. Jiménez-Mesa, J. E. Arco, M. Valentí-Soler, B. Frades-Payo, M. A. Zea-Sevilla, A. Ortiz, M. Ávila, D. Castillo-Barnes, J. Ramírez, T. del Ser-Quijano, C. Carnero-Pardo, and J. M. Górriz, "Automatic classification system for diagnosis of cognitive impairment based on the clock-drawing test," in *Artificial Intelligence in Neuroscience: Affective Analysis and Health Applications*, pp. 34–42, Springer International Publishing, 2022 (**chapter 10**)

2 | NEUROIMAGING FUNDAMENTALS

In this chapter, the fundamental aspects of working with neuroimaging in CAD systems will be discussed. Firstly, the utility of neuroimaging will be explored in section 2.1, highlighting its crucial role in providing valuable insights into the structure and function of the brain. Next, the various imaging techniques used in neuroimaging will be examined in section 2.2. Understanding the strengths and limitations of each technique is essential for choosing the most appropriate imaging modality for different research or clinical scenarios. Additionally, the typical preprocessing steps required for medical images to be effectively used in CAD systems will be explored in section 2.3. Preprocessing ensures the enhancement and normalisation of images, removing artifacts and noise to optimise subsequent analysis. The quality of preprocessing directly impacts the accuracy and reliability of CAD results, making it a crucial stage in the neuroimaging workflow. Lastly, the medical applications that have been investigated in the context of this thesis will be presented in section 2.4.

2.1 Introduction to Neuroimaging

As previously mentioned, neuroimaging is a discipline that focuses on studying the brain and nervous system using various non-invasive imaging techniques. These techniques allow for the visualisation of the brain's structure and function in real-time, providing valuable information about its organisation, activity, and connectivity.

From neuroimaging data, various aspects of the brain's structure can be analysed, including brain anatomy, where structures like the cerebral cortex, hippocampus, or amygdala can be examined, and features related to them, such as volume or thickness, can be measured. Brain asymmetry can also be explored, revealing differences in brain structure between the left and right hemispheres. Additionally, brain connectivity, particularly white matter (WM) connectivity, has emerged as one of the most recent areas of study. Among the most investigated aspects is the complex patterns of cortical folding. Neuroimaging allows the visualisation of such intricate patterns in the brain, characterised by sulci (grooves) and gyri (ridges), as illustrated in Figure 2.1, providing unique insights into brain development and individual variability [42]. Specifically, sulcal patterns offer intriguing information, as they typically form during the fetal third trimester and early life and remain largely unchanged throughout adulthood. This contrasts with the fact that the cerebral cortex is constantly changing throughout life. Therefore, sulcal patterns potentially contain valuable information about the early development of an individual, including the fetal and infant environment.

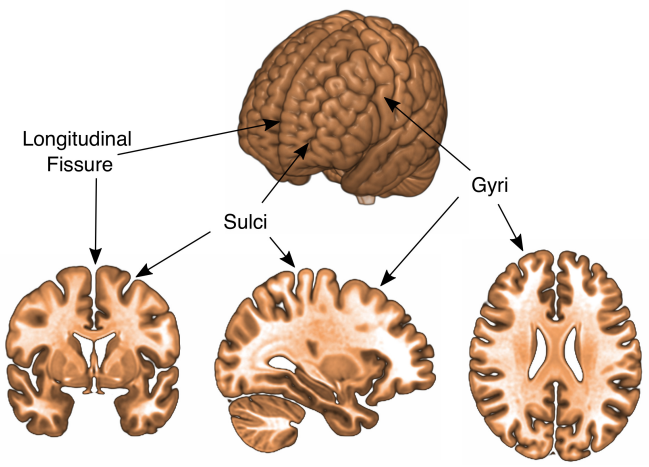


Figure 2.1: Cerebral cortex: sulci and gyri.

In clinical studies, structural brain analysis enables the detection and localisation of brain lesions. It also facilitates longitudinal studies to detect changes in brain volume, which are valuable to assess disease progression or treatment effects.

The functionality of the brain is another fascinating area of study. Neuroimaging data allows for the analysis of brain activation patterns during various tasks or condi-

tions, the assessment of functional connectivity between different brain regions even in the absence of specific tasks, and the study of brain responses to specific stimuli, substances, or events. These studies help identify the activation of regions involved in specific cognitive, motor, or sensory processes, as well as how these regions communicate and interact with each other, shedding light on how the brain operates in specific cognitive and emotional processes.

2.2 Neuroimaging Techniques

Neuroimaging techniques are necessary to conduct the above mentioned research. Some of the most common neuroimaging techniques include Magnetic Resonance Imaging (MRI), also known as structural MRI (sMRI), which provides detailed images of brain anatomy and detects structural changes; functional MRI (fMRI), which measures changes in blood flow related to brain activity; Positron Emission Tomography (PET) and Single Photon Emission Computed Tomography (SPECT), which provide information about brain metabolism and function; and electroencephalography (EEG) and magnetoencephalography (MEG), which record the brain's electrical and magnetic activity in real-time; among others. In this thesis, only those used directly in the studies detailed in Part II will be further described.

2.2.1 Magnetic Resonance Imaging

MRI is one of the most widely used imaging techniques, commonly employed not only in neuroimaging but also in various other medical applications. Its non-invasive nature and absence of ionising radiation make it a preferred choice over techniques such as Computed Tomography (CT) [43]. MRI provides exceptional anatomical detail of internal body structures (soft tissues, organs, blood vessels, etc.), allowing for accurate diagnosis and evaluation. Thus, MRI is particularly valuable in diagnosing a wide range of conditions and evaluating various diseases, injuries, and abnormalities, such as brain tumors, stroke, joint injuries or cancer, among others.

MRI uses a combination of a strong magnetic field, B_0, and radio waves to produce high-resolution images of the body's tissues and organs. The technology behind MRI relies on the behavior of hydrogen atoms, as they are one of the most abundant elements in the human body, constituting approximately 70% of its mass. Moreover, they allow to obtain a strong Signal-to-Noise Ratio (SNR).

When a patient is placed inside the MRI scanner, the magnetic field causes the hydrogen nuclei to align in a specific direction. The alignment of the protons creates a net magnetic moment in the direction of the magnetic field. After the initial alignment, radiofrequency pulses are applied to the patient's body. These pulses are carefully tuned to the resonant frequency of the hydrogen nuclei, known as the Larmor frequency. The Larmor frequency is a specific frequency at which the protons in the magnetic

field precess, a phenomenon similar to the rotation of a spinning top, which expression is as follows:

$$f_{\text{Larmor}} = \frac{y}{2\pi} B_0 \tag{2.1}$$

where y is a parameter dependent on the nuclei (y_H = 42.6 MHz/T). This frequency match results in perturbing the alignment of the nuclei, temporarily tilting their magnetic moments away from the magnetic field. This process is called *excitation*. Once the radiofrequency pulse is interrupted, the tilted protons start to relax back to their original alignment with the magnetic field. As the atoms return to their original alignment, known as *relaxation* time, they emit radio signals that are detected by the MRI scanner's receiver coils. During that time, two different relaxation times can be set: T1 and T2. T1 relaxation time is the time it takes for the protons in a tissue to return to their equilibrium state, i.e to be realigned to the longitudinal plane. T2 relaxation time is the time it takes for the protons in a tissue to lose their phase coherence, i.e. to be realigned to the transversal plane. The emitted signals contain information about the local environment and relaxation properties of the protons in different tissues.

These properties give rise to various MRI modalities, including T1-weighted, T2-weighted, and proton density images. T1-weighted images emphasise short T1 relaxation times, making them valuable for evaluating anatomy and detecting certain pathologies with high contrast between grey matter (GM) and WM. T2-weighted images highlight short T2 relaxation times, aiding in the detection of inflammation and edema by emphasising cerebro-spinal fluid (CSF). Proton density images focus on the abundance of protons in tissues. An illustrative example of a T1-weighted brain scan can be seen in Figure 2.2.

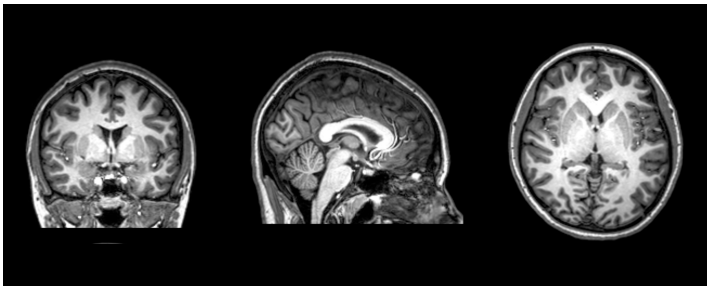


Figure 2.2: Example of a T1-weighted MRI brain scan. From left to right: coronal, sagittal and axial planes.

Additionally, specialised MRI techniques, such as Diffusion-Weighted Imaging (DWI), Magnetic Resonance Angiography (MRA), and fMRI, offer additional insights into tissue microstructure, blood flow, and brain activity, respectively.

2.2.2 Single Photon Emission Computed Tomography

SPECT is a nuclear medicine imaging technique used to evaluate the function and blood flow of organs and tissues within the body. It finds particular application in

neurology and cardiology, playing a crucial role in diagnosing, treating, and monitoring various diseases, such as Parkinson's Disease (PD), epilepsy, cardiac ischemia or brain tumors.

SPECT involves the injection of a radioactive tracer, known as a radiopharmaceutical, into the bloodstream. The radiopharmaceutical is a molecule that contains a radioactive isotope (radioligand), typically a gamma-emitting radionuclide. The choice of radiopharmaceutical and the timing of image acquisition are critical to ensure optimal imaging of the specific function or process being studied. The range of radiopharmaceuticals available for use is indeed diverse. For instance, in this thesis, the radiopharmaceutical applied is the $I^{[123]}$-Ioflupane radioligand. This particular radioligand exhibits a high binding affinity for dopaminergic transporters in the brain, allowing for a quantitative measurement of dopaminergic neuronal loss. Consequently, ^{123}I-FP-CIT SPECT brain scans are extensively employed in diagnosing PD, often referred to as SPECT-DaTSCAN due to the use of this specific radiopharmaceutical. An illustration of a ^{123}I-FP-CIT SPECT scan is shown in Figure 2.3.

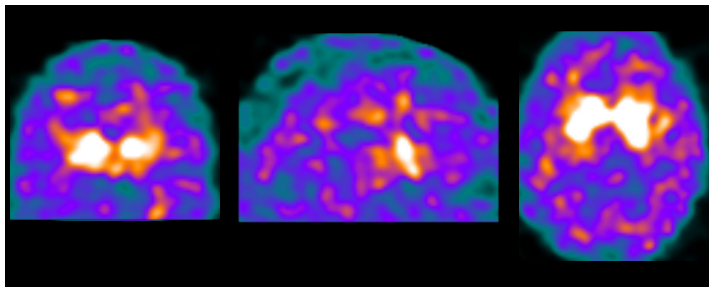


Figure 2.3: Example of a ^{123}I-FP-CIT SPECT scan. From left to right: coronal, sagittal and axial planes.

When the radiopharmaceutical is injected, it starts to circulates throughout the body and accumulates in the target tissues or organs. Once inside the body, the radiopharmaceutical emits gamma rays as a result of radioactive decay. These emitted gamma rays are detected by a specialised gamma camera. The detected intensity of these gamma rays is directly related to the concentration of the radiopharmaceutical in the tissues. During the imaging process, the scanner rotates around the patient, capturing multiple 2D images from various angles. Each 2D image represents a projection of the radiopharmaceutical distribution in the body from a specific angle. These projections are then combined and processed by a computer to create a three-dimensional representation of the distribution and activity of the radiopharmaceutical in the body.

SPECT imaging has limitations, including lower spatial resolution compared to other imaging modalities such as CT or MRI. However, it remains a valuable and non-invasive tool in specific clinical scenarios where functional information is crucial for diagnosis and treatment planning. Moreover, recent advances in neuroimaging are enabling multimodal imaging by combining, for example, SPECT-CT scans, to generate images that take advantage of the benefits of both modalities and reduce their limitations [19].

2.2.3 Electroencephalography

EEG is a non-invasive technique used to measure and record the electrical activity of the brain. It involves placing electrodes on the scalp to detect and amplify the electrical signals produced by the brain's neurons, providing valuable insights into brain activity patterns, including neuronal oscillations and event-related potentials [44]. EEG is a safe and relatively inexpensive method for assessing brain activity in real-time. These characteristics made EEG a widely used technique in neuroscience and clinical practice. In neuroscience research, EEG helps investigate brain dynamics during various tasks, such as attention, memory, and language processing, or even sleep disorders. In clinical settings, EEG is usually employed for diagnosing and monitoring neurological conditions. For example, it is the primary tool for diagnosing epilepsy, as it can detect abnormal electrical activity indicative of seizures.

The electrical signals recorded by EEG are typically displayed as waveforms, known as EEG traces or EEG waves. These waves represent the collective activity of millions of neurons firing in synchrony, whose electrical impulse is sufficient to be detected by the electrodes placed on the scalp. This synchronised activity is produced in different frequency ranges [45]. A typical frequency spectrum in which to divide the different EEG waves are: alpha, beta, delta, gamma, or theta waves. All of these bands correspond to different states of brain activity and can provide information about cognitive processes, sleep stages, and abnormal brain patterns.

The number of electrodes placed on the scalp to detect signals depends on several factors, such as the clinical purpose, the type of study and the equipment used. Such a number can go from 20 electrodes in a routine EEG study to 256 for more precise studies. For example, there are international standards for electrode placement, such as the 10-20 system [46], which uses specific locations based on distances proportional to 10% or 20% of certain head dimensions. An example of electrode placement scheme can be observed in Figure 2.4 (left). It shows the recording electrodes placed throughout the head, as well as others depicted externally representing the reference electrode and the ground electrode. The former allows the difference between the measured signal and the fixed signal to be established as the EEG waves, and the latter matches the amplifier potential to that of the subject's body to reduce artifacts that may occur. These EEG waves for each recording electrode can be seen in Figure 2.4 (middle).

Once the signals are preprocessed, mainly filtered, it is possible to compute the evoked potential linked to the analysed condition. An example is shown in Figure 2.4 (right), where increased activity is observed in the frontal region of the brain.

EEG has certain limitations, as it does not offer precise details on the exact location of neuronal activity and primarily measures activity on the cortical surface of the brain, making it challenging to detect signals from deeper structures. Additionally, EEG can be affected by artifacts, and accurately identifying neural sources is a complex task. Nevertheless, recent advancements in EEG technology, including high-density electrode arrays and advanced signal processing techniques, have led to improvements

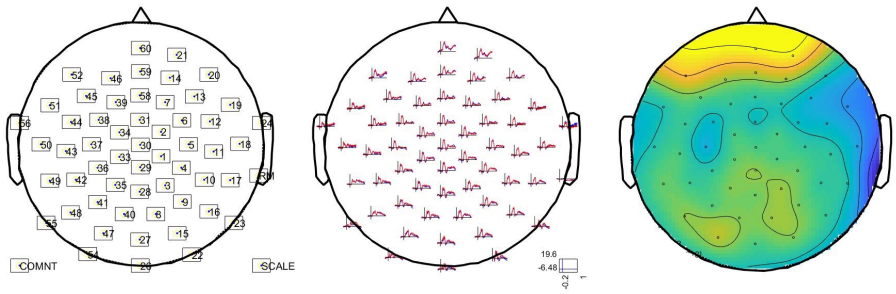


Figure 2.4: Layout of a typical EEG distribution. Average EEG signal distribution given a condition. Topographical frequency.

in the spatial and temporal resolution of EEG recordings. This progress enables more precise location of brain activity and a better understanding of brain networks and connectivity.

2.3 Preprocessing in Neuroimaging

The preprocessing of medical images is of utmost importance in CAD systems as it addresses various challenges associated with these images. These challenges include noise reduction, image standardisation, artifact removal, and image registration. By effectively dealing with these issues, preprocessing techniques significantly enhance the accuracy and reliability of the subsequent analysis performed by CAD systems.

Among the crucial steps in preprocessing is the spatial and intensity normalisation of the images. These processes ensure that the images become comparable with each other, allowing the algorithms implemented in the CAD systems to operate under the best possible conditions. Therefore, normalisation leads to more consistent results, enhancing the overall performance of the CAD system in medical image analysis.

2.3.1 Spatial Preprocessing

Spatial preprocessing refers to all the techniques and transformations applied to the images to align them to a common reference frame and enhance their comparability. In spatial preprocessing, geometric transformations are applied to the images to align them to a common reference framework. This involves the registration and correction of images to address any variations in orientation, position, or scale that may have occurred during acquisition. By performing spatial normalisation, anatomical and geometric differences are removed, enabling a more accurate and reliable comparison

of anatomical structures across different individuals or longitudinal studies.

One of the most complex processes is related to MRI, as it involves more than just registering all the images. Typically, the images are first aligned to a reference template, then (optionally) smoothed, and finally segmented to work with GM or WM maps. This segmentation process therefore results in skull stripping. An overview of these steps can be seen in Figure 2.5.

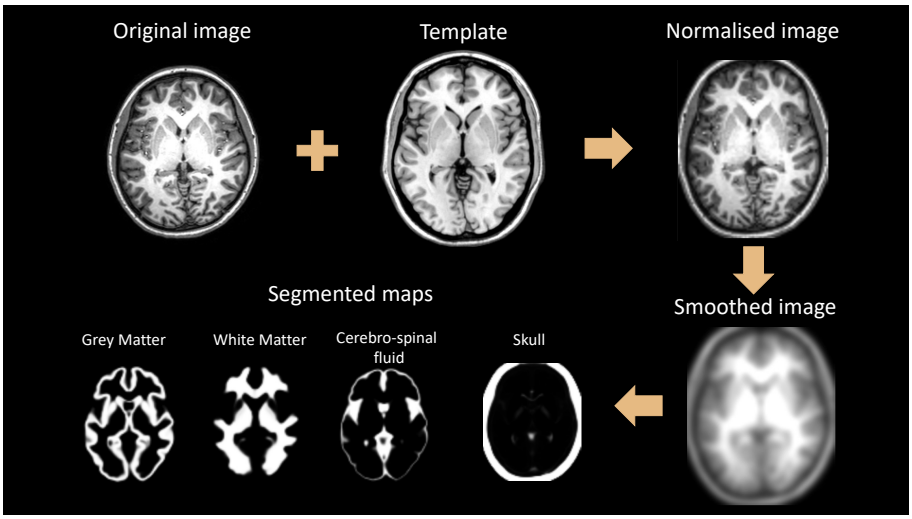


Figure 2.5: Typical steps involve in spatial preprocessing of MRI scans.

Registration or Spatial Normalisation

Registration is a general term used to describe the process of aligning two or more images or datasets in a spatially consistent manner. Registration can be rigid, affine, or non-rigid, depending on the degree of transformation allowed. In rigid registration, only translation, rotation, and scaling are permitted. Affine registration allows for additional shearing and stretching. Non-rigid registration, on the other hand, permits more complex deformations, enabling the alignment of images with more substantial spatial differences, such as different shapes or deformations due to anatomical variations or pathology. Registration techniques can align images acquired at different times or from different modalities, allowing CAD systems to track changes, identify abnormalities, and facilitate longitudinal analysis.

Spatial normalisation is a specific type of registration that involves transforming multiple images to a common coordinate system or reference space. The process aligns different images so that corresponding anatomical or functional regions in each image match spatially. The most common application of spatial normalisation is to align individual subject images to a standard anatomical template or atlas. This allows for direct comparison of different subjects or datasets and facilitates group-level statistical analysis.

Co-registration

Co-registration goes beyond simple image registration, as it involves the process of aligning multiple image modalities. This occurs when two or more neuroimaging techniques are acquired simultaneously, such as MRI and SPECT scans, in a multimodal context. In this scenario, the first step is to register the lower-resolution images (e.g., SPECT) with respect to the higher-resolution images (e.g., MRI). Subsequently, the high-resolution image is normalised to a template, and finally, the parameters and warping obtained from the higher-resolution modality are applied. This multi-step co-registration approach allows for accurate integration of information from different modalities, enabling researchers and clinicians to leverage the complementary strengths of each imaging technique effectively.

Segmentation

Segmentation refers to the process of delineating and classifying different anatomical structures or ROIs within brain images. This technique involves partitioning the image into distinct and meaningful regions based on intensity, texture, or other image features. Typically, the segmentation process refers to isolating the different brain tissues, generating GM, WM or CSF maps. This process is crucial for various neuroimaging applications, including brain morphometry, lesion detection, and functional localisation in both research and clinical settings.

Another widely used technique in neuroimaging is parcellation. It involves dividing the brain into different regions or parcels based on certain characteristics or criteria, often using atlas templates. The main objective of parcellation is to create more manageable and meaningful units within the brain, which facilitate its analysis. Each resulting region or parcel may correspond to functional areas, specific anatomical structures, or brain circuits with particular functions.

2.3.2 Intensity Normalisation

Intensity normalisation is a process that adjusts the voxel intensities of images to a common scale or range. This allows for the comparison of intensity values across several images from different subjects or time points, enabling more accurate and consistent quantitative analysis. These differences in intensity arise due to variations in acquisition settings, equipment, or patient characteristics. It should be noted that intensity normalisation is particularly crucial in functional images such as SPECT or PET. In these modalities, variations in intensity values can significantly impact the analysis and interpretation of results, as intensity levels directly relate to the underlying biomarkers being studied. On the other hand, in structural images, such as MRI, intensity differences are generally less relevant as these images are often considered unitless.

Various methods exist for intensity normalisation. The most common approach is

to rescale the image intensities using linear transformations, as shown below:

$$I = \frac{I_0}{I_p} \tag{2.2}$$

where I represents the new intensity value, I_0 is the original value, and I_p is a constant parameter set for each of the images to be normalised to rescale the intensity while preserving its fundamental information. This parameter can be estimated in different ways. A widely accepted option is to apply normalisation to the maximum approach [47, 48]. This method estimate I_p as the average of the highest $k\%$ intensity values in the image, where k us usually choosen as 5%.

Another type of widely used methods are those based on a general linear transformation as follows:

$$I = aI_0 + b \tag{2.3}$$

where a is a scaling factor and b is an offset. This transformation allows for more control over the intensity rescaling process. In this case, some of the proposed methods are based on histograms [49], the α-stable distribution or the gaussian distribution (a particular case of the latter) [50]. From the α-stable distribution, the linear transformation can be computed as:

$$I = \frac{\gamma^*}{\gamma} I_0 + \left(\mu^* - \frac{\gamma^*}{\gamma}\mu\right) \tag{2.4}$$

where γ^* represents the mean of γ (dispersion) parameters from all the input scans, and μ^* is calculated as the average of μ (location). As depicted in Figure 2.6, this procedure reduces the differences between scans due to external factors such as the amount of radioligand injected to each patient, their absorption rate or the calibration of the acquisition equipment, among others.

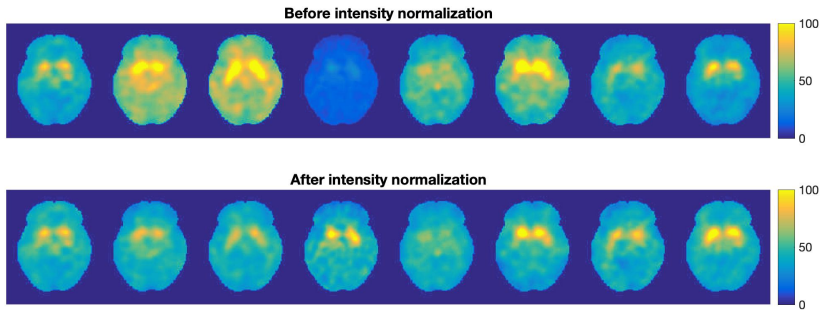


Figure 2.6: Examples of differences between ^{123}I-FP-CIT SPECT scans before/after intensity normalisation using α-stable distributions.

2.4 Medical Applications

Throughout the previous sections, it has been evident that medical applications of neuroimaging are highly diverse. Consequently, CAD systems are gaining widespread

recognition as valuable tools for clinical assessment [51, 52] and there is abundant literature on their research. In this section, the focus will be on the description of the three diseases that have been investigated throughout this thesis: Alzheimer's Disease, Parkinson's Disease, and Schizophrenia.

2.4.1 Dementia and Alzheimer's Disease

Dementia is a syndrome characterised by a deterioration of cognitive function as a result of a variety of disorders that affect the brain. It encompasses a range of cognitive and behavioral symptoms, including memory impairment, difficulties in thinking, and challenges in performing daily activities. According to the World Health Organization (www.who.int), more than 55 million people worldwide suffer from dementia, with nearly 10 million new cases reported each year. This has a direct impact not only on the psychological, social and economic aspects for people living with dementia, but also for their relatives, caregivers and society in general.

An early diagnosis of dementia is crucial to slow down its progression and enhance the quality of life for affected CI individuals. Cognitive assessment tests, such as the Mini-Mental State Examination (MMSE), the Clinical Dementia Rating (CDR) and the CDT [53, 54, 55], are commonly employed in this context. These tests are often complemented by brain imaging studies to aid in the diagnostic process. In this regard, CAD systems [56, 57] can serve as valuable clinical tools, assisting healthcare professionals in accurately identifying and monitoring cognitive impairments.

Alzheimer's Disease (AD) is the most prevalent cause of dementia, constituting around 60-80% of all dementia cases. In the United States alone, it is estimated that the number of individuals aged 65 and older with AD reaches 6.7 million in 2023 [58]. It is a progressive neurodegenerative disorder that primarily affects the brain, leading to a decline in memory, thinking abilities, and overall cognitive function. This disease is characterised by the accumulation of abnormal protein (β-amyloid and phosphorylated τ) deposits in the brain. These deposits disrupt the communication between neurons and ultimately lead to the death of brain cells (neurodegeneration), resulting in the gradual decline of cognitive function [59].

In the early stages of AD, individuals may experience subtle memory loss and have difficulty recalling recent events. As the disease progresses, symptoms may include confusion, disorientation, mood and behavior changes, language problems, and challenges with problem-solving and decision-making. In later stages, individuals often require assistance with daily activities such as eating, dressing, and personal hygiene. Due to its medical importance and societal implications, there has been consistent interest in assisting the early diagnosis of AD within the medical imaging community [60]. Distinguishing between AD and related neurological disorders, including its prodromal stage Mild Cognitive Impairment (MCI), is particularly challenging during the early stages of the disease from a clinical evaluation perspective.

The exact cause of AD remains incompletely understood, but it is thought to

result from a complex interplay of genetic, environmental, and lifestyle factors. Age represents the most significant risk factor, as the majority of cases are observed in individuals aged 65 and older, which can be named as Late Onset AD (LOAD). However, there are rare forms of AD, such as Dominantly Inherited Alzheimer's Disease (DIAD), characterised by specific genetic mutations that trigger symptoms at a much younger age, typically between the ages of 30 and 50 [20].

There is no cure for AD yet either. Nevertheless, clinical resources employed for its diagnosis, treatment, and monitoring are continually improving. Significant advancements in medical technology and neuroimaging techniques have enhanced early detection and provided valuable insights into disease progression. Moreover, ongoing research is focused on gaining a deeper understanding of the underlying mechanisms of AD. This includes studying genetic factors, cellular processes, and the role of specific proteins in the brain. Regarding the neuroimaging techniques used for studying AD, the most widely employed one is MRI, due to its ability to detect brain atrophy in both WM andGM that occurs during the development of the disease, especially the loss of GM in regions such as the hippocampus and parahippocampal gyrus [61]. Additionally, nuclear imaging techniques like PET or SPECT are also applied, as AD is characterised by reduced brain activity in regions such as the precunei, lateral-parietal, and posterior temporal cortex [62].

2.4.2 Parkinson's Disease

Parkinson's Disease is another significant cause of dementia. Approximately 3.6% of dementia cases are attributed to PD, while 24% of individuals with PD will develop dementia at some point during the course of the disease [58]. This neurodegenerative disorder results from a progressive loss of dopaminergic neurons in the nigrostriatal pathway, with causes that are still unclear, but both genetic and environmental factors are believed to play a role [63].

As dopamine is a neurotransmitter involved in regulating movement and coordination, this disease primarily affects movement. The main symptoms of PD include tremors (involuntary shaking), rigidity (stiffness of muscles), bradykinesia (slowness of movement), and postural instability (impaired balance and coordination). Additionally, other non-motor symptoms may be present, such as cognitive changes, mood disorders, sleep disturbances, and autonomic dysfunction [64]. While it is more commonly seen in older individuals, typically appearing after the age of 60, early-onset cases can also occur in younger people.

As with AD, there is currently no cure for PD, making early diagnosis essential for implementing optimal treatments to control symptoms. Functional imaging techniques are commonly employed in this disorder to detect the level of dopamine transporter uptake. Specifically, the most frequently used imaging modality for this purpose is SPECT-DaTSCAN, as its radiopharmaceutical binds to the dopamine transporters in the striatum. In the case of PD, where the amount of dopamine transporters is reduced,

individuals exhibit smaller and more irregular patterns in the striatum compared to healthy subjects, resulting in brighter and more uniform patterns in the scan.

2.4.3 Schizophrenia

Schizophrenia is a chronic mental disorder related to an altered perception of reality (psychosis) that affects how a person thinks, feels, and behaves. It is characterised by a range of symptoms that can include hallucinations, delusions, disorganised thinking and speech, social withdrawal, and impaired cognitive function. These symptoms can vary in severity and may emerge gradually or suddenly and can be categorised as positive symptoms, those that involve the presence of abnormal experiences or behaviors (e.g. hallutinations); and negative symptoms, those related to the absence or reduction of normal behaviors (e.g. decreased emotional expression). The prevalence of this disorder is approximately 1% of the population [65].

The exact cause of SCZ is unknown, but it is believed to result from a combination of genetic, environmental, and neurochemical factors. Moreover, it is a hereditary condition. Imbalances in certain brain chemicals, such as dopamine and glutamate, may contribute to the development of the disorder [66]. The onset of the disease usually begins in early adulthood. However, it is a complex and heterogeneous disorder, and its onset can vary among individuals. Therefore, an early detection and appropriate treatment are crucial in helping individuals with this condition and improving their quality of life.

Early detection of SCZ using MRI has been a subject of ongoing research in the field of neuroimaging. Brain structural abnormalities have been found in individuals with SCZ, which are already present in the early stages of disease development [67]. In terms of GM, there is evidence of an overall reduction of cortical folding, being noteworthy the cortical reduction in temporal-parietal-occipital regions [67, 68]. Moreover, the reduction in the superior temporal gyrus is related to positive symptoms [69] and the reduction in the prefrontal area to negative symptoms [70]. Differences in sulci have also been found between case-control groups; for example, a shorter paracingulate sulcus and shallower superior temporal sulcus, which are related to hallucinations [71].

On the other hand, functional imaging techniques, such as PET and SPECT, play a crucial role in expanding the understanding of the etiology of this mental disorders and improving current treatments. Among the studies conducted, particular emphasis is placed on those evaluating dysregulations in dopamine levels and glutamatergic neurotransmission [65].

3 | STATISTICAL INFERENCE IN NEUROIMAGING

In neuroimaging, analysing the entire population to investigate a condition is impractical due to the high number of individuals and associated time and costs. Instead, a population sample is studied to draw specific conclusions, which are then used to derive general conclusions with some level of risk. Thus, statistical inference allows for the generalisation of findings from a smaller sample and provides an assessment of the associated risk. This chapter includes the techniques most commonly used in neuroimaging to perform such statistical inference.

3.1 Hypothesis testing

Hypothesis testing is the process of inferring from a sample whether or not a given statement about the population appears to be true [72]. The statement, known as the *hypothesis*, consists of the *alternative hypothesis* or H_1, which is the statement to be proven, and the *null hypothesis* or H_0, which is its negation and the one being tested. For example, in neuroimaging, H_0 typically states that there is no significant difference

between a certain condition and a control group, while H_1 implies the presence of a significant difference. The test statistic serves as an indicator of agreement or disagreement with the null hypothesis. A decision rule is then applied to determine whether to accept or reject the null hypothesis based on the values of the test statistic. This decision is typically based on a predetermined significance level, denoted as α, which represents the maximum allowable probability of erroneously rejecting the null hypothesis. If the calculated test statistic falls within the critical region, determined by the significance level, the null hypothesis is rejected in favor of the alternative hypothesis.

It is important to consider the possibility of making incorrect decisions in hypothesis testing. If the null hypothesis is true, it can be erroneously rejected, resulting in a *Type I error*. This error occurs when a significant effect or difference is concluded even though none exists in the population. Similarly, a *Type II error* can occur when the null hypothesis is false, but it is not rejected. This error arises when independence is declared or a genuine effect or difference in the population is not identified.

These two error types have associated with them certain probabilities of the errors being made, such as the significance level (α), which is already mentioned. It is related to the p-value [73], which represents the probability of observing a test statistic as extreme as, or more extreme than, the observed value assuming the null hypothesis is true. If the p-value is less than alpha, the null hypothesis is rejected. β is the probability of committing a Type II error, which is failing to reject the null hypothesis when it is false. The complement of beta is the power of the test (1-β), which represents the ability to correctly detect a true alternative hypothesis. The complementary of α (1-α) represents the confidence level in the decision made by not rejecting the null hypothesis. A visual summary of these concepts can be observed in Table 3.1.

		Decision	
		Accept H_0	Reject H_0
Situation	H_0 is true	Correct decision probability 1-α	Type I error probability α (*significance level*)
	H_0 is false	Type II error probability β	Correct decision probability 1-β (*power*)

Table 3.1: Hypothesis testing decision and error probabilities.

3.1.1 Two-sample testing

On the basis of a binary classification problem, an analysis of differences between conditions included in a dataset (e.g. AD vs HC) can be established as an hypothesis test of a two-sample problem. Given $\mathrm{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ as a dataset where $\mathbf{x}_i \in \mathbb{R}^N$ are the observed features and $y_i \in \{0, 1\}$ the class labels, the null hypothesis implies that there is independence between data ($\mathbf{x}_i$) and its label (y_i) in terms of conditional probability

distributions:

$$H_0 : p(\mathbf{x}_i, y_i) = p(\mathbf{x}_i)p(y_i) \quad \text{vs.} \quad H_1 : p(\mathbf{x}_i, y_i) \neq p(\mathbf{x}_i)p(y_i) \tag{3.1}$$

A rejection of H_0 would mean that there is a dependency between data and labels, i.e., the distribution of $\mathbf{x}$ is conditional on the value of the class labels y. In terms of functional neuroimaging H_0 implies that there is no effect in $\mathbf{x}$ given y.

Traditionally, classical statistics have been used to analyse the difference of the population means within two-sample distributions [74]. Nevertheless, more and more studies are opting for non-parametric approaches [75] due to the implications and assumptions of the former [32].

3.2 Statistical Tests

Some of the statistical tests commonly used in data analysis, which are applied in this thesis, are the ones described below. Although they serve different purposes, they are all used to assess and analyse data for hypothesis testing and comparisons. These tests can be categorised into two groups: those evaluating normality, or those conducting group comparisons.

3.2.1 Assessing Normality

One of the most typical assumptions when working with parametric tests is that the sample (data) follows a normal distribution. However, this is not always true, so it is good practice to apply tests to check the distribution of the data under analysis. The Shapiro-Wilk test is one of the most commonly used tests of normality [76].

The test evaluates the null hypothesis that the sample $(x_1, x_2, ..., x_n)$ is normally distributed, i.e. come from a normally distributed population. Its associated test statistic is:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.2}$$

where $\bar{x}$ is the sample mean, $x_{(i)}$ is the th-order statistic, i.e. $x_{(1)}$ is the smallest value of the sample and $x_{(n)}$ is the largest one, and a_i is a coefficient related to the covariances, variances, and means of a normally distributed sample [76] .

3.2.2 Comparing groups

Once the distribution of the sample is known, further tests can be applied to the sample to detect differences between groups or classes.

If the sample does indeed follow a normal distribution, the most commonly used test is the two-sample t-test, usually known as Student's t-test [77, 78]. This test

compares the means of two independent groups to determine if there is a significant difference between them. Whereas H_0 implies that both means are equal, H_1 can be either two-tailed (different), left-tailed (mean of the first group greater than that of the second group), or right-tailed (the opposite). In either case, it assumes that the data is normally distributed. Depending on the sample size of the groups (balanced or unbalanced) and whether or not the same variance is assumed in both groups, the expression for the test statistic is different. In the case of considering the variances of both groups similar, irrespective of the sample size, the expression is as follows:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{3.3}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the means of group 1 and group 2, respectively, n_1 and n_2 their sample sizes, and s_1^2 and s_2^2 are their variances.

If it is not certain that the sample follows a normal distribution, non-parametric tests such as the Mann-Whitney U test can be applied, which is also known as the Wilcoxon rank-sum test [79, 80]. This test compares the medians of two independent groups. It does not assume any specific distribution and is applicable when the data is ordinal or skewed. It tests the null hypothesis that the two groups come from the same population against the alternative hypothesis that they come from different populations. The test statistic of this test is defined as the smaller of:

$$\begin{aligned} U_1 &= n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \\ U_2 &= n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 \end{aligned} \tag{3.4}$$

where R_1 and R_2 represent the sum of the rank in each group.

3.3 Statistical Methods for Analysis

Other options that allow for examining relationships between different variables to detect those of interest are linear models such as linear regression or Analysis of Variance (ANOVA). In terms of neuroimaging, these techniques help identify patterns of brain activation associated with variables of interest and control for factors that could influence the results, thereby enhancing the interpretation of the studies.

Linear regression is widely used to investigate the relationship between a dependent variable of interest, y, and m predictor variables $\mathbf{x} = \{x_1, ..., x_m\}$ for n samples. This statistical technique provides information about the strength and direction of the association between the variables and allows for the prediction of the dependent variable based on the values of the predictor variables. In the context of neuroimaging, these variables may represent measures of brain activity and clinical or demographic features. The linear regression model seeks to find the best-fitting straight line that

describes the linear relationship between the variables. Its expression could be defined as:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m + \epsilon \tag{3.5}$$

where β_0,β_1,...,β_m are the model parameters, being β_0 the intercept (constant term) and β_i ($i \geq 1$) the coefficients corresponding to the predictor variables, and ϵ an error term.

ANOVA is a statistical technique used to compare means between groups or conditions. Although both ANOVA and the t-test serve the same purpose of comparing means between groups, they are applied under different circumstances. ANOVA is used when there are three or more groups or conditions to compare, while the t-test is specifically designed for comparing means between only two groups. An example of an application of ANOVA in neuroimaging is to examine differences in brain activity (features) between different experimental conditions or groups of subjects. The most common type of ANOVA is one-way ANOVA which involves a single independent variable with three or more groups. To evaluate the null hypothesis of there is no significant difference among the means of three or more groups being compared, the F-statistic is applied, which is related to the ratio of the computed variance between means to the within-sample variance as follows:

$$F = \frac{V_{between}}{V_{within}} = \frac{\sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x})^2/(K-1)}{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2/(N-K)} \tag{3.6}$$

where $V_{between}$ and V_{within} represent the between-group and within-group variabilities, respectively, K is the number of groups, N denotes the sample size, $\bar{x}_i$ and n_i are the mean and number of samples of the i-th group, $\bar{x}$ denotes the sample mean of a feature, and x_{ij} is the j-th observation of the feature in the i-th group.

To conduct these methods, some assumptions must be considered. For linear regression, these assumptions include linearity, independence, homoscedasticity, and normality of the residuals. Similarly, when performing ANOVA, it is important to take into account the assumptions of independence, normality of the dependent variable within each group, homogeneity of variance, and random sampling. Therefore, techniques for feature selection based on ML are more suitable to be applied in neuroimaging (see section 4.2).

3.4 Group-level analysis

When analysing neuroimaging data to compare multiple subjects within different groups or conditions, researchers have several options available. The previously mentioned tests and methods mainly facilitate univariate analysis, wherein each feature is analysed independently, and their relevance is compared, often based on the obtained p-value for each feature. However, some of them (e.g. linear regression) also enable the analysis of causality, effects, or correlations between multiple variables, leading to multivariate analyses, which will be explored further in chapter 4 through ML

techniques. In this section, the primary focus will be on using images as features for analysis. Within this context, each voxel comprising the image can be treated as an individual feature (voxel-wise inference). Alternatively, they can be aggregated, for instance, based on regions defined by an atlas (region-wise inference), or by grouping significant voxels into clusters and then evaluating the statistical significance of the entire cluster size (cluster-wise inference). Voxel-based analysis has traditionally been the most commonly implemented approach, providing a comprehensive examination of brain activity or structural differences at a fine-grained level, making it particularly suitable for detecting localised brain changes or activations.

3.4.1 The General Linear Model

One of the bases for voxel-wise analysis is the GLM. The GLM is a versatile and widely used statistical framework that allows modelling the relationship between a dependent variable (observation) and one or more independent variables (explanatory variables or predictors) through a linear combination of these variables, along with the potential inclusion of other terms such as intercepts, interactions, and covariates. For instance, when there is only one independent variable and one dependent variable, the GLM reduces to a simple linear regression.

In the context of neuroimaging, the GLM can be defined for a between-subject comparison for each one of the voxels as:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \tag{3.7}$$

where $\mathbf{y}$ is the $N \times 1$ observational vector (e.g. voxel intensity), being N the number of samples, $\mathbf{X}$ is commonly denoted as design matrix ($N \times M$) and it contains the M explanatory variables (e.g. experimental conditions or group membership), β represents the $M \times 1$ vector of coefficients that quantify the relationship between the explanatory variables and the dependent variable, and ϵ is the $N \times 1$ error (residual) vector, representing the variability not explained by the model.

To estimate the coefficients that best fit the data, i.e. that minimise the difference between the observed values and the predicted value, the ordinary least squares estimation is usually applied [81, 82], assuming that the errors are independent and identically distributed. Then, the estimation of β is given by:

$$\hat{\beta} = \left(X^t X\right)^{-1} X^t y \tag{3.8}$$

The GLM also allows the construction of more complex models which include covariates, additional independent variables that account for sources of variability or confounding factors. Covariates can be continuous variables (e.g., age, MMSE score) or categorical variables (e.g., sex, clinical diagnosis). By including covariates, the GLM enables the examination of specific effects while controlling for other factors that may influence the dependent variable.

Once, the GLM is constructed, the significance of the independent variables and their effects on the dependent variable can be assessed by means of statistical tests. These tests can determine whether there are significant differences between groups, significant associations between variables, or significant main effects and interactions in factorial designs.

3.4.2 Statistical Parametric Mapping (SPM)

SPM constitutes an statistical approach used in neuroscience and neuroimaging to analyse data obtained from brain images. It is primarily used in studies involving functional imaging, such as fMRI (timeseries) or PET (static image), for which it was first proposed by [83]. Currently, it is suitable for a wide range of imaging modalities, such as EEG or MRI. SPM enables statistical analysis of brain image data to identify significant correlations and differences in brain activity between different conditions or study groups. To do so, it applies inference techniques based on hypothesis testing and an optimal GLM is formulated that effectively describes the variations present in the data. For sMRI, this methodology is commonly referred to as Voxel Based Morphometry (VBM), where each voxel is used to compare volume or density between groups in order to identify structural variations [84].

In the SPM software [85], various tests, including the massive univariate t-test and ANOVA, can be utilised. These tests involve employing a design matrix that specifies a contrast based on t (for t-test) or F (for ANOVA). Once the statistics are estimated, SPM converts them to Z-scores, which represent the number of standard deviations an observation deviates from the mean, with a positive or negative sign indicating its position above or below the mean, respectively. Once the test is computed voxel-wise, statistical maps that represent the probability of finding differences among the studied conditions are created. These maps provide information about brain regions that exhibit significant changes in neuronal activity associated with a specific task or condition. Such significance can be concluded by estimating p-values. These p-values are associated with the likelihood of obtaining Z-scores that are as extreme or more extreme than the one observed.

There are different thresholds that can be used to obtain the statistical map. For example, p-value corrected for multiple comparisons (see section 3.4.3) is usually selected. Another option is to use an uncorrected p-value, i.e. each voxel is evaluated separately. Nevertheless, the latter generates much less conservative significance maps with a high FP rate [34]. Thus, the general recommendation is to apply a corrected p-value or add an extent threshold to limit the amount of significant voxels by requiring a minimum number of voxels be clustered together. The latter can be referred to as *cluster-wise analysis*. Typically, the significant threshold chosen in a voxel-wise analysis is $\alpha = 0.05$, i.e. any voxel with a p-value smaller than 0.05 ($p < 0.05$) is considered statistical significant and the null hypothesis can be rejected in such voxel. This value indicates that, under the assumption of repeating the experiment multiple times, only

5% of the instances would yield a result as extreme or more extreme than the one observed.

The process described above is illustrated in Figure 3.1. In the upper right corner an example of resulting statistical map can be visualised. This enables the visual examination of significant brain regions, which can also be displayed over an anatomical reference, as it is shown in the bottom right corner. Figure 3.1 also includes a preprocessing step, which, although not part of the method itself, has become a widespread practice covered by the software. This includes procedures such as realignment, segmentation, normalisation or smoothing, among others.

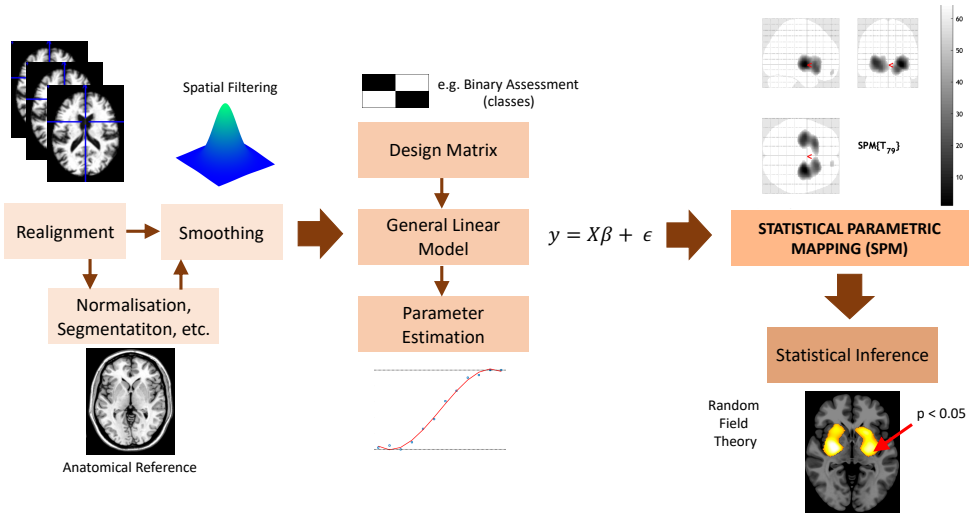


Figure 3.1: Summary of the process performed by SPM. First, several preprocessing procedures are applied to the data. Then, GLM is estimated given a design matrix. Finally, statistical maps, derived from SPECT scans, are obtained, which reflect the most significant regions or voxels.

Therefore, the current version of the software Statistical Parametric Mapping v.12 (SPM12) is widely used in neuroscience and clinical research to investigate patterns of brain activation, identify areas related to specific cognitive functions and explore differences in brain activity between study groups, such as case-control studies.

3.4.3 The Multiple Comparisons Problem

As it has been already commented, the SPM analysis presents certain challenges, especially when voxel-wise inference is applied. As each voxel is independently tested for statistical significance (and an image can contain millions of voxels), the likelihood of FP is high even at a reduced significance level. For example, when applying a statistical test to an image with 100000 voxels at a significance level of 0.05, it could potentially result in 5000 FP. Therefore, to effectively control the FP rate, it is necessary to consider the multiple comparisons problem and apply measures of FP risk, such as

Family Wise Error (FWE) rate or False Discovery Rate (FDR).

3.4.3.1 Family-Wise Error Rate

In imaging analysis, when the statistic map is obtained, it can be defined as an image of test statistics $T = \{T_i\}$, where T_i is the test statistic related to the i-th voxel of the original image of V voxels. In this scenario, the null hypothesis at each voxel, $H_{0,i}$, states that there is no effect in such voxel i. To test whether $H_{0,i}$ is rejected given a significance level α and a significant threshold u, the p-value is assessed as follows:

$$P\{T_i \geq u | H_{0,i}\} \leq \alpha \tag{3.9}$$

Let define H_0 as the non-existence of effect in any voxel, which can be named as *omnibus hypothesis* [86]. Therefore, the question is moving from the analysis of *independent* voxels to a *family* of voxels (or *volume* of values), assuming an FP risk known as the FWE rate [87]. It refers to the probability of making at least one false positive (Type I error) [88]. Let then H_0 be denoted as the family-wise null hypothesis. The family-wise null hypothesis testing examines whether there is any significant effect present among the group of related statistical tests, all while ensuring that the overall risk of making FPs is controlled. Thus, if any p-value related to the test statistics, $P = \{P_i\}$, satisfies that $p_i \leq \alpha$, H_0 is rejected.

Therefore, it is necessary to adjust for each T_i a significance threshold u to control the overall probability of obtaining FPs. This ensures that significant results are more reliable and that the differences found between conditions or groups are more robust and less likely to be a result of random variability. Various multiple comparison correction techniques (or statistical thresholds) could be applied to keep the FWE rate under control, including the Bonferroni method and Random Field Theory (RFT).

Bonferroni Correction The Bonferroni method is a simple and conservative approach to control the FWE rate. The method is derived from the Bonferroni inequality, which involves a truncation of Boole's formula [89]. It adjusts the significance level of each T_i by dividing it by the total number of tests conducted, e.g. the number of voxels, V. Then, Equation (3.9) would be modified as:

$$P\{T_i \geq u | H_{0,i}\} \leq \frac{\alpha}{V} \tag{3.10}$$

This comes from considering that all test statistics are drawn from the null distribution and whose p-values have a probability α of being greater than the set threshold. Therefore, as FWE is the probability of at least obtained a p-value greater than α, and α is small, the following equation can be derived:

$$\text{FWE} = 1 - (1 - \alpha)^V \leq V\alpha \tag{3.11}$$

On this basis, given a constrained (controlled) value for FWE, the threshold associated with each p-value must be:

$$\alpha = \frac{\text{FWE}}{V} \tag{3.12}$$

For example, in an image of 100000 voxels (and therefore 100000 test statistics), if a value of FWE below 0.05 is imposed as a constraint, the threshold u to compare with each T_i is the one related to $\alpha = 0.05/100000 = 5 \times 10^{-7}$. This approach reduces the likelihood of obtaining FPs, but it may lead to a decrease in statistical power. The Bonferroni correction is considered conservative because it treats all voxels as independent tests, disregarding any spatial smoothness that might exist.

Random Field Theory Another FWE correction method that avoids the latter limitation is RFT, which is the one applied in SPM [85]. RFT is a specific technique for neuroimaging analysis that takes into account the spatial structure of brain images and controls the FWE rate by considering the relationship between intensity values in neighbouring voxels (correlation). Therefore, in this approach instead of focusing on each voxel, *resels* (resolution elementents) [90] are analysed, which are blocks of voxels of the same size of the smoothing kernel. The number of resels in the image depends on the smoothness and its volume.

In order to determine the threshold for a smooth statistical map that meets the desired FWE rate, this method relies on the Euler Characteristic (EC), which can be defined as the number of clusters or blobs that will be above a certain threshold [87]. Then, the expected EC, denoted as $\mathbb{E}[\text{EC}]$, can be thought of as the probability of finding a cluster above the specified threshold in the statistical map, i.e. FWE $\approx \mathbb{E}[\text{EC}]$ (only when high thresholds are applied). A detailed mathematical development is available in [88]. Thus, using the appropiate RFT equation and knowing the number of resels in the image, $\mathbb{E}[\text{EC}]$ can be calculated for any given threshold. Moreover, the threshold u could be estimated given a specific value for $\mathbb{E}[\text{EC}]$, e.g. $\mathbb{E}[\text{EC}] = 0.05$.

3.4.3.2 False Discovery Rate

The FDR is a more tolerant measure of FP risk than FWE. It can be defined as the expected proportion of FP among the voxels identified as significant [91], which can be denoted as $\mathbb{E}[V_{FP}/V_s]$. In other words, FDR allows for a certain level of false positives while identifying relevant finding, which is useful when seeking a balance between discovery of significant findings and control of error.

Several variants of FDR procedures has been proposed [88], although the most commonly applied it the one described by Benjamini and Hochberg [91]. This procedure involves ranking the p-values obtained from the V statistical tests in ascending order as follows:

$$p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(i)} \leq \ldots \leq p_{(V)} \qquad \forall i = 1 \ldots V \tag{3.13}$$

Once this is done, a critical threshold, denoted as q, is selected, which represents the desired level of the FDR, tipically $\mathbb{E}[V_{FP}/V_s] = 0.05$. Let k be the largest i for which:

$$p_{(i)} \leq \frac{i}{V} q \tag{3.14}$$

then, all the null hypotheses corresponding to p-values $p_{(1)}, p_{(2)}, ..., p_{(k)}$ are rejected, while the remaining null hypotheses with p-values $p_{(i+1)}, p_{(i+2)}, ..., p_{(V)}$ are not rejected. Therefore, $p_{(k)}$ is the significance level, α, that ensures the FDR is controlled, and $T_{(i)}$ is the corresponding critical threshold u.

3.5 Permutation test

Unlike parametric testing methods seen in the previous sections that rely on specific assumptions about the data distribution, a permutation test [92, 93] is a non-parametric approach that does not assume any particular distribution and can be applied to obtained p-values. However, permutation methods require the assumption of exchangeability (except in randomised experiments), which means that samples can be permuted without altering the joint probability distribution [94].

The goal of a permutation test is to assess whether there is a significant difference between two groups or conditions using empirical evidence from the observed data [95]. It compares the observed differences in a statistic of interest (such as mean, median, or proportion) with differences that would occur by chance under the null hypothesis that there is no real difference between the groups (see Equation 3.1). To do this, the sample is resampled without replacement, for example, performing label permutation [96], and the statistic of interest is calculated for each permutation. Then, the observed statistic (in the non permutated sample) is compared to the distribution of statistics generated from the random permutations (the null distribution) to determine if it is statistically significant. To generate the null distribution, ideally more than 1000 permutations must be performed, and typically thousands of them are conducted, which is computationally expensive.

In the specific scenario of statistical maps, Holmes et al. [97] proposed the implementation of this method to address the multiple comparisons problem. In this proposed procedure, the null distribution is generated by the maximal voxel statistic across the volume under analysis. This means that in each permutation only the maximum statistic value obtained across all voxels is retained. Subsequently, the observed test statistic in the original image is compared with this null distribution, and *corrected* p-values are determined based on the fraction of permutations in which the null distribution is greater or equal than the original test statistic. Where p-value is below the significance level α, the null hypothesis is rejected. In terms of statistical power, this method demonstrates performance similar to other methods, such as RFT [86].

The permutation test is particularly useful when assumptions of normality or

equal variances are not met, or when the data is of ordinal or categorical nature. Due to its non-parametric approach, the permutation test provides a flexible and robust alternative for hypothesis testing in various statistical situations.

4 | MACHINE LEARNING IN NEUROIMAGING

In neuroimaging studies, the conventional perspective has treated each voxel or feature individually, as mentioned in the previous chapter. However, with the growing implementation of AI, an approach has emerged known as Multivariate Pattern Analysis (MVPA). MVPA involves analysing patterns of activity or interactions across multiple brain regions, voxels, or features to decode or classify information, rather than examining individual responses [98, 99]. This approach often applies ML techniques, a branch of AI that provides systems with predictive analytics capabilities.

When implementing a CAD system, ML techniques can be categorised into four major blocks: preprocessing, feature selection, feature extraction, and classification. This chapter provides a comprehensive description of each block, including the validation methods and metrics used to evaluate system performance. Furthermore, this chapter introduces the XAI techniques that will be applied in Part II.

4.1 Data Preprocessing

In order to enhance the quality and usefulness of available data, CAD systems typically incorporate a data preprocessing stage. This section provides descriptions of some of the most common preprocessing techniques applied.

Features normalisation Feature normalisation, also known as standarisation, adjusts the features of the dataset to a specific scale in order to prevent one feature from dominating others due to their absolute values. The most popular method is Z-score normalisation [100], which subtracts the mean of the feature and divides it by its standard deviation:

$$\hat{x} = \frac{x - \mu_x}{\sigma_x} \tag{4.1}$$

Missing values A typical problem in neuroimaging is the incomplete availability of all features for all samples in a database, resulting in missing values. This affects the performance of the model and requires the use of methods such as removing rows or columns with missing values, replacing them with means or medians, or using more sophisticated techniques [101, 102].

Resampling of unbalanced data Another common scenario is when the database contains different classes that are imbalanced, which reduces the reliability of the model as it fails to learn from the different classes under analysis in the same proportion. In such cases, resampling techniques [103] such as oversampling (increasing the sample of the minority class) or undersampling (reducing the sample of the majority class) can be applied to balance the classes and improve the model's performance. However, it is important to consider in the first method that the generated data is synthetic, reducing its clinical applicability.

One-hot encoding Sometimes, the database may contain variables (features) that are not numerical but categorical. To handle this situation and process all the variables together, it is common to represent them as binary vectors [104]. This transformation is typically performed when applying ML algorithms that require numerical features.

4.2 Feature Selection

The problem of the curse of dimensionality in neuroimaging has already been mentioned. To reduce the feature-to-sample ratio (d/n), it is common to apply feature selection and extraction techniques to enhance classification performance while preserving system complexity. In this section, some of the methods associated with feature selection are described. These methods allow for the removal of irrelevant features from the sample, which can also facilitate interpretation. They can be broadly categorised into three primary methods: filtering, wrappers, and embedded techniques [105].

Filtering approach This approach is based on evaluating features independently using a ML model. Statistical or information measures, such as correlation, information gain, or chi-square test, are used to assign a score to each feature. Then, features with the highest scores are selected. A common example is selecting the top-k features using variance or correlation.

Wrapper approach In this approach, a specific ML model is used to evaluate different combinations of features and select the optimal subset that maximises the model's performance. This process is more computationally expensive than the previous one but can lead to better feature selection. Popular examples include Recursive Feature Elimination [106] and Forward/Backward Stepwise Selection [107].

Embedded approach These methods incorporate feature selection directly into the model training process. Unlike filter and wrapper approaches, where feature selection is performed independently of the model, embedded approaches adjust the ML model in such a way that it automatically selects relevant features during training. These embedded approaches often use algorithms that have built-in feature selection capabilities. During training, these algorithms automatically evaluate and weigh the features based on their contribution to learning patterns in the data. Popular examples of embedded approaches include Regularized Linear Regression, such as LASSO [108], and Random Forests [109].

4.3 Feature Extraction

These additional methods related to dimensionality reduction aim to identify meaningful multivariate feature sets and transform the high-dimensional space into a lower-dimensional representation while preserving the important aspects of the original data. This process helps in eliminating noise and redundant information, leading to improved processing and enhanced performance of ML algorithms. Three widely used

methods, namely Principal Component Analysis (PCA), Partial Least Squares (PLS), and Autoencoder (AE) are described in this section.

4.3.1 Principal Component Analysis

Principal Component Analysis [110, 111] transforms a high-dimensional dataset into a lower-dimensional representation by identifying the principal components, which are orthogonal linear combinations of the original features that capture the maximum variance in the data. The first principal component explains the largest possible variance, followed by the second principal component, and so on. Each subsequent principal component is uncorrelated with the previous ones.

Given a matrix of features, $\mathbf{X}_{n\times p}$, where n is the number of samples and p the number of features (with mean value normalised to zero), PCA is performed on the basis of the eigenvectors, w extracted from the covariance matrix $Cov(\mathbf{X})$, which can be seen as a $W_{p\times p}$ matrix or p-dimensional vectors of weights. Thus, $\mathbf{X}_{n\times p}$ can be mapped to a new matrix of principal components of a reduced dimension l ($l < p$) as:

$$\mathbf{T}_{n\times l} = \mathbf{X}_{n\times p}\mathbf{W}_{p\times l} \tag{4.2}$$

where the first l eigenvectors are used to reduce dimensionality while preserving as much variance as possible.

4.3.2 Partial Least Squares

Partial Least Squares [112] is a supervised method which allows dimensionality reduction while retaining the patterns for higher separability of the classes. Given a matrix of features, $\mathbf{X}_{n\times m}$, where n is the number of samples and m the number of features, and a vector of labels $\mathbf{Y}_{n\times 1}$, PLS generates a matrix of loadings $\mathbf{X}_l$, which is related to the initial data by the following linear combination:

$$\mathbf{X} = \mathbf{X}_s\mathbf{X}_l^T + \mathbf{E} \tag{4.3}$$

where $\mathbf{X}_s$ is the score matrix and $\mathbf{E}$ the assumed error matrix. The reduced d-dimensional space desired comes from the dimensions of $\mathbf{X}_l$ ($m \times d$), as $m > d$. This new reduced space contains the original information of $\mathbf{X}$.

4.3.3 Autoencoders

An Autoencoder is a type of NN commonly used for unsupervised learning and dimensionality reduction [113, 114]. It consists of two main components: an encoder $e(x)$ and a decoder $d(x)$. The encoder reduces the dimensionality of the input data to a lower-dimensional representation known as the latent space or Z-layer, while the decoder reconstructs the original dimensionality from the low-dimensional data. In

other words, given a sample set with a dimension of M where $\mathbf{x_i} \in \mathbb{R}^M$, the encoder $e(\mathbf{x})$ provides a representation of these samples with a dimension of Z where $\mathbf{z_i} \in \mathbb{R}^Z$. The presence of this Z-layer makes autoencoders a powerful configuration for feature extraction, as they allow for associating input features with reduced-dimensional space features without significant information loss [28].

Due to the nature of the AE, its goal is to minimise the reconstruction error, which is the difference between the encoder input $\mathbf{x}$ and the decoder output $\hat{\mathbf{x}} = d(e(\mathbf{x}))$. For the computation of this error, Mean Squared Error (MSE) algorithm is usually applied:

$$MSE = \frac{1}{N} \sum_i (\mathbf{x_i} - d(e(\mathbf{x_i})))^2 \tag{4.4}$$

4.4 Classification methods

Once features are processed, they are fed into classifiers to detect patterns or differences between the conditions under analysis. This section introduces the classifiers employed in this thesis, including three based on traditional ML algorithms and two others using DL architectures. It is worth noting that while these classifiers are highlighted, there exists a wide range of classifier options available [104].

4.4.1 K-nearest Neighbors

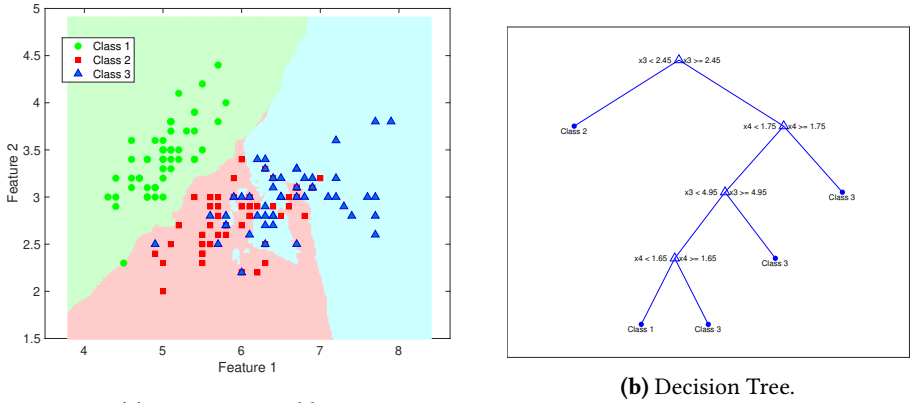
K-Nearest Neighbours (KNN) is a non-parametric algorithm that classifies or predicts a data point (sample) by considering its K nearest neighbors in the training set. The K represents the number of neighbors to consider. For classification, the majority class among the K neighbors is assigned to the sample under analysis, see Figure 4.1a. For regression, the average or weighted average of the K neighbors' values is used as the prediction.

The algorithm works based on the assumption that similar data points tend to have similar labels or values. To determine the nearest neighbors, the algorithm calculates the distance between the new data point and all the training data points using measures like Euclidean distance. The K closest neighbors are then selected.

KNN is widely used due to its simplicity and interpretability. It can be effective in situations where the decision boundaries are complex or where there is no explicit underlying model. However, it may not perform well with high-dimensional or sparse data.

4.4.2 Decision Trees

Decision trees are models that make decisions or predictions by following a tree-like structure. Each node represents a feature, and the branches represent the possible



(a) K-nearest Neighbors.

(b) Decision Tree.

Figure 4.1: Examples of KNN and Decision Tree algorithms for a multiclass problem.

values of that feature. The leaves of the tree represent the final decisions or outcomes. This scheme can be seen in Figure 4.1b. Decision trees have advantages such as interpretability and the ability to handle various types of features. They can capture nonlinear relationships and are robust to outliers. However, they can be prone to overfitting.

There are different algorithms for building decision trees, such as ID3, C4.5, and CART [115, 116]. These algorithms use measures like information gain, Gini index, or MSE to determine the best features and splits. Pruning techniques can be applied to avoid overfitting.

4.4.3 Support Vector Machine

Support Vector Machine (SVM) is a classification algorithm that estimates the maximum margin hyperplane to separate the existing classes in a dataset. There are several types of kernels that can be applied in this algorithm [117]. The choice of kernel plays a crucial role in shaping the hyperplane and ultimately affecting the classifier's performance. In this thesis, the linear kernel has been applied [118]. In a linear binary problem, this hyperplane could be described as the sets of points $\mathbf{x}$ that meet:

$$\mathbf{w}^T\mathbf{x} - b = 0 \tag{4.5}$$

where $\mathbf{w}$ represents the normal vector to the hyperplane and b represents the error. There are two parallel hyperplanes associated with the main one to maintain the largest possible distance between the two classes:

$$\begin{aligned} \mathbf{w}^T\mathbf{x} - b &= 1 \\ \mathbf{w}^T\mathbf{x} - b &= -1 \end{aligned} \tag{4.6}$$

Thus, elements above the first hyperplane are considered to be of one class, and those below the second hyperplane are considered to be of the other class. The space distribution of this classifier can be seen in Figure 4.2. In neuroimaging, the use of SVM as a classification algorithm is widely adopted when a small sample set is involved [35, 119].

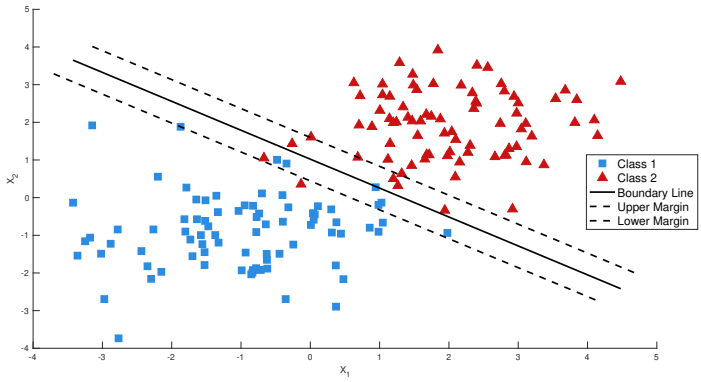


Figure 4.2: Example of a SVM classifier with a linear kernel on a binary problem.

4.4.4 MultiLayer Perceptron

The Multilayer Perceptron (MLP) is a feedforward artificial NN composed of fully-connected layers [120]. Fully-connected layers are those layers with i perceptrons each one, where the connections go in forward direction and there are no connections within a layer. The former is the cause of one of the disadvantages associated with these layers, since as all perceptrons are connected with all those in the next layer, the number of parameters increases exponentially, which is inefficient. The equation associated with each perception is:

$$y_i^n = f\left(w_i^n \cdot \mathbf{y}^{n-1} + b_i^n\right) \tag{4.7}$$

where $f(\cdot)$ is the activation function applied to the i-th perceptron of the layer n, w_i^n is the weight vector that multiplies the activations of the previous layer ($\mathbf{y}^{n-1}$) and b_i^n is the associated bias.

4.4.5 Convolutional Neural Network

The Convolutional Neural Network (CNN) [120] is an architecture that has become the standard one in image processing. The application of CNN to neuroimaging has revolutionised the field, addressing problems in a more efficient way, such as in brain's tumor detection [121] or in the identification of patterns associated with Autism [122]. CNNs are usually formed by several layers, from the first related to the extraction of

informative patterns to the last ones whose purpose is perform classification. This architecture can be used individually or as a part of a more complex network, such as U-net [123], DenseNet-121 [124] or Mobilenetv2 [125].

In contrast to the MLP, the CNN is composed of convolutional layers in addition to linear layers. Convolutional layers have a higher complexity, allowing the application of other different algorithms (for example, pooling or transpose convolutions) to images, which results in outstanding results especially in image analysis and processing. Similar to equation (4.7), an equation for convolutional neurons can be posed:

$$y_i^n = f\left(w^n \cdot y_i^{n-1} + b_i^n\right) \tag{4.8}$$

where only two differences exist, the w^n term indicates that the weight matrix is shared by all the neurons in layer n, which allow the convolution, and y_i^{n-1} means that for the neuron i of layer n only a part of the outputs of the previous layer is used, known as kernel size.

4.5 Validation procedure

The validation process is used to assess the performance and generalisation ability of a ML model. The main objective of validation is to determine how well the model performs on new and unseen data that was not used during training. It helps identify overfitting issues, where the model becomes overly tuned to the training data and does not generalise well to new data.

To assess such performance, the commonly used measure is the accuracy (or its associated error). Given a dataset S and a learning algorithm L, it is possible to calculate the estimated error of the algorithm, $\hat{E}_L$, i.e. the fitted classifier's error. Two types of errors associated with classification must be defined, (1) the empirical error, E_{emp}, which is the one associated with the training set, and (2) the actual error, E_{act}, the one obtained with samples not used for training (test set).

4.5.1 Cross-Validation

Several validation methods allow the estimation of the actual error. The most popular one is K-fold cross-validation (CV) [126]. The estimation of the actual error obtained from each of the K partitions, or folds, is defined as:

$$\hat{E}_L^{KCV} = \frac{1}{card\left(S^k\right)} \sum_{i \in S^k} I\{L_{S^{(k)}}(x_i) \neq y_i\} \tag{4.9}$$

where L_S is the function obtained by the learning algorithm L given the dataset S, $I(\cdot)$ is the indicator function, the k-th partition S^k is the test set and the remainder of the data $S^{(k)}$, the training set. That is, the actual error obtained is the average of the different errors obtained with the K folds when used as test set.

Typically the most preferred configuration is $K = 10$ because of the trade-off between variance and bias [126]. Other well-known configuration is K=N, where N is the number of samples in the dataset, which is named Leave-One-Out (LOO) [127]. The latter, although more unbiased, generates a large variance as all samples are considered one by one, requiring a higher computational burden.

In neuroimaging, CV encounters a limitation due to the small sample size problem. When dividing the dataset into folds, the resulting groups are reduced, leading to increased variability compared to scenarios with larger sample sizes. This limited sample size can impact the reliability and generalisability of the results.

4.5.2 Resubstitution with upper bound correction

Another well-known validation method is resubstitution. It tends to be discarded because it is applied in the training set, i.e. it generates an empirical rather than a real error, resulting in an over-optimistic nature [128]. Nevertheless, it can be considered as optimum in scenarios with low sample sizes [129]. This empirical error could be described as:

$$\hat{E}_L^{resub} = \frac{1}{n} \sum_{i=1}^{n} I\{L_S(x_i) \neq y_i\} \tag{4.10}$$

Furthermore, there are proposals that allow estimating the actual error by applying resubstitution. Once the empirical error of the classifier is calculated, the actual error under the worst case with probability $1 - \eta$ can be estimated by means of an upper bound [130, 131]. The upper bound can be seen as the difference between empirical and actual errors given a fitted learning algorithm, $\mu \geq |E_{act}(L_S(x)) - E_{emp}(L_S(x))|$. In this thesis, this is denominated as resubstitution with upper bound correction (RUB) [36]. Thus, the actual error is defined as:

$$\hat{E}_L^{RUB} = \hat{E}_L^{resub} + \mu \tag{4.11}$$

where μ represents the upper bound. This upper bound could be seen as a theoretical classification limit, which allows the use of all accessible data to establish the metrics of interest. Besides, accuracy, sensitivity and specificity can be limited by this value considering that its associated errors are partial errors of the classification one.

Different upper bounds are described in the literature. The most well-known is based on the Vapnik-Chervonenkis (VC) dimension as proposed by V. Vapnik [132, 133], which is defined as:

$$\mu_{VC} \leq \sqrt{\frac{h\left(\ln\left(\frac{2n}{h}\right) + 1\right) - \ln\left(\frac{\eta}{4}\right)}{n}} \tag{4.12}$$

where η is the significance level, n is the size of the training set, d is the features dimension and h is the VC dimension [130]. This dimension is equal to $d + 1$ for linear functions, such as SVM, while for DL algorithms a lively debate exists about their value [134].

Other model-free upper bound of the actual risk is presented in [31], which is based on the assessment of concentration inequalities (training set distributed in general position, i.g.p, in $\mathbb{R}^d$) and it is only applicable to linear classifiers, e.g. SVM with linear kernel. Its expression is:

$$\mu_{i.g.p} \leq \sqrt{\frac{1}{2n} \ln \frac{2\sum_{k=0}^{d-1} \binom{n-1}{k}}{\eta}} \tag{4.13}$$

where n is the size of the training set, and d is the feature's dimension. Compared to Vapnik's bound (Equation (4.12)), this bound is less restrictive in linear scenarios, as shown in Figure 4.3.

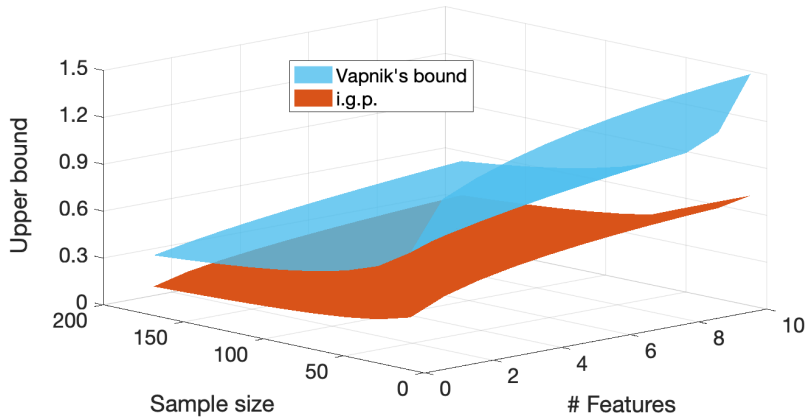


Figure 4.3: Values of the upper limit as a function of the sample size and the number of features under analysis. The blue surface represents Vapnik's upper bound for linear algorithms, while the red surface represents the i.g.p. bound.

The implementation of Probably Approximately Correct (PAC)-Bayesian bounds is another interesting proposal. In thesis, a dropout bound [135] is also applied. This bound considers a dropout rate, $\alpha \in [0, 1]$, which reduces the complexity cost of the function. The effect of this dropout is stronger the closer its value is to 1. The expression of this bound in the scenario proposed in this work is:

$$\mu_{PAC-bayes} = \min_{1 \leq i \leq k} \left(\frac{1}{1 - \frac{1}{2\lambda_i}} - 1 \right) \hat{E} + \frac{1}{1 - \frac{1}{2\lambda_i}} \left(\frac{\lambda_i E_{max}}{n} \left(\frac{1-\alpha}{2} \|\Theta\|^2 + \ln \frac{k}{\eta} \right) \right) \tag{4.14}$$

where k different values of the parameter λ, which is set to $1/2 \leq \lambda \leq 10$, are evaluated to minimise the bound. The estimated value of the loss function to be bounded, i.e., the error of the classifier, is $\hat{E}$. Its maximum value, E_{max}, which must be a real number, is 1 in this case. Finally, Θ is the classifier's parameter set.

Furthermore, the analysed trade-off between empirical error and actual error in RUB can be extrapolated in a very similar way to CV. According to the theory related to generalisation errors, irrespective of the validation method, it must be satisfied that $E_{act} = \mu + E_{emp}$ in the worst-case. Ideally, such an upper bound would be zero, as this means that both errors are equal, and therefore, the classifier is able to generalise. In mathematical terms, this can be easily observed in the following expression:

$$\frac{\mu}{E_{emp}} = \frac{E_{act}}{E_{emp}} - 1 \tag{4.15}$$

where a positive μ/E_{emp} implies a poor generalisation ($E_{act} > E_{emp}$) since $\frac{E_{act}}{E_{emp}} \propto \frac{\mu}{E_{emp}}$, whereas if the ratio becomes negative, it would be even better than the ideal situation, as that means that the actual error is lower than the empirical error. Nevertheless, in order to be able to analyse this pattern, the bound should be constant, because if it is not, there is no effective learning during training as the two errors would not be related to each other.

In summary, the advantages of resubstitution over CV, such as lower computational cost and reduced variability [31, 136], are extended by RUB eliminating the inherent bias related to resubstitution. This is because, based on the conditions of a given classification scenario, it is able to set a limit to the classification capability of the model [31, 132]. Furthermore, it has been found that the relationship established between E_{emp} and E_{act} is not only related to resubstitution, but is also valid for CV. Indeed, while in a resubstitution scenario the upper bound is constant, during CV this need not be the case, leading to scenarios of low effective learning capacity. All this makes RUB is an optimal option due to the correction applied, especially for small sample sizes (common in neuroimaging), as it has been already tested [9, 35, 37].

4.6 Performance Evaluation Metrics

Once the validation process is completed, a performance evaluation is computed. Performance of the classifiers is evaluated through metrics extracted from the confusion matrix, which provides an overview of the classifier's outcomes (Figure 4.1). In neuroimaging, the typical convention is that the positive condition refers to the condition that is the focus of study or interest, while the negative condition serves as the control or reference point for comparison.

		Predicted Condition	
		Positive (P)	Negative (N)
Actual Condition	P	True Positive (TP)	False Negative (FN)
	N	False Positive (FP)	True Negative (TN)

Table 4.1: Confusion matrix in a binary problem. Total Population: P+N.

The performance metrics employed for evaluating the results in this thesis include accuracy (and balanced accuracy), specificity, sensitivity and F1-score. Their equations are:

$$\begin{aligned} Acc &= \frac{TP + TN}{P + N} \\ Bal\ Acc &= \frac{1}{2}\left(\frac{TP}{P} + \frac{TN}{N}\right) \\ Spec &= \frac{TN}{TN + FP} \\ Sens &= \frac{TP}{TP + FN} \\ F1\text{-}score &= \frac{2TP}{2TP + FP + FN} \end{aligned} \tag{4.16}$$

The Receiver Operating Characteristics (ROC) curve is also an interesting metric of performance. It is a graphical representation used to assess the performance of a binary classification model. It displays the trade-off between the TP rate (sensitivity) and the FP rate (1-specificity) for different classification thresholds. The closer the curve is to the top-left corner of the plot, the better the model's performance. The area under the ROC curve (AUC) is a commonly used metric to quantify the overall predictive performance of the model, with values ranging from 0.5 (random performance) to 1 (perfect performance) [137, 138]. An example of this metric can be seen in Figure 4.4.

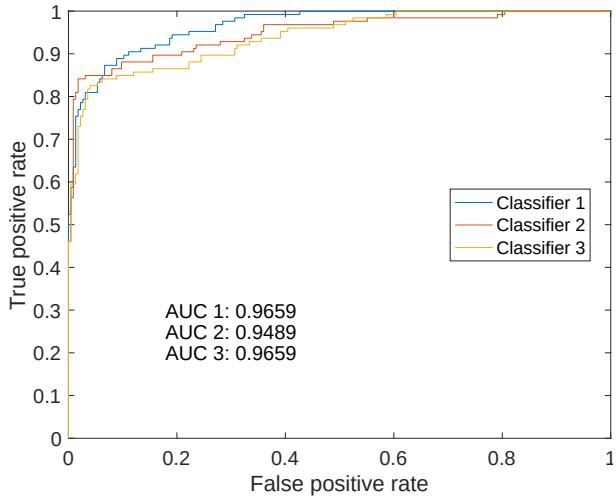


Figure 4.4: Example of a ROC curve and its AUC value for three different classifiers.

4.7 Explainable Artificial Intelligence

Finally, this section included the techniques that aim to enhance the transparency and interpretability of ML models. These algorithms give a qualitative understanding

of performance, making them more understandable to humans. This emerging field is referred to as XAI (already mentioned). Here, only the techniques applied in this thesis are described. First, algorithms that assess feature relevance in classifiers are included, such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). Then, techniques specifically designed to identify areas of interest in images are described, such as Saliency Maps and Guided Gradient Class Activation Map (Grad-CAM).

4.7.1 Local Interpretable Model-agnostic Explanations (LIME)

LIME [139] focuses on providing explanations of individual predictions of the classifier model. To do so, it makes a local approximation to an easily interpretable model. Given the type of data used in this study, LIME highlights the most relevant, both positively and negatively, sulcal features during classification. In other words, LIME shows if a high value of a feature brings the sample closer to a class (acts positively) or reduces the likelihood of the sample belonging to that class (acts negatively).

This algorithm is able to explain any prediction model f locally. This means that LIME provides explanations for a particular sample x, since globally faithful explanations are still a challenge for complex models [139]. To do this, the algorithm selects an explanation model $g \in G$, where G is a class of potentially interpretable models. The selection is made according the following objective function related to the faithfulness of the explanation model:

$$\xi = \operatorname{argmin} \ \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{4.17}$$

where interpretability and local fidelity is ensured by minimising the trade-off between the loss related to the discrepancy between g and f given the local kernel π_x, and the complexity of g, measured by $\Omega(g)$.

4.7.2 SHapley Additive exPlanations (SHAP)

SHAP [140] is a model-agnostic algorithm which can explain any classification model. SHAP assigns the relevance of each feature by means of Shapley values, a concept from game theory [141].

Given the set of features S, the contribution of each feature s is estimated on the basis of its average marginal contribution to all subsets of features $T \subseteq S$, which do or do not include the feature s. Let the prediction of the model given a particular sample and a subset of features be denoted as $f_x(T)$. The marginal contribution of the feature s is estimated as the difference in predictions when applying or not applying such a feature, $[f_x(T \cup s) - f_x(T)]$. So, the Shapley value, ϕ_s, is computed considering all possible subsets $T \subseteq S \setminus \{s\}$:

$$\phi_s(f, x) = \sum_{T \subseteq S \setminus \{s\}} \frac{|T|!\,(|S| - |T| - 1)!}{|S|!} \left[f_x(T \cup s) - f_x(T) \right] \tag{4.18}$$

SHAP values are the solution to Equation (4.18); i.e., they are Shapley values of a conditional expectation function of the original model which satisfy properties such as local accuracy, missingness and consistency [140]. Several approximation methods to compute SHAP values are proposed, since its exact computation is difficult to achieve. The one applied is this work was Kernel SHAP, a model-agnostic approximation which combines Shapley values and linear LIME (local linear regression) to estimate the importance of each feature. To do this, the solution of Equation (4.17) are Shapley values; i.e., local accuracy, missingness and consistency must be satisfied. Inherently, LIME does not meet all these properties by choosing its parameters heuristically.

4.7.3 Saliency Map

Saliency map is one of the oldest and more common interpretation method [142, 143]. A saliency map represents the parts of the image that contribute most to the network's decision. Given an image I and a class score function $S_c(I)$, which depends on the vector weights and bias of the model, a saliency map is computed by obtaining the derivative w calculated via backpropagation at a given point p [144]:

$$w = \left. \frac{\partial S_c}{\partial I} \right|_p \tag{4.19}$$

Then, the saliency map is finally obtained by rearranging the elements of w, i.e according to the pixels distribution in the final score.

4.7.4 Guided Gradient Class Activation Map (Grad-CAM)

Grad-CAM is another of the most commonly used methods of visual interpretation. It was originally designed as an improvement of the CAM algorithm [145] and it can be applied to networks that include fully-connected layers. Once the class c under analysis is selected, Grad-CAM computes the gradient of the score for c, y^c, according to the activations maps of the final convolutional layer. Then, gradients flowing back are global-average-pooled to obtain the neuron importance weights α_k^c [146]:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{4.20}$$

where A_{ij}^k represents the activation map k in the convolution layer over the indexes i and j related to width and height, respectively. The first part of the equation represents the global averaged pooling. Once the importance weights are computed, they are

multiplied by its associated activation map and all are summed. Finally, the final heatmap is obtained after applying the Rectified Lineal Unit (ReLU) nonlinearity.

$$\text{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right) \tag{4.21}$$

5 | DATASETS

A total of seven datasets covering a range of modalities and conditions have been used in the development of this thesis. Table 5.1 provides a concise summary of these datasets, followed by a comprehensive description of each one.

Acronym	Entity	Modality	Disorder	Other Information
ADNI-AD	ADNI	MRI	AD	
KAGGLE-AD	Kaggle	MRI	AD	MRI, demographical and clinical features
DIAN-AD	DIAN	Several	AD	Imaging and non-imaging biomarkers
CDT-AD	CIEN, FIDYAN	Drawings	CI	
PPMI-PD	PPMI	SPECT, MRI	PD	^{123}I-FP-CIT SPECT scans
UGR-COG	CIMCYC	EEG		Healthy participants
SHG-SCZ	SMHC	MRI	SCZ	Sulcal measurements extracted

Table 5.1: Overview of the datasets used in this thesis.

5.1 Alzheimer's Disease-related datasets

5.1.1 ADNI-AD, the Alzheimer's Disease Neuroimaging Initiative

The Alzheimer Disease Neuroimaging Initiative (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 MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. The data collected by ADNI is made available to the scientific community and has been used for numerous studies, contributing to the advancement of knowledge in the field of AD. Additionally, ADNI has established standards and assessment protocols that have been adopted by other studies and projects related to AD. For more information, please visit: adni.loni.usc.edu.

One of the datasets used in the preparation of this thesis was obtained from ADNI database. The dataset is composed of sMRI scans acquired at 1.5T from 229 Healthy Controls (HC) and 188 AD participants. Demographic information is shown in Table 5.2.

Group	N	Sex (M/F)	Age ($\mu \pm \sigma$)	MMSE ($\mu \pm \sigma$)
HC	229	119/110	75.97 ± 5.00	29.00 ± 1.00
AD	188	99/89	75.36 ± 7.50	23.28 ± 2.00

M: Male, F: Female

Table 5.2: Demographic details of the ADNI-MRI dataset.

All scans were processed using SPM12 software [83] by generating a processing pipeline that combine realignment, coregistration, spatially and intensity normalisation. In addition, SPM12 was used to segment the images obtaining GM and WM maps [147]. In total, 417 GM maps of dimensions $121 \times 145 \times 121$ were used in this work as features, normalised to the intensity range $[0, 1]$.

5.1.2 KAGGLE-AD, a Kaggle multiclass dataset

Kaggle is a virtual community comprising data scientists and practitioners specialised in ML which was first launched in 2010. Potential uses include finding and sharing datasets, developing and analysing models within a web-based data-science environment, collaborating with fellow data scientists and ML engineers, and participating in competitions to solve data science problems. In fact, the database to be described was provided for the International challenge for automated prediction of MCI from MRI data (https://inclass.kaggle.com/c/mci-prediction) [148].

The subjects in the dataset were categorised into four classes according to their diagnosis: HC subjects, AD patients, MCI subjects whose diagnosis did not change in the follow-up and converter MCI (cMCI) subjects that progressed from MCI to AD in the follow-up of the disease. MRI scans were selected from the ADNI and preprocessed by

Freesurfer (v5.3) [149, 150]. In total 429 demographical, clinical as well as cortical and subcortical MRI features were available for each subject.

The database was composed of two different sets, one for training and one for testing the proposed methods for automated prediction of MCI from MRI data. The training dataset consisted of 240 ADNI real subjects (60 HC, 60 MCI, 60 cMCI and 60 AD). The testing dataset comprised a total of 500 subjects, out of which 160 were real subjects, whereas the 340 remaining subjects were artificially generated from the real data. Demographic information is shown in Table 5.3 (training set and testing set). It should be noted that demographics of the test set only shows information of the 160 real patients excluding 340 dummy subjects. When both subsets are combined into a single set of 400 participants, the demographic information is as shown in Table 5.3 (real set).

	Group	N	Sex (M/F)	Age ($\mu \pm \sigma$)	MMSE ($\mu \pm \sigma$)
Training set	HC	60	30/30	72.34 ± 5.67	29.15 ± 1.11
	MCI	60	28/32	72.19 ± 7.42	28.32 ± 1.56
	cMCI	60	35/25	72.96 ± 7.20	27.18 ± 1.87
	AD	60	29/31	74.75 ± 7.31	23.43 ± 2.11
Test set	HC	40	18/22	74.88 ± 5.48	29.00 ± 1.10
	MCI	40	23/17	72.40 ± 8.04	27.65 ± 1.87
	cMCI	40	25/15	71.75 ± 6.23	27.58 ± 1.80
	AD	40	23/17	73.11 ± 8.05	22.68 ± 1.98
Real set	HC	100	48/52	73.47 ± 5.76	29.09 ± 1.11
	MCI	100	51/49	72.27 ± 7.71	28.05 ± 1.73
	cMCI	100	60/40	72.47 ± 6.89	27.34 ± 1.86
	AD	100	52/48	74.10 ± 7.70	23.13 ± 2.10

M: Male, F: Female

Table 5.3: Demographic details of the KAGGLE-MRI dataset.

5.1.3 DIAN-AD, the Dominantly Inherited Alzheimer Network

The Dominantly Inherited Alzheimer Network (DIAN) is an initiative launched in 2008 which is focused on studying dominantly inherited forms of AD [151]. It was established to better understand and accelerate the research on familial or autosomal dominant AD by monitoring the evolution of people at risk of such mutation. The primary goals of DIAN are to identify biomarkers for early detection, track disease progression, and facilitate the development of new therapeutic strategies for DIAD. DIAN has established a large network of participating research centers worldwide, collaborating to collect and analyse data from affected individuals and their family members. Each participant had standardised longitudinal assessments which includes clinical, cognitive, neuroimaging, CSF and plasma tests. For more information, please visit: dian.wustl.edu.

The images that compose this dataset were obtained from baseline assessments of

the DIAN Observational Study Data Freeze 14. Its baseline assessments were composed of a total of 1219 samples. The criteria for excluding subjects from the database for this thesis were as follows. Firstly, only data from the initial visit of the subjects was considered, which reduced the number of samples from 1219 to 534 samples. Then, 29 samples were excluded due to being diagnosed with at least one of the following diseases: cerebral stroke (3 samples), transient ischemic attack (2 subject), dementia by alcoholism (4 samples), PD (1 samples), traumatic brain injury with chronic deficit/dysfunction (3 samples), dementia with Lewy bodies (1 samples), vascular dementia (1 samples) and dementia by unknown causes (5 samples). Besides, in order not to increase the heterogeneity in symptomatic subjects, LOAD cases in the DIAN study were also discarded (15 samples).

Technical criteria were also applied, the criterion regarding missing values was to select those features that were fulfilled for at least 80% of the samples and to exclude the remaining features as well as incomplete subjects for the chosen features. Once this was done, the final set consisted of 333 samples with 722 features each of them. Such 722 features are categorised as follows: 188 MRI features and 520 Fluorodeoxyglucose (FDG) F18 PET features. The remaining 14 features were non-imaging biomarkers obtained from CSF and blood plasma, where protocols INNO, xMAP, PL_xMA and Lumipulse were used for measurements [152]. Specifically, fibrillar amyloid-β (Aβ) depositions and CSF τ protein were measured as markers: Aβ_{40}, Aβ_{42}, Aβ_{40}:Aβ_{42} ratio, τ and p-τ. An apolipoprotein E (APOE) genetic test was also undertaken.

The partition of subjects between non-carriers (NC) and mutation carriers (MC) sets was 123 subjects in the first group and 210 in the MC group. Due to the large difference in the number of subjects in both groups, they were balanced in 123 subjects in each group, reducing the MC set randomly. Thus, the total number of samples selected was 246. Demographic information is shown in Table 5.4, where global CDR scale data are provided [54].

Group	N	Sex (M/F)	Age ($\mu \pm \sigma$)	CDR ($\mu \pm \sigma$)
NC	123	49/74	38.13 ± 11.88	0.00 ± 0.00
MC	123	78/45	37.55 ± 9.93	0.26 ± 0.50

M: Male, F: Female

Table 5.4: Demographic details of the DIAN-PD dataset.

5.1.4 CDT-AD, a Clock Drawing Test dataset

This dataset is composed of drawings made to assess the CDT, which will be explained in detail in chapter 10. In short, it is a widely used paper-and-pencil test for cognitive assessment in which an individual has to manually draw a clock on a paper. The drawings were collected from volunteers in the Multidisciplinary Unit of CIEN Foundation (Madrid, Spain) and the Department of Neurology of FIDYAN Neurocenter

(Granada and Málaga, Spain).

The Centro de Investigación de Enfermedades Neurológicas – Centre for Research in Neurological Diseases (CIEN) was established in 2008 as a nonprofit institution affiliated with the Spanish Carlos III Institute of Health and associated with the Biomedical Research Networking Center for Neurodegenerative Diseases (CIBERNED) dedicated to the research and study of neurological diseases, with a particular focus on AD and other dementias. Its main objective is to conduct high-quality scientific research in the field of neurological diseases to contribute to the advancement of knowledge, diagnosis, and treatment of these conditions.

FIDYAN Neurocenter is a neurological clinic with more than 10 years of experience specialised in neurological pathology, both in adults and children. Its aim is to offer evaluation and intervention in pathologies of neurological origin.

All the individuals whose clinical information and CDT drawings were contained in the dataset had provided an informed consent to use their data in clinical research. The Vallecas Project of the CIEN Foundation, as the framework of this study, was approved by the Ethics Committee of the Carlos III Institute of Health. Cognitive status of every participant was diagnosed by consensus of a team of experienced neurologist and neuropsychologist, taking into account their age, functional status, clinical data and performance in an extensive neuropsychological battery. The criteria from the National Institute on Aging-Alzheimers Association (NIA-AA) [153], and from the fourth edition, text revised, of the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV-TR) [154] were used to diagnose MCI and dementia, respectively.

This dataset consists of 7009 CDT drawings; 5368 of them were drawn by individuals with normal cognition (HC), and 1641 by individuals with CI including MCI or dementia. The average age of the participants is 73.30 years, whereas 51.73% of them have superior education (high school or more). All the information regarding demographics is summarised in Table 5.5 (complete dataset), where statistical information is also included. As relevant demographic information is available in this dataset, statistical significance tests are applied.

Since the number of HC is much higher than the CI patients and the sample of cases is not recruited as a population-based cohort, a balanced version of the dataset where the condition has an a priori probability of 50% has been mainly used in this thesis. Thus, the number of drawings in this set was 3282 and Table 5.5 (balanced dataset) shows the demographic information.

Regarding the process of the clock drawings collection, participants were given an A4 size paper and a pencil and asked to draw a clock with the clock hands pointing to ten past eleven. Once the clock drawing was finished, physicians assigned a score to the resulting drawing from 0 to 7, according to standard rules [155], such as the shape of the clock face, the way the numbers are arranged or the position of the clock hands. The participant got the maximum score when a perfect clock was drawn, which usually means that the person does not suffer any relevant cognitive impairment. By

	CI	HC	Total	T/χ^2 statistic	p-value
Complete dataset					
N	1641	5368	7009		
Age ($\mu \pm \sigma$)	74.36 ± 8.21	72.98 ± 6.06	73.30 ± 6.65	7.38	< .0001a
Education (S/NS)	671/970	2955/2413	3626/3383	8.58	.003b
Sex (M/F)	709/932	2104/3264	2813/4196	8.41	.004b
Centre (CIEN/FIDYAN)	580/1061	4909/459	5489/1520	2329.40	< .0001b
Balanced dataset					
N	1641	1641	3282		
Age ($\mu \pm \sigma$)	74.36 ± 8.21	72.99 ± 5.51	73.68 ± 7.02	5.60	< .0001a
Education (S/NS)	671/970	914/727	1585/1697	72.05	< .0001b
Sex (M/F)	709/932	660/981	1369/1913	3.01	.08b
Centre (CIEN/FIDYAN)	580/1061	1510/131	2090/1192	1139.42	< .0001b

S: Superior education, NS: non-Superior education, M: Male, F: Female

The p-values were obtained using: two-sample t-test (a) or Chi-Square test (b)

Table 5.5: Demographic details of the CDT-AD dataset.

contrast, a score of 0 indicates that the subject is unable to draw the clock, and it is highly likely that he/she suffers a severe CI. Figure 5.1 illustrates the diversity of the drawings in the dataset. The associated scores from these drawings range from the lowest score (0, clock in the left) to the highest (7, clock in the right).

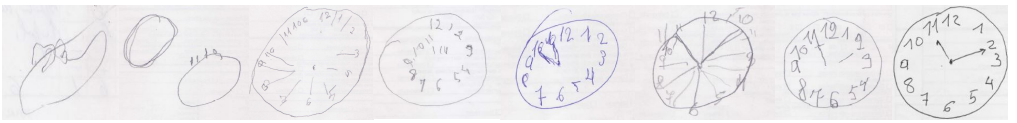


Figure 5.1: Examples of drawings made by the subjects of the CDT-AD dataset. From left to right, their associated scores range from the lowest (0) to the highest (7) possible score.

5.2 PPMI-PD, the Parkinson's Progression Markers Initiative

The Parkinson's Progression Markers Initiative (PPMI) is a public-private partnership funded by The Michael J. Fox Foundation for Parkinson's Research and funding patners, including Abbot, Biogen Idec or F. Hoffman-La Roche, among others. The full list of funding partners can be found on ppmi-info.org/about-ppmi/who-we-are/study-sponsors. It was launched in 2010 as a global research initiative aimed at identifying and validating biomarkers for measuring PD progression. Its goal is to improve the diagnosis, monitoring, and treatment of this neurodegenerative disease. PPMI collects clinical, imaging, and biological longitudinal data. For more information, please visit: ppmi-info.org.

For this dataset, information of 80 participants have been extracted from the PPMI database. It is a balanced dataset that includes a total of 40 HC subjects and 40 patients with PD of which both MRI and PET scans are available. In all cases, the time between MRI and ^{123}I-FP-CIT SPECT imaging acquisitions were no longer than 15 days. Demographics of the participants are depicted in Table 5.6.

Group	N	Gender (M/F)	Age ($\mu \pm \sigma$)
HC	40	28/12	62.22 ± 10.56
PD	40	23/17	61.22 ± 7.89

M: Male, F: Female

Table 5.6: Demographic details of the PPMI-PD dataset.

All scans have been spatially normalised (non-linear transformations) to a reference space defined by the Montreal Neurological Institute (MNI) using SPM12. For the MRI scans, each subject's scan has been spatially registered to the MNI152 template included in SPM12 [156]. In the case of the ^{123}I-FP-CIT SPECT scans, a functional template was first generated following [157], and then registered each sample to this new template (in the same position as the MNI152 template). Once spatially normalised, both MRI and ^{123}I-FP-CIT SPECT scans presented the same size ($121 \times 145 \times 121$ voxels) and a voxel-size of $1.5 \times 1.5 \times 1.5$ mm. An intensity normalisation was also applied for functional images using α-stable distributions (see Section 2.3.2).

5.3 UGR-COG, a cognitive analysis dataset

This dataset was obtained from the Centre for Mind, Brain and Behaviour Research (CIMCYC) of the University of Granada. It is related to an experiment on selective attention and perceptual expectations from performing a computer-based visual preparation task [158]. The EEG data analysed come from a subset of the blocks, named "Localiser", which were designed to obtain clear perceptual data associated with visual stimuli. Visual stimuli were preceded by auditory cues that predicted either faces or names with 75% validity. To ensure the participants' engagement in the task, 10% of trials showed an inverted stimulus (face or name) to which participants responded with a key press. These trials were discarded from the sample. Each trial, see Figure 5.2, consisted on the presentation of a tone (unrelated to the task) lasting 200 ms, an inter stimulus interval of 1500 ms and the presentation of the face or name stimulus for 100 ms. Trails were separated by 1500 ms intervals. Raw EEG data was preprocessed using the same steps applied in [159]. In this thesis, only the visual representations of target faces vs name stimuli have been considered.

EEG data from 48 HC participants (mean age= 22.06, range = 18 – 31; 29 women, 18 men, 1 non-binary) comprise the dataset. They were all native Spanish speakers, right-handed with normal or corrected vision, and signed informed consent prior to participation. Visual stimuli used consisted on 160 male and female faces (50% each,

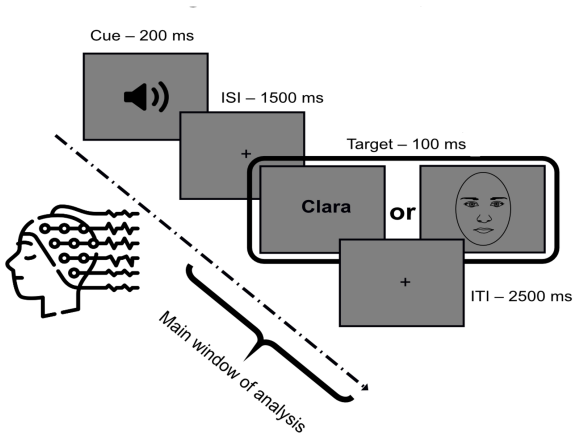


Figure 5.2: Behavioral task and design: example trial of the UGR-COG dataset. Participants were cued about an incoming target stimulus (a face or a name).

with ~ 6^ox9^o visual angle, extracted from The Chicago Face Database [160] plus 160 unique Spanish male and female names (50% each, with ~ 8^ox2^o visual angle). Auditory stimuli were four different tones (250, 300, 350 and 400 Hz).

5.4 SGH-SCZ, a schizophrenia MRI dataset

Participants included in this dataset were recruited from the Shanghai Mental Health Centre (SMHC). It is one of the largest and prominent psychiatric hospitals in China, which offers a wide range of mental health services, including diagnosis, treatment, rehabilitation, and prevention of mental disorders. Moreover, the SMHC also participates in research projects and conducts clinical studies. This data set is from one such study [161]. All participants provided a written informed consent. The study was approved by the Ethics Committees of the SMHC and the Institute of Psychology, the Chinese Academy of Sciences.

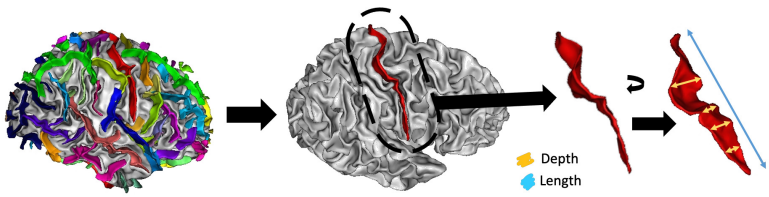


Figure 5.3: Example of a brain with sulci regions automatically labelled by BrainVISA using Morphologist 2021 pipeline (right). The central sulcus is highlighted (middle) and indicates how length and depth are measured in a region (left).

The dataset consists of MRI scans from 65 (27 females) Han Chinese patients with SCZ and 57 (24 females) HC. Participants underwent structural neuroimaging as well as

a clinical evaluation. High-resolution T1-weighted structural images were acquired on a 3T MRI scanner with 1mm isotropic voxel size. Details of the acquisition parameters are given elsewhere [161]. In this thesis, no non-linear spatial normalisation has been applied to the scans to avoid possible bias generated from shape deformations of the sulcal patterns [162, 163]. The MRI scans have been processed using BrainVISA 5.0.4 [164] to extract sulcal features by means of the Morphologist 2021 pipeline [165, 166]. Information is obtained from 62 areas per hemisphere (123 in total; the sulcus of the supra-marginal gyrus is only defined in the left hemisphere). In each region, the features measured in Talairach space [167, 168] are length, depth (average and maximum), fold opening, medial surface of the cortical folds and GM thickness [169, 170]. The average and maximum depth and length are calculated as features. Other available features are associated with morphological parameters rather than surface topology, and have therefore not been considered. Figure 5.3 shows an example of an automatically labelled brain by BrainVISA and the features extracted from a specific region.

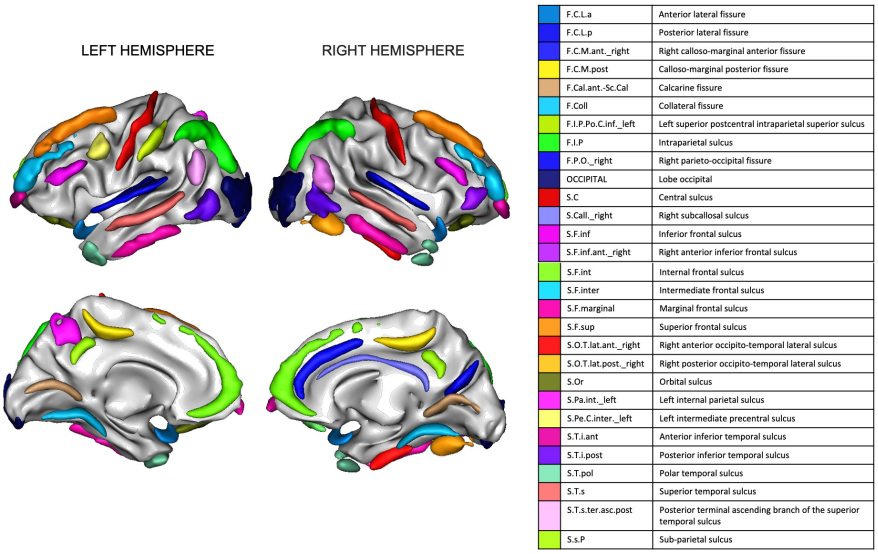


Figure 5.4: The forty-nine regions from the BrainVISA sulcal atlas used in the SHG-SCZ dataset. All other regions were excluded due to sulcal misdetection.

In some cases, a particular sulcus could not be identified or was misdetected by the Morphologist 2021 pipeline. Therefore, samples with more than 18 of these events (15% of the total number of regions) have been excluded from the dataset. Insula (left and right) regions have been also excluded because of the high possibility of being misdetected due to its peculiar shape. After these exclusions, any region that still has at least one misdetection across remaining participants has been excluded. Finally, features have been normalised to zero mean and standard deviation 1. Individuals with any feature with values greater than 6 times the standard deviation have been removed. These exclusion criteria result in 49 remaining areas for analysis, which are shown

in Figure 5.4. The final number of individuals (samples) is 114, the demographics for whom are shown in Table 5.7. As relevant demographic information is available in this dataset, statistical significance tests are applied. It can be seen that the sample set was matched for size, sex and age, with a sample size of 58 SCZ patients and 56 HC.

	SCZ	**HC**	**Total**	t/χ^2 **statistic**	p**-value**
N	58	56	114		
Sex (M/F)	35/23	29/27	64/50	0.85	.357
Age ($\mu \pm \sigma$)	22.95 ± 5.64	24.79 ± 7.36	23.85 ± 6.57	1.20	.233
IQ ($\mu \pm \sigma$)	93.18 ± 18.40 (N=55)	116.41 ± 14.38	105.30 ± 19.91	7.16	< .0001*
Edu ($\mu \pm \sigma$ years)	12.41 ± 2.91	13.40 ± 2.54	12.90 ± 2.77	1.93	.056
Hallu (Yes/No)	20/38	0/56	20/94	76	< .0001*

M: Male, F: Female, IQ: Intelligence quotient, Edu: Education, Hallu: Hallucinations

p-values were obtained using: two-sample t-test or Chi-Square test, * when $p < .05$

Table 5.7: Demographic details of the SGH-SCZ dataset.

Part II

CONTRIBUTIONS OF THIS THESIS

6 APPLICATION OF MACHINE LEARNING IN MULTICLASS CLASSIFICATION

6.1 Introduction

The detection of AD in its early stages is crucial for patient care and drugs development. Distinguishing between AD and its related neurological disorders, including its prodromal stage MCI, is very challenging from the clinical evaluation point of view, predominantly in the early stages of the disease. Motivated by this fact, the neuroimaging community has extensively applied ML techniques to the early diagnosis problem with promising results despite the fact that discrimination between MCI and AD has been shown to be a difficult task [51, 171, 172, 173]. Machine Learning applications in neuroimaging have become an indispensable tool for brain image analysis and CAD systems, producing a prolific area of research [60]. However, the lack of standardised datasets hinders direct comparisons of approaches, and the identification of their virtues.

The open data policy and the creation of big databases have facilitated the organisation of competitions for improving CAD systems for AD diagnosis, such as CADDementia [174] and TADPOLE (https://tadpole.grand-challenge.org/), among others. This has helped the community to address different raised problems and to standardise the approaches to the problem. It also facilitates the reproducibility of results, a problem that is becoming of central interest in neuroimaging research [16].

In this chapter, data from an international challenge for automated prediction of MCI from MRI data is analysed to address the multiclass classification problem. This will allow to analyse the different stages associated with a CAD system, as well as the most common ML techniques.

The objective of the competition was the development of CAD systems for the multiclass classification of 4 classes: HC, MCI subjects, cMCI subjects, and AD patients. The challenge provided with preprocessed MRI data of the different classes to allow participant proposals of optimised CAD systems, based on the finding that combining multiple anatomical measures improves classification of early diagnosis of AD [61]. The results of the challenge were published in the special issue [175] on the Journal of Neuroscience Methods. The winner proposal used a random forest ensemble with feature extraction methods [176], yielding a 61% accuracy in the multiclass classification problem. Therefore,

Ensemble methods have been successfully applied to neuroimaging problems [177, 178, 179, 180]. It has also been proven that multiclass approaches using binary classifiers can be a competitive solution, such as those based on one-versus-one or one-versus-rest approaches [181, 182]. This study uses and compares ensemble methods and aggregation methods by binary classifiers, as those of highest rated approaches in the challenge [148].

To optimise the multiclass classification through combination of anatomical features, not only classifier aggregations are necessary, but also feature extraction techniques [183]. Different approaches to feature selection and extraction reported high relevance in the literature, with high accuracy and also a correspondence between automatic cortical and subcortical region selection and clinical findings [177]. Concretely, brain atrophy has been found to be relevant for AD diagnosis in WM cortical and subcortical regions, as well as hippocampal volume, cortical thickness, and grey matter density, thus making feature extraction a reasonable preprocessing step (see [61] and references therein). One of such successful methods for feature selection and combination is PLS, linearly transforming the data into a space maximal separation between classes [184, 173, 185, 7].

Through the extensive use of feature extraction and one-vs-one feature selection and classification, a CAD system for identification of early stages of AD and MCI is studied in this chapter. The system aims to optimise the combination of multiple anatomical measures of brain atrophy to improve classification performance, which yields an accuracy of 67%, outperforming the winning proposal in the competition.

6.2 Methodology

The methodology followed in this study aims at optimising the binary classification of the different classes, HC, MCI, cMCI and AD in the multiclass classification problem, so that the overall classification performance is increased. To that aim, the process is

divided in four steps as depicted in Figure 6.1. A first preprocessing step is applied to discard outliers and standardise the data. Secondly, based on the observation of the existence of irrelevant or redundant data, a filter is applied to eliminate unimportant features. Once a set of features is selected, a combination of statistical tests and PLS techniques are used to extract features at binary level for each one vs. one classification (HC vs. MCI, HC vs. cMCI, etc..). Finally, binary classifiers are trained on these data, and an aggregation method is proposed for achieving a final multiclass decision.

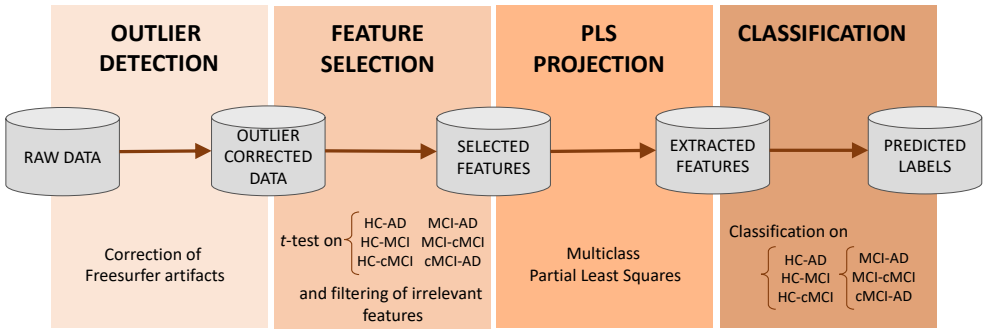


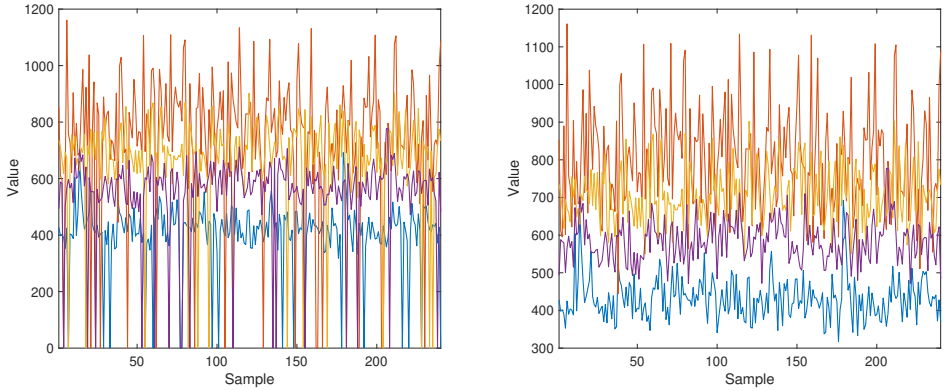
Figure 6.1: Flowchart of the proposed method.

The KAGGLE-AD dataset is utilised in this study, which is described in section 5.1.2. Specifically, the two subsets of training and test provided by the challenge are the ones used, of 240 and 500 samples, respectively. Each of the samples has 429 features available. No information about the class labels of the test set was available during the competition. The test set was half split into public and private test sets and only the accuracy score on the public dataset was available for competitors until the challenge ended. Once the challenge finished, class labels for the subjects on the test set were provided to the competitors. The accuracy score on the real subjects of the testing set was used as the figure of merit in the competition.

6.2.1 Preprocessing

The presence of outliers is usually an undesirable source of instabilities for ML applications. In neuroimaging, outliers are specially challenging as they are frequently found due to acquisition, scanner differences, preprocessing artefacts or resulting from large intrinsic inter-subject variability, having a dramatic effect on the statistical based analysis [186].

A carefully analysis of the data reveals a high abundance of outliers on each of the 429 data features. A common preprocessing step in ML consists of centering the data to zero mean and one standard deviation values, usually known as z-score values (for more information, see section 4.1). However, as Figure 6.2a shows, data contain high



(a) Uncorrected values.

(b) Corrected values.

Figure 6.2: Training data visualisation using four features for corrected and uncorrected values. These features are standard deviations of cuneus, entorhinal, inferior temporal and postcentral thickness in the left hemisphere.

salt-and-pepper type noise. This effect could be attributed to a miss-transformation of the data format, coming from Freesurfer software [150], as described by the challenge organisers. For this reason, a different preprocessing process has been adopted prior to centering the data. The outlier correction algorithm is described in Algorithm 6.1. The algorithm results are shown in Figure 6.2b. Correcting this format defect reveals a different data structure, with high redundancy, justifying posterior steps. Concretely, the features sets 1-35, 45-73, 71-139, 140-277, 278-347, 348-413, 413-429, seem to contain very low inter-patient variability, suggesting that the feature space dimension can be highly reduced. For example, feature space dimension can be optimally reduced to a value below 20, as will be justified later.

Data: Raw data matrix D of r features and s subjects
Result: Clear outlier values
Compute median values $M(s)$ of D for each feature ;
for *i from 1 to r* **do**
  for *j from 1 to s* **do**
    if $D(i,j) > 50 * M(j)$ **then**
      Replace outlier value by $D(i,j)/1000$;
    else if $D(i,j) < 50 * M(j)$ **then**
      Replace outlier value by $D(i,j) * 1000$;
    end
  end
end

Algorithm 6.1: Outlier elimination algorithm

6.2.2 Feature selection and extraction

The preprocessing step is followed by the elimination of irrelevant features and the extraction of features for classification. The former is a filter in a one vs. one approach. The features are sorted according to a specific criteria, thus eliminating the features with the lowest relevance. The latter is achieved under a multiclass PLS transformation of the selected features, reducing the feature space dimension [187].

The sorting criteria to detect irrelevant features is based on binary comparisons between classes: HC vs. MCI, HC vs. cMCI, HC vs. AD, MCI vs. cMCI, MCI vs. AD, and cMCI vs. AD. For each binary comparison a t-test is performed for each feature, and the features f_i are sorted according to their value of the t statistic, t_i, $i = 1, 2, ..., 429$. A 6x429 matrix S of sorted features is generated in this process. From this matrix S, a submatrix T is constructed by eliminating the n last columns, ordered by decreasing value of the t statistic. The number of times m a feature appeared in the matrix T was calculated for each feature. The parameter m is a significance measure for each feature and is constrained: $0 \leq m \leq 6$. All the features with a value of m under a fixed threshold R where filtered out, resulting in a feature selected set S containing the most relevant features for all the individual comparisons:

$$S_R = \{f_i : f_i \in T \ \& \ m_i > R\} \quad i = 1, 2, ..., 429 - n \tag{6.1}$$

The parameter n was fixed by CV, and the parameter R has six possible values, being $R = 3$ a reasonable compromise between very restrictive and non-existent filter.

The feature set S_R selection is followed by a PLS-based feature extraction (see section 4.3.2). This last feature extraction step produces a transformed data matrix D_t. The PLS transformation maximises the separation between classes in the new space, and can also be used to reduce the feature space dimension by selecting a reduced number of PLS components.

Apart from this last technique, the use of an AE (section 4.3.3) was tested for the same purpose as PLS, in order to use the data generated at the encoder output as low-dimensional data input in the classifier. However, its explanation will not be extended since it is not finally used in the CAD pipeline due to worse results than PLS for this dataset.

6.2.3 Classification

A simple solution to the multiclass classification problem is to build binary classifiers and combine them. Classical aggregation techniques of binary classifiers in multiclass problems are usually based on the Error Output Correcting codes (ECOC) [188, 189]. The idea is to construct an output code matrix, where each column represents a binary classifier trained to distinguish between two classes. During prediction, the binary classifiers are applied to the input, and the output code matrix is used to decode their decisions and determine the predicted class.

Given the multiclass classification problem on N classes, the simplest example is the one-vs-rest model, where the output code is generated by N binary classifiers that exhaust all possible one class versus the $N - 1$ rest of the classes classifications. A more refined option is to employ a one-vs-one methodology. In this case, there are $K_s = N(N-1)/2$ binary classifiers. Each classifier uses only two classes at a time, chosen from the entire set of classes. This ensures that all possible combinations of 2 classes are covered without any repetition. Table 6.1 presents the ECOC-style representation of both approaches. Once the selected approach is implemented, a decoding algorithm is used to assign a final class to each generated output code. Considering the output as a length N codeword, the decoding algorithm can be modelled as a communication problem, where the class information is being transmitted [189].

	one vs. rest				**one vs. one**					
Class	f_1	f_2	f_3	f_4	f_1	f_2	f_3	f_3	f_4	f_5
1	1	-1	-1	-1	1	1	1	0	0	0
2	-1	1	-1	-1	-1	0	0	1	1	0
3	-1	-1	1	-1	0	-1	0	-1	0	1
4	-1	-1	-1	1	0	0	-1	0	-1	-1

f_i: binary classifier

Table 6.1: ECOC coding in a multiclass classifier (4 classes), for both one-vs-rest and one-vs-one approaches.

In this study, the following optimised approach to the multiclass classification is considered: a binary classifier is trained on the one-vs-one individual classification tasks using the transformed data matrix D_t, producing a six-bit codeword output for each sample. This output is aggregated to produce a final prediction on the test sample, defined as a decoding process in ternary ECOC algorithms, taking values on the four possible classes: HC, MCI, cMCI and AD. The Hamming decoding [188] is used to map each possible codeword into a single output class as $HD(x, y_i) = \sum_{j=1}^{K_s} 1/2(1-\text{sign}(x^j y_i^j))$. The justification of this choice lies on the fact that the classes are nested. Concretely, the MCI class is considered as an early stage of AD, although not free of controversy[59, 190]. In any case, the class cMCI is an early stage of AD, and thus can be considered as a subclass of the AD class. For this reason, a ternary ECOC with three possible symbols allows for reduction of the non-relevant class influence in the codeword coding and decoding, and thus managing the possible errors arising from the difference on binary classification accuracies.

Different classifiers, which are described in section 4.4, are used to perform the individual binary tasks: SVM, KNN and decision trees, using different ensemble techniques: bagging [191], boosting [192] and random forest [109]. The DL algorithms used include MLP and CNN, whose main purpose in using them is for reference and comparison to other published results of the challenge [193, 176].

Apart from this, RUB validation technique is considered in this study along with a 10-

fold CV strategy to estimate the performance of the proposed method. For description and comparison of these procedures, see section 4.5.

6.3 Results

To perform the parameter fitting of the proposed method a 10-fold CV strategy was employed. Once the parameters were optimised, the test set was used to estimate the accuracy, recall and F1-score. Regarding the parameters of Equation (6.1) and the number of PLS components, a grid search strategy was employed. Figure 6.3 shows the accuracy results on the training set for each pair *(number of PLS components, number of features)*, where the number of features is selected by order from the pool of S_R, affected by the value of n. It can be claimed that a wide range of values around 10 PLS components and 10 selected features produce competitive classification results, whereas a choice of PLS components above 3 and below 20 is also a good compromise independently from the number of features selected. This can be related to the robustness of the method. In the end, the number of selected features was 22 from which 13 PLS components were obtained. These features cover 19 regions, the MMSE score, sex and age. Figure 6.4 illustrates the selected regions.

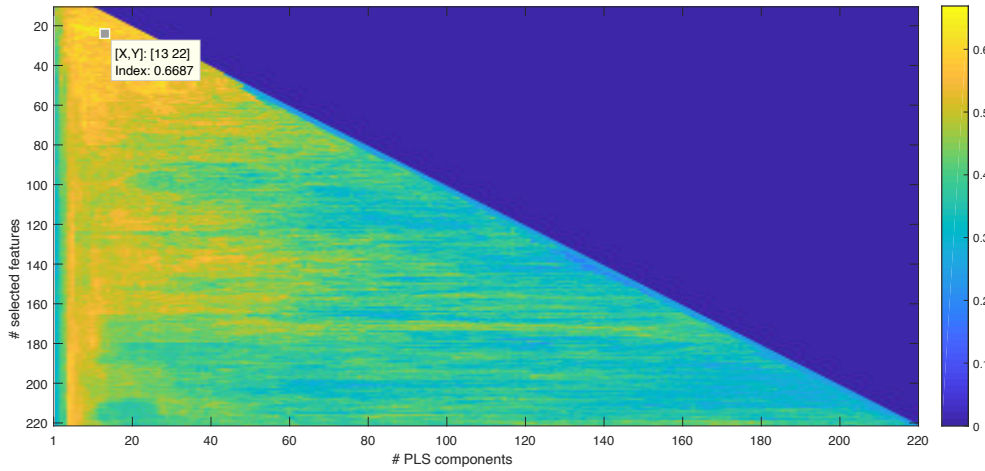


Figure 6.3: Accuracy values for each pair number of PLS components and number of selected features. The final selected values, 22 features and 13 PLS components, are indicated, as well as their associated accuracy.

Table 6.2 summarises the classification results on the training and the test set of 160 samples excluding the dummy subjects. SVM outperforms every other classifier, and linear kernel provides slightly better performance than non-linear kernels. However, even the simplest 1-NN (KNN) classifier provides very competitive results if related to the challenge results, which are summarised in Table 6.3. The competitive performance

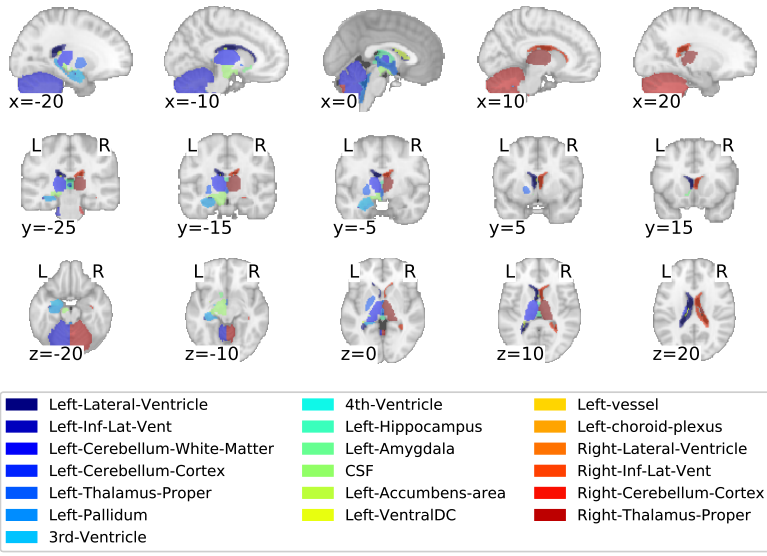


Figure 6.4: Selected regions after one-vs-one *t*-test feature selection.

of every classifier is a sign of the preprocessing importance, revealing that feature selection and extraction provide a very relevant set of features. Furthermore, the fact that more complex techniques are the ones with the lowest performance, such as CNN and DL algorithms in general, is consistent. It is mainly due to the use of such a low number of subjects and features, since these techniques are especially focused on problems with a large and high-dimensional dataset.

		Training			**Test (without dummies)**		
Ensemble	**Classifier**	Accuracy	Recall	F1-score	Accuracy	Recall	F1-score
-	SVM lineal	0.48	0.47	0.47	**0.67**	**0.52**	**0.66**
-	SVM RBF	0.47	0.47	0.48	0.67	0.52	0.63
LogitBoost	Decision Tree	0.48	0.44	0.51	0.64	0.46	0.60
Random forest	Decision Tree	0.48	0.44	0.52	0.64	0.51	0.57
AdaBoost	Decision Tree	0.50	0.42	0.47	0.60	0.43	0.52
-	5-NN	0.52	0.43	0.44	0.60	0.44	0.46
-	1-NN	0.52	0.39	0.42	0.58	0.39	0.44
-	3-NN	0.51	0.40	0.39	0.58	0.41	0.40
-	MLP	0.56	0.57	0.56	0.60	0.60	0.59
-	CNN	0.55	0.55	0.54	0.48	0.47	0.48

NN stands for nearest neighbours.

Table 6.2: Performance results for selected features using different classifiers.

Table 6.4 summarises the linear SVM classification results obtained following the proposed aggregation method. F1-score and recall are also reported during the training

Team	Ranking (Partial)	Team	Ranking (Final)
Stavros Dimitriadis – Dimitris Liparas	0.35999	Stavros Dimitriadis – Dimitris Liparas	0.61875
GRAAL	0.35599	SiPBA-UGR	0.5625
Bari Medical Physics Group	0.34799	Sørensen	0.55
BrainE	0.34399	Bari Medical Physics Group	0.55
gogogo	0.336	GRAAL	0.54375
DevinAnnaWilley	0.336	Jean-Baptiste SCHIRATTI	0.54375
fengxy	0.332	Neuroimage Division – CIFASIS – ARG	0.54375
ChaseCowart	0.328	Salvatore C. | Castiglioni I.	0.5375
BoyX	0.328	Loris Nanni	0.53125
SiPBA-UGR	0.324	BrainE	0.525
Jean-Baptiste SCHIRATTI	0.324	utaphys	0.525
utaphys	0.32	gogogo	0.525
Salvatore C. | Castiglioni I.	0.31199	ChaseCowart	0.51875
Sørensen	0.30399	agrickard	0.50625
Neuroimage Division – CIFASIS – ARG	0.30399	fengxy	0.5
agrickard	0.30399	JocelynHoye	0.5
JocelynHoye	0.30399	DevinAnnaWilley	0.46875
Loris Nanni	0.29199	BoyX	0.4625
Webiolab	0.276	Webiolab	0.2125

Table 6.3: Partial and final results of the challenge by group, using the whole test set and the test set without dummies, respectively. The accuracy scores are given as they appear in the challenge.

and test phases. The results are detailed for each class: HC, MCI, cMCI and AD. As expected, AD and cMCI are the classes with highest recognition values, whereas MCI report recognition rates slightly over random classification during training, but improved values on test. Overall, recognition rates are several percentage points over the challenge winner approach, outperforming every proposal of the challenge in the partial and final rankings (Table 6.3).

	Training			**Test**			**Test (without dummies)**		
Class	Accuracy	Recall	F1-score	Accuracy	Recall	F1-score	Accuracy	Recall	F1-score
HC	0.43	0.41	0.42	0.43	0.39	0.41	0.80	0.67	0.72
MCI	0.23	0.27	0.25	0.23	0.39	0.29	0.45	0.62	0.52
cMCI	0.50	0.48	0.49	0.46	0.31	0.37	0.62	0.59	0.61
AD	0.75	0.70	0.73	0.38	0.44	0.41	0.80	0.78	0.79
mean	0.48	0.47	0.47	**0.38**	0.39	0.37	**0.67**	0.67	0.66

Table 6.4: Performance results by class using the selected and extracted features and the one-vs-one classification scheme with SVM.

Apart from this, a study about the control of the FWE rate in the proposed CAD system was performed based on the resubstitution estimation. A HC dataset of 100 samples, 60 from the training set and 40 from the test set (without dummies), was used (see Table 5.3). The dataset was randomly divided into two subsets of 50 subjects each throughout 1000 iterations. Then, the resubstitution estimation was evaluated under the null hypothesis that the actual risk was equal to 0.50 (no group difference in the feature set should be true), where the number of PLS components (dimensions) was chosen equal to 1. The resubstitution accuracy obtained was equal to 0.612 with a

standard deviation of 0.037. The upper bound associated to this configuration is equal to 0.136, when applying a significance level of 0.05 in Equation (4.13). Thus, the actual accuracy was at most 0.523 ± 0.037. As a conclusion, the null-hypothesis could not be rejected in the test, so the method had a good control of the FWE rate.

On the other hand, if a 10-fold CV strategy was tested instead of RUB, an accuracy of 0.583 ± 0.060 was obtained with 13 PLS components. Although the null hypothesis could be rejected with the current test, the null hypothesis could have not been rejected with a confidence interval of 0.10, a value higher than the usual one of 0.05, but interesting in the neuroimaging field [34].

The last verification of the significance of the selected features was assessed using RUB as validation approach in classification. Following Equation (4.13), the upper bound considering the final 13-dimensional dataset in each one-vs-one classification was estimated. The sample size in each comparison was 200 that is, by combining both training and test (without dummies) sets, a total of 200 subjects were considered in a two-class analysis. With a significance level of 0.05, this upper bound was equal to 0.343 in all cases. Table 6.5 shows empirical and actual errors of each one-vs-one comparison according to the upper bound. HCvsMCI and MCIvscMCI actual risks were above 0.50 at the worst case, which means that the selected features cannot be accepted as significant to classify these conditions at the given significance level. Nevertheless, the difficulty of separating these conditions is well-known, thus in general terms a high relevance of the selected features is observed.

Classifier	Empirical error	Upper bound	Actual error
HC vs. MCI	0.320	0.343	**0.663**
HC vs. cMCI	0.135	0.343	0.478
HC vs. AD	0	0.343	0.343
MCI vs. cMCI	0.240	0.343	**0.583**
MCI vs. AD	0.065	0.343	0.408
cMCI vs. AD	0.130	0.343	0.473

Table 6.5: Actual risk associated to each one-vs-one classifier using the selected (22) and extracted (13) features and SVM applying RUB with 200 samples and a significance level of 0.05.

The relationship between the actual error and dimensionality used in each classifier could be detected. Figure 6.5 shows that the actual risk was never less than 0.50 in the HCvsMCI comparison, whilst in the MCIvscMCI classification the number of PLS components needed for achieving that condition was less or equal to 6. Nevertheless, the use of 6 PLS components would have decreased the overall accuracy down to 60%.

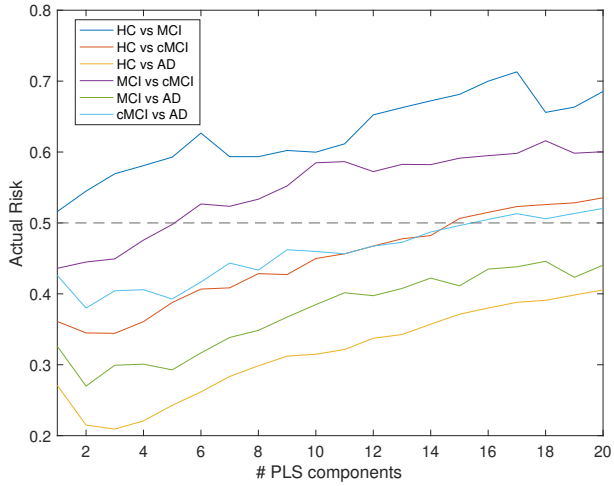


Figure 6.5: Estimates of actual risk in each one-vs-one classifier for several dimensions using a sample size of 200 subjects and a significance level of 0.05.

6.4 Discussion

The results presented reveal the importance of feature selection and extraction in the CAD pipeline of this challenge problem, which applies to any other CAD system. Concretely, the default sorting of the 429 cortical thickness values provided by the organisers of the challenge seem to contain already important information for classification. The best results are obtained by removing some of the original set of default sorted features and by applying the proposed minimum filter. After the feature selection by means of the filter in Equation (6.1), it is relevant to emphasise that there is not an equilibrium between right and left hemisphere regions. Specifically, there is a dominance of left-sided hemisphere regions, which is coherent with recent findings in CAD diagnosis of AD [194]. If compared with other competent CAD proposals of the challenge, such as the winner Dimitriadis-Liparas proposal [176], there is a significant overlap with the feature extraction results. Their approach also results in a left-hemisphere regions predominance. Therefore, a successful feature selection and extraction method is critical for optimal performance.

The use of t-tests as sorting criteria for filtering features and the upper-bound tests for assessing the feature relevance are justified on the basis of the Kolmogorov-Smirnov (KS) and the upper-bound tests [37]. Moreover, the KS test quantifies the departure of the empirical distribution function of the features from a cumulative distribution function of a particular statistical distribution. In this case, the assumption underlying a t-test implies that the feature values follow a normal distribution, which is an acceptable assumption in the light of the results of the KS test presented in Figure 6.6. It is important to stress that direct comparisons between different t_i values at different

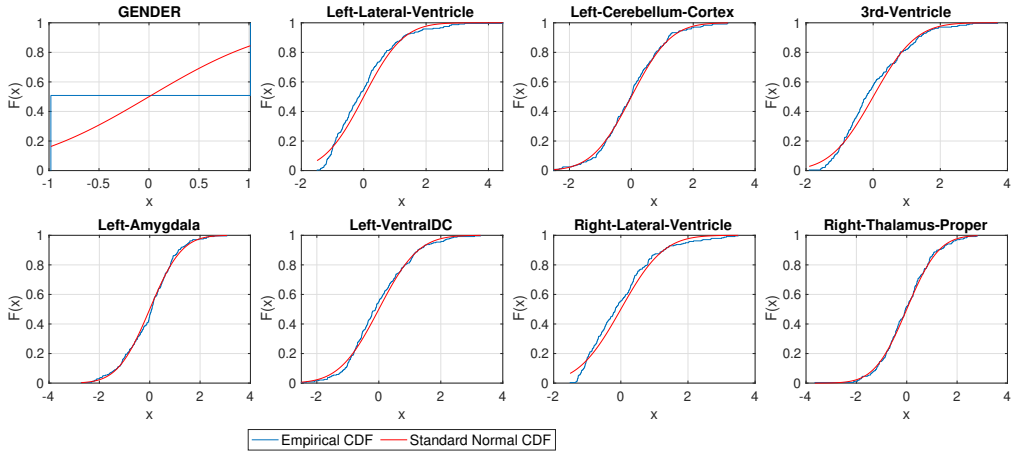


Figure 6.6: Results of the Kolmogorov-Smirnov test for some of the selected features.

tests are never performed, but the values of t_i are used for feature sorting. In addition, the novel RUB approach for testing relevance is applied in a set of features. It is a data-driven approach (agnostic or free-parameter model) based on the resubstitution error estimate and the theoretical upper-bound of the empirical errors that provide a confidence interval for performing hypothesis testing.

The present methodology can be applied in other multiclass classification problems, in which there is a hierarchy and overlap between classes. Furthermore, the computation of the final selected algorithm is fast, which is an advantage over tested DL techniques, which require a longer processing time associated with network training.

Concerning the limitations on the present study, the preprocessing of outliers reveal a high redundancy on the original data. Therefore, the preselection of brain regions for the challenge, and the extraction of cortical thickness affects the maximum achievable performance in several ways. Firstly, the limited number of training samples makes statistical estimations prone to bias, a widely known-problem in medical imaging [16, 31]. The CV technique used in this study for performance estimations can be considered as "pessimistic" [126], and therefore some mismatches between training fitting and final test estimations can be expected, limiting the capabilities of the system for reaching its highest performance at test level.

Even though the proposed CAD was evaluated using the test set labels, which were not available during the challenge competition, the robustness of the method would have led to the best competition results with just a few submissions. Table 6.2 and Figure 6.3 illustrate how the method is robust against small variations on the optimal parameters and classifier choice, providing with accuracy values over 60% for a wide range of combinations.

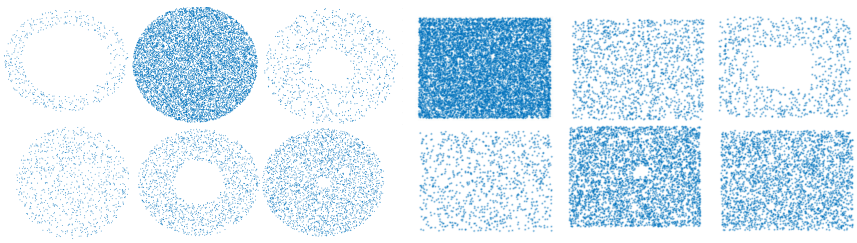
7 | A NON-PARAMETRIC STATISTICAL INFERENCE FRAMEWORK

7.1 Introduction

The previous chapter examined the ML techniques involved in the different stages of a CAD system, with particular emphasis on the importance of feature selection and extraction methods. This chapter will delve into another crucial aspect of CAD system implementation: model validation.

In order for CAD systems to be considered valuable tools for clinical assessment, the information they provide must, above all, be reliable. To detect this, during the development process of the systems, the performance of the implemented ML techniques is evaluated. Under optimal conditions, i.e. with a large data set, in which the generalisation error is small, one can assume a robust system has been obtained. Nonetheless, it may happen that the model implemented generates such a small generalisation error even with corrupted data, e.g. by deliberate permutation of data labels [96]. This is

especially true for models based on DL due to the high label-adaptability of NN [195]. Therefore, the capacity of NN to generalise is currently under discussion [196, 197, 198], and methods to measure and detect such generalisation would be advantageous [199]. For example, looking at Figure 7.1, it would be logical to think that it is easy to separate circles from squares. However, it could be the case that the classifier is not focusing on the main shape but on the holes within the shapes. it could be the case that by permuting the samples, the classification would still be effective by classifying the shapes according to whether they have a hole or not. This problem usually occurs when the number of samples is significantly lower than the dimensionality of the accompanying features, which is particularly common with medical data since a MRI can contain millions of voxels.



(a) Circles. **(b)** Squares.

Figure 7.1: Synthetic database of class *circles* and class *squares*.

Prior work has analysed the reliability of classification results in neuroimaging [200, 201]. They tested if the estimated accuracy is significantly better than that obtained by chance. To this end, the most widely used approaches are permutation methods, see section 3.5. In addition, methods for analysing statistical power and type I error in ML have been also proposed in the literature [202, 203].

Furthermore, the estimation of the classifier error through traditional approaches, such as cross-validation, which are used in conjunction with a variance-based bounds, enable calculation of empirical confidence intervals. These intervals are not always effective estimators of the true error for small sample sets [204]. Even the popular and recommended K-fold CV method [126] cannot properly work under these unstable conditions [31, 15]. Thus, the predictive power of the classifier can be controversial.

The idea underlying this chapter is to analyse the generalisation capacity of a given classifier by means of proposals for the above-mentioned problems. A two-sample test based on classification accuracy is used for statistical significance analysis of the results. The null distribution associated with the test is performed by permutation test. Although not a novel method a priori [128], the use of RUB as a validation method eliminates most of the disadvantages previously associated with the method [203]. It is proved that permutation test can be used on small databases and validation techniques based on resubstitution need not to generate a positive bias (tendency not to be centered around a half). The method implemented in this study allows its use in scenarios of small sample sizes, common in neuroscience, for balanced, unbalanced, binary and

multiclass datasets. Thus, this non-parametric framework assesses the statistical power and Type I error control by comparing the results obtained using randomly labelled data and the ground truth.

Two scenarios are analysed. First, a two-conditions design comparing two different sample distributions is considered, e.g. AD and HC samples. As labels on these samples are permuted, no group detectable differences should be apparent, which is reflected in the classification accuracy. Second, a one-condition task is analysed with HC data and a putative task design that generates results that should control the FWE.

7.2 Methodology

A non-parametric methodology for analysing the prediction certainty in DL-based systems is proposed. This statistical analysis of the model validation is based on a hypothesis test, since a classification problem can be seen as a study of differences between conditions (see section 3.1). In this case, a test statistic grounded in data-driven SLT is applied. Specifically, the out-sample prediction error, or actual error, associated with classification is the chosen test statistic.

Therefore, a random-effects framework based on a permutation test is proposed to conduct such statistical analysis. Given H_0, as it is defined in Equation (3.1), the test statistics that will be applied are the actual error using K-fold and RUB, denoted as $\mathcal{T}^{KCV}$ and $\mathcal{T}^{RUB}$, respectively. The null distribution Π of the statistic is obtained by a permutation test (distribution of permuted errors). Each of the M errors obtained with permutations in the dataset is denoted as $\mathcal{T}_\pi$. This procedure, described in section 3.5, is preferred to parametric inference because, as previously mentioned, neuroimaging experiments usually involve high dimensional databases of small sample size (*curse of dimensionality*) where central limit approximations tend to be poor [200]. Thus, the statistical significance of K-fold CV and RUB is calculated from statistical power and Type I error by redefining the baseline H_0.

7.2.1 Assessment of Statistical Power

The stated hypothesis associated with this study is:

$$H_0 : \mathcal{T}_\pi = \mathcal{T} \quad \text{vs.} \quad H_1 : \mathcal{T}_\pi > \mathcal{T} \tag{7.1}$$

Thus, the null hypothesis assumes that the classification error associated with a permuted dataset is equal to the actual error of the original one, $\mathcal{T}$, since there is independence between $\mathbf{x}$, the observed features, and y, the class label. The alternative hypothesis is that there is an effect in $\mathbf{x}$ given y, and therefore the error obtained by permutation is larger than the actual error. This is related to the Type II error, which occurs when independence is declared when a dependency exists, see section 3.1.

The evaluation of the results is made by comparing the error obtained with the original database with the errors obtained by permutation and with itself. Consequently, it is used as mathematical expression to compute its p-value:

$$p_{value} = \frac{card\{\mathcal{T}_\pi \leq \mathcal{T}\} + 1}{M + 1} \tag{7.2}$$

where $card(.)$ denotes the cardinality and M the number of permutations. The addition of 1 to the numerator and denominator is equivalent to including the fixed statistic value in the test, which allow to obtain a valid test [30].

7.2.2 Assessment of Type I error control

An evaluation of Type I error control of the inference method is also conducted by testing one-condition distributions, randomly split in two groups of the same size. Thus, the null hypothesis of independence between samples and labels is true by construction and the proportion of false positives detected should be close to a given significance level, α.

In order to compute the proportion of analyses (or iterations) that give rise to any significant result, an Omnibus test is applied. First, a similar expression to Equation (7.2) is used for the computation of the p-values of dimension M:

$$p_{value,m} = \frac{card\{\mathcal{T}_\pi \leq \mathcal{T}_{\pi,m}\}}{M} \tag{7.3}$$

where $\mathcal{T}_{\pi,m}$ represents the m-th error from Π, and which is compared to all values included in such distribution (including itself). In this case, there is no need to add 1 to the expression as only the statistical values obtained from the M iterations of the permutation test are considered. Once the M p-values are obtained, they are compared with α. The number of FP is calculated as the number of these values that are less than or equal to the significance level. The estimated FWE rate, or FP rate, is the number of false positives divided by the number of permutations:

$$\text{FWE } rate = \frac{card\{p_{values} \leq \alpha\}}{M} \tag{7.4}$$

From a parametric perspective, this analysis could be performed with a contrast of means, where the confidence interval obtained should include 50%, since the dataset is randomly split into two sets of the same distribution.

7.2.3 Summary of the procedure

Considering the background described above, a graphical scheme of the proposed process is illustrated in Figure 7.2. The summary of the procedure would be as follows:

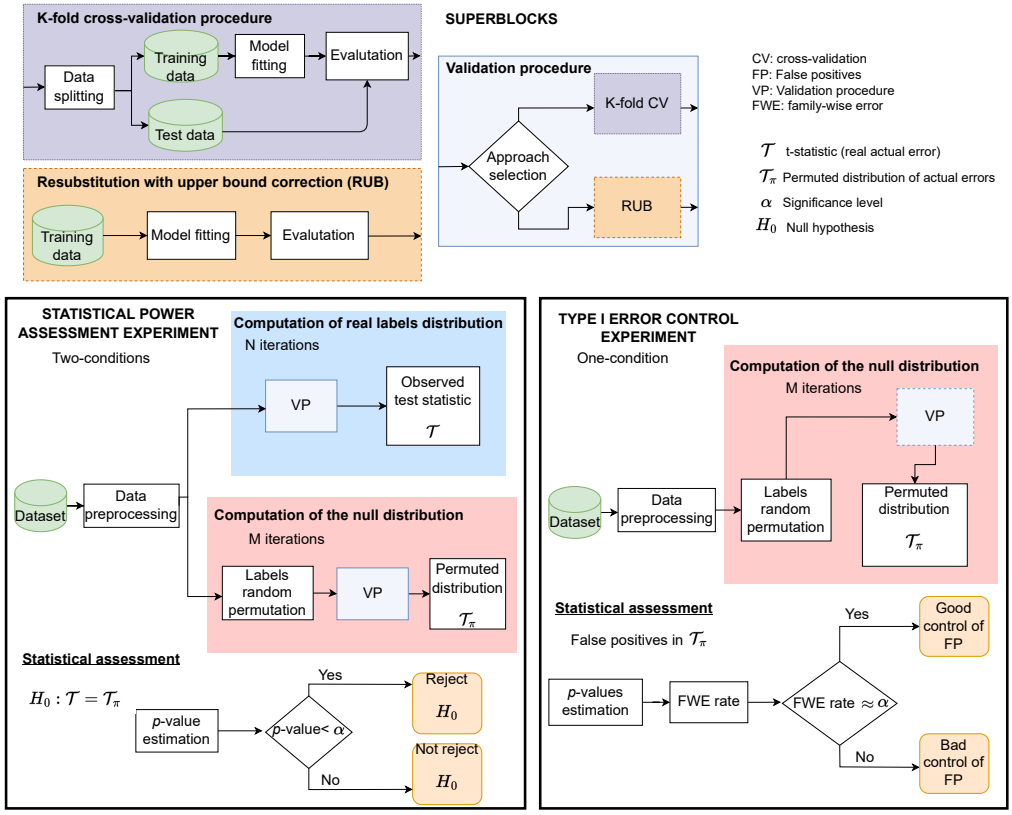


Figure 7.2: Flowchart of the proposed method for statistical power assessment (left) and type I error control analysis (right). The recurrent processes in both experiments were divided into blocks depicted in the upper part. The statistical power assessment experiment needs a dataset of two different conditions: the statistic related to the real labeling, $\mathcal{T}_\pi$ is compared to the label permuted distribution $\mathcal{T}_\pi$ in order to reject or not reject H_0 given a significance level α. The type I error control analysis needs a one-condition dataset and it divides the sample into two random sets, thus obtaining a permuted distribution $\mathcal{T}_\pi$. The values are compared with each other to detect whether the number of false positives is similar to the set α.

1. Determine the test statistic related to the original dataset, i.e. $\mathcal{T}^{KCV}$ or $\mathcal{T}^{RUB}$. To this end, data preprocessing, classifier fitting and evaluation of performance metrics are steps previously required. The observed test statistics are derived from N iterations for both approaches, with the aim of obtaining the most accurate estimation of the actual errors. At each iteration, the dataset labels were shuffled. Note that instead of dealing with the error, it is possible to work with its associated accuracy.

2. Computation of $\mathcal{T}_\pi$. The resulting null distribution contains M error or accuracy values. In case of using RUB, the number of required iterations in the permutation test is M, while for K-fold CV, O iterations are required depending on K ($O \times K \approx M$). Considering each fold independently allows a computation of the total

variance of the method instead of the average variance generated by the K folds.

3. Statistical assessment. The trade-off between statistical power and Type I error is explored.

    - Statistical power. The p-value is calculated according Equation (7.2). If the p-value obtained is below the significance level, α, the null hypothesis H_0 is rejected, otherwise it is accepted. Such a rejection implies that data and labels are not independent.
    - Type I error control. Once the M p-values are computed according Equation (7.3), the FWE rate is estimated using Equation 7.4. Note that in this case only one distribution from the dataset must be used. In terms of neuroimaging, the distribution chosen should be the control one, since given a confusion matrix it tends to be related to the negative condition [34, 37].

As only a finite number of permutations are performed due to computational restrictions, the standard deviation related to the Monte Carlo approximation of the p-value is estimated in the study following the expression $\sqrt{\frac{p(1-p)}{n}}$, where p is the p-value obtained and n is the number of samples [205].

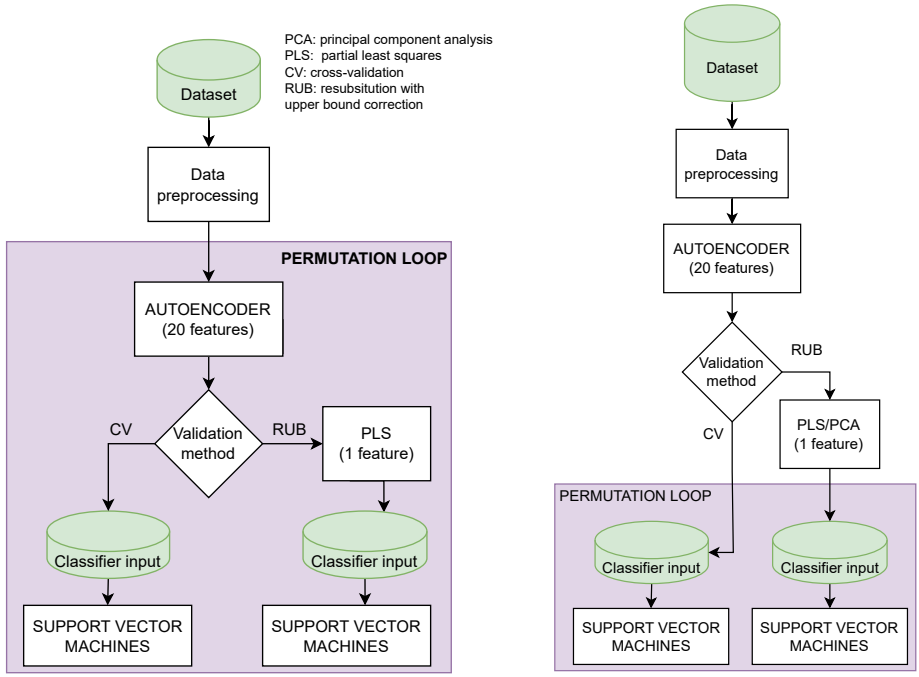
To demonstrate the effectiveness of this procedure, several well-known datasets related to Alzheimer's Disease are applied, which include balanced, unbalanced, binary and multiclass scenarios. These datasets are ADNI-AD, KAGGLE-AD and DIAN-AD, which are described in section 5.1.

7.2.4 Classification framework

The classification model chosen to conduct the experiments was based on a combination of AE and SVM architectures. The AE was used for feature extraction while the SVM (linear kernel) was used as classifier. Both algorithms are widely used in medical imaging [206, 207, 208, 209]. The reasons for selecting these architectures include: i) good performance in neuroimaging tasks [35], ii) the possibility of using a DL architecture to reduce dimensionality in a low-sample size scenario [28], and iii) linearity in the final classification level for applying RUB [31]. Thus, this modelling allows the capabilities of neural networks to be harnessed as a feature extractor [210]. Then, instead of using a DL model with classification layers, SVM is implemented. Apart from the computational cost, this also has a direct implication on the use of RUB, as upper bounds for DL are still under development and analysis as they are much more complex, even impossible, to implement [196, 211, 212, 134].

In addition to linearity, for the application of RUB it is desirable that the number of input features to the classifier is as low as possible (ideally only one feature) since there exists a high dependency between the estimation of the upper bound and features

dimension, see Equation (4.13). Therefore, when using RUB as a validation method the dimensionality of the features extracted from the Z-layer was further reduced by applying PLS [213]. In the following experiments, features were reduced to 20 in the Z-layer. In the case of applying 10-fold as validation method, these features were introduced as input to the SVM linear classifier. Nevertheless, in order to benefit from the advantages of RUB, these were reduced to 1 PLS component and then introduced to the classifier. This scheme is illustrated in Figure 7.3a.



(a) Inside the permutation loop.

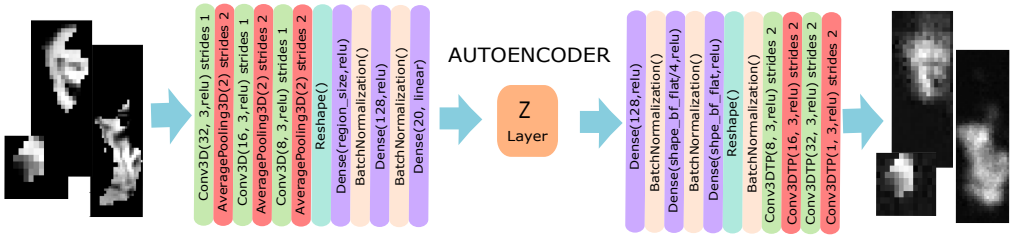
(b) Outside the permutation loop.

Figure 7.3: Flowchart of the classification procedure when the feature extraction step occurs inside the permutation loop (main experiments) and outside the permutation loop an specific approach for analysing the variability in the feature extraction step.

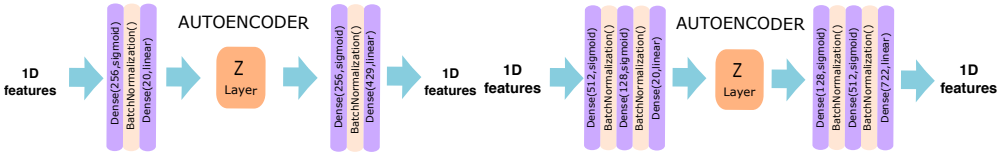
Moreover, to study how feature extraction is affected by using a permuted database and the generalisation ability of DL models as feature extraction methods, an experiment where the feature extraction step is performed only once and on the original (non-permuted) database is also conducted. Therefore, the permutation test is applied directly to the low dimensionality data as is illustrated in Figure 7.3b. This implies that the permutation test only affects SVM classification, rather than feature extraction and classification as in the main framework proposed. For this study, the same criteria and parameters are applied as in the previous studies, the only difference is the use of PCA, described in section 4.3.1, instead of PLS for dimensionality reduction in the subset of control subjects, as PLS could not be applied using a single class.

The AE architectures implemented as a feature extractor are shown in Figure 7.4.

Batch normalisation is applied to reduce overfitting [214]. The configuration parameters of the AEs are shown in Table 7.1. The main difference between the architecture for ADNI-AD and the other two databases is that the latter are one-dimensional rather than three-dimensional data, which is why instead of using convolutional layers, fully-connected layers are chosen.



(a) ADNI-AD (axial slices of the 3D MRI scans).



(b) KAGGLE-AD. **(c)** DIAN-AD.

Figure 7.4: AE model configuration for the different datasets.

Dataset	AE dimension	# Layers	Optimiser	Learning rate	Loss function	# Epochs
ADNI-AD	3D	9	Adam	0.0001	MSE	250
KAGGLE-AD	1D	2	Adam	0.001	MSE	50
DIAN-AD	1D	3	Adam	0.001	MSE	80

#: "number of", MSE: mean squared error

Table 7.1: Summary of the parameters associated to the different AE configurations used in the experiments. The number of layers is the one of the encoder part.

7.3 Results

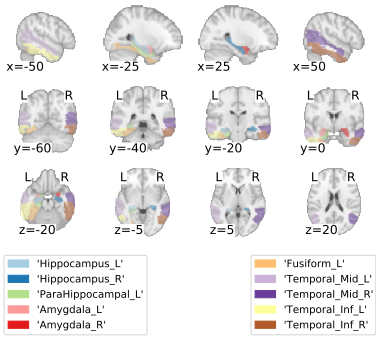
The aim of this section is to prove the validity of the methodology presented in Figure 7.2 for statistical analysis, and to examine the performance of the trained models. Specifically, it seeks to analyse the trade-off between statistical power and type I error and the advantages offered by the use of RUB over K-fold.

In the following experiments, K-fold CV with $K = 10$ was applied. According to the scheme proposed in section 7.2.3, the selected number of iterations performed were $N = 20$, $M = 1000$ and $O = 100$ since $K = 10$ (100×10). The level of significance α was set as 0.05.

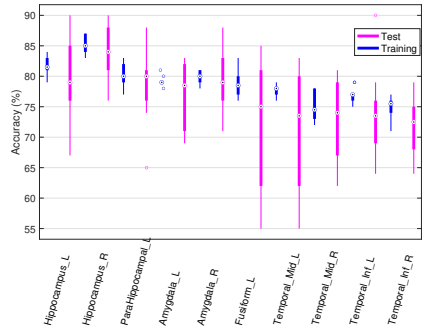
The scripts generated for the experiments were based on Python language (v.3.7) and Keras library over Tensorflow on a Nvidia Tesla V100-SXM2.

7.3.1 ADNI-AD: a three-dimensional experiment

For this experiment, an ensemble approach was implemented for classification. Instead of using the whole brain GM map, 10 specific regions of the Automated Anatomical Labeling (AAL) atlas [215] were used: hippocampus (left/right), parahippocampal (left), amygdala (left/right), fusiform (left), middle temporal gyrus (left/right) and inferior temporal gyrus (left/right). These regions were selected according to their importance to AD diagnosis [37] and are shown in Figure 7.5a. The use of a limited number of regions reduces the computational cost making the ensemble methodology tractable, and the results more reliable overall. The classification model implemented was trained using GM maps of each region as input and the final label selection was made considering the probabilities given by the individual classifiers. See Appendix A.1 for more information on the ensemble methodology.



(a) Selected brain regions.



(b) Training and test accuracies per region.

Figure 7.5: Brain regions of interest to work with, extracted from the AAL atlas in ADNI-AD dataset, and their training and test accuracies as independent input features applying the architecture proposed and 10-fold CV.

Following the framework shown in Figure 7.2, the original (non-permuted) dataset was used to train and test the model proposed for 20 iterations. The mean scores of this procedure are shown in Table 7.2 in the “Original dataset" rows where it can be deduced that the actual errors (i.e. test statistics) were $\mathcal{T}^{KCV} = 0.1681$ and $\mathcal{T}^{RUB} = 0.2344$ using 10-fold CV and RUB as validation methods, respectively. An example of accuracy rates related to each region is also shown in Figure 7.5b for one 10-fold cross-validation iteration. Note that to apply the RUB approach it was necessary to calculate the upper bound associated with the experiment by means of Equation (4.13). The value of the upper bound in this case is 0.0665 when the significance level is set at 0.05.

		Accuracy	Sensitivity	Specifity	p-value
10-fold CV					
Original dataset	Training	**0.8548 [0.0094]**	0.9166 [0.0094]	0.7794 [0.0163]	-
	Test	**0.8319 [0.0541]**	0.8989 [0.0627]	0.7504 [0.0932]	-
Permuted dataset	Training	**0.4791 [0.0864]**	0.4359 [0.0993]	0.5219 [0.1026]	-
	Test	**0.4987 [0.0753]**	0.4550 [0.1187]	0.5425 [0.1180]	0.0010 [0.0010]
Resubstitution					
Original dataset		**0.8321 [0.0059]**	0.9033 [0.0075]	0.7455 [0.0117]	-
Permuted dataset		0.5380 [0.0339]	0.4854 [0.1608]	0.5904 [0.1539]	0.0010 [0.0010]
Upper-bounded resubstitution (μ = 0.0665)					
Original dataset		**0.7656 [0.0059]**	0.8368 [0.0075]	0.6790 [0.0117]	-
Permuted dataset		0.4715 [0.0339]	0.4189 [0.1608]	0.5239 [0.1539]	0.0010 [0.0010]

Symbol "-" indicates that values were not computable and μ stands for upper bound.

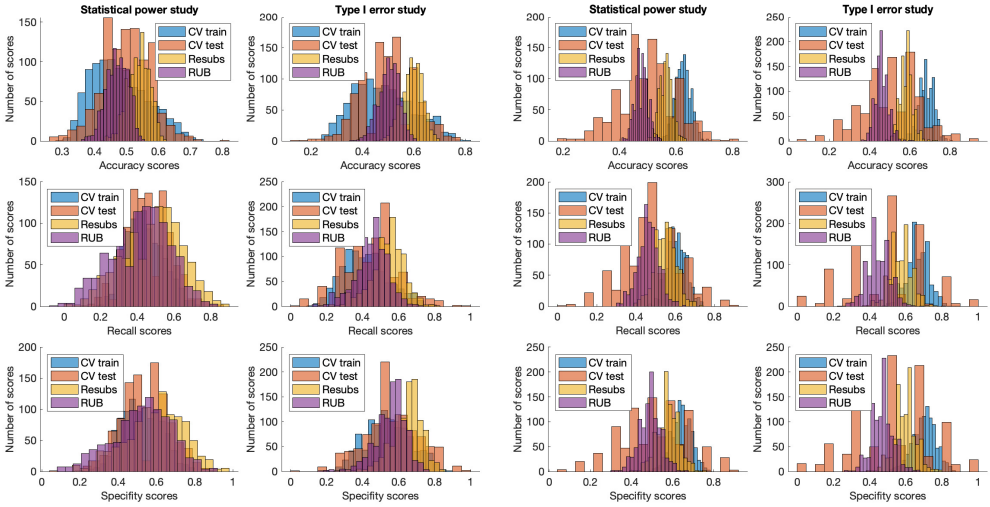
Table 7.2: Results related to the statistical power experiment using ADNI-AD original and permuted datasets. Validation methods applied to the permuted dataset were 10-fold CV (100 iterations, high computational cost, top), resubstitution (1000 iterations, high computational cost, middle) and RUB (by applying the upper bound, low computational cost, bottom). Original dataset scores were obtained from 20 iterations (medium computational cost). The significance level of the test was 0.05. Bold type indicates that the value is commented in section 7.4.

7.3.1.1 Randomisation on HC vs. AD

In order to conduct the statistical power analysis, the next stage is the derivation of the null distribution by means of permutations of the complete dataset. The results of the permutation test are shown in Table 7.2 in the "Permuted dataset" rows. It can be observed that both test statistics are considerably lower than the mean errors associated with the permutation distributions (0.5013 and 0.5285 for 10-fold and RUB, respectively). Moreover, Figure 7.6a (left) illustrates the distribution of the several metrics reported in Table 7.2 over the iterations of the permutation test. It should be remembered that although 100 iterations were performed with the 10-fold method, there were actually 1000 values when all the folds are considered together. With this information, the p-values obtained are equal to 0.0010 for both CV and RUB methods. Indeed, no error belonging to the null distribution was lower than the statistic test in any case. As both p-values were smaller than the significance level imposed in the test (0.05), the null hypothesis was rejected. Therefore, there was no independence between conditions (HC, AD) and samples.

7.3.1.2 Randomisation on HC

A subset of the HC sample was used to evaluate the Type I error control of the model. Thus, the number of subjects in this study was 229, which makes the value of the upper bound 0.0897 using Equation (4.13). The distribution of the scores of the permutation test is shown in Figure 7.6a (right). A summary of the results is indicated in Table 7.3, with mean actual errors of 0.4988 and 0.4966 using 10-fold CV



(a) ADNI-AD. **(b)** DIAN-AD.

Figure 7.6: Distribution of scores among the samples of the permuted datasets (1000 iterations) in the statistical power study (left) and type I error study (right). In the RUB study, the bounds applied are $\mu = 0.0665$ and $\mu = 0.0897$ in the two-condition and one-condition experiments, respectively, for ADNI-AD and $\mu = 0.0866$ and $\mu = 0.1225$ for DIAN-AD dataset, respectively.

and RUB, respectively. The FWE rates obtained were 0.0410 using CV and 0.0440 using RUB. Hence, the number of false positives obtained was consistent with the level of significance, 0.05.

		Accuracy	Sensitivity	Specifity	FWE rate
10-fold CV					
Permuted dataset	Training	0.4790 [0.1152]	0.4151 [0.1273]	0.5423 [0.1259]	-
	Test	0.5012 [0.1007]	0.4455 [0.1484]	0.5568 [0.1497]	0.0410 [0.0063]
Resubstitution					
Permuted dataset		0.5931 [0.0406]	0.5390 [0.0921]	0.6466 [0.0865]	**0.0440 [0.0065]**
Upper-bounded resubstitution ($\mu = 0.0897$)					
Permuted dataset		0.5034 [0.0406]	0.4493 [0.0921]	0.5569 [0.0865]	**0.0440 [0.0065]**

Symbol "-" indicates that values were not computable and μ stands for upper bound.

Table 7.3: Results related to the Type I error experiment using ADNI-AD permuted HC subset. Validation methods applied to the permuted dataset were 10-fold CV (100 iterations, high computational cost, top), resubstitution (1000 iterations, high computational cost, middle) and RUB (by applying the upper bound, low computational cost, bottom). The significance level of the test was 0.05. Bold type indicates that the value is commented in section 7.4.

7.3.2 DIAN-AD: when distributions are similar

The particularity of this dataset is the average age of the participants (see Table 5.4), which is considerably younger than in the other datasets analysed. Most of the subjects were young people with hardly any symptoms of Alzheimer's Disease but with or without a mutation connected to AD. It is therefore particularly interesting to analyse the statistical power of the classifier in this dataset.

7.3.2.1 Randomisation on MC vs. NC

Following the framework, assessing statistical power began by using the original database to train and test the model to obtain the test statistics. Mean results based on 20 iterations (with data shuffled on each iteration) are shown in "Original dataset" rows of Table 7.4. The test statistics, i.e. the mean actual errors, were $\mathcal{T}^{KCV}$ = 0.3714 and $\mathcal{T}^{RUB}$ = 0.4565. To compute the null distribution, all samples were permuted and shuffled as in the previous experiment. Figure 7.6b (left) illustrates the distribution of the several metrics obtained. The upper bound in this case is 0.0866. The metrics generated by the permutation test using a stratified 10-fold CV are shown in the first rows of Table 7.4, with a corresponding p-value of 0.0819. Using RUB (bottom of Table 7.4), the p-value was 0.0020. Therefore, the null hypothesis was rejected (data and labels are dependent) when tested with RUB, but not CV.

		Accuracy	**Sensitivity**	**Specifity**	**p-value**
10-fold CV					
Original dataset	Training	0.7325 [0.0225]	0.7901 [0.0364]	0.6748 [0.0359]	-
	Test	0.6286 [0.0962]	0.6802 [0.1360]	0.5777 [0.1560]	-
Permuted dataset	Training	0.6260 [0.0271]	0.6078 [0.0483]	0.6442 [0.0497]	-
	Test	**0.4963 [0.0975]**	0.4785 [0.1433]	0.5141 [0.1477]	**0.0819 [0.0087]**
Resubstitution					
Original dataset		0.6301 [0.0355]	0.7301 [0.0693]	0.5301 [0.0552]	-
Permuted dataset		0.5641 [0.0228]	0.5500 [0.0538]	0.5783 [0.0529]	0.0020 [0.0014]
Upper-bounded resubstitution (μ = 0.0866)					
Original dataset		0.5435 [0.0355]	0.6435 [0.0693]	0.4435 [0.0552]	-
Permuted dataset		**0.4775 [0.0228]**	0.4634 [0.0538]	0.4917 [0.0529]	0.0020 [0.0014]

Symbol "-" indicates that values were not computable and μ stands for upper bound.

Table 7.4: Results related to the statistical power experiment using DIAN-AD original and permuted dataset. Validation methods applied for the permuted dataset were 10-fold CV (100 iterations, medium-low computational cost, top), resubstitution (1000 iterations, medium-low computational cost, middle) and RUB (by applying the upper bound, low computational cost, bottom). Original dataset scores were obtained from 20 iterations (low computational cost). The significance level of the test was 0.05. Bold type indicates that it is commented in section 7.4.

		Accuracy	Sensitivity	Specifity	FWE rate
10-fold CV					
Permuted dataset	Training	**0.6830 [0.0378]**	0.6645 [0.0620]	0.7013 [0.0616]	-
	Test	**0.5028 [0.1395]**	0.4799 [0.2070]	0.5260 [0.2101]	0.0300 [0.0054]
Resubstitution					
Permuted dataset		0.5831 [0.0314]	0.5707 [0.0611]	0.5953 [0.0592]	**0.0440 [0.0065]**
Upper-bounded resubstitution (μ = 0.1225)					
Permuted dataset		0.4606 [0.0314]	0.4482 [0.0611]	0.4728 [0.0592]	**0.0440 [0.0065]**

Symbol "-" indicates that values were not computable and μ stands for upper bound.

Table 7.5: Results related to the Type I error experiment using DIAN-AD permuted NC subset. Validation methods applied for the permuted dataset were 10-fold CV (100 iterations, medium-low computational cost, top), resubstitution (1000 iterations, medium-low computational cost, middle) and RUB (by applying the upper bound, low computational cost, bottom). The significance level of the test was 0.05. Bold type indicates that it is commented in section 7.4.

7.3.2.2 Randomisation on non-carriers

Results of the permutation test to assess type I error control are given in Table 7.5, whilst in Figure 7.6b (right) the distribution of several metrics is illustrated. In this case, the model was fitted and evaluated using HC. The number of participants was 123, but the parameters remained otherwise unchanged from the preceding experiment leading to an upper bound of 0.1225. The mean actual error related to 10-fold CV was 0.4972 after 100 iterations (or 1000 values) while 0.5394 was the value associated to RUB after 1000 iterations. The p-values obtained were 0.0300 for CV and 0.0440 for RUB, again p-values were close or equal to α (0.05). Therefore, the number of false positives was the number expected at that significance level.

7.3.3 How relevant are the findings?

Apart from the already analysed outcomes, very similar results were obtained by applying KAGGLE-AD dataset, even in multiclass scenarios. The detailed study is provided in Appendix A.2, while its statistical results are presented in Table 7.6 together with those of the other experiments. It is also included the p-values and FWE rates related to Vapnik's upper bound correction using Equation (4.12). Furthermore, the trend of FP rates for several significance levels, like 0.001, 0.01, 0.05 and 0.1, in the Type I error experiment was analysed in Figure 7.7. Looking at the ideal performance, one can observe that the FP rate is always lower than the line of identity and that a performance closer to the ideal tends to occur using RUB.

To detect possible overfitting in classifiers using K-fold CV, the performance of cross-validation in terms of upper bounds is analysed. The capacity of generalisation can be analysed by studying the value and variability of the ratio μ/E_{emp} associated with each of the iterations performed (see Equation 4.15), both in original and permuted databases using 10-fold CV. The upper graph in Figure 7.8 is a comparison of the

	Permutation test: statistical power			Permutation test: Type I error		
	10-fold	RUB	R-Vapnik	10-fold	RUB	R-Vapnik
ADNI-AD	0.0010 [0.0010]	0.0010 [0.0010]	0.0010 [0.0010]	0.0410 [0.0063]	**0.0440 [0.0065]**	**0.0440 [0.0065]**
KAGGLE-AD (4C)	0.0040 [0.0020]	0.0010 [0.0010]	0.0010 [0.0010]	**0.0480 [0.0068]**	0.0330 [0.0056]	0.0330 [0.0056]
KAGGLE-AD (2C)	0.0010 [0.0010]	0.0010 [0.0010]	0.0010 [0.0010]	-	-	-
DIAN-AD	**0.0819 [0.0087]**	0.0020 [0.0014]	0.0020 [0.0014]	0.0300 [0.0054]	**0.0440 [0.0065]**	**0.0440 [0.0065]**

Symbol "-" is used to express equality with the row above it. R-Vapnik indicates that Vapnik's bound was used.
4C: four clases, 2C: two classes.

Table 7.6: Summary of the obtained p-values (statistical power study) and FWE rates (Type I error study). The level of significance was 0.05 in all experiments. Bold type indicates that the value exceeds the imposed significance level, either the value itself or its possible maximum value.

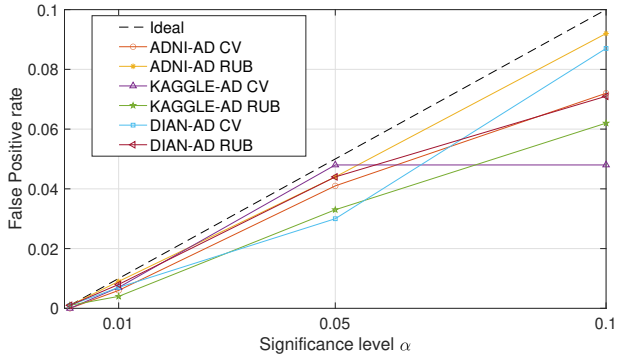


Figure 7.7: Estimated FP rate derived from the Omnibus test of the analysed methods at given significance levels (α=[0.001,0.01,0.05,0.1]). Data used are those already calculated in the Type I error control experiments, applying the significance level under analysis.

μ/E_{emp} ratios vs. the empirical risk obtained for the original and permuted datasets in the statistical power experiment. The lower graph shows the equivalence in the experiment related to false positives control.

Figure 7.8 reveals that the lowest empirical errors were associated to the non-permuted binary databases, in line with experimental results, e.g. in Table 7.2 or Table 7.4. Furthermore, ADNI-AD ratios were closer to zero and negative values, which means that empirical and actual errors were close to each other, which coincides with what is stated in Table 7.2. Regarding the binary KAGGLE-AD dataset, a large variation in the value of the ratio μ/E_{emp} was observed, which may indicate the existence of overfitting. In the permuted databases, the cloud of points associated with the ADNI-AD is particularly noteworthy with ratios that encompass negative values. This implies that the actual errors were lower than the empirical errors, and therefore the implemented model had good generalisation with permuted data, which is not ideal. With respect to the subsets of controls in the other two databases, their ratios were mostly positive, so it was not possible to declare generalisability.

In view of all the previous analyses, Table 7.7 summarises which validation method would be most relevant for each of the experiments. In statistical power, slightly

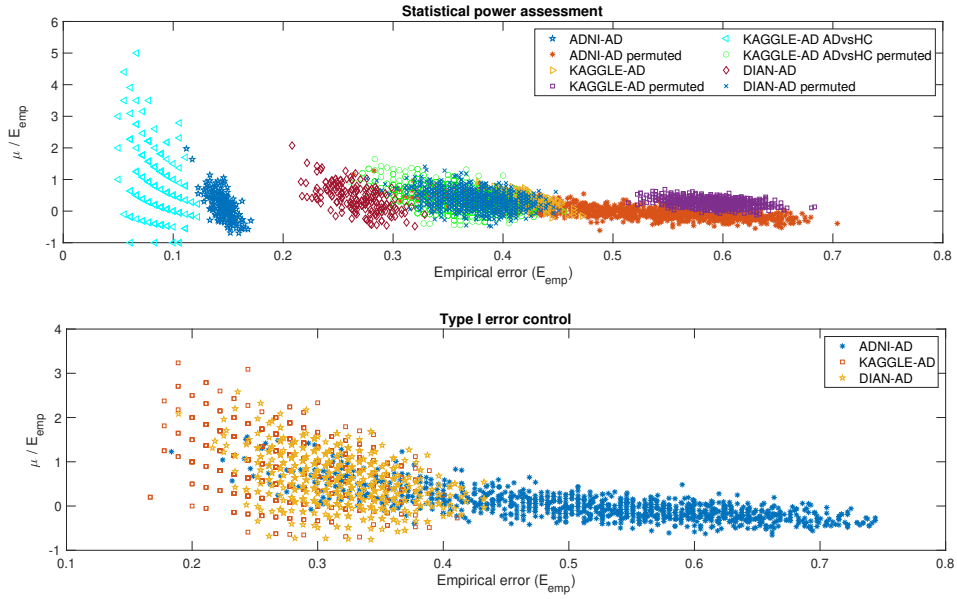


Figure 7.8: Ratio upper bound versus empirical risk, μ/E_{emp}, using 10-fold CV in the original datasets (20 iterations, 200 values) and permuted datasets (100 iterations, 1000 values). Values related to statistical power assessment (top) and type I error control (bottom) studies.

better results were observed using RUB, especially for heterogeneous data such as the multiclass experiment or with DIAN-AD. Regarding Type I error control, both validation methods were closely aligned, with a more stable trend and closer to the significance value set using RUB, see Figure 7.7. The larger the sample size, the more reliable results 10-fold CV generated, and with less variability. For example, as ADNI-AD sample is larger and more complex (3D), a better adaptation of the DL model using CV was observed. However, the increased complexity of the network also made it more adaptable to corrupted data.

Dataset	Experiment	Data Nature	Sample size	Average CC	Best VA
ADNI-AD	Power	3D	417	High	CV
	Type I error	3D	229	High	RUB
KAGGLE-AD	Power (4C)	1D	400	Medium-Low	RUB
	Power (2C)	1D	200	Low	RUB
	Type I error	1D	100	Low	RUB
DIAN-AD	Power	1D	246	Low	RUB
	Type I error	1D	123	Low	CV

4C: four classes, 2C: two classes, CC: computational cost, VA: validation approach.

Table 7.7: Identification of the best validation method based on the performance of the experiments. It includes parameters of interest in the evaluation.

7.3.4 Variability in feature extraction

The last analysis is to determine the influence of the feature extraction process on the classification capability of the model. Table 7.8 indicates the results obtained following the scheme described in Figure 7.3b. In general, statistical power and FWE rate were both similar to those obtained in the prior experiments, even lower values are obtained. This is consistent with the fact that by applying feature extraction to the original database and subsequently working in low dimensionality, feature extraction was not influenced by false labels and samples being shuffled, thus decreasing variability. The reader is referred to Appendix A.3 for more detailed results, where, for example, it can be seen that the accuracies associated with permutations were effectively centered on the 50%.

	ADNI-AD		**KAGGLE-AD**			**DIAN-AD**	
	Two-condition	One-condition	Four-condition	Two-condition	One-condition	Two-condition	One-condition
	***p*-value**	**FWE rate**	***p*-value**		**FWE rate**	***p*-value**	**FWE rate**
10-fold CV	0.0010 [0.0010]	0.0450 [0.0066]	0.0040 [0.0020]	0.0010 [0.0010]	0.0390 [0.0061]	0.1179 [0.0102]	0.0150 [0.0038]
Resubstitution	0.0010 [0.0010]	0.0490 [0.0068]	0.0010 [0.0010]	0.0010 [0.0010]	0.0390 [0.0061]	0.0010 [0.0010]	0.0340 [0.0057]
RUB	μ = 0.0665	μ = 0.0897	μ = 0.0679	μ = 0.0960	μ = 0.1358	μ = 0.0866	μ = 0.1225
	0.0010 [0.0010]	0.0490 [0.0068]	0.0010 [0.0010]	0.0010 [0.0010]	0.0390 [0.0061]	0.0010 [0.0010]	0.0340 [0.0057]

Table 7.8: Alternative scheme. Statistical power results (p-values) and Type I error results (FWE rate). Validation methods applied were 10-fold CV (100 iterations, medium computational cost, top), resubstitution (1000 iterations, low computational cost, middle) and RUB (by applying the upper bound, μ, low computational cost, bottom). The significance level of the test was 0.05.

7.4 Discussion

A non-parametric statistical inference framework based on permutations was applied to evaluate the performance of various neuroimaging scenarios, both binary and multiclass, as well as balanced and unbalanced. In order to assess such performance, statistical power and type I error control were analysed. In addition, an opportunity was provided to explore which validation method was more suitable: cross-validation or validation based on resubstitution.

In general terms, the architectures implemented allow a high statistical power, since in most of the experiments the p-value was 0.0010. Nevertheless, the null hypothesis of independence could not be rejected using CV in DIAN-AD (p-value=0.0819); 81 out of the 1000 actual errors associated with the null distribution were lower than the actual error of the original database. Thus, a tendency that statistical power is slightly lower using CV than RUB was observed. This particularly occurred in cases where the accuracy rates related to the original datasets were not excessively high, as in DIAN-AD or KAGGLE-AD 4-condition datasets. The main reason is the heterogeneity of these datasets and therefore, by using subsets of the sample instead of the whole dataset, a higher variability appears.

Good control of false positives was also observed in all experiments, as the FWE rate was never greater than the significance level, 0.05. Nevertheless, if the standard deviation associated with the p-value was considered, its maximum could be greater than 0.05. This occurred using CV in KAGGLE-AD 4-condition dataset, and ADNI-AD and DIAN-AD using RUB, see Table 7.6. Previous studies [15] indicated the need for a high sample size in order to apply K-fold and obtain a low FWE, which is confirmed in this study since 100, 123 or 229 are small sample sizes. For example, the highest FWE rate was 0.0480 obtained in the case of having only 100 samples (KAGGLE-AD).

Both, the control of Type I error and the statistical power remain unaffected by performing the permutation loop after the feature extraction step. Nevertheless, it was observed a reduction in the standard deviation by keeping the feature extraction phase out of the permutation test. Therefore, despite the fact that AE are a unsupervised learning technique that does not depend on the labels, by not including AE training in the permutation loop, variability influenced by, e.g, the position of the samples in the set or other optimisation parameters, has been eliminated. The variability generated in feature extraction is precisely one of the points that wanted to be highlighted in this work. The analysed scenario, where a reduced dimensionality was used directly, does not represent the current scenario in neuroscience, which is orientated more towards classifications using high complexity classifiers in complex spaces than on the use of linear SVM and low dimensional classifiers. Thus, if some variability is observed in a low dimensional space as in this case, what can happen in today's highly complex classification scenarios? There is a growing need for adequate instruments to assess the reliability of ongoing classification systems, and RUB could be one of them.

Regarding the use of CV, resubstitution or RUB as validation methods, similar patterns were found in the accuracy distributions associated with CV test results and resubstitution when real labels were used. For example, in ADNI-AD statistical power experiment, the CV test accuracy is 0.8319 [0.0541] and it was obtained 0.8321 [0.0059] using resubstitution, see Table 7.2. Logically, the standard deviation is larger in CV as different subsets of the sample are used in each iteration. The situation differs for permuted datasets. In both studies (statistical power and Type I error) similarities were found between the mean values related to the CV test subset and RUB, with the standard deviation being smaller in the distributions obtained by RUB. For example, in DIAN-AD statistical power study (see Table 7.4), when the real labels were used the mean accuracy using RUB was 0.4775 and 0.4963 using CV. Therefore, the use of RUB, far from generating positively biased results, estimates a worst-case accuracy similar to or even more conservative than that obtained with the CV test subset, with the advantage of using the complete set (more reliable).

Exploring cross-validation in more detail, the possibility of overfitting that may potentially arise was analysed by means of the generalisation error, i.e. the discrepancy between the accuracy associated with the training subset and the test subset. Apart from the observation of such discrepancy, this can be done by analysing the variation in the ratio μ/E_{emp}, see Equation (4.15). For example, this variability is clearly seen in

the one-condition set of DIAN-AD in Figure 7.8, which is consistent with the separation between the empirical error (0.6830) and the actual error (0.5028) shown in Table 7.5. It follows that deeper configurations, as the one related to ADNI-AD, can develop a greater capacity for generalisation by increasing its complexity and learning capacity of the data. Thus, although it is a positive factor when the database is large and well labelled, one should be careful with the choice of data being trained as the results may be satisfactory even when they should not be, detecting differences when they should be absent.

In a nutshell, it was observed that in small sample size scenarios, resubstitution has a similar performance to cross-validation. Moreover, it was verified that the use of the former to establish a two-sample test based on classification errors is equally or more valid than using CV, as indicated by [128]. Furthermore, the application of RUB eliminated the positive bias associated with resubstitution when performing the permutation test. Therefore, it could be believed that the use of RUB in small sample size scenarios, as is the case in neuroimaging, is highly recommended. The problem of splitting the set into different subsets is avoided, which tends to generate imbalance and increased variability. In addition, the computational cost is also greatly reduced and the bias associated with resubstitution is suppressed. A disadvantage of applying RUB to analyse DL models is that the final classifier should be ideally linear, as the computation of the upper bound becomes more complex in other circumstances. Future work is expected to be able to compute these more complex bounds in classification neural networks.

8 | STATISTICAL AGNOSTIC MAPPING

8.1 Introduction

In the previous chapters, the usefulness of ML, which includes DL techniques, in pattern recognition problems related to the diagnosis and understanding of brain conditions has been observed. Hence, data-driven approaches are gradually replacing classical statistics to address neuroimaging problems [182, 216]. These problems include a low signal-to-noise ratio, difficulty in estimating effect sizes and small sample sizes. These challenges can lead to a lack of generalisation of results between different datasets [15] and instability across reports, regarding the numerous parameters that need to be considered before performing analyses [217]. For example, such generalisation capability has been analysed in chapter 7, where RUB is proposed as a potential validation method of choice for databases with small sample sizes. This proposal

remains an essential element of this chapter, where the application of ML in brain mapping is explored.

Statistical brain mapping is a simple approach to detect ROIs when comparing two groups of subjects, i.e. HC and AD subjects, and detect relevant patterns between them. A group analysis can be helpful to delineate the ROIs of the brain for a specific condition/pathology [218]. Thus, when an exploratory analysis is required, statistical inference maps based on null-hypothesis (H_0) are commonly used (see section 3.4). Nevertheless, more and more studies are proposing new approaches due to its dependence on a variety of assumptions [34, 219]. Data-driven SLT would be one of those proposals [128]. Although, ML approaches were not originally designed to test hypotheses in brain mapping, a detailed discussion can be found at [30, 32, 33, 220, 221], they are theoretically grounded to provide confidence intervals in the classification of image patterns (protected inference) that can be seen as maps of statistical significance. This can be achieved by assessing the upper bounds of the actual error in a binary classification problem (a confidence interval), and by using simple significance tests of a population proportion within it. On this basis, a data-driven approach based on concentration inequalities is described, named as Statistical Agnostic Mapping, which was published in [37].

8.2 Statistical Agnostic Mapping (SAM)

The main objective of this framework is to generate significance maps based on the data rather than on classical statistics [85]. The significant areas derived from Statistical Agnostic Mapping (SAM) correspond by construction with those regions having an empirical error E_{emp} that, under the worst case scenario, has associated an actual error E_{act} greater than the random guess accuracy $\pi = 0.5$. Confidence intervals derived from the concentration inequalities allow to bound the worst case at the "upper" border of the confidence interval, providing a protective inference. Thus, within this confidence interval, a significance test can be used to make an inference about whether the accuracy value for a specific region differs from the null-hypothesis of the random proportion $\pi = 0.5$. Therefore, the statistical significance of any region is assessed, in combination with confidence intervals, by evaluating the p-value of any ROI at a given significance level, i.e. $\alpha = 0.05$.

In terms of classical statistics such significant regions are derived as the following. Given a set of regions $j = 1, \dots, l$ the accuracy for each region is evaluated under the worst case using Equation (4.13) by the following hypothesis test:

$$\begin{aligned} H_0 &: Acc^j > \hat{Acc}; \quad \text{region j is significant} \\ H_1 &: Acc^j < \hat{Acc}; \quad \text{region j is not significant} \end{aligned} \tag{8.1}$$

where Acc^j is the estimated actual accuracy in the classification of region j with probability $1 - \delta$, and $\hat{Acc}$ is the averaged proportion of subjects correctly classified in all regions within the confidence interval.

In more detail, the chosen test for a proportion based on prevalence to achieve population inference is as follows. Let denote $\hat{\pi}$ the sampling distribution of empirical errors E^i_{emp}, for $i = 1, \dots, m$, then the null hypothesis test about the population proportion within the confidence interval has the form:

$$H_0 : \hat{\pi} = \pi_0 \quad ; \quad H_1 : \hat{\pi} > \pi_0$$

where π_0 denotes a particular proportion value between 0 and 1, i.e. 0.5. The test-statistic in a population proportion is:

$$z = \frac{\hat{\pi} - \pi_0}{\sigma_0} \tag{8.2}$$

where $\sigma_0 = \sqrt{(\pi_0(1 - \pi_0))/m}$. For large samples, i.e. for $\pi_0 = 0.5$ at least $m = 20$, if H_0 is true, the sampling distribution of the z test statistic is the standard normal distribution. It should be noted that other kind of tests could be applied as well, like those described in [222].

Concerning the classification scenario implemented in SAM, a linear decision function is selected for two reasons. The first is that such functions have have already been successfully used in neuroimaging [185, 209, 223]. The second is that the upper bound of inequality is dependent on the complexity of the classifier [131]. Therefore, with linear functions the upper bound can be minimised. Thus, the classification algorithm adopted in this implementation is SVM with linear kernel, obtaining Equation (4.13) for the upper bound (see section 4.5.2).

Moreover, a feature extraction step is also applied in first place. In this case, the chosen method is PLS. PLS methods have demonstrated its utility in describing the relationships between brain activity and experimental design or behaviour measures within a multivariate framework [213, 224]. This reduces the ratio between feature dimension and sample size (*curse of dimensionality*) and generates smaller upper bounds for the same sample size (see Figure 4.3).

Therefore, this methodology can be seen as a region-wise analysis where a fitted linear SVM classifier in a multivariate feature space is implemented. The motivation for a multivariate framework in assessing the areas of relevance is analogous to other proposed techniques for addressing the multiple comparison problem in functional imaging, e.g. RFT for neuroimaging analysis [225], random/mixed/conjunction analyses in multiple p-value maps [95] or the classical p-value corrections for multiple comparison after null-hypothesis testing [85]. In general, only those voxels (or ROIs) showing a tight association, i.e. high performance in terms of accuracy, should be considered as relevant maps or patterns in that particular condition with probability $1 - \eta$.

8.2.1 Summary of the procedure

The different steps of the procedure already described can be summarised as follows, which is illustrated in Figure 8.1 (middle and right columns):

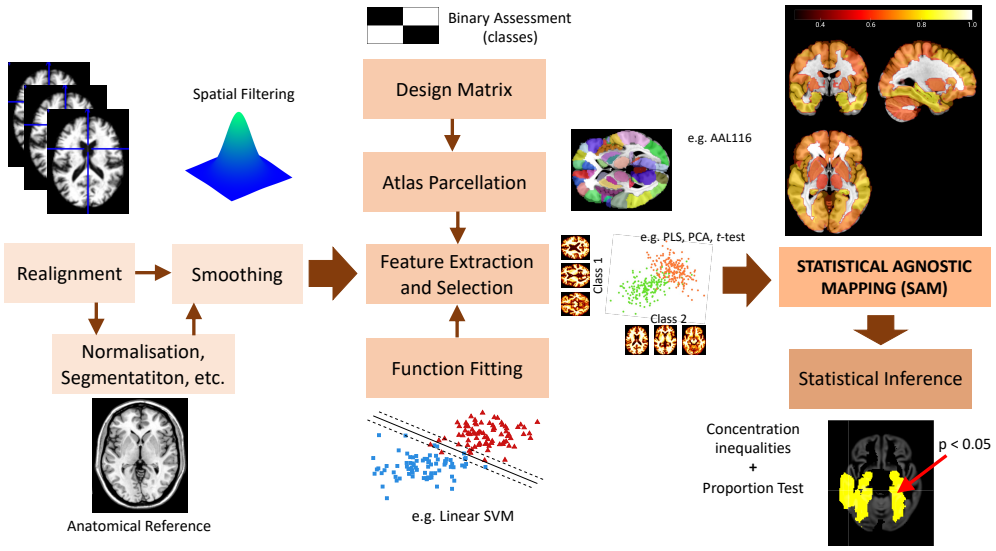


Figure 8.1: Complete diagram of SAM including typical preprocessing steps in SPM for different modalities (left column of blocks), classification fitting and feature extraction and selection for actual risk estimation (middle column) and inference to derive the statistical map (right column).

- **Step 1: Data Preparation and parcellation:** Design the group comparison (design matrix) and select the regions (ROI) to be analysed across subjects. By default, the AAL116 atlas [215] is used in the implementation.
  Then, *for* each ROI do:

- **Step 2: Training feature set:**

  - Apply a feature extraction and selection stage, by default based on PLS, to the ROI and obtain $\mathbf{Z^n}$ (the feature space)
  - Train a linear SVM by empirical risk minimisation to obtain the function that best fits the data (resubstitution estimation).

- **Step 3: Assessment of concentration inequalities:**

  - Compute empirical error (or accuracy) (Equation (4.10)).
  - Determine the actual accuracy Acc^j under the worst case with probability $1 - \eta$ (Equation (4.11))

  end *for*

- **Step 4: Statistical assessment of the accuracies:** Calculate the z-test statistic for each actual accuracy in $\{Acc^j\}$ (Equation (8.2)) for testing significance.

This procedure allows for working with small sample sizes, detecting large and subtle relevant effects in studies of clinical and experimental conditions using neuroimaging techniques, including structural and functional resonance imaging (fMRI/sMRI), SPECT or PET [37, 40, 226]. In this chapter three case studies applying SAM are detailed. The first one analyses structural imaging (sMRI) using ADNI-AD dataset. The second one analyses functional imaging (SPECT) and sMRI using PPMI-PD dataset. Finally, an EEG study using UGR-COG dataset is included to standardise the procedure in temporal studies.

8.3 A structural MRI study: ADNI-AD

The aim of this section is to experimentally validate the proposed procedure, as well as to compare its performance with the accepted framework used by the neuroscience community based on the SPM analysis (see section 3.4.2). In contrast to SPM, SAM is based on data-driven approach based on concentration inequalities. Therefore, the main difference with the former is that the latter is a non-parametric method.

In this case, a two-group comparison AD vs. HC is considered for establishing a clear framework for comparing both statistical paradigms (SAM and SPM), as the observable differences in structural imaging between the two classes are well-established in literature [190, 227, 228].

8.3.1 Methodology

In this experiment, statistical maps are generated from the MRI scans of the ADNI-AD dataset (see section 5.1.1). To obtain statistical agnostic maps, first of all, the proposed methodology try to fit in an optimal way a linear SVM classifier in the feature space obtained after applying PLS as a feature extraction approach. Then, the empirical error is estimated for each of the 116 standardised regions in which the brain volume is parcellated. This estimation is corrected (actual error) by the use of upper bounds drawing a novel set of accuracy values (proportions) and a confidence interval. To do so, in addition to Equation (4.13), Vapnik's bound (Equation (4.12)) is also evaluated. Then, all those relevant regions for the characterisation of AD based on absolute values are heuristically identified in terms of hypothesis testing within confidence intervals.

The methodology employed when using SPM for comparative purposes is as follows: first, a first-level analysis to derive the GLM for the dataset under assessment (a design matrix for group comparisons) is conducted. Then, in the 2nd-level analysis, the contrast images are fed into a GLM for implementing the statistical test. Two types of two-sample t-test-based inferences are performed using FWE and uncorrected p-value to finally obtain the statistical parametric maps.

In order to thoroughly assess the performance of both methods, a comparative analysis is conducted not only in terms of their capability to detect significant regions

but also in terms of FP control, FWE rate, and TP rate. This evaluation encompasses the consideration of various sample sizes and confidence levels within the different approaches being analysed. This allows to obtain a comprehensive understanding of the methods' performance in small sample size scenarios.

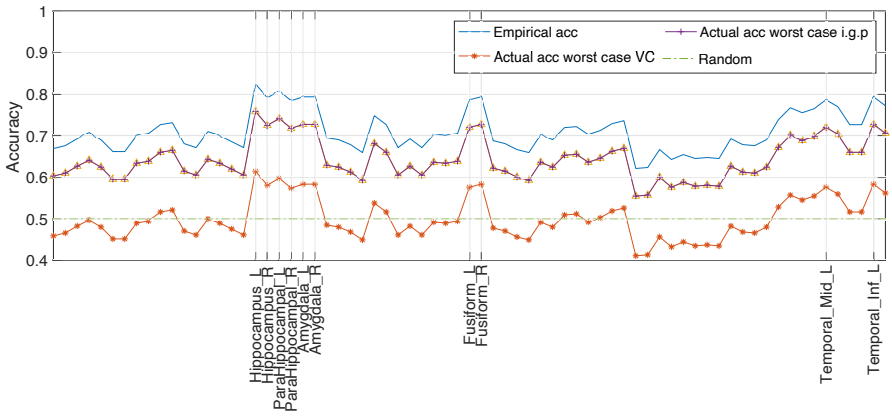
8.3.2 Results

For the SAM approach, the number of PLS components chosen to fit the linear SVM classifier was 1, i.e, the first PLS component extracted by this regression analysis is applied. This PLS score for each subject can be conceptualised as the representation of the subject into a multi-dimensional reference system as described in [229]. Figure 8.2a shows the empirical error obtained for each of the 116 standardised regions (solid blue line). To improve the visibility of the plot, only regions #30 to #90 from the AAL116 atlas are shown. The actual error is estimated by applying the upper bounds related to Equation (4.13) and the VC approach (Equation (4.12)). For the computation of these bounds, the number of samples (n = 417), the feature's dimension (d = 1) and a confidence level of 95% were considered. Thus, the lower accuracies in Figure 8.2a corresponds to the worst cases as considered by the selected concentration inequalities. Accuracy values of several regions are highlighted, which are relevant in the biological definition of AD, i.e. Hippocampus, Temporal, Amygdala and Parahippocampal regions, corresponding to peaks of these curves. Moreover, it can be observed how the VC approach is more pessimistic than the one based on Equation (4.13) [31].

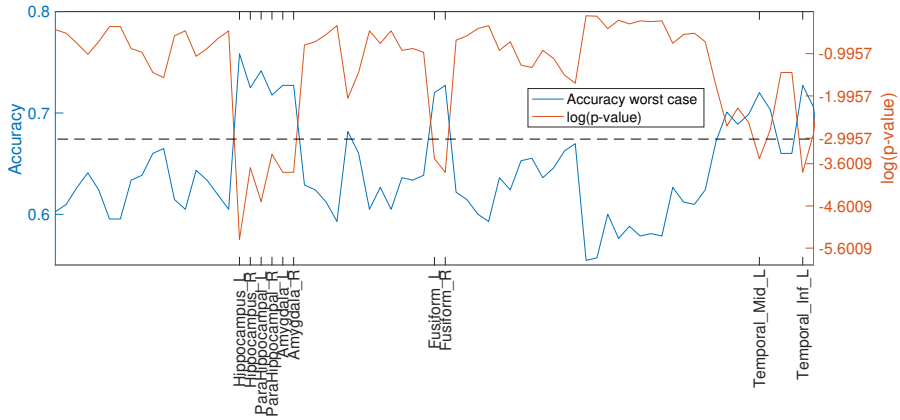
Figure 8.2a shows the next step of the proposed methodology, the statistical assessment of the accuracies by means of a significance test for a proportion. Thus, all those relevant regions for the characterisation of AD were identified based on absolute values (accuracy). To estimate this, the same p-value as the one of confidence intervals using concentration inequalities is applied (α = 0.05). Note that it is shown the probability of observation of the set of accuracy values under H_0, i.e. random distribution. The resulting significant regions are the same as those highlighted in Figure 8.2a.

A direct comparison with the SPM approach is shown in Figure 8.3 and Figure 8.4, in terms of the sample-size analysis and the relevant regions determined by both methods. In these scenarios, a voxel-wise inference using a standard two-sample t-test with FWE p-value= 0.05 (and null extent threshold -voxels-) is computed in SPM12. Key to this comparison is the different working operations, i.e. SAM includes the spatial structure of data at the first feature extraction ans selection stage, whilst SPM do it at the final stage, by means of RFT. For this reason, SPM is more specific (voxel-wise) but widespread comparing to SAM. The number of identified ROIs conforming the SPM increases as the number of sample increases, unlike the proposed approach, which provides the same volumetric differences for $n = 200, 417$.

Figure 8.4 shows that main regions identified by SPM are included in the ROIs deployed by SAM-based approach. In addition, the number of "activated" voxels in SPM is associated with sample size and these voxels are widespread across several anatomical



(a) Classification results in standardised ROIs.

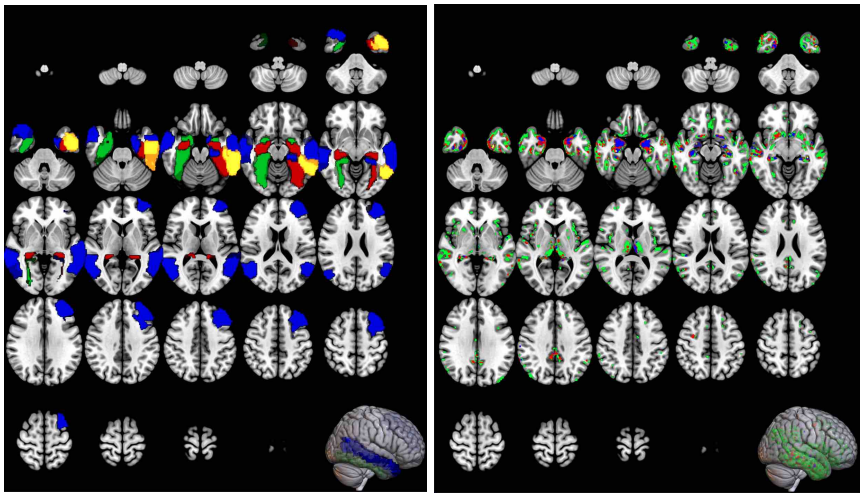


(b) Significance test for a proportion.

Figure 8.2: (a) Accuracy values and upper bounds in 116 standardised regions of interest (only significant regions from #30 to #90 are shown) for two methods based on concentration inequalities (Equation (4.13)) and Equation (4.12)). The confidence interval is drawn in the space between the solid blue line and the colored lines. (b) Accuracy values in the worst case (Equation (4.13)) and the set of probabilities (log(p-values)) within the confidence interval. The regions of interest (p< 0.05) are detected using a significance test for a proportion π.

regions. The number of voxels in ROIs obtained by SAM is almost independent on the sample size, except for the extreme case $n = 50$, and given the magnitude of the effect being sought in the HC-v-AD comparison.

Moreover, it should be analysed the ability of the proposed method for controlling the FWE rates for voxel, clusterwise or regionwise inferences as shown in [34]. To this purpose two groups of subjects ($n = 50, 100, 228$) are randomly drawn from a relatively large ($n = 228$) group of HC from the dataset, where the null hypothesis of no group difference in brain activation should be true.

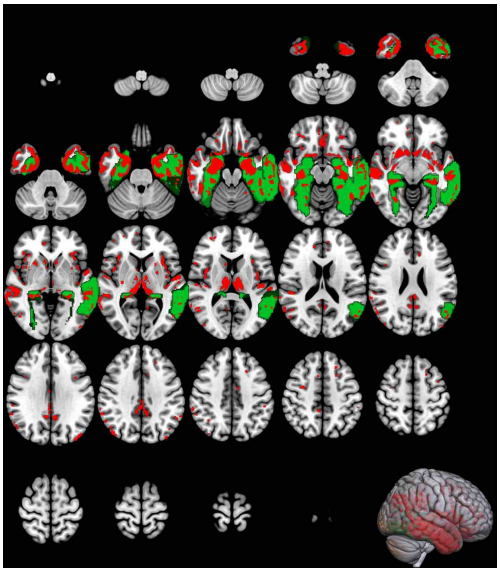


(a) SAM. **(b)** SPM.

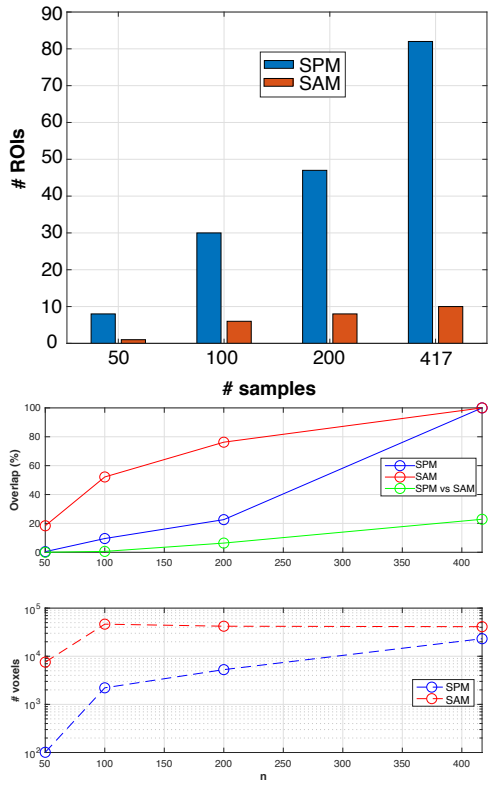
Figure 8.3: Statistical comparison of brain volumes using SAM and SPM in the ADNI-AD dataset. Green area corresponds to the whole dataset while the rest of colors (red, blue, yellow) are linked to data subsets, which are plotted in increasing n (opacity of representations is preserved for clarity reasons). The ROIs selected for increase $n = 50, 100, 200, 417$, satisfy $S_j \subset S_{j+1}$ except for $n = 50$ where an additional region "Frontal Mid L" is selected. It is worth mentioning that all the ROIs extracted in different sample-size configurations were included in the confidence interval and with probability slightly higher than the significance level ($\alpha = 0.05$).

A total of 48k random group analyses were performed, following the same steps as in the previous section (parcellation, feature extraction and selection, function fitting, etc.) to compute the empirical FP rates of SAM and SPM (in cluster and voxelwise inferences). Regarding the SPM study, two types of two-sample t-test-based inferences were performed using FWE and uncorrected p-value. Each statistic map was first thresholded using a p-value= 0.001 (uncorrected for multiple comparisons). The degree of FP to compute the FWE rate was finally estimated as the number of significant voxels within any of the 116 atlas regions, meaning a voxelwise inference. Furthermore, a conservative clusterwise inference is applied by using uncorrected p-values, where the surviving clusters are then compared with a cluster extent threshold-based criteria in regions (at least 25% of activated voxels in any of the 116 regions of the atlas). The estimated FWE rates are simply the number of analyses with at least a significant result divided by the number of analyses (1.000).

Figure 8.5 illustrates similar results as those described in [34] that were obtained using a conservative voxelwise inference and invalid clusterwise inference for the two-sample t-test used in these simulations. SAM provided a conservative operation mode when the Vapnik's bound (Equation (4.12)) was applied to the empirical data (often falling below the significance level, e.g. 5%). It also performs a very realistic and competitive approach when the estimation of the upper bound in Equation (4.13) is used, named as i.g.p. (in general position), as shown in [31]. In particular, FWE



(a) SPM vs SAM.



(b) ROIs detection and overlap analyses.

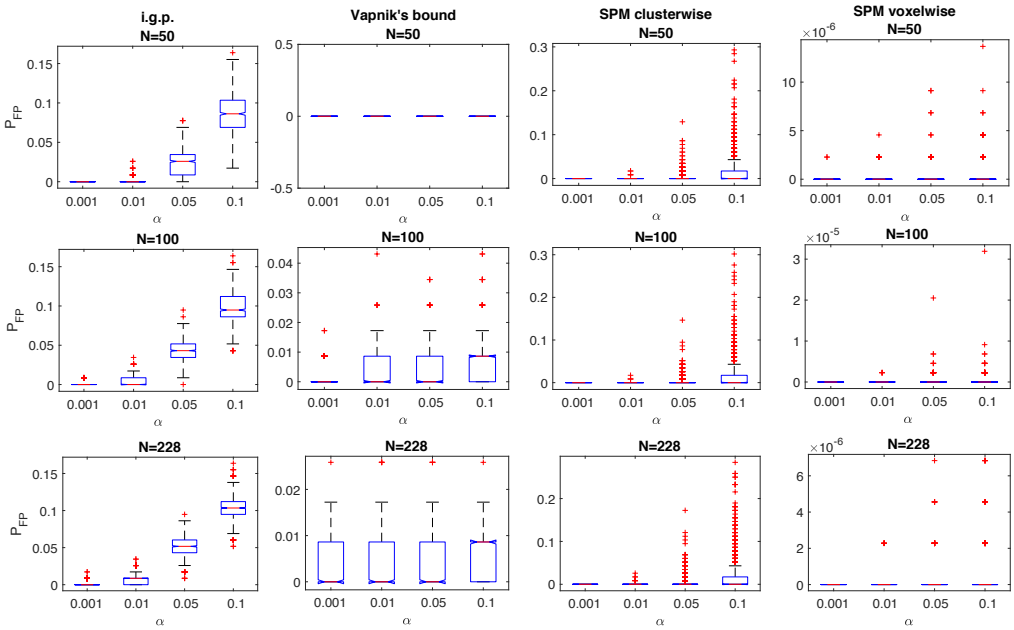
Figure 8.4: (a) SPM (red) over SAM (green) using the complete ADNI-AD dataset ($n = 417$). (b) Number of regions of interest vs. sample size (top) and overlap analysis vs. sample size (bottom). Observe how the SPM activation map linearly increases with n and is located on more than 80 standardised regions with the whole dataset (although part of these isolated activation voxels could be removed from the map using the extent threshold).

rates for clusterwise inference far exceed their nominal level using SPM (using a p-value= 0.001 uncorrected for multiple comparisons), despite the surviving clusters were then compared with an cluster-extent threshold based criterion of 25% of activated voxels in that region for a fair comparison with SAM. On the other hand, parametric voxelwise (SPM) and regionwise (SAM, e.g. Equation (4.12)) based inferences are valid but over conservative, often falling below the predefined levels of significance, $\alpha = 0.001, 0.01, 0.05, 0.1$. However, estimated FWE rates based on corrections described in [31] are close to the predefined levels, and are almost independent on the number of subjects that were randomly drawn in the simulation.

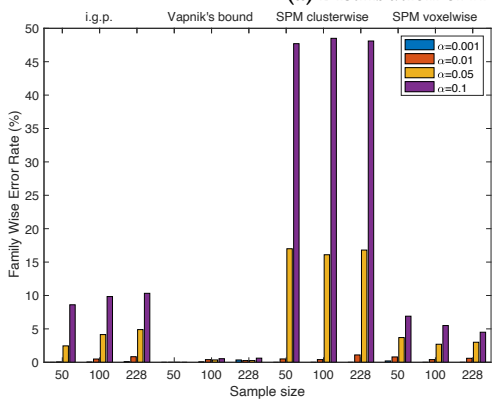
Finally, the TP rate[1] (i.e. the null hypothesis of no group differences in brain activation should be false) may also be assessed on balanced groups of subjects ($n = 50, 100, 150, 300$) that are randomly drawn from a relatively large ($n = 229 + 188$) group

[1]The estimated TP rates are simply the probability of a region to be activated, e.g in voxelwise SPM $P_{TP} = \frac{\#TP}{116}$

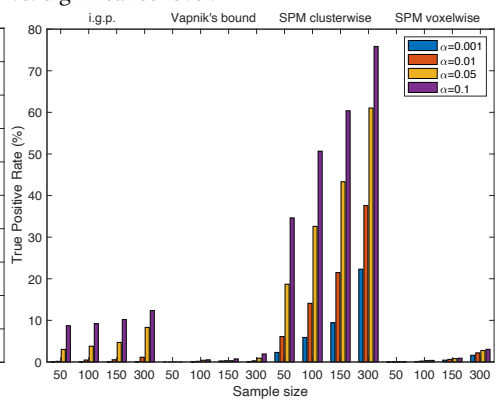
of HC and AD from the ADNI-AD dataset (making a total of 64K analyses). At a given significance level, the TP rate should be almost constant for different sample sizes (i.e. the method provides the same number of significant regions in each experiment). As clearly shown in Figure 8.5c, the SAM methodology provides almost constant TP rate with different sample sizes, unlike SPM TP rate in cluster and voxel-wise inferences, which clearly increases with sample size.



(a) Distribution of FP vs. significance level.



(b) FWE rate vs. sample size.



(c) Experimental TP rate vs. sample size.

Figure 8.5: Results for two-sample *t*-test and ad-hoc clusterwise/voxel inference in regions, showing estimated FWE rates and TP rates for four different activity paradigms (FWE SPM, uncorrected SPM, Vapnik SAM, and i.g.p. SAM). These results were generated using {50; 100; 228} subjects in each group analysis for FWE rate and (N = 50, 100, 150, 300) for TP rate. Note: unc. SPM: clusterwise inference, FWE SPM: voxelwise inference.

8.3.3 Discussion

As seen in the results obtained, in general the SAM is a very robust method, in terms of sample size, to find relevant standardised areas, and a stable framework which contains those regions defined as relevant by the SPM, with sufficiently large sample size. Thus, this data-driven approach, mainly devoted to classification problems with limited sample sizes, is able to derive statistical model-free (agnostic) mappings. Although the latter is *not designed for testing competing hypothesis or comparing different models* in neuroimaging, the SAM is derived assuming the existence of classes (H_1), at voxel or multi-voxel level. It is worth mentioning that SAM employs the concept of prevalence [219] to derive the activation maps, since it is the result of the classification performance in terms of accuracies or proportions in the feature space.

The behaviour of the analysed methods depends on the size of effect of interest. In the seek of subtle effects, such as the ones found in AD or Autistic patterns, and provided that hypothesis tests cannot separate important, but subtle, and actually trivial effects [33], SAM focus on standardised ROIs to avoid the presence of false positives in the sought maps. In this sense, SPM is more specific and can detect, within these regions sought by SAM, which substructures are responsible for the discrimination between classes. Nevertheless, this voxelwise analysis could be carried out as well using this framework, e.g. by assessing the PLS-maps derived at the feature extraction and selection stage as shown in [229]. However, it was additionally found that using the SPM univariate approach i) small effect sizes in a heterogeneous population with a limited sample size fail to be detected whilst with larger samples sizes their detection overshoots, and ii) in large sample sizes it can yield highly significant p-values even when effect sizes are so small that they become trivial in practical terms.

On the other hand, when large effects are bound to be found, SAM is a suitable method in their detection since, with a few amounts of samples, it provides similar results than the ones obtained with complete datasets (Figure 8.4). This is in line with the main idea derived from [220] that when an effect is found in small datasets is more than likely to be extrapolated in large samples. On the contrary, only in small datasets with small – but meaningful – effects that are missed, missing data, sampling bias, etc. the absence of replication is found, i.e. across data collecting sites [33].

Finally, it has been seen the usefulness of the confidence intervals derived for the SLT based on concentration inequalities to achieve a confidence framework beyond sharp null-hypothesis testing. Key to this methodology in the field of SLT is that it is based on in-sample estimates (a similar procedure in exploratory analysis using hypothesis testing), unlike the out-sample estimates in CV procedures, which usually subdivide the (small) datasets for an estimation of the actual error. In this way, an analytical bound depending on sample size (n) and number of predictors (d) defines a "worst-case" operation point. Nevertheless, the experiments showed the application of a systematic hypothesis test for the selection of significant empirical errors which conforms the highlighted regions in the SAM. Only in this case, a model is assumed in

the set of accuracies, but it has been demonstrated to be in accordance with the nature of the one-dimensional data and sufficiently accurate for the purposes.

8.4 Statistical Mapping on Parkinson's Disease

Traditionally, studies about PD have made use of functional SPECT images. Nevertheless, to avoid some of its disadvantages with special focus on its high cost, the unreliable supply of radiotracer, and potential adverse effects like disorders of the heart beat (including cardiac arrest) [230], in the last years many studies [231, 232, 233] have tried to use another imaging alternatives such as MRI scans able to evaluate subtle changes in the GM tissue. This study presents a qualitative analysis of ROIs when using MRI for PD by the computation of statistical significance maps. Moreover, it allows to assess the reliability of the results generated by neuroimaging software tools to identify the most relevant ROIs when comparing HC and PD patients using MRI and ^{123}I-FP-CIT SPECT images.

8.4.1 Methodology

In this experiment, the PPMI-PD dataset is used (see section 5.2) to generate statistical maps in order to compare structural (sMRI) and functional (SPECT) imaging for the detection of ROI in PD. To conduct this comparison, significance maps will be generated using both parametric and non-parametric procedures, specifically the standard SPM12 package and SAM, respectively.

To generate and evaluate the significance maps associated with each statistical mapping tool, three experiments are proposed. The main one is a two-group analysis comparing HC vs. PD whereas the other two experiments confronted samples within the same class arbitrarily separated into two groups: i.e. HC vs. HC and PD vs. PD. While in the first case, the number of input samples per class is 40 (balanced classes), in the other two experiments, the set is randomly divided into two balanced subgroups of 20 samples.

8.4.2 Results

Once the images were preprocessed as indicated in section 5.2, they were given as inputs for the statistical mapping tools. When using SPM12, a standard two-sample t-test was conducted. Firstly, a factorial design along its design matrix was generated. Then, the parameters estimation of the GLM was performed. Finally, a t-contrast test (HC>PD) was defined and applied to the data. Here, two different inference models were studied: a voxelwise inference approach, where a FWE p-value$\leq$ 0.05 was applied; and a clusterwise approach where an uncorrected for multiple comparisons p-value $\leq$ 0.001 along with an extent threshold of 10 voxels was applied. In case of using SAM,

Experiment	Method	# Voxels	# ROIs	List of ROIs
HC vs. PD	SPM12 (voxel)	4583	12	$Insula^{L}$,$Insula^{R}$,$Hippocampus^{L}$,$Hippocampus^{R}$,$Amygdala^{L}$,$Amygdala^{R}$, $Putamen^{L}$, $Putamen^{R}$,$Pallidum^{L}$,$Pallidum^{R}$,$Thalamus^{L}$,$Thalamus^{R}$
	SPM12 (cluster)	20662	46	$Frontal\_Sup\_Orb^{L}$,$Frontal\_Inf\_Oper^{L}$,$Hippocampus^{L}$, $ParaHippocampal^{R}$, $Amygdala^{R}$, $Caudate^{R}$,$Putamen^{R}$ or $Thalamus^{R}$, among others
	SAM	12933	6	$Insula^{R}$,$Putamen^{L}$,$Putamen^{R}$,$Pallidum^{L}$,$Pallidum^{R}$,$Thalamus^{L}$
HC vs. HC	SPM12 (voxel)	0	0	
	SPM12 (cluster)	327	7	$Frontal\_Sup\_Orb^{R}$,$Frontal\_Mid\_Orb^{R}$,$Sup\_Motor\_Area^{L}$, $Cingulum\_Post^{L}$,$Calcarine^{L}$,$Precuneus^{L}$,Vermis_4_5
	SAM	24526	5	$Frontal\_Inf\_Tri^{L}$,$Fusiform^{R}$,$Precuneus^{R}$,$Caudate^{L}$,$Thalamus^{L}$
PD vs. PD	SPM12 (voxel)	0	0	
	SPM12 (cluster)	53	2	$Calcarine^{R}$,$Thalamus^{R}$
	SAM	12202	2	$Frontal\_Inf\_Tri^{L}$,$Cerebelum\_Crus1^{R}$

Table 8.1: Summary of the statistical maps obtained using SPM12 and SAM for ^{123}I-FP-CIT SPECT images. Experiments highlight regions of interest among different or same classes (HC and PD). The names of the regions are based on the AAL116 atlas nomenclature.

the images were parcellated according to the AAL atlas of 116 regions. The feature selection method chosen was *t*-test and the feature extraction method was PLS using 1 component. Then, the features were classified using a linear kernel SVM classifier and its accuracy was determined under the worst case with probability 95%. Finally, significance of each region was tested by means of a z-test statistic applied to each actual accuracy (see Equation (8.2)). In contrast to the previous methods, this would be a region-wise inference. Table 8.1 and Table 8.2 illustrate the results obtained for each experiment and image modality.

Experiment	Method	# Voxels	# ROIs	List of ROIs
HC vs. PD	SPM12 (voxel)	0	0	
	SPM12 (cluster)	98	10	$Frontal\_Sup^{R}$,$Frontal\_Mid^{R}$,$Frontal\_Inf\_Tri^{L}$, $Fusiform^{R}$,$Postcentral^{R}$,$Parietal\_Sup^{R}$,$Precuneus^{R}$, $Putamen^{R}$,$Cerebelum\_Crus1^{R}$,$Cerebelum\_6^{R}$
	SAM	31525	4	$Precentral^{R}$,$Precuneus^{R}$,$Temporal\_Inf^{L}$,$Temporal\_Inf^{R}$
HC vs. HC	SPM12 (voxel)	0	0	
	SPM12 (cluster)	83	11	$Frontal\_Inf\_Oper^{L}$,$Frontal\_Inf\_Tri^{L}$,$Olfactory^{R}$,$Insula^{R}$, $ParaHippocampal^{L}$,$Cuneus^{L}$,$Lingual^{L}Occipital\_Mid^{R}$, $Paracentral\_Lobule^{L}$,$Temporal\_Sup^{R}$,$Temporal\_Mid^{R}$
	SAM	0	0	
PD vs. PD	SPM12 (voxel)	0	0	
	SPM12 (cluster)	102	10	$Frontal\_Sup^{R}$,$Frontal\_Mid^{L}$,$Frontal\_Mid^{R}$,$Supp\_Motor\_Area^{L}$, $Cingulum\_Mid^{L}$,$Cingulum\_Mid^{R}$,$Precuneus^{L}$, $Paracentral\_Lobule^{R}$,$Temporal\_Sup^{L}$,$Cerebelum\_Crus1^{R}$
	SAM	42163	5	$Precentral^{R}$,$Lingual^{L}$,$Postcentral^{R}$,$Precuneus^{L}$,$Temporal\_Mid^{L}$

Table 8.2: Summary of the statistical maps obtained using SPM12 and SAM for GM MRI scans. Experiments highlight regions of interest among different or same classes (HC and PD). The names of the regions are based on the AAL116 atlas nomenclature.

Figure 8.6 depicts the significance maps generated by the proposed inferences for the ^{123}I-FP-CIT SPECT scans. The colour blue is associated to the significance map obtained using SAM. Colours red and green are related to the clusterwise and voxelwise inferences conducted in SPM12. The same range of colours is used in the

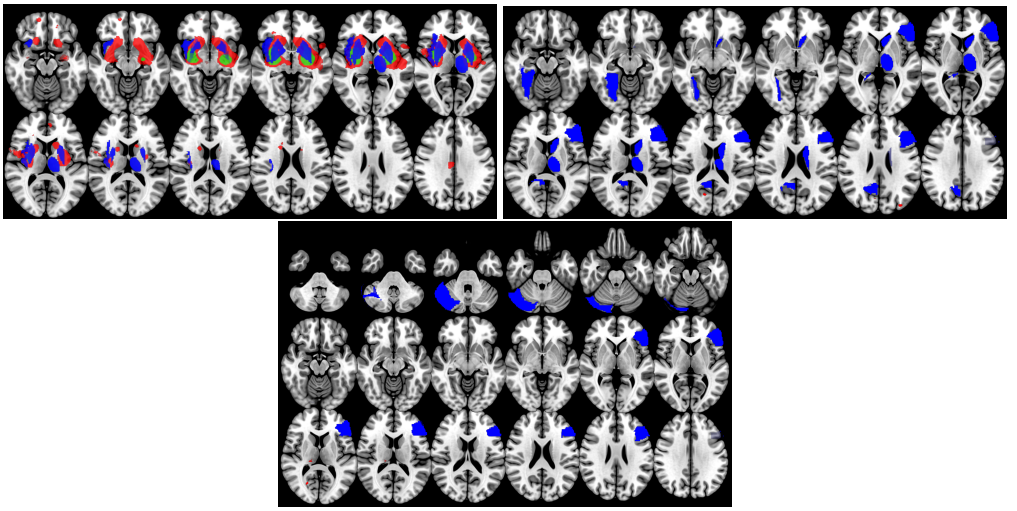


Figure 8.6: Significance maps obtained for ^{123}I-FP-CIT SPECT scans. Voxelwise SPM is represented in green, clusterwise SPM in red and SAM is blue. HC-vs-PD experiment (n = 80) is illustrated on the up left, HC vs. HC (n = 40), on the up right and PD vs. PD (n = 40) is located at the bottom. Voxelwise SPM is only non-null in the experiment HC vs. PD. The ROIs obtained are overlapped over 59.50% between SAM and voxelwise SPM12 and 57.67% between SAM and clusterwise SPM12.

three experiments and in Figure 8.7, where the significance maps for the GM MRI scans are illustrated. It should be noted that the significance maps of experiment HC vs. HC were practically null for all the inferences analysed, so its inclusion was discarded.

8.4.3 Discussion

In this section, a qualitative analysis of significance maps of Parkinson's Disease is conducted. To do so, two different approaches were performed, a parametric approach (SPM) and a non-parametric approach (SAM).

The main experiment (HC vs. PD) assesses the possibility of characterisation of PD by the imaging modality under study. This kind of experiment indicates relevant regions to perform a two-class (binary) classification. In MRI scans, results do not suggest that there are any highly relevant ROIs to address the classification problem. As seen in Table 8.2 and Figure 8.7 (left), SPM12 significant maps are practically empty. The only region that appears in both SPM clusterwise inference and SAM was the Precuneus (right hemisphere). Nevertheless, when using SAM, the Inferior Temporal Gyrus appears as ROI as described in [234]. Further study is needed to understand the implications of these regions for Parkinson's Disease, especially as they are more closely associated with the dementia that patients may develop.

Contrary to MRI, significance maps related to ^{123}I-FP-CIT SPECT clearly states the relevance of striatum region as ROI for PD. Both Putamen, Pallidum and Thalamus

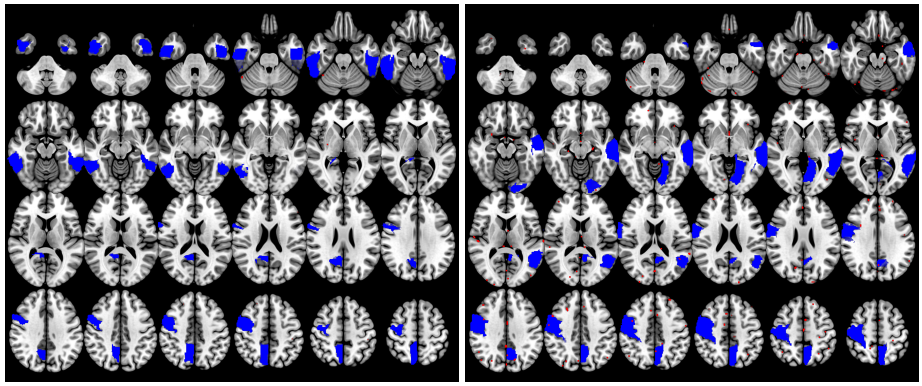


Figure 8.7: Significance maps obtained for GM MRI scans. Clusterwise SPM12 is presented in red and SAM is blue. Voxelwise SPM is null in both experiments. HC-vs-PD experiment (n = 80) is illustrated on the left, PD vs. PD (n = 40), on the right. In the HC-vs-HC experiment (n = 40), only clusterwise SPM12 is non-null but with very few voxels. No relevant overlapping was detected.

appear in all the inference maps as seen in Table 8.1. Indeed, the importance of this area can also be appreciated in Figure 8.6 (up left) where significance maps of SAM and voxelwise SPM have an overlap of 59.50% and the ones of clusterwise SPM and SAM, 57.67%. Similar results are obtained in [37], where SPECT scans from PPMI and a dataset from Virgen de la Victoria Hospital (Malaga, Spain) were analysed.

At this point two more experiments were conducted selecting samples of the same class arbitrarily divided into two groups to analyse their contrast, i.e. HC vs. HC and PD vs. PD. These experiments allowed to deduce the reliability of classification: occurrence rates of ROIs indicate if input data is highly heterogeneous and even if it could be separated among them. In any of the experiments involving SPM voxelwise inference maps point out several ROIs for both image modalities. Though the clusterwise also does, its number of detected voxels is small and they seem to have been randomly choosen (see Table 8.1 and Table 8.2). Similarly, for the SAM approach it also seems that detected regions have no relevant implications for the study. Only the results derived from HC-vs-HC experiment using ^{123}I-FP-CIT SPECT should be further analysed (especially if it is focused on the Fusiform Gyrus).

Regarding the methods analysed, it is clear that the use of FWE p-values generates highly conservative significance maps [34]. In fact, when a true effect appears, it can be seen in Figure 8.6 (top left) how the least conservative significance map is the one that uses an uncorrected p-value for multiple comparisons. However, this method increases the number of FP. Therefore, its use is not advisable [34]. An intermediate term would be the significance maps generated using SAM. Indeed, as many other works in the current state of art where the small sample-size problem appears, the classical statistics creates a challenging scenario where the generalisation of the results is unclear. Thus, this other example to suggest that proposals like SAM should be stated as gold standard due to its reliability when working with small sample sizes.

After analysing our results, it can be stated that MRI scans is not a much reliable source of information for PD diagnosis even despite the classification results reported in related works such as [231]. Though few other studies have proposed the use of MRI imaging markers as main inputs in CAD systems for PD, they might be falling into some inadvertencies. For example, they work in conjunction with SPECT and MRI, whereby the former is the main contributor to a high accuracy (biased) [232], or the feature selection is done outside the CV loop [233].

In any case, to evaluate the use of MRI for PD, research projects using larger sample sizes are needed. And even then, it will be necessary to evaluate the effect of sample size on the inference maps and to analyse the effect of FP.

8.5 Standarisation of Agnostic Learning Techniques: EEG

The aforementioned different neuroimaging problems are particularly noticeable in the field of cognitive neuroscience, where sample sizes rarely exceed fifty or so participants, and results rely on subtle brain activity changes that are locked to particular events. Classification methods on this field are used to tell apart different conditions across the same participant. That is, validation methods include training and testing the same brain in different points in time, which again poses the challenge of inter-subject generalisation, and potential Type I and Type II errors.

In this study, the methodology underlying SAM to perform spatial detection of ROIs in neuromaging modalities, such as MRI, has been applied to a temporal EEG study within the field of cognitive neuroscience, where the use of MVPA has been growing exponentially [98, 99, 235]. Instead of estimating ROIs by means of a significance maps, the time points where trials are most significant are highlighted. Temporally resolved techniques, such as EEG and MEG, also common in neuroimaging research, have been less frequently analysed through classification approaches (but see [236, 237, 238] as examples). Therefore, the aim is to determine whether the validation model on which SAM is based is also useful in time-resolved EEG data obtained from a cognitive neuroscience task designed to compare the representational information associated with predicted and unpredicted target stimuli.

8.5.1 Methodology

The MVPA implemented consisted of three steps: feature extraction, feature selection and classification. To conduct the experiment, the MVPAlab Toolbox [159] was used, which was adapted and complemented using SAM. The UGR-COG dataset was analysed, which is described in section 5.3. Each participant had associated an amount n of trials (observations), consisting of the raw potential measured in p electrodes over time (features). Each participant's trials were categorised according to two conditions or classes (seeing faces or seeing names on screen).

At the feature extraction stage, the features associated with each trial were normalised with mean 0 and standard deviation 1. Then, sets of t trials (within the same condition) were averaged to increase the SNR and reduce the computational burden. In the feature selection stage, a dimensionality reduction of the features was performed. PCA was applied to project the sensor space features onto a reduced number of features. Finally, SVM with linear kernel was implemented as classifier. Each participant was analysed independently, considering as accuracy the mean of the accuracy obtained for each subject at each time point.

For the evaluation of the classification model, RUB was selected as validation approach. However, the study was also conducted using K-fold CV for comparative reasons.

The last step was to modify the spatial study implemented in SAM to perform a temporal study of the significance of the results obtained. In SAM, once the accuracy of each region of the brain (given an atlas) is calculated, the significance of these values is analysed by means of z-test stadistics to detect ROIs and generate the significance map. Thus, the spatial study was altered to a temporal one by analysing time points instead of regions. Time Point of Interests (TPOIs) were those moments in time that were considered to be significant. In this case, the significance selection criterion was that its associated accuracy was higher than 0.50.

8.5.2 Results

To conduct the study, the number of trials was reduced by applying a factor of 8 for generating supertrials. The times considered for the experiment ranged from −100 ms to 1400 ms, with a total of 385 time points during that time. The feature vector generated by the 64 electrodes used was reduced to 1 feature by using PCA.

Results are illustrated in Figure 8.8. The solid blue line indicates the accuracy values over time associated with using resubstitution as validation approach (empirical accuracy). This value was reduced after applying the upper bound under the worst case with probability 95% (green line, actual accuracy). The accuracy obtained by using K-fold CV with $K = 5$ is also included (red line). The areas including the standard deviations of the hits considering the 48 participants are also included around the mean value. The horizontal lines shown indicate the TPOIs obtained depending on the approach. For the estimation of the significance of the time points, uncorrected resubstitution and K-fold results were calculated with a t-test against chance, while for estimating the ones related to RUB, accuracies above chance were considered significant. These time points were from 97 ms to 335 ms, approximately, using RUB.

8.5.3 Discussion

In this work, the relevance of SAM tool on EEG to analyse the temporal significance of stimuli was tested. The effect of visual perception was analysed through a binary

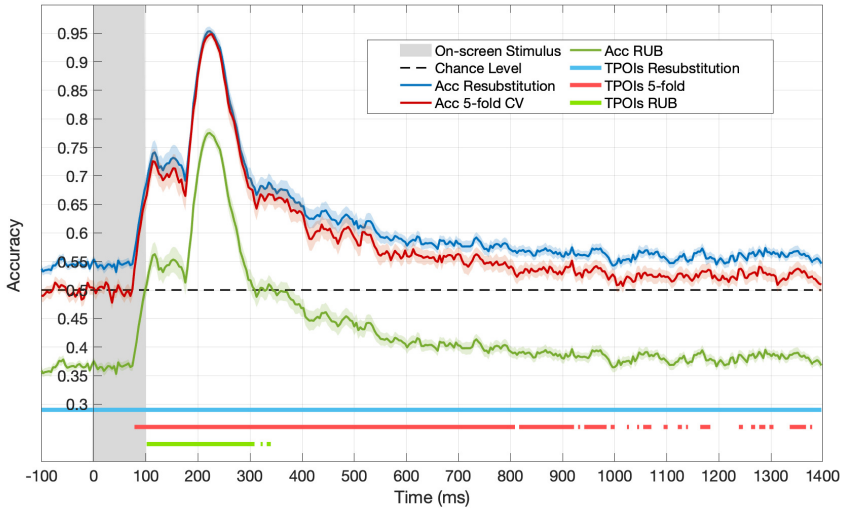


Figure 8.8: Accuracy values obtained. Blue indicates the accuracy values over time associated with using resubstitution as validation approach (empirical error). This value is reduced after applying the upper bound under the worst case with probability 95% (actual error, green). Red shows the accuracy obtained by using 5-fold CV. The grey area indicates stimulus presence onscreen (0-100 ms). Horizontal colored lines indicate temporal significance.

classification task. Although the effect has been theorised to be large and persistent in time [239, 240], accuracy values derived from prediction tests were narrowly above the nominal value of statistical significance, i.e. 50%, in the whole stimulus interval. When performing an identical MVPA using K-fold cross-validation instead of RUB, the obtained results are less conservative and closer to the expected outcome. However, the empirical (uncorrected) error obtained was slightly higher than that obtained by CV, as expected [36]. Importantly, the resulting pattern of accuracy values was strikingly similar between resubstitution and K-fold, even if the actual values were biased upwards for resubstitution, especially given the variability associated with K-fold [15]. Thus, the results obtained suggest that this approach can also be applied in EEG. It will be necessary to establish a significance test more adjusted to this type of data, which generates less conservative TPOIs. That is, the temporal range of significance should more closely resemble that obtained with K-fold. For this, different bounds could provide an accurate connection between actual and empirical errors.

Regarding the implementation of SAM as a temporal analysis software, it would be advisable to extend its utility to other well-established applications of MVPA, such as classification across time and conditions, or sensor space analysis. Future applications and developments of this method of analysis should test how different methods of dimensionality reduction (e.g. PCA or PLS) affect the results.

9 | EXPLORING RELEVANT FEATURES: INTERPRETABILITY OF MACHINE LEARNING MODELS

9.1 Introduction

So far, the thesis has primarily focused on exploring the reliability of CAD systems across different scenarios. Starting from this chapter onwards, the emphasis will shift towards the interpretation of results. This aspect becomes particularly crucial in clinical studies where experts need to comprehend the findings in order to draw meaningful conclusions. In the field of neuroimaging, this situation is commonly encountered due to its multidisciplinary nature, where the focus lies not only on the implementation of innovative reliable techniques but also on assessing the utility of the CAD system as clinical support tool or for investigation of complex biological patterns. In this chapter, an in-depth analysis of sulcal patterns extracted from structural brain images is undertaken to broaden our current understanding of the etiology of Schizophrenia due to the early development of the cerebral sulci (see section 2.1).

Sulcal information has proven to be useful in the study of a wide range of conditions; for example, in AD [241, 242], Parkinson's disease [243, 243], and anorexia [244, 245]. SCZ has a rich and well-replicated literature establishing patterns of cortical change [67, 246, 247, 248]. Whilst there has been some work on both overall and specific sulcal information in SCZ [249, 71, 250], no information has been found on exploring sulcal patterning as a way of classifying individuals with SCZ from unaffected controls.

As mentioned on previous chapters, one of the main problems often encountered in conducting this type of study is the limited number of samples available. This is of particular concern when the number of features associated with each sample is very high (*curse of dimensionality*) [187]. This is also a problem when applying classical statistics which make strong assumptions based on the sample conforming to the normal distribution. When the number of samples is small, it is not easy to accurately determine the distribution from which they are sampled and sometimes invalid techniques are implemented or inaccurate results are obtained [251, 34]. For this reason it is useful to consider other methods, such as data-driven approaches based on ML [37, 221]. A key benefit is to obtain insights similar to those obtained by parametric statistical approaches but without requiring the dataset to satisfy certain conditions. Furthermore, the black box problem, whereby there is no easy interpretation of the biological meaning of a classification result or understanding of the underlying decision-making process, is now being addressed with XAI algorithms [252, 253, 38].

In this study, the capacity of measurements of sulcal patterns to discriminate between patients with schizophrenia and controls is explored. To do so, features relevant to this classification problem are identified by traditional univariate statistical methods and by MVPA. Both approaches are then compared by means of ML classifiers of varying complexity. XAI techniques are also deployed to give a richer description of the pattern of case-control differences observed.

9.2 Methodology

For this study, the SHG-SCZ dataset was used, which is detailed in section 5.4, including a comprehensive description of the preprocessing steps involved. The dataset comprises a total of 114 samples, consisting of 58 subjects with SCZ and 56 HC. Each sample has associated 49 specific brain areas (sulci) which are represented by the sulcal length and maximum and mean sulcal depth (147 features per sample in total). Thus, the number of features is a value larger than the number of samples, 114. This is an undesirable, but very common situation in neuroimaging. It is important to acknowledge that various methods exist in the literature for detecting, labeling, and characterising sulci [254, 255, 256, 257, 258, 259, 241], each with its unique strengths and limitations [260]. In this study, the BrainVISA software is employed, as described in section 5.4.

The methodology followed aimed to finding relevant sulcal patterns in a classi-

fication scenario SCZ vs. HC, i.e. a binary classification problem. For this purpose, the process was divided in several stages as depicted in Figure 9.1. In summary, two different scenarios were implemented: feature selection (see section 9.2.1), which highlighted the relevance of sulcal features, and feature extraction see section 9.2.2), which generated a reduced set of features to make the best possible use of the information extracted from the original data. From the features highlighted or generated by both approaches, a classification stage followed where ML and DL models were applied using various validation methods, as described in section 9.2.3. Lastly, the classification model obtained was analysed by means of XAI techniques, which are described in section 9.2.4. At each stage, alternatives were compared in terms of their performance in the context of a small sample.

9.2.1 Feature analysis and selection

Once the dataset was preprocessed, sulcal length and maximum and mean sulcal depth were tested by univariate statistical methods to identify features important to classification. Both parametric and non-parametric techniques were considered.

Regarding parametric techniques, the Shapiro-Wilk test was initially applied to identify which features obeyed a normal distribution, since the null hypothesis is that samples come from a normally distributed population. For those features where a normal distribution was followed, a two-sample t-test was applied to detect the relevance of the feature to distinguish between schizophrenia and control participants. To compare the importance of features, the p-values associated with the tests were used. Those features that did not follow a normal distribution were assessed with the Mann-Whitney U test, and the corresponding p-value used. These tests are described in section 3.2.

The importance of a feature to classification was also evaluated by means of an data-driven based approach: SAM (see chapter 8), i.e. a non-parametric approach. First, each feature was independently fed into a supervised classification model. Then, accuracies obtained for each feature were sorted based on a proportion test and its test-statistic, see Equation (8.2). Once the z-statistics were calculated, the p-value of each statistic was estimated. For this, the null hypothesis was considered to be true and therefore the test statistic follows a standard normal distribution. From this p-value, the most relevant features of the study were determined.

9.2.2 Feature extraction

Along with feature selection, features were also processed to generate more compact information and reduce the dimension of the feature vector. To do so, PLS was applied (see section 4.3.2) and the extracted components are those used in the classification step.

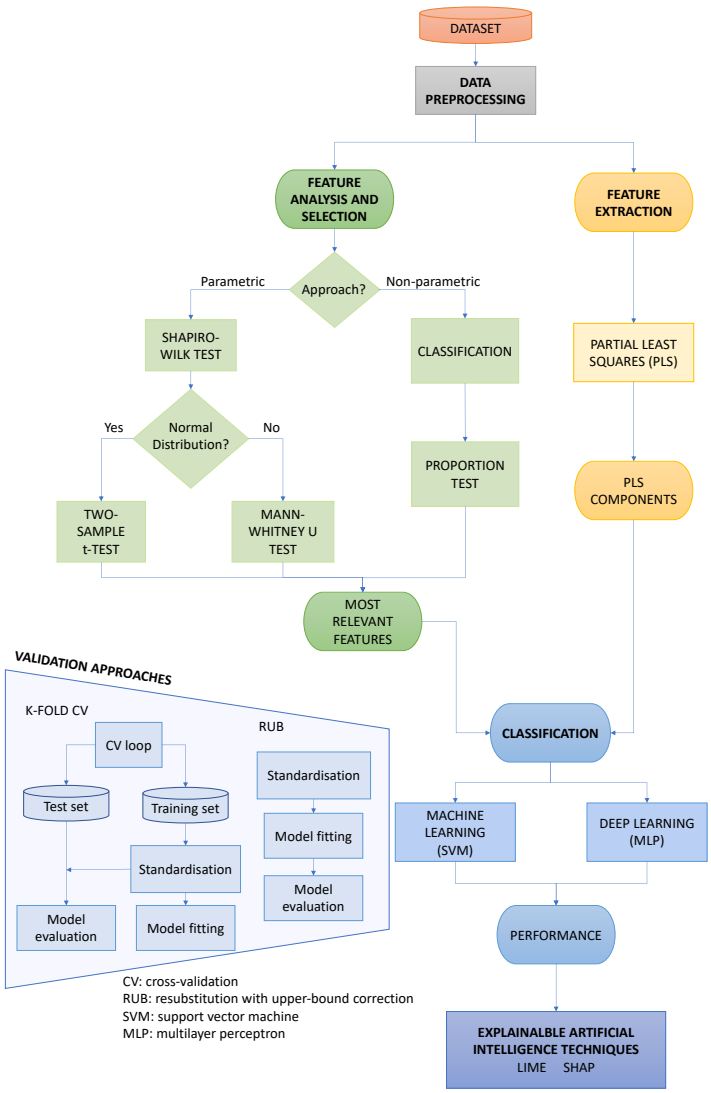


Figure 9.1: Flowchart of the study. After preprocessing the data, two independent feature selection and feature extraction analyses were conducted. The information extracted from both was fed into ML and DL classifiers. Two validation methods were applied. Finally, the classifiers' performance was analysed by means of XAI techniques.

9.2.3 Classification

Once the features to undertake the classification were selected (either the selected or extracted ones), the next stage was classification. For the binary classification problem posed in this study, both ML and DL methods were applied.

The ML algorithm implemented in this study was a SVM classifier with linear kernel, see section 4.4.3. This combination was chosen for its easy explainability as well as its

propensity to generate excellent results in neuroimaging [35, 261, 262].

Given the small number of samples and features, an MLP architecture was chosen for the study as DL approach. Furthermore, the use of a one-dimensional feature vector makes MLP more suitable than CNN, which is designed to extract patterns in higher dimensions. The network configuration implemented is shown in Figure 9.2. The number of epochs involved in the training was 18, with a batch size of 1. The optimiser selected was Adam [263] with a learning rate of 0.001, and the stopping criterion computed as the cross entropy loss with balanced weights.

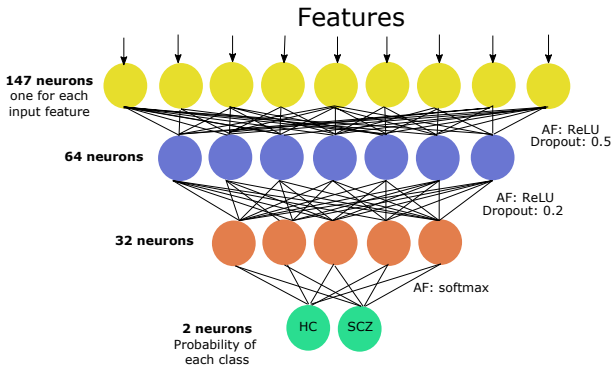


Figure 9.2: Scheme of the MLP composed of four layers: input layer, two hidden layers and the output layer. Note that AF: activation function.

Two validation methods were used to assess the performance of the classifiers. First, a 10-fold stratified CV scheme was applied, which guaranteed independence between training and test samples. For the computation of the performance metrics, the mean and standard deviation of the values obtained in the ten iterations were used.

The second validation was RUB. Thus, the entire database was used as the training set for the classifier, i.e. resubstitution was performed, and then the actual accuracy was obtained by means of the upper bound. This could be considered a theoretical classification limit that allows the use of all accessible data to compute the metrics of interest. In addition to accuracy, other metrics such as sensitivity or specificity can also be of upper-bounding value since their errors are related to the classification error. In this study, the upper bounds applied were the one based on the assessment of concentration inequalities (Equation (4.13)), as a linear classifier was also implemented, and a PAC-Bayesian bound defined in Equation (4.14).

Performance of the classifiers was evaluated through metrics extracted from the confusion matrix, where the positive class was SCZ. These metrics were balanced for accuracy, specificity and sensitivity, see Equation (4.16). The ROC curve was also constructed and the AUC evaluated the ability of the model to differentiate between the two classes, see section 4.6.

9.2.4 Explainable Artificial Intelligence

In the final step of the procedure, XAI techniques were applied to detect the relevance of features to classification. Therefore, apart from considering the performance of the classifier, two model-agnostic approaches were used to analyse the influence of features on the decision making by the classifiers. The XAI techniques applied were LIME and SHAP, a description of which can be found in section 4.7. Whereas LIME focuses on generating local explanations for individual samples, SHAP is used to obtain more general information about the overall performance of the classifier and the contribution of features in general. It was therefore considered appropriate to use both and compare their outcomes.

9.3 Results

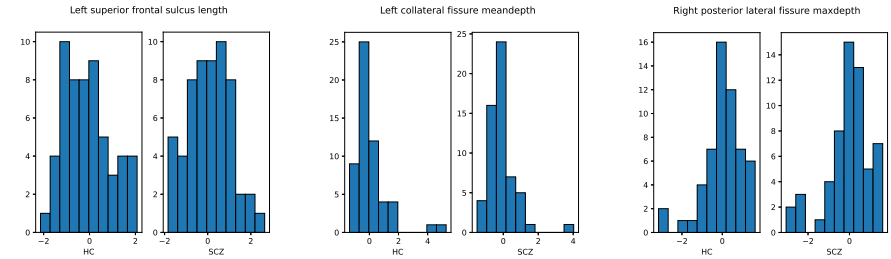
9.3.1 Feature selection

To analyse the relevance of the features, firstly parametric approach was followed. The Shapiro-Wilk test for normality determined that among the 147 features 125 followed a normal distribution, while the remaining 22 did not. Figure 9.3a shows examples of histograms of three features that did not follow a normal distribution. It can be seen that the main reason for this was the long tails skewing the distribution. By visual inspection selecting eligible samples, the number identified was adequate for a two-sample t-test with all the features. The Mann-Whitney U test was used for non-normally distributed features.

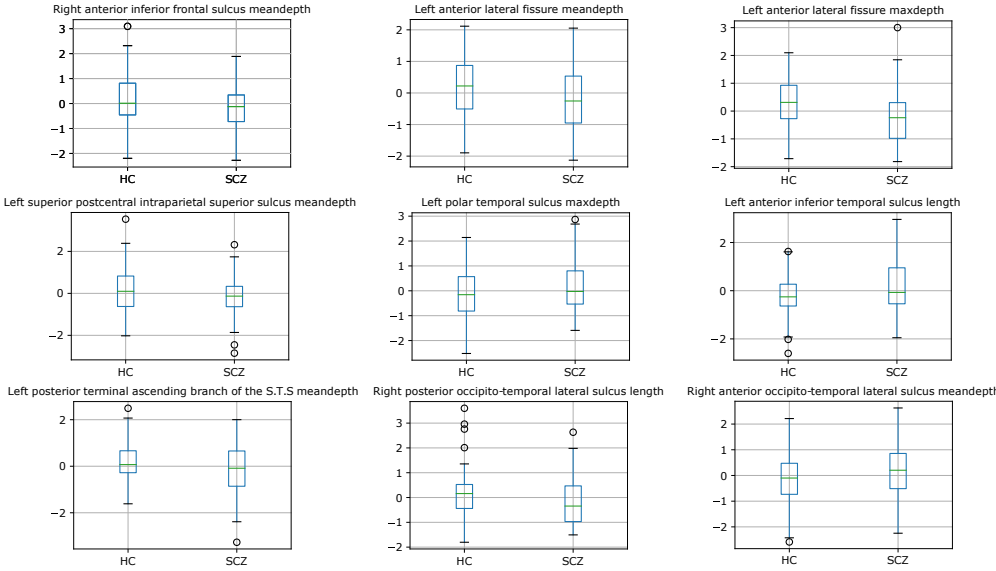
The significance of each feature was assessed with the p-value obtained in their respective tests. Figure 9.3b shows a boxplot of the nine most relevant features according to the tests applied. Only the first five had a p-value < 0.05, while ten of them had a p-value < 0.1.

Using the SVM algorithm with linear kernel and upper bounding resubstitution-based validation (non-parametric approach), 1000 permutations were applied by randomly modifying the position of the samples in the set and calculating the mean accuracy for each feature. For accuracy estimation, a balanced estimator with class weights balancing was applied during the classifier training. The p-value associated with each feature was assessed with a significance test for a proportion. The nine most relevant features obtained are shown in Figure 9.4.

To compare the two approaches (parametric and non-parametric), Figure 9.5 shows the most relevant features ranked by their p-value in both approaches. Relevant features had a p-value < 0.05, identifying nine features. Three features appeared with both approaches, together with 2 from the parametric and 4 from the non-parametric approach. These include both depth-related and length-related features.



(a) Non-normal distributed features.



(b) Relevant features from the parametric approach.

Figure 9.3: Statistical features analysis. (a) Histograms related to non-normal distributed features. (b) Boxplots of the nine most relevant features according to the two-sample t-test and the Mann-Whitney U test, depending on whether the feature follows a normal distribution or not. These features are arranged from Frontal lobe to Occipital lobe (from left to right and from top to bottom).

9.3.2 Use of reduced dimensionality in classification

Instead of analysing the relevance of the features independently, it is possible to analyse the relevance of the feature set for the case-control classification. To do this, a feature extraction stage was implemented by applying PLS to the original data. The reduced feature dimension was then classified with a SVM classifier with a linear kernel. Performance was analysed using cross-validation, resubstitution and RUB validation procedures. Figure 9.6a shows results how the classifier's performance varied according to the PLS components used. The upper bound applied in RUB, Equation (4.13), depends on the number of features, the fixed number of samples (114) and significance level (0.05). Although there were no discernible trends in performance as a function of number of PLS scores using K-fold, a decreasing trend was observed applying RUB and

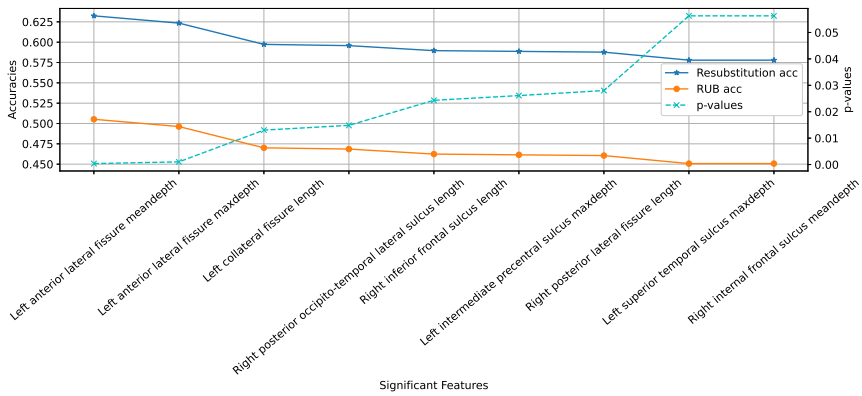


Figure 9.4: The nine most significant features obtained by a classification approach (non-parametric approach). Their related accuracy was estimated as the mean value of 1000 permutations shuffling the samples and using a SVM with lineal kernel classifier and RUB as a validation approach. The p-values related to each region were estimated using a test of a proportion.

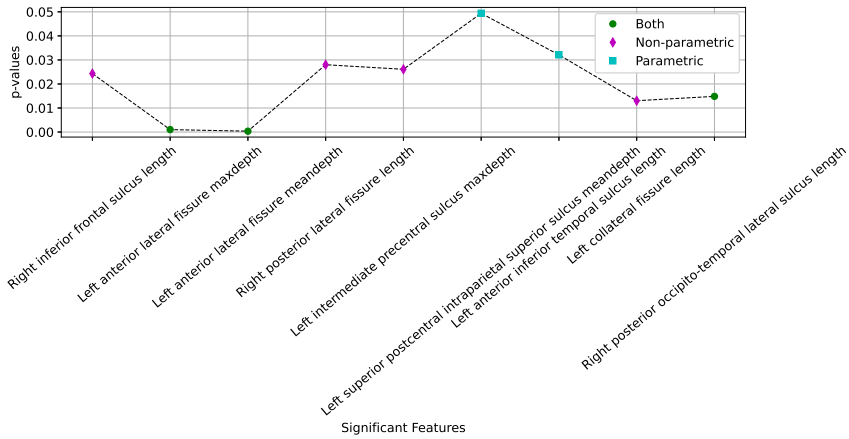
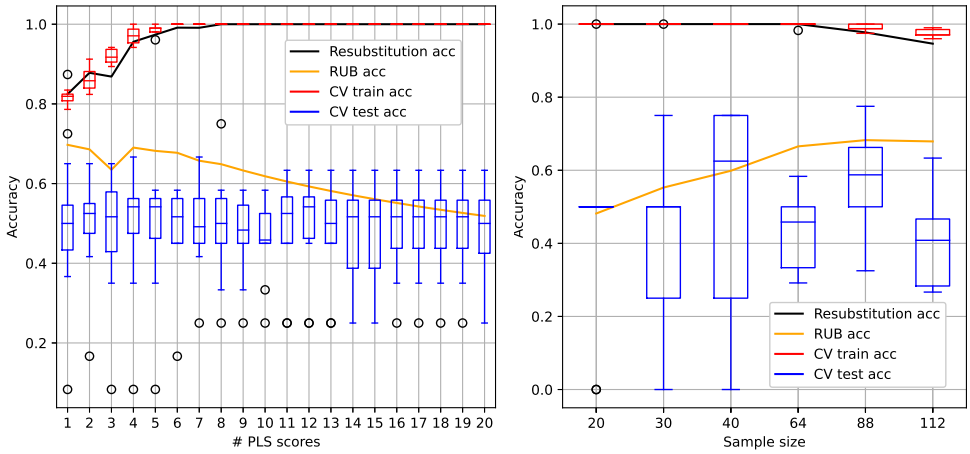


Figure 9.5: Features under analysis with a p-value < 0.05 in any of the parametric and non-parametric tests. These features are arranged from Frontal lobe to Occipital lobe. The significant features under the parametric analysis are coloured cian, non-parametric analysis are coloured magenta, or if both they are coloured green.

higher accuracies were obtained using fewer components (less than 6). Conflating both approaches, 4 PLS components were selected.



(a) Accuracy vs. PLS components.

(b) Accuracy vs. sample size.

Figure 9.6: Performance of the SVM classifier along with PLS as the feature extraction technique. Results are shown for a wide range of PLS components (1-20) and using 4 PLS components for several balanced samples sizes (20,30,40,64,88 and 112). In both cases: resubstitution (black line), RUB (orange line) and 10-fold CV (box-plots).

In addition, to understand the effect on performance of sample size, 4 PLS components were chosen as input features for the classifier. The results are shown in Figure9.6b. Theoretically, accuracy should increase as the sample is enlarged, but this was not the case with K-fold. The upper bound applied in RUB changes according to the number of samples, with a fixed number of features, 4, and a significance level, 0.05.

9.3.3 Various classification scenarios

Case-control classification was undertaken with the features selected with the parametric, non-parametric and ensemble approaches as well as PLS features. The testing of the classification results were obtained by performing 1000 permutations of the dataset, which are shown in Table 9.1. Note that for the computation of the upper bound of Equation (4.13), the RUB validation approach took into account the number of samples, 112 (as the data was balanced in each iteration), the number of features (9 or 4, depending on the case), and the significance level (0.05). This gave values for the upper bound of 0.3695 per unit or 36.95% for 9 features and 0.2675 (26.75%) for 4. The reader is reminded that these values must be subtracted from the accuracy rate obtained in order to determine the actual worst-case accuracy rate. While K-fold CV worked reasonably well with the extracted features, especially with those obtained with the parametric method, RUB had improved performance with the PLS components

due to fewer input features, and thus a tighter upper bound. This is especially true when extracting the main components of the full set of features.

		Parametric	Non-parametric	Both	PLS (all)	PLS (both)
10-fold training	Acc (%)	73.59 ± 0.96	71.25 ± 0.83	71.19 ± 0.78	97.18 ± 0.56	70.82 ± 0.82
	Sens (%)	67.98 ± 1.31	66.21 ± 1.42	67.72 ± 1.38	97.37 ± 0.67	67.80 ± 1.32
	Spec (%)	79.20 ± 1.08	76.28 ± 1.50	74.68 ± 0.98	96.98 ± 0.75	73.85 ± 1.10
	AUC	0.80 ± 0.01	0.76 ± 0.01	0.76 ± 0.01	1.00 ± 0.00	0.76 ± 0.01
10-fold test	Acc (%)	66.26 ± 2.37	62.32 ± 2.72	64.45 ± 2.30	49.64 ± 3.04	63.12 ± 2.48
	Sens (%)	62.07 ± 2.80	59.65 ± 3.50	61.94 ± 3.03	50.08 ± 4.39	62.00 ± 3.07
	Spec (%)	70.44 ± 3.48	64.97 ± 4.02	66.92 ± 3.30	49.23 ± 4.05	64.25 ± 3.66
	AUC	0.73 ± 0.02	0.67 ± 0.03	0.68 ± 0.03	0.48 ± 0.03	0.66 ± 0.03
RUB	Acc (%)	34.38 ± 1.44	34.96 ± 1.41	33.34 ± 1.22	68.88 ± 1.15	43.77 ± 1.60
	Sens (%)	27.51 ± 1.79	29.82 ± 1.86	30.88 ± 2.37	69.77 ± 1.48	40.64 ± 2.30
	Spec (%)	41.25 ± 1.52	40.11 ± 2.86	35.81 ± 2.48	67.99 ± 1.54	46.90 ± 2.43
	AUC	0.43 ± 0.01	0.39 ± 0.01	0.38 ± 0.01	0.73 ± 0.00	0.49 ± 0.01

Table 9.1: Performance of the SVM classifier using the nine extracted features in the parametric, non-parametric and both analyses after 1000 permutations. Results using 4 PLS components as input to the classifier are also included when they are extracted from all 147 and the 9 globally significant ones. Upper bounds related to this analyses were 0.3695 (9 features) and 0.2675 (4 features) for a significance level of 0.05.

In addition to these classification performances, a PAC-Bayes upper bound was applied under the same experimental conditions to test its performance against the upper bound based on concentration inequalities. As this different bound depends on the dropout rate, see Equation (4.14), several values of dropout were applied: 0, 0.25, 0.5, 0.75 and 0.95. The results are shown in Figure 9.7, where the dashed horizontal lines represent the accuracy shown in Table 9.1 with the RUB approach.

9.3.4 Examining predictions with XAI

The same classifier was tested using the 144 features as input features and k-fold CV as validation approach. A summary of the performance results are shown in Table 9.2. Accuracy values were below 50%. With the same features as input, the MLP achieved a 58.83% accuracy on the test set by applying CV. When using all the features as input, it is possible to apply XAI techniques to reveal the internal process of the algorithm. Due to the better performance obtained using MLP, the subsequent results are associated with this classification model.

LIME allowed us to identify qualitative patterns on the most relevant features according to a classifier that distinguished case and control classes. Four examples of individual explanations are shown in Figure 9.8. These examples are related to correctly classified HC (Figure 9.8a) and SCZ (Figure 9.8b) test samples by the MLP. For this analysis, the 10 most relevant features in the classification for each sample were selected and displayed sorted from most to least importance according to LIME.

All four major lobes of the brain appear in this analysis, although the Temporal

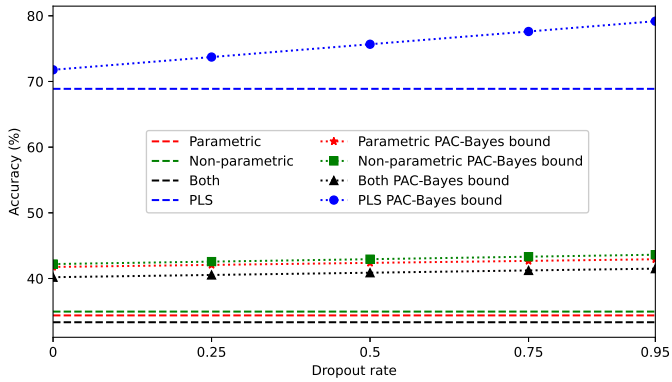


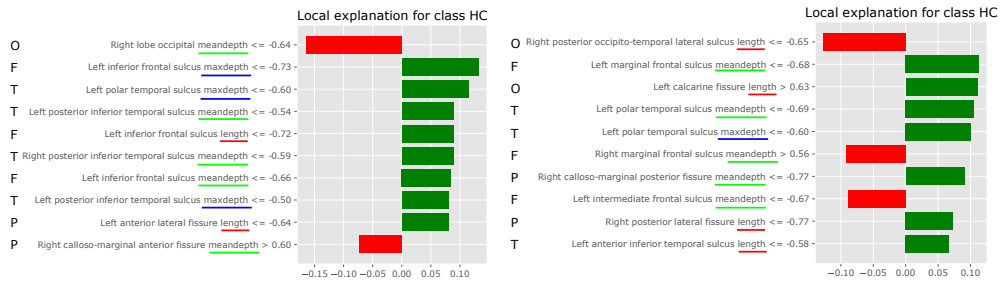
Figure 9.7: Accuracies obtained with the RUB approach using two different upper bounds. The dashed horizontal lines are the accuracies obtained with the upper bound based on concentration inequalities (Equation (4.13)). Accuracies with markers are those with the PAC-bayes bound (Equation ((4.14)). The classifier applied was SVM using the nine extracted features in the parametric, non-parametric and both analyses, and 4 PLS components. Accuracies shown are the mean values after 1000 permutations.

		SVM	MLP
10-fold training	Acc (%)	100 ± 0.00	67.94 ± 8.90
	Sens (%)	100 ± 0.00	67.04 ± 26.06
	Spec (%)	100 ± 0.00	68.83 ± 21.01
	AUC	0.40 ± 0.49	0.71 ± 0.07
10-fold test	Acc (%)	49.50 ± 9.72	58.83 ± 6.28
	Sens (%)	54.00 ± 17.50	57.00 ± 27.87
	Spec (%)	45.00 ± 14.47	60.67 ± 28.43
	AUC	0.45 ± 0.13	0.56 ± 0.10

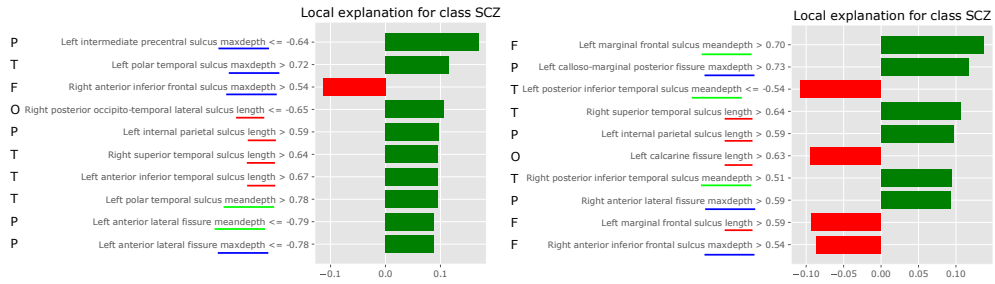
Table 9.2: Classification performance of models based on SVM and MLP when the 147 features (the complete set) were fed as input of the classifier. Cross-validation was used as validation approach (10-fold CV).

and Frontal lobes have greater representation. The same applies to the three types of features related to length and depth. Several features included in Figure 9.5 as the most relevant features according to parametric and non-parametric approaches, were also relevant in this analysis. One prominent example was the length of right posterior occito-temporal lateral sulcus. In the top right sample, a low value of sulcal length decreased the probability of being associated to HC class, while in the bottom right sample, a similar low value increased the chance of being classified as a SCZ patient.

Regarding the implementation of SHAP, Figure 9.9 shows the graphs that this technique returned from the MLP model. Figure 9.9 (top left) is a summary graph of the impact of the features during the classification process. The ten most relevant features in the classification are displayed. High SHAP values were associated with the SCZ, while low values were associated with HC. Colour blue in the instances of the test sample indicates low values of the feature whereas a pink value indicates the opposite.



(a) HC samples.



(b) SCZ samples.

Figure 9.8: Local explanations extracted from LIME for the SCZ and HC classes, all of them are correctly classified samples. Features in green represent values that increase the chance of being classified as the class under analysis. Features in red reduce it. To improve comprehensibility, length, mean depth and maximum depth are underlined in red, green and blue, respectively. On the left side, the letters F, T, P and O represent the feature belonging to Frontal, Temporal, Parietal or Occipital lobe, respectively.

Overall, mean and max depth of specific sulci are notable in their contribution to the classification. In this analysis, the impact of length is minor. As examples, a high value of the maximum depth of left intermediate precentral sulcus was associated with the HC class, since it was associated with lower SHAP values, whilst a high value of marginal frontal sulcus mean depth was associated with the SCZ class.

Another type of graph is a dependency plot. Dependency plots illustrate the relationship between the SHAP value and the magnitude of the feature. A second feature reflected in the colour of the samples is included, which may indicate some dependency between features. For example, in Figure 9.9 eight different dependency plots are illustrated. They all include features that appear in the summary plot. At the bottom, it is observed that higher values of left inferior frontal sulcus maximum depth, i.e. deeper values, bring the sample closer to the SCZ class (higher SHAP value). In the last graph, which includes the comparison between left anterior lateral fissure maximum and mean depth, there is a correlation between the two features, since samples with low values of the maximum depth also have a lower mean depth.

A summary plot including the impact on the classification model of the 147 features is illustrated in Figure 9.10. The order selected was from Frontal lobe to Occipital lobe,

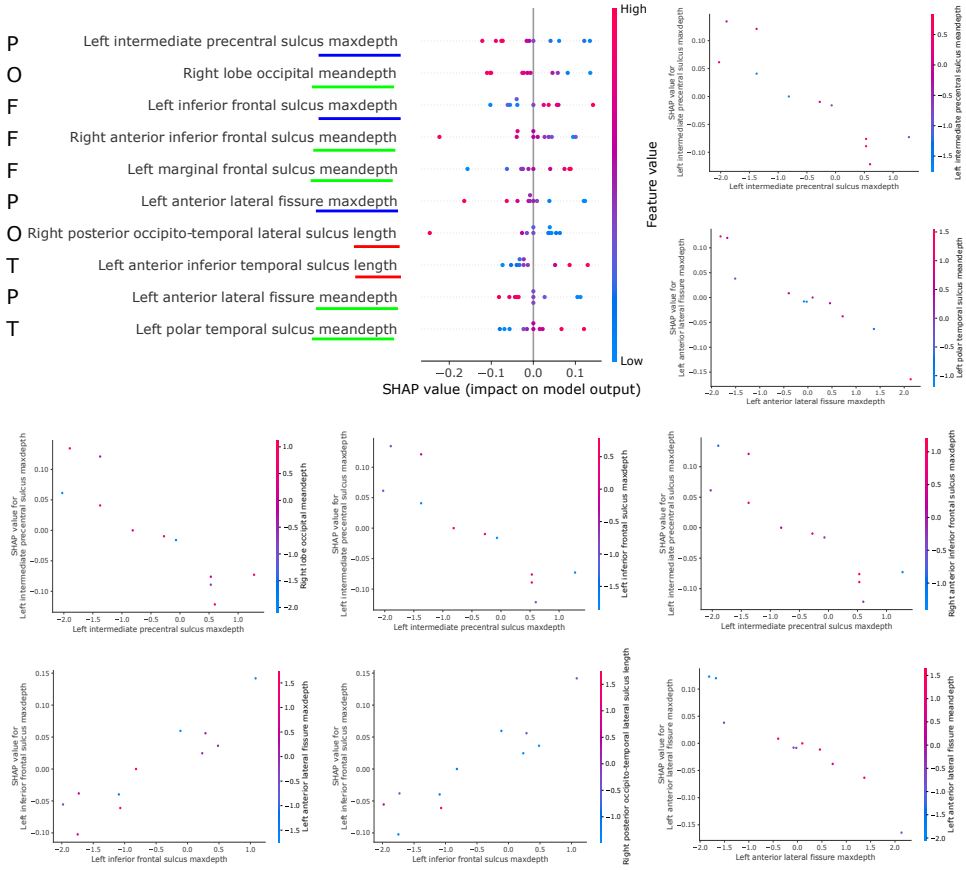


Figure 9.9: SHAP charts, where each point represents an instance of the test sample. Top left: Summary plot of features importance in the classification decision; the ten most relevant are shown. To improve comprehensibility, length, mean depth and maximum depth are underlined in red, green and blue, respectively. Letters F, T, P and O represent the feature belonging to Frontal, Temporal, Parietal or Occipital lobe, respectively. Top right and bottom: Dependence plots of some relevant regions according to their SHAP values. Colour in the graph corresponds to the value of a second feature for that same sample. The positive class is SCZ.

which are arranged from from left to right and from top to bottom on the illustration.

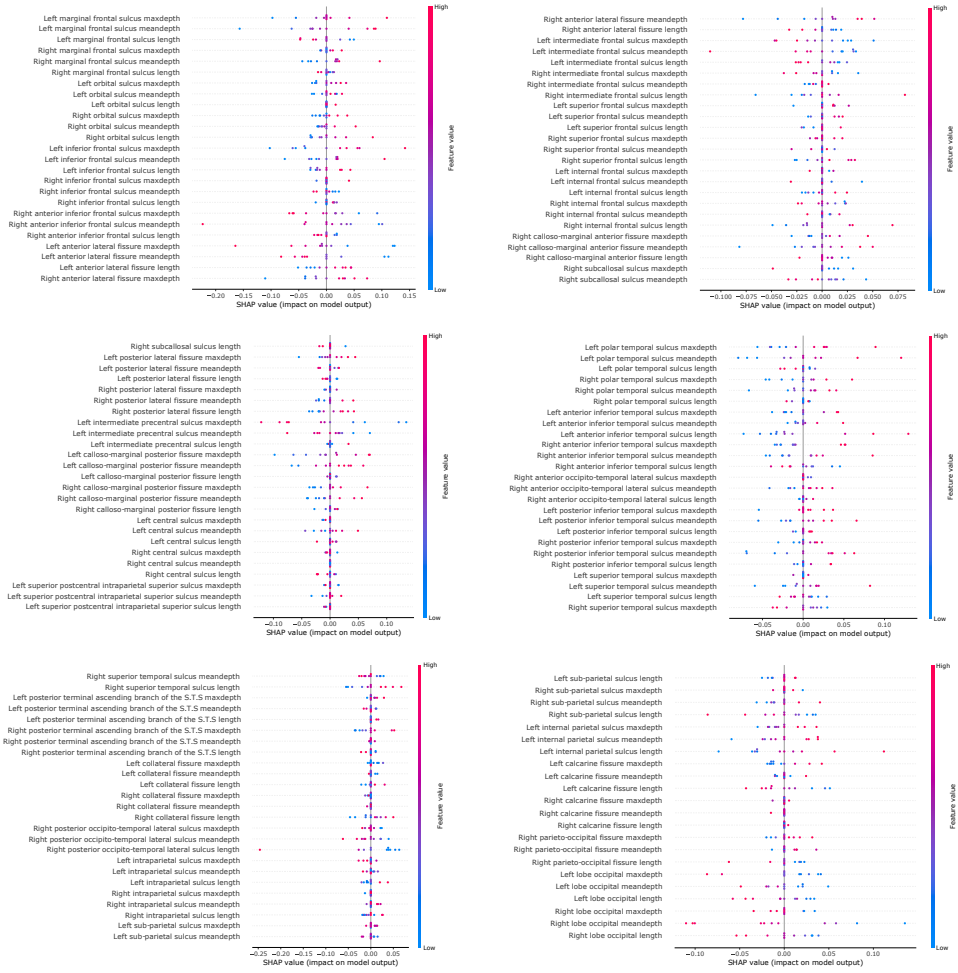


Figure 9.10: Summary plot of features importance in the classification decision. These features are arranged from Frontal lobe to Occipital lobe (from left to right and from top to bottom). The positive class is SCZ. Each point represents an instance of the test sample.

9.4 Discussion

In this study, a staged approach of statistical, ML, and DL techniques were applied to perform an analysis of sulcal patterns in a case-control comparison of SCZ. Feature calculations were performed by BrainVISA, where a 3D U-Net CNN was implemented to the labeling of sulci [166]. Subsequently, sulcal length and depth were selected as features. These features were standardised and independently tested with parametric (*t*-test) and non-parametric (data-driven) approaches. Machine and deep learning algorithms were applied to classify SCZ patients from HC, and its predictions are

evaluated by XAI techniques.

Unlike most work on sulcus patterns, the features applied are extracted fully automatically and encompass the entire cerebral cortex. The sulcal dectection processes remain in their early development, as it is still very difficult to correctly label all the sulcal patterns, especially those that are small or peculiarly shaped [242]. In the dataset used, the amount of detection failures obtained was high, thus reducing the number of sulci and the number of subjects finally included in the study (see section 5.4). This made it impossible to study some high-interest regions such as the left hemisphere paracingulate sulcus [71], while the right hemisphere pair is represented in Figure 5.3 as right calloso-marginal anterior fissure. Moreover, it was impossible to include data from other centres because even with standardised MRI scans, similar values for the extracted features were not achieved. For these reasons, the literature includes a large number of works which combine automatic extraction and manual revision [264, 247, 249], apply manual segmentation [265] or reduce the study to a concrete number of regions of interest [170, 266].

Most significant features obtained in this study reflect a similar importance of length and depth, as it can be seen in Figure 9.3, Figure 9.4 and Figure 9.5, albeit slightly higher in the case of depth. Regarding the relevance of using the maximum or mean depth, their occurrence in the most significant features is practically identical. However, given the same feature, both are not necessarily equally relevant. According to the upper right graph in Figure 9.9, while the correlation between maximum depth of the intermediate precentral sulcus and its effect on classification is inversely proportional, the mean depth has no direct relationship with maximum depth.

The hemisphere most represented in these findings is the left hemisphere, which is consistent with other studies [71, 162, 247, 267]. Both the length and depth of the sulci in this hemisphere tended to be smaller in SCZ subjects, as observed previously [162]. For example, in line with previous studies, Figure 9.9 (top left) shows a negative correlation between the intermediate precentral sulcus and the disease [268, 269]. Nevertheless, differences were also found in the right hemisphere, which is aligned with hemispheric symmetry previously discussed in the literature [250], and as can be seen in Figure 9.8, where both left and right values were relevant in the classification. Nevertheless, there is something noteworthy in the relevance of the temporal region and that is that there was no decrease in the length values of this region for those from the SCZ class. There was a decrease in the value of maximum depth in the superior temporal sulcus in SCZ patients, which is consistent with previous work [71]. In fact, this feature is one of the most relevant obtained in the non-parametric approach, see Figure 9.4.

Several other features associated with the temporal cortex can be seen in Figure 9.3. Of these, only the posterior terminal ascending branch of the superior temporal sulcus (referred to as S.T.s.ter.asc.post in the atlas) had a lower average value for SCZ samples. This is also seen in Figure 9.9 by the association of high values of inferior temporal sulcus features with the SCZ class. On the contrary, the length of posterior

occito-temporal lateral sulcus was associated with smaller values for the SCZ class, see Figure 9.8.

As mentioned above, one of the most important regions for the study of schizophrenia is the medical surface of the brain around the cingulate sulcus [270, 266, 271]. However, it is impossible to draw clear conclusions about this area in this study due to the elimination of most of its features at the preprocessing stage by the failure of the sulcal detection software. It is possible that this occurred because the surface morphology of this region varies greatly from one subject to another making it difficult to classify automatically. For this reason, the literature tends to undertake manual detection of this sulcus [270, 71].

The limitations of this study include the reduced number of samples available. With a larger sample size, the results obtained could be strengthened and subtle changes in sulcal dimensions could be analysed in more detail. This is especially important when applying Deep Learning, as shown in its performance in Table 9.2. When introducing the 147 features, the network, although not excessively complex, was not able to obtain robust classifications due to a lack of samples. Therefore, in order to optimise the information extracted from the available data and to avoid the *curse of dimensionality*, in addition to the widely used cross-validation, resubstitution with upper-bound correction was also adopted [130, 36].

This approach allows better performance to be obtained in small sample sizes, especially when the number of features is very small (ideally 1) [37, 36, 9]. This is because it takes advantage of all the samples in the set to fine-tune the classification approach (resubstitution), adjusting the results a posteriori without bias (upper bounding). This is clearly seen in the PLS component and sample size studies in Figure 9.6. For example, while the performance using CV was very similar for different PLS values, RUB managed to improve performance when the number of components was small. With the sample size used in this study something similar happened. RUB managed to improve the results with increasing sample size, while CV remained inconsistent for any sample size. The former was expected, since by increasing the set, the classifier's learning should theoretically improve.

The contrast between validation approaches is also seen in Table 9.1. In this case, by working with a slightly larger number of features, 9, the upper bound obtained to apply in RUB was large, and therefore better results were obtained by applying K-fold. Conversely, when the number of features was 4 (PLS column), the best performance was again achieved by using RUB, irrespective of the upper bound applied, see Figure 9.7. In this figure, when using K-fold, the generalisation capacity of the algorithm was lost. RUB managed to maintain results close to those obtained with the most relevant features in the first columns. Consistently better results were obtained using only the most relevant features compared to dimensionality reduction techniques. This is because in the feature extraction process, all analysed regions were included.

The results in Table 9.1 also indicate better results when using the features selected by parametric rather than non-parametric methods. The difference in accuracy is less

than 4%, so both methods were feasible to use. This suggests that in the absence of normal distributions or with a reduced sample size, non-parametric techniques are a tempting option. However, Figure 9.5 shows how both methods report relevant features.

Future work should expand the analysis to include the interaction between features, as well as comparisons between sulcal and gyral morphological features. For this purpose, further processing tools should be tested, such as calcSulc and Freesurfer with a multidisciplinary working group, in order to be able to analyse in detail all the results obtained. It would also be useful to expand the database to be able to verify the results obtained on an independent dataset. It would even be highly interesting if such an extension could include databases from different regions in order to be able to detect the environmental impact on schizophrenia. Moreover, more specific studies could be conducted, such as the identification of patterns in those who suffer from hallucinations.

10 | EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR IMAGING

10.1 Introduction

This chapter will explore and discuss the application of existing XAI techniques in the field of imaging, with a particular focus on their implementation and effectiveness in analysing 2D drawings related to the study of Alzheimer's Disease. An early diagnosis of AD is crucial for slowing its progression and reducing its impact on the patient's quality of life. Cognitive tests play a significant role in this early diagnosis by assessing various cognitive domains, including memory, attention, language, visuospatial ability, and executive functions. These tests enable healthcare professionals to identify potential cognitive deficits associated with AD and determine the extent of CI.

The Clock Drawing Test is a common paper-and-pencil screening tool for the identification of cognitive changes related to visuospatial functions, frontal lobe execution or memory, among others [272]. During the test, patients are said to draw a clock including the numbers from 1 to 12 and a specific position of the clock hands: ten past eleven. After that, a physician evaluates the resulting drawing and establishes a score, which reflects the patient's cognitive status and detects an eventual CI. This test is widely employed given its simplicity and high sensitivity [273], which can be improved by including additional cognitive tests such as the Mini-Cog [274, 275], to assess memory and other cognitive domains [55]. However, the CDT scoring task

performed by the physician is manual, time-consuming and based on a subjective decision.

The use of CAD systems for the classification of medical imaging is widespread [276, 277, 278, 279, 280], and large part of them are focused on the study of Alzheimer's disease [281, 282, 283, 284]. All these works have in common that they use direct measures from the brain (such as different image modalities or data from EEG) to establish the diagnosis of a specific disorder. Although complexity of the classification task mainly depends on the differences between the two groups to be classified, patterns extracted from direct measures are usually more informative, which means that classification based on them should be easier than when it relies on indirect measures, such as behavioral or test measures. However, the use of indirect measures can also be extremely interesting because of their reduced cost compared to direct ones, as the topic of this work shows. The paper-and-pencil test is much less expensive than acquiring an MRI or a PET scan, which means that it would be extremely relevant to develop a method for detecting cognitive impairment or dementia. In fact, recent works have demonstrated the importance of behavioral data as an indicator of a specific disorder [285, 286]. Therefore, the CDT is a valuable instrument for the study of AD, for which ML and DL techniques have already been used. In fact, the number of works involving ML [287, 288, 289] or DL [290, 291, 57] has increased substantially in recent years for the automatic evaluation of this cognitive test. One of the first studies was conducted by Z. Harbi et al. [288], which established the basis for applying ML methods to automatically interpret and segment CDT drawings.

Most of the published works were focused on a digital version of the CDT, in which a digital ballpoint pen was used instead of a pencil. This allows the acquisition of additional information such as pressure on surface or air-time during drawing [287], which is in line with the current trend towards the use of more advanced technologies to collect and store a larger amount of data more easily. According to previous studies, the digital version of the CDT provides higher diagnostic accuracy than the standard one [292]. Moreover, according to J.Y.C. Chan et al. [293], the pooled sensitivity and specificity obtained screening cognitive impairment in previous studies using digital CDT (dCDT) is higher than CDT. Specifically, the former is associated with a sensitivity and specificity of 0.86% and 0.92%, and the latter with 0.63% and 0.77%, respectively. Despite this boost in performance, the digital version of the test requires expensive equipment compared to the standard one, in which only a pencil and a paper is needed. This can be problematic in scenarios where this technology is not available. Thus, it would be quite interesting to find a methodology that performs similarly to a dCDT but applies only on the classical version of the CDT.

At present, the highest accuracy obtained using data from the digital CDT was published by S. Chen et al. [290], with an accuracy of 96.65%. Nevertheless, the classification problem they posed was not between diagnoses, but between passing or failing the test, using only patients with a diagnosed disease. Thus, cannot be taken as a benchmark. This study was also aimed not only at classifying two conditions

but also at establishing an automatic score for the drawings, where they obtained an accuracy of 72.20%. Another important issue to analyse in previous work on the CDT is the sample size of the studies. Most of them have a very small or unbalanced sample size. For this reason, some apply data augmentation [294] and some transfer learning techniques [57, 290].

In this chapter, an alternative for automatically identifying patients with CI using the classical version of the CDT is proposed. Specifically, this proposal relies on the use of a preprocessing to isolate the regions of interest from all images that are subsequently entered into a CNN. The model is trained in order to find the relationship between the drawings and the diagnosis established by the experts for each patient, i.e. whether or not CI is present. Therefore, the main aim of this method is to identify spatial patterns in the drawings that are relevant in the diagnosis of cognitive impairment.

10.2 Methodology

The system implemented in this study was composed of all the stages of a CAD system: preprocessing, feature extraction and selection and classification. It employed a preprocessing pipeline in which the clock was detected, centered and binarised to decrease the computational burden (section 10.2.1). Then, the resulting image was fed into a CNN to identify the informative patterns within the CDT drawings that were relevant for the assessment of the patient's cognitive status, as it is explained in section 10.2.2. Moreover, XAI methods were applied to identify the most relevant regions during classification, since finding these patterns is extremely helpful to understand the brain damage caused by CI. In order to compare the performance, a ML-based system is implemented as described in section 10.2.1.

For evaluating the performance of the system, the CDT-AD dataset, which is described in section 5.1.4, was used. This dataset comprises a total of 7009 CDT drawings, and when considering a balanced sample size, it was reduced to 3282 drawings. The drawings were drawn by CI and HC individuals.

10.2.1 Image preprocessing

The paper-and-pencil draws of all patients were scanned in order to obtain a digital version of the images. In addition to the drawing of the clock, there is the possibility that the sheet may contain non-relevant information, such as previous drawing attempts, comments from the clinicians and numerical identifiers related to the subject. For example, the existence of a circle smaller than the clock can be seen in the second clock depicted in Figure 5.1 (section 5.1.4), which is a previous attempt. For this reason, a preprocessing process was applied in order to isolate the region of interest (the clock drawing), eliminating the non-relevant information for further analysis.

The left side of Figure 10.1 summarises the different stages of the preprocessing pipeline. First, the original scanned images were converted to grayscale, as the colour (RGB) of the drawing is indifferent. After that, the resulting images were binarised to isolate the pixels contained in the clock from those that form the background. Then, an edge filling process [295] was applied to detect the objects contained in the image and to identify if they belong to the region of interest. This algorithm properly recognises elements even when they are drawn outside the clock face, which is not unusual for numbers 12, 3, 6 and 9, as it can be seen in the sixth clock in Figure 5.1. Finally, the images were cropped and downsampled to a final size of 224x224 to reduce the computational burden while preserving their quality. The resulting images were binary, which means that the intensity of the pixels was 1 for the informative ones and 0 for the rest.

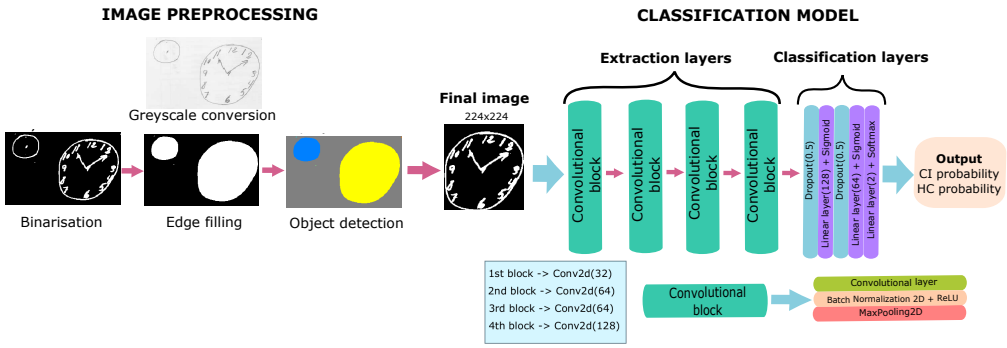


Figure 10.1: Framework of the work. First, the preprocessing of the images are applied. The original image is converted to greyscale to apply binarisation (a manually selected threshold equal for all images), filling of existing elements and detection of objects. Finally, the image is cropped to the clock only and its dimensions are standardised (224x224). The latter is fed into the classification algorithm, a CNN. The architecture consists of four convolutional blocks, including a convolutional layer, bach normalisation and maxpooling, as well as fully-connected layers for the classification stage with dropout.

10.2.2 Deep learning approach

After preprocessing the images, the resulting versions were entered into a deep learning model based on a CNN as it is depicted in the right side of Figure 10.1. There, the architecture of the CNN implemented is illustrated, which includes four 2D convolutional blocks: convolutional layer, batch normalisation, ReLU activation function and a maxpooling layer; and three fully connected layers. Dropout [296] was applied in combination with the linear layers to prevent overfitting, whereas a final softmax layer was added to the model to predict the probability of each sample belonging to the two classes under analysis (CI and HC). Regarding the hyperparameters associated with the CNN, a dropout of 0.5 and an Adam optimisation algorithm with a learning rate of 0.001 were employed. Besides, a Binary Cross-Entropy loss function was used,

whereas the system was trained during 70 epochs employing a batch size of 1.

Visual explainability

Despite the great performance that CNNs offer, a clear disadvantage is that they work as black boxes, which means that it is not easy to explain what the network is basing its decisions on. This is especially problematic in the biomedical field, since any CAD system to be implemented in a clinical environment must be understandable by clinicians and apply trustworthy criteria [297]. For this reason, it is extremely important to provide models that are interpretable in order to widen our knowledge about the reasons why the different classes can be distinguished. To do so, backpropagation-based saliency maps [144] and the Grad-CAM algorithm [146] were used in order to identify which areas of the patients drawings are more relevant in classification. Both approaches are explainable methods that assist in the interpretation of the CNN's predictions, which are described in section 4.7. Neither of them require configuration changes or re-training.

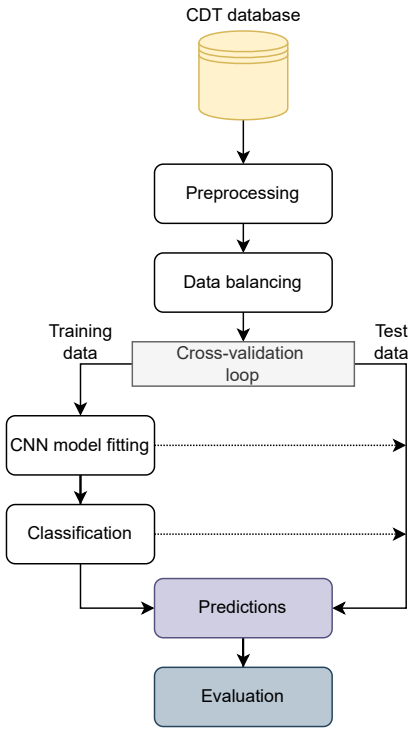
10.2.3 Machine Learning approach

The preprocessed images were also entered into an alternative based on ML that was used as baseline, i.e. a performance to compare with. Following the usual pipeline in classification contexts [35, 298], a method based on PLS [112] was employed to reduce the dimensionality of the input data while extracting informative patterns [187]. The resulting d features (number of components) , $d = 5$ in this work, were then used as input of a SVM classifier with a linear kernel [118].

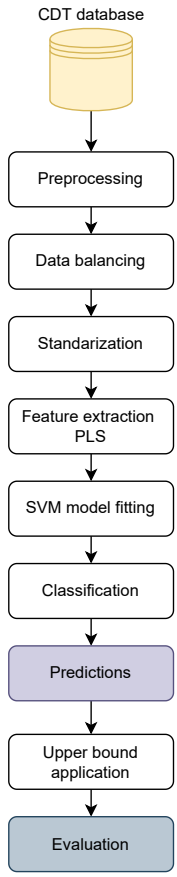
10.2.4 Validation procedure

To assess the reliability of the results, two different validation methods were applied. In the CNN-based approach a 5-fold stratified CV scheme was employed in order to guarantee the independence between the samples used to train the classification model and the ones used for estimating its generalisation ability. As the database was split randomly over $K = 5$ iterations, in each iteration 80% of the database was used as the training set. The remaining 20% was used as the test set, each time using a different fold as a test set. This flowchart is shown in Figure 10.2a. The mean and Standard Deviation (std) of all performance metrics were calculated from the values obtained in the five iterations.

In contrast to the previous scenario, an upper bound-corrected resubstitution was used as a validation method for the ML approach. The flowchart related to this scenario is depicted in Figure 10.2b. The five main components extracted by PLS were used as the input features of the classifier. Regarding the RUB method, Equation (4.12) was applied with a significance level of 0.05.



(a) Cross-validation (DL).



(b) Upper-bounding resubstitution (ML).

Figure 10.2: Flowchart of the analysed models. (a) Flowchart associated with the DL-based model, which is based on CV (K-fold); (b) Flowchart associated with the ML-based model, based on RUB as validation procedure.

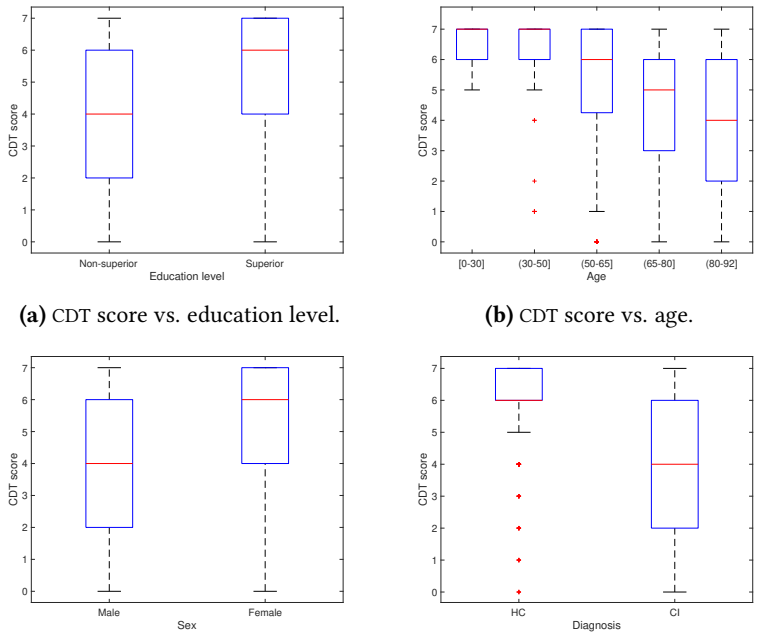
The performance metrics employed for evaluating the results include accuracy, specificity (TN rate) and sensitivity (TP rate), as indicated in Equation (4.16). Moreover, two additional metrics were included: the Positive Predictive Value (PPV) or precision, and the Negative Predictive Value (NPV), as follows:

$$PPV = \frac{T_P}{T_P + F_P} \qquad NPV = \frac{T_N}{T_N + F_N} \tag{10.1}$$

The relevance of these two metrics is that while PPV and NPV depend on the prevalence of the condition in a specific population, whereas sensitivity and specificity depend on the test conducted. Additionally, the AUC was employed as an additional measure for evaluating the ability of the model to identify the different classes [137, 138].

10.3 Results

As a preliminary analysis to the classification, the connection between the quality of the drawing made by the subject (CDT score) and different demographic characteristics was explored. Figure 10.3 contains such analysis using educational level, gender and age as demographic characteristics. For this, only the 1520 samples from FIDYAN Neurocenter were used, data in Table 5.5 (complete dataset), as this information (CDT scores) was not available for the other samples.



(a) CDT score vs. education level.

(b) CDT score vs. age.

(c) CDT score vs. sex.

(d) CDT score vs. diagnosis.

Figure 10.3: Relationship between CDT score and several demographic characteristics given the data subset from FYDIAN neurocenter (1520 samples) of CDT-AD dataset. The '+' marker symbol represents outliers.

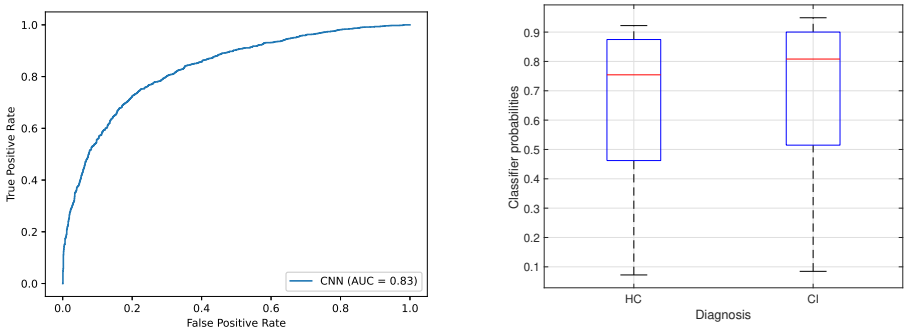
Classification results obtained by each of the methods described in previous sections are shown in Table 10.1. In the CNN scenario, the values for PPV and NPV are 76.86 ± 1.36 and 74.66 ± 1.84, respectively. The ROC curve obtained is depicted in Figure 10.4a with an AUC value of 0.8337. Moreover, the difficulty of classifying the samples according to their class is illustrated in Figure 10.4b by the representation of the CNN output probabilities. In the RUB approach, the accuracy obtained was 72.01% when all the dataset is trained and evaluated, whereas the mean accuracy obtained using the same training subsets as in the CV scenario was 73.37%. In the first case, the upper bound applied was 0.1263 per unit since the sample size was 3282. In the latter case, the upper bound was 0.1394 since the number of samples is lower. In addition, the full unbalanced database was analysed in the CV scenario, where the accuracy obtained was

70.04% with an AUC value of 0.8322. Table 10.2 provides a summary of the performance metrics obtained in recent works focused in classification systems for diagnosis of CI based on the CDT and our results applying the DL approach.

	CNN	SVM+PLS	
Validation	5-fold CV	Resubstitution with upper bounding	
Dataset	Test set	All	CV training set*
Acc (%)	75.65 ± 1.10	72.01	73.37 ± 0.22
Spec (%)	77.82 ± 2.13	70.67	72.11 ± 0.76
Sens (%)	73.49 ± 2.98	73.35	74.65 ± 0.58
PPV (%)	74.66 ± 1.85	72.97	74.36 ± 0.46
NPV (%)	76.86 ± 1.36	71.11	72.46 ± 0.58
AUC	0.8337 ± 0.0143	0.8013	0.8074 ± 0.0029

*Set composed of the same 2625 subsets of samples than in the training set (5-fold experiment). Its upper bound is 0.1394.

Table 10.1: Classification results obtained using CNN and SVM with their different validation methods.



(a) ROC curve.

(b) Output class probabilities.

Figure 10.4: Metrics of interest obtained by the CNN model in conjunction with 5-fold CV. (a) The ROC curve along with the AUC value; (b) Distribution of the output class probabilities of well-classified test samples in its corresponding class, HC or CI patient, in the first CV fold.

In order to verify whether theNN's learning about the distribution of the drawings is correct and makes sense, saliency maps and Grad-CAMs were obtained given an image. The left part of Figure 10.5 shows an example of drawing along with its associated saliency map, illustrated as a heat map (top left) and the different Grad-CAMs depending on whether the label under analysis is that of normal or cognitively impaired subjects (bottom left and right, respectively). In addition, the saliency maps of the correctly classified images have been averaged, separating them into normal and cognitively impaired samples. This was repeated for different sample sizes, 50 and 1250 samples and illustrated in the right part of Figure 10.5.

	Reference	Is the face clock preprinted?	Methodology	Patients (CI/HC)	Accuracy	AUC
Digital Clock Drawing Test	[287]	No	ML methods (best SVM)	2169 (1763/406)	-	0.9100
	[291]	No	Neural networks	198 (163/35)	83.69	-
	[57]	No	Pretrained MobileNet V2	3423 (160/3263)	95.50	0.8130
	[289]	No	Random forest	231 (56/175)	90.48	0.8976
Clock Drawing Test	[290]	Yes	Pretrained DenseNet-121	1315*	96.65	-
	[294]	No	CNN	747 (293/454)	77.37	-
	Our method	**No**	**CNN**	**3282 (1641/1641)**	**75.65**	**0.8337**
	Our method	**No**	**CNN**	**7009 (1641/5368)**	**70.04**	**0.8322**

*Labels in this work were "pass" or "fail" the test, all subjects had at least one positive diagnosis.

-: unknown information.

Table 10.2: Summary of previous works focused on CDT automatic classification in addition to the performance metrics obtained.

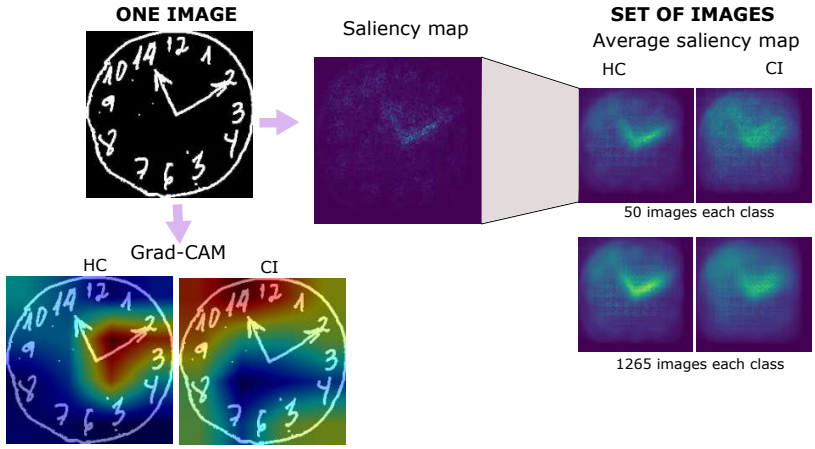


Figure 10.5: Activation maps related to the trained CNN model. In the left part, a particular image of the sample is shown and its Grad-CAM and saliency map. Grad-CAM display the locations where the CNN is focused on detecting each of the classes, HC and CI. The saliency map is represented like a hot map. On the right, it is displayed the average saliency map obtained from several samples (50 up and 1265 down). The images used for averaging are those correctly classified in both the training and test set in the fourth fold of the CV approach.

10.4 Discussion

In this work, a classification framework for the automatic classification of CDT is proposed. This approach is based on a preprocessing pipeline where images are cropped, centered and binarised. Once these images are standardised, they are entered into a CNN model that identifies the most relevant features of each individual class. The performance in the CDT is evaluated to differentiate between patients diagnosed with CI and HC. Besides, activation maps were employed in order to visually verify

that the training of the DL model was performed correctly. Moreover, it was proved whether the results were reliable by implementing a ML model based on theoretical limits with similar results. The underlying hypothesis for this classification is that there are differences between the drawings made by HC and those with CI. As shown in Figure 10.3d, those with CI tend to score lower on the test. Furthermore, this score tends to decrease with age (Figure 10.3b), which is closely related to the decline of cognitive abilities [299, 300].

The results obtained indicate that the methodology proposed outperforms the expected performance according to J.Y.C Chan et al. [293] when the paper-and-pencil CDT is analysed: a mean sensitivity and specificity of 0.63% and 0.77%, respectively. Moreover, results from our previous work have been surpassed by more than 7 points in accuracy [41]. The reason for this improvement is the increase in the number of samples in the database from less than 1000 to more than 3000 samples in the balanced case. Previous studies [291, 301, 302] have developed systems for the automatic diagnosis of CI from the CDT with better results than those obtained in this work, as it can be seen in Table 10.2. Nevertheless, these works rely on the use of a digital version of the CDT. This leads to a higher variety of features to be employed, resulting in a considerable increase in performance. It should be highlighted that these devices are not common in clinical centres, and even unaffordable for hospitals of some regions. Therefore, it is also necessary to continue the development of automatic classification systems for the original CDT. With respect to the results obtained in other studies using the paper-and-pencil CDT, it should be pointed out that this method led to a lower accuracy than the one obtained in Y.C. Youn et al. [294]. However, one of the limitations of that study was the small unbalanced sample size, 747, whereas our sample size is 3282. Therefore, these results have a more reliable generalisability. In addition, as far as we know the implemented study is the one with the largest number of real samples so far, 7009, although the accuracy rate is lower due to the high imbalance.

Another relevant aspect of this study is that it was not provided a preprinted face clock to patients to make them fill the rest of information [288, 290]. This has as an important consequence that all the resulting drawings have a common part: the circumference used as the face clock. When trying to learn the relevant aspects of a clock, using a preprinted circumference decreases the variability between the different drawings. This has two main consequences: first, the differences (if so) between the drawings of both classes (CI and HC) must be inside the face clock. Second, the model can detect easier the presence of the clock since all of them have the same structure. However, the way patients draw the face clock can also contain vital information about their cognitive state. The main drawback is that modeling and identifying the informative patterns associated with each individual drawing is not a straightforward task, since the variability between clocks is much higher. The large sample size and the performance obtained in this work manifests the ability of the implemented method for accurately extracting the drawing pattern of a person with CI, regardless of individual differences in the way the face clock is drawn.

Comparing the results obtained in our previous work [41] with those obtained in this study, it can be seen that the accuracy has increased using the CV approach (from 68.62% to 75.65%) while with the RUB-based approach the performance has slightly decreased (from 74.25% to 73.37%). This is due to the increase of the sample size. On the one hand, a methodology based on DL requires the use of a large sample size in order to learn and generalise well [196]. Therefore, increasing the database has improved the generalisation ability of the DL model. On the other hand, the ML model applied is a linear classifier, so it is logical that the extension of the database does not lead to an improvement in the results, since the classification ability of linear kernels can be limited. Therefore, the linear approach allowed to establish a theoretical range of performance, which was overcome by enlarging the sample. Moreover, it also demonstrates the advantages of applying a simpler structure in conjunction with a resubstitution-based method when the sample size is very small, which is common in the field of neuroimaging [36]. It offers the advantage of being able to use the complete dataset for the evaluation of the results. The upper bound correction implies not to consider the empirical error obtained but the actual one, setting an upper limit on the theoretical accuracy the classifier is able to perform. In this work, Vapnik's bound (see Equation (4.12)) was chosen because it is the best known, but there is a high number of bounds proposals that can be applied, as the ones described in section 4.5.2.

From a visual perspective, the results reflect what can be expected from the drawings. Figure 10.5 clearly shows that the NN focuses on the position of the clock hands during classification. This relatively differs from the clinicians' criteria, as they focus on more diverse elements. Nevertheless, it is still valuable information as it allows for the identification of patterns in a large number of subjects (right part of Figure 10.5), or even highlighting the cognitive importance of setting the hands of the clock correctly. The patterns of each class indicate that, while in HC the relevant features are located in central positions, in CI subjects this information is around the edges. This is due to the high variability between subjects in these areas, especially those who do not draw a clock correctly. This is supported by Figure 10.5, where the clock hands are easily identified in the average activation map of HC subjects but in the map associated with CI patients the zone of interest is much imprecise. Moreover, the hand-clock zone becomes more intense as the sample size increases. This can also be connected to the probabilities shown in Figure 10.4b. As the network is especially focused on the clock hands, if a HC subject makes the drawing more irregular, the output probability of belonging to the HC class is lower.

The fact that a subject is healthy does not imply that their drawing is perfect. This can be observed in the preliminary study shown about the CDT scores in Figure 10.3d. For example, educational level leads to a better score (Fig 10.3a), or even the sex (Figure 10.3c). The latter may be due to the fact that historically women have tended to be more involved in some educational activities, such as drawing or reading, while men are more devoted to physical and social activities from an early age [303, 304, 305]. This is supported by the statistical differences shown in Table 5.5. For example, the CI class tends to have a lower level of education [306].

Therefore, the model presented in this chapter offers reliable results that would allow the CAD system to be implemented as a method to help specialists in clinical tasks. The method performs both the preprocessing and the classification stages and has been tested using a large sample size. Moreover, the samples used were real drawings, which provides valuable information. This made it possible to analyse drawing patterns based on patients' condition instead of focusing only on the accuracy rate. Besides, the large performance obtained in the analysis of the paper-and-pencil based clock test demonstrates that it is a reliable and cost-effective method for being used as an aid for clinicians in any hospital and research centre. This will require the implementation of a low-cost tool, e.g. through a web-based app on smartphones, that allows clinicians to take a picture of the drawing and to analyse it, from the preprocessing phase to the extraction of a diagnostic conclusion. Although the conclusion would not be ultimate, it could show the probabilities returned by the network for each class. In this way, the clinician can be aware of the usefulness of the tool in each specific case.

Future work could be focus on the implementation of more sophisticated classification systems based on ensemble frameworks [179, 279, 122] in order to obtain a similar performance than when using digitised versions of the test. The aim is to be able to establish a helpful tool for clinicians.

Part III

GENERAL DISCUSSION AND CONCLUSIONS

11 | GENERAL DISCUSSION AND CONCLUSIONS

11.1 General Discussion

The contributions presented in this thesis have already been extensively discussed in their respective chapters within Part II. In this final chapter, the focus will be on evaluating the impact of this work on the field of neuroimaging, particularly in relation to the application of Machine Learning in CAD systems. Additionally, findings related to the various brain disorders analysed in this thesis will be discussed, with emphasis on their relevance and implications.

11.1.1 Discussion on the Algorithms

Machine Learning techniques have experienced significant expansion in our daily lives, revolutionising various fields, including medicine. In the context of medicine, ML, which includes DL algorithms, has emerged as a powerful tool for analysing complex data and making accurate predictions. In particular, in the field of neuroimaging, ML has opened new avenues for understanding the brain and improving patient care. This kind of algorithms can assist in the early detection and diagnosis of brain disorders, such as Alzheimer's Disease or Parkinson's Disease, identifying subtle abnormalities or biomarkers that may indicate the presence of a particular condition. This early detection can significantly impact patient prognosis by enabling timely intervention and treatment. Furthermore, ML has facilitated the prediction of clinical outcomes in neuroimaging studies, such as disease progression, treatment response, or even cognitive decline. In addition to diagnosis and prediction, ML has also improved image

analysis methods in neuroimaging (segmentation, brain mapping, noise reduction, etc.). Nevertheless, applying ML in neuroimaging faces several challenges and limitations.

Firstly, obtaining high-quality and well-annotated neuroimaging data can be a challenging task. Neuroimaging datasets often suffer from limited size, posing obstacles for training and validating ML models that should rely on a sufficient amount of data to yield reliable and generalisable results. Additionally, the absence of standardised protocols in medical centers leads to variability in imaging acquisition procedures and data processing. Consequently, comparing and combining data across different studies becomes arduous, ultimately affecting the quality and reproducibility of results in the field of neuroimaging.

Secondly, the interpretability of MLmodels in neuroimaging is a significant concern. This is especially true for DL models, which are known for their black-box nature, making it difficult to understand the underlying features and decision-making processes. This lack of interpretability can hinder the trust and acceptance of ML methods, especially in multidisciplinary work where the ultimate goal is clinical application, where transparency and explainability are crucial for clinical decision-making.

Another challenge is the need for computationally efficient algorithms. Neuroimaging data is typically large and complex (e.g. millions of voxels), requiring substantial computational resources for processing and analysis. Developing efficient algorithms that can handle the scale and complexity of neuroimaging data is crucial for practical and real-time applications.

Lastly, the ethical considerations and privacy concerns associated with neuroimaging data should not be overlooked. Protecting patient privacy and ensuring the ethical use of data are essential when applying ML techniques in neuroimaging research and clinical practice.

This thesis is focused on the first two challenges mentioned above, which are addressed in the contributions presented in the previous chapters. It was already stated in **chapter 1** that the aim of this thesis was to explore different approaches to achieve reliable and interpretable CAD systems in neuroimaging. To this end, different methods have been analysed and proposed to address the sample size problem and to be able to increase the generalisability of the classifier. Moreover, XAI algorithms have been applied in the CAD systems implemented to provide the system with greater clinical interpretability.

A first approach to CAD systems is provided in **chapter 6**, which is published in [35]. In this chapter, the application of ML techniques in a multiclass scenario is assessed through the performance achieved in an international contest for automated prediction of MCI from MRI data. Here, the main challenge was the efficient optimisation of a system that simultaneously involve several classes. Nevertheless, while working through the system, the problematic nature of the neuroimaging data also became apparent. Neuroimaging datasets (e.g. MRI scans) often contain high-dimensional data with millions of voxels or data points. In this case, the starting point was preprocessed

data consisting of cortical and subcortical MRI features rather than the scans themselves, and yet typical failures were detected in preprocessing procedures that generated a significant amount of outliers (see Figure 6.2). Not to mention that even when reducing the dimensionality of the data from millions of voxels to 417 features per sample, the problem of small sample size still existed, and the *curse of dimensionality* was still present as the ratio of features (417) vs. samples (250, 60 per class) was still high. Therefore, the main concern in the development of the CAD system for the competition was considered to be the processing of the data. Feature selection and dimensionality reduction techniques were employed to extract relevant and informative features from the data. This made it possible to implement efficient algorithms and computational techniques, improving model performance, as as shown in the research group's ranking (SiPBA-UGR) in the competition (Table 6.3).

After a thorough analysis of the features, the selected method was based on a one vs. one approach for feature selection (t-tests as sorting criteria for filtering), PLS feature extraction and classification. Such methodology was capable of identifying the most relevant features for a multiclass classification by a sorting-and-filtering method, and was evaluated using different parameters and classifiers, see Table 6.2. Moreover, the final selected algorithm computation was fast unlike other DL-based methodologies tested in this study, addressing the problem of computational efficiency in neuroimaging.

The final results outperformed all the proposals submitted to the challenge by more than 5 percentage points in accuracy, see Table 6.4. The method was also coherent with recent findings in CAD of AD (dominance of left-sided hemisphere regions), and can be applied to other multiclass classification problems. Furthermore, the control of the FWE rate in the proposed system was evaluated based on the resubstitution estimation using the HC subset. The null hypothesis that no group difference in the feature set should be true could not be rejected, indicating that the method has a good control of the FWE rate. Another verification of the significance of the selected features was performed using RUB. This approach returned undesirable high actual errors when classifying HC vs. MCI and MCI vs. cMCI (Table 6.5). This implies that the identified features cannot be accepted as statistically significant in classifying these conditions at the specified significance level (0.05). However, it is widely recognised that distinguishing these conditions is a challenging task, and in general, relevance is observed in the other classifications.

Looking at the sample size rather than the features themselves, one of the limitations of the study was the limited number of training samples. When using K-fold CV this is more pronounced, since some mismatches between training fitting and final test estimations can be expected, limiting the generalisability of the model.

This is precisely the topic covered in **chapter 7**. The chapter includes a comprehensive comparison of K-fold and RUB as validation methods in various scenarios (balanced, unbalanced, binary, and multiclass datasets), all of which share the small sample size problem. The comparison was conducted using a non-parametric statistical

inference framework based on permutations, which was published in [36]. The primary objective of this framework was to assess the performance of the proposed models (and its capacity for generalisation) in typical neuroimaging scenarios, focusing on statistical power and type I error control. This process made it possible to determine the most suitable validation method in each case.

Overall, a similar performance was found using resubstitution and K-fold in small sample size scenarios. Moreover, the implementation of an upper limit on resubstitution (RUB), resulted in the elimination of the positive bias related to resubstitution in classification. Therefore, as explored extensively in this chapter and discussed more broadly in other chapters, the use of RUB as a validation method presents certain advantages in the field of neuroimaging. For example, the most relevant would be not to have to split the dataset into several subsets, since it is already limited in size. Such segregation tends to generate imbalance and increased variability. In contrast, using the whole sample set at once increases the capacity of generalisation of the model by having more samples to learn from. In addition, the computational cost is also greatly reduced.

A summary of the results and conclusions reached in each analysed scenario can be found at Table 7.7. For example, in the DIAN-AD dataset, a tendency that statistical power is slightly lower using CV than RUB was observed. The main reason is that this sample set is not easily classifiable since the samples are highly heterogeneous as they are young subjects, most of them without signs of dementia at the time of data collection. Therefore, the division of the sample set into subsets only increases the difficulty of obtaining good generalisability. Moreover, in the statistical power experiment, e.g. using the ADNI-AD dataset, the CV test accuracy and resubstitution accuracy were similar (Table 7.2), which differs from the accuracies obtained using the permuted datasets, where similarities were found between the mean values related to the CV test subset and RUB. Hence, employing the RUB method yields a worst-case accuracy estimation that is comparable to or even more conservative than the accuracy obtained using the CV test subsets, with the advantage of using the complete set (more reliable).

Good control of false positives was also observed in all experiments. Both validation approachs, CV and RUB, obtained a FP rate very close to the significance level for any input dimension. However, several high standard deviations were observed in the results (able 7.6), which is consistent with existing literature which indicates that a high sample size is needed in order to obtain a low FWE [15].

Furthermore, when analysing the variability generated by including or not including feature extraction step in the permutation process, a certain decrease in the standard deviation was detected when such step was not included (see Figure 7.3b). This variability is an important aspect to highlight. The current neuroscience scenario involves using high complexity classifiers in complex spaces, unlike the linear SVM and low-dimensional classifiers used in this analysis. Therefore, if variability is observed in such low-dimensional spaces, it raises concerns about the reliability of classification

in highly complex scenarios. Moreover, in general, the models demonstrated optimal data-fitting capability by not adapting to corrupted data. Nevertheless, it was observed that this capacity to adapt to corrupted data increased as the DL architecture became more sophisticated. All of these findings highlight the growing need for reliable assessment instruments of ongoing classification systems, and one potential solution could be modifying the validation process by incorporating methods such as RUB. However, currently, the use of RUB as a validation method faces a significant limitation. The theoretical limits established so far cannot be applied, particularly in the case of DL-based algorithms, due to their inherent computational complexity. Future research aims to develop methods that can compute these more complex bounds for NNs classification [196, 212].

In the meantime, this validation method can not only be used for classification, but can also be an element of image analysis enhancement, in particular brain mapping, as it is explored in **chapter 8**. The generation of statistical maps based on data rather than classical statistics, especially in those studies involving VBM, increases the reliability of such maps since the latter rely on assumptions that are frequently violated [32, 33]. Moreover, In terms of classical statistics (SPM) the small sample size problem derives in a challenging scenario that constrains the generalisation of the results from small datasets to new unseen samples.

The proposed data-driven approach, presented in [37], is mainly devoted to classification problems with limited sample sizes, to derive statistical model-free (agnostic) mappings. It is important to note that this approach is not specifically designed for testing competing hypotheses or comparing different models in the field of neuroimaging. However, SAM was derived based on the assumption of the existence of classes (H_1) at the voxel or multi-voxel level. The analysis of the worst case considers the upper bounds of the actual risk, under suitable theoretical conditions and a selection of regions with a highly-corrected empirical risk, according with a test for significance on a population proportion.

Based on various experiments conducted, both reported in this thesis and others mentioned [37, 221, 226], it was observed that SAM addressed the issue of instability in limited sample sizes when determining relevance maps for neurological conditions such as AD or PD. It proved to be a competitive and complementary method to the widely accepted SPM framework in the neuroimaging community. For instance, SAM was less influenced by sample size when detecting ROIs, as depicted in Figure 8.4b. Regarding the control of FWE rate, a comparison was made between SAM using Equation (4.13) for the upper bound and different SPM approaches (cluster and voxel-wise inferences). It was found that FWE rates for SPM cluster-wise inference exceeded the nominal level, while SPM voxel-wise inference was valid but overly conservative. On the other hand, the region-wise (SAM) inference with corrections closely approached the predefined levels and was nearly independent of the number of samples.

Therefore, the proposed method is applicable to neuroimaging modalities such as MRI, SPECT, or PET, which are more focused on spatial brain mapping. It is of interest

to verify whether the procedure can be extrapolated to other modalities, such as EEG, which are oriented towards temporal brain studies. In Section 8.5, this conversion is analysed, and although the results have room for improvement (see Figure 8.8), it is confirmed that the method can indeed be used in other types of neuroimaging studies. It would be advisable to establish a significance test more adjusted to this type of data, more closely resembling the results obtained with K-fold, for example, by using alternative upper bounds.

The analysis of the suitability of using parametric and non-parametric techniques in neuroimaging continues in **chapter 9**, where the relevance of sulcal features was explored to detect patterns of case-control differences in SCZ. Additionally, in this chapter, XAI techniques are applied, which are becoming more and more important for CAD systems as they offer a straightforward means to enhance the interpretability of the classification process.

Current methodologies for the extraction of sulcal features have made it possible to work with this type of data, which has been use in the literature but still offers significant room for expansion and improvement. Due to the need for enhancing these methods, a standard gold method, similar to SPM for VBM, has not yet been established. In this study, BrainVISA was used for automatically extracting values of sulcus depth and length. However, several sulci were misdetected or not even identified, which limited the exploitation of the entire dataset. Consequently, only a portion of the sulci (specifically, 49 of them) could be analysed. To assess the relevance of these areas, a feature selection step using parametric and non-parametric approaches. Shapiro-Wilk test and t-test were the parametric approaches applied, while Mann-Whitney U test and a classification based on SAM (RUB and proportion test) were used as non-parametric methods to detect the relevance of each feature. Figure 9.5 illustrated the most relevant features to differentiate SCZ from HC subjects, and some of them coincide in both procedures. In fact, the performance is very similar when applying the two approaches in classification, see Table 9.7, although slightly better results are obtained when using the features selected by parametric rather than non-parametric methods (less than 4%). These findings suggest that non-parametric techniques are an appealing alternative when statistical assumptions (e.g. normal distributions) cannot be considered or the sample size is limited. Precisely for the problem of the small sample size, the study conducted in the feature extraction phase revealed how the validation method influences the generalisability of the classifier. The results indicated that RUB exhibited more consistent behavior compared to K-fold. Figure 9.6b demonstrated that the K-fold method was unable to improve classifier performance as the sample size increased, when the opposite is expected.

The XAI techniques were employed in the proposed DL model due to the insufficient accuracy obtained with the ML model, as shown in Table 9.2. However, the results obtained with the NN were also suboptimal due to the limited sample size. Consequently, despite the network's moderate complexity, it faced challenges in achieving reliable classifications when incorporating the 147 features. The clinical implications of these

findings will be discussed in the subsequent section. It is worth noting that increasing the sample size could potentially improve the results and enable a more detailed analysis of subtle variations in sulcal dimensions.

This limitation was not faced in the study described in **chapter 10** and presented in [38], where more than 7000 samples were available. Such amount of samples (real drawings) made possible to efficiently implement a CNN model that identified the most relevant patterns to differentiate between patients diagnosed with CI and HC. In this case, image-oriented XAI techniques were used to detect patterns in the CDT drawings that were of interest in the classifier's decision-making. Both the saliency maps and Grad-CAM algorithm revealed expected patterns of interest, which will be discussed more extensively in the following section. These techniques provide useful insight into the behaviour of the classifier, especially in interdisciplinary works where approval from both technicians and clinicians is necessary. Notably, the results obtained demonstrate that cognitive tests not only hold clinical utility as initial diagnostic tools but can also be effectively leveraged to automate the diagnostic process and facilitate assessment by specialists.

The achieved performance aligns with the expected outcomes when using an analogical version of the CDT, even surpassing the mean values of sensitivity and specificity reported in the literature [293]. While it is true that some works have achieved better results, they either relied on the dCDT (which provides better features for classification) or had small, unbalanced sample sizes (making them less reliable). A summary of these works can be found in Table 10.2. In order to assess the reliability of the obtained performance, a comparison was made with other approaches, both for classification (SVM) and validation (RUB). This alternative configuration helped establish the theoretical accuracy that the classifier could achieve.

An interesting discovery was that the performance improved compared to a previous study [41], which had a smaller sample size. However, this improvement was observed only in the case of the DL model (CNN and K-fold), whereas the SVM and RUB approach did not exhibit significant enhancements with the increased dataset size. This finding underscores the importance of having a larger sample size for effective learning and generalisation in DL methodologies, as they heavily rely on sample data for optimal performance [196]. In contrast, linear classifiers like SVM are less likely to benefit significantly from enlarging the database. Thus, it is evident that simpler structures in conjunction with resubstitution-based methods prove particularly advantageous when dealing with limited sample sizes, as these approaches use the entire set for learning. On the other hand, when a sufficient sample set is available, more complex techniques can be employed, and the samples can even be divided into subsets for classification.

In a nutshell, this thesis has analysed in detail the methods used in each of the stages of a CAD system. The importance of applying selection and feature extraction techniques has been proven. Different approaches for the validation of results have been compared, especially taking into account the typical peculiarities that appear in a small

sample size scenario, which is so common in neuroimaging. Finally, various techniques have been implemented for different data modalities to improve the interpretability of the outcomes of a CAD system, which is so necessary in the multidisciplinary field of neuroimaging.

11.1.2 Discussion on the Disorders

It is precisely because neuroimaging is such a multidisciplinary field that this thesis goes beyond technical results by exploring their implications for the brain disorders studied: Alzheimer's Disease, Parkinson's Disease and Schizophrenia.

In the study of Alzheimer's Disease, the most significant contribution has been the detection of patterns using the drawings made for the CDT. These drawings tend to be done more poorly as the person ages, which is consistent with the cognitive decline that occurs with ageing, see Figure 10.3b. Likewise, these drawings allow the detection of patterns that differentiate healthy individuals and those with CI, often associated with the development of AD. In this study, the analysis involved hand-drawn clock drawings, with the notable aspect that even the clock face was not preprinted. This ensured the inclusion of the entire spectrum of variability. Consequently, this approach captures all available information necessary for detecting the cognitive state of the patient.

The results obtained in this work reflected a high relevance in classification in the position of the clock hands. It was found that, while the relevant patterns for HC subjects were located in central positions of the clock, in CI subjects this pattern were observed around the edges. These edges were related, for example, to the fact that these individuals sometimes could not even draw the clock face correctly. This observation is clearly depicted in Figure 10.5, where the average activation map of HC subjects highlights the identification of clock hands, while the map corresponding to CI patients exhibits a less precise ROI. These findings contrast with the criteria commonly employed by clinicians, who typically consider a broader range of factors, including numbers, clock size, and the clock face, among others. However, it provides valuable information to know that when considering such a large number of samples, a noteworthy pattern is detected in the placement of the clock hands.

One of the main advantages of conducting a study on the paper-and-pencil-based clock test is its applicability in hospitals and medical clinics globally, as it does not require expensive resources. Additionally, this type of analog test is easier to perform than tests involving the use of ballpoint pens. Therefore, it is valuable to continue enhancing the diagnostic capacity of this test.

To a lesser extent, Parkinson's Disease has been studied. In this case, the feasibility of using other types of imaging, different from SPECT, has been analysed in order to avoid the associated disadvantages. The results indicate that by applying statistical brain maps, ROIs are not obtained from MRI scans, whereas with SPECT these regions are statistically significant. This can be seen by comparing Figure 8.6 (SPECT) and

Figure 8.7 (MRI). Therefore, it can be stated that MRI scans are currently not a reliable source of information for the diagnosis of PD. This contradicts several studies in the literature, where their use in CAD systems is proposed. However, these studies have been analysed and some weaknesses have been found, such as bias [232] or an incorrect CV process [233]. Thanks to the development of new and better biomarker proposals, in the future it will be feasible to use other more efficient and advantageous methods for the study of PD.

Finally, this thesis has investigated the morphological alterations in the brain associated with Schizophrenia. Sulcal features of the entire cerebral cortex were automatically extracted to perform a case-control comparison. At the time of writing this thesis, no previous studies have reported the use of sulcal features from the entire cortex in a CAD system similar to the one used. Discovering patterns at this morphological level is particularly intriguing since these patterns develop during the prenatal period and early years of life, making it highly valuable to enhance our understanding of these changes.

The achieved performance reveals potentially interesting features, such as the collateral fissure or the superior postcentral intraparietal superior sulcus, which have not been previously reported in terms of their length and depth. In addition, the expected results have been obtained. For example, the left hemisphere has been identified as predominant, both in the feature selection stage and in the analysis of the classifier decision process using XAI techniques. Similar expected results have been observed in the temporal and precentral areas. Figure 9.9 (top left) demonstrates a negative correlation between the intermediate precentral sulcus. Furthermore, a decrease in the maximum depth value in the superior temporal sulcus has been observed in SCZ patients. Precisely, this pattern makes this feature one of the most relevant findings obtained in the non-parametric approach, as shown in Figure 9.4.

Overall, advancements in brain mapping and the analysis of sulcal features present an opportunity for a more comprehensive morphological investigation of not only these specific diseases but also any brain disorder that induces alterations in the cerebral cortex. Such advancements hold the potential to enhance our understanding of the etiology and progression of brain disorders and enable earlier diagnoses, ultimately improving the quality of life for patients.

11.2 Conclusions and Future work

In recent years, the field of neuroimaging has increasingly embraced the use of ML techniques as a valuable tool for diagnosing and predicting neurological and psychiatric disorders. However, many neuroimaging studies have predominantly focused on observational and mechanistic approaches, aiming to identify differences in brain structure and function among different groups, rather than solely focusing on classification tasks. This thesis aimed to demonstrate the utility of ML techniques not

only in classification scenarios but also in brain mapping and the detection of relevant features. Additionally, the thesis examined the reliability and interpretability of CAD systems and explored ways to improve these aspects. In summary, the conclusions drawn from the contributions made in this thesis are as follows:

- The application of an upper-bounded resubstitution throughout the various chapters has demonstrated its viability in neuroimaging as validation method. This theoretical estimation of the error or risk associated with a classification algorithm enhances its capacity of generalisation by leveraging information from all available samples. This becomes particularly valuable in scenarios with small sample sizes, which are commonly encountered in neuroimaging studies.

- Another approach to enhance reliability is by considering the stages of feature selection and extraction. These stages play a vital role in identifying any inconsistencies in the data and optimising the performance of the classifiers. Additionally, these steps aim to reduce computational burdens, which can be significantly high when analysing high-dimensional data in neuroimaging.

- In fact, a CAD system is proposed in chapter 6 for *A Machine learning neuroimaging challenge for automated diagnosis of Mild Cognitive Impairment* in which the feature extraction and selection step was of utmost relevance. The detailed study of this stage allowed a considerable improvement in performance compared to all the proposals submitted to the challenge. This post-competition method was capable of identifying the most relevant features for a multiclass classification by a sorting-and-filtering method, and was evaluated using different parameters and classifiers. The method was also coherent with recent findings in CAD of AD, and could be applied to other multiclass classification problems.

- The most comprehensive comparison between RUB and K-fold CV was conducted in chapter 7 using a non-parametric methodology to analyse the prediction certainty in CAD systems. To this end, the trade-off between statistical power and Type I error was explored. Both validation approaches, CV and RUB, obtained a FP rate very close to the significance level for any input dimension. Moreover, both approaches offered acceptable statistical power, although it was slightly lower using CV. Nevertheless, the generalisation ability could be optimised by using RUB as validation method instead of CV as it was observed that RUB is less influenced by overfitting when a correction is applied. All this with the advantage of using the whole sample set. Thus, considering also that the computational cost per iteration is lower using RUB than CV, its use for statistical studies is recommended.

- The SAM methodology, proposed in chapter 8, is a data-driven approach based on RUB and a test of proportions. It offers a feasible approach for deriving statistical model-free (agnostic) mappings. Through experiments conducted in various experimental frameworks and datasets, this multivariate approach has

demonstrated its capability to establish a novel model-free method for assessing significant changes across brain volumes. It has been rigorously developed and tested in scenarios with varying sample/dimension ratios and effect sizes ranging from large to trivial. It exhibited effective control over FPs and yielded consistent results across different sample sizes. Consequently, it serves as a highly competitive and complementary alternative to the widely accepted SPM framework in the neuroimaging community, which relies on parametric assumptions that are often difficult to satisfy.

- The current stage of development for SAM primarily focuses on providing an alternative method to SPM. However, there is potential for the method to be adapted to handle other types of contrasts or factorial analyses, taking advantage of the numerous advancements in ML in recent years. In fact, this thesis explored the possibility of extrapolating spatial functionality to temporal EEG studies. Nevertheless, in order to enhance its performance, refinement of the method is necessary. This would involve modifying the selection process for TPOIs and adopting a less conservative upper bound that better accommodates this type of data.

- The study in chapter 9 revealed the potential of combining various statistical, ML, and DL techniques with sulcal features to address a novel SCZ case-control classification task. It sheds light on the challenges encountered when using novel features derived from datasets with limited sample sizes in classification tasks. The methods employed in the study showcased the efficacy of feature extraction and selection techniques, as well as validation methods like RUB, in effectively addressing the inherent difficulties associated with limited data. Furthermore, the study confirmed the value of XAI techniques in enhancing interpretability and gaining insights into the importance of each feature in the classification process.

- From a clinical perspective, the findings obtained in this study were highly intriguing as no previous research has automated the analysis of sulcal features across the entire cerebral cortex. This process had enabled the replication of findings from prior studies in temporal and precentral areas, while also providing new insights into the underlying mechanisms in SCZ. These results offered valuable information that contributes to a deeper understanding of the condition.

- Finally, the CAD system proposed in chapter 10 aimed to automate the diagnosis of CI using the classical version of the CDT. The achieved performance aligned with expectations when employing an analogical version of the test, and the use of a large number of real samples ensured the reliability of the results. Additionally, graph-theoretical methods were applied to enhance interpretability and identify relevant patterns of information. Specifically, the position of the clock hands was detected as ROI. These findings demonstrated the suitability of the system for implementation in hospitals and medical clinics worldwide,

particularly in regions with limited resources where the analog version of the test is used.

Additionally, considering the strengths and limitations encountered during the development of this thesis, there are several promising directions for future research. The work presented is limited to the application of the upper bound associated with linear classifiers. Future work may focus on implementing and testing other bounds on more complex classifiers, for example, in NNs and comparing their performance with other validation methods. Another appealing line of research would be to extend the analysis of sulcal features to investigate their interrelationships and establish comparisons between sulcal and gyral morphological features. This avenue of investigation should not be limited to SCZ but could also be explored in other disorders such as AD or PD. By doing so, a broader understanding of the role of cortical morphology in various neurological conditions could be gained.

Part IV

APPENDIX

A | SUPPLEMENTARY MATERIAL FOR THE NON-PARAMETRIC STATISTICAL INFERENCE FRAMEWORK

This chapter includes supplementary material related to chapter 7. Such material includes further information from the analyses performed and examples, such as the multiclass study.

A.1 ADNI-AD: class selection

An ensemble methodology was used with the ADNI-AD dataset using several brain regions instead of the whole brain. The method of label estimation after the classification was conducted was as follows. To obtain the label for a sample y, posterior probabilities returned by SVM for each region are necessary, so that for each sample the total probability of belonging to a class, C, depends on the L regions:

$$p^c_{total}(y) = \sum_l \frac{p^c_l(y)}{n_{regions}} \tag{A.1}$$

where $p^c_{total}(y)$ is the probability of a sample y being of class c, and the term $p^c_l(y)$ is the probability of a sample y being of class c in a specific region l. Then, the maximum value among the probabilities for each class is chosen as a label:

$$Label(y) = \arg\max_y \left(p^c_{total}(y)\right) \tag{A.2}$$

A.2 A multiclass experiment: KAGGLE-AD

This section includes the complete study developed with the KAGGLE-AD multiclass dataset. The main difference between this dataset and those analysed in chapter 7 is the fact that there are more than two classes in the dataset. Thus, this experiment was conducted from three perspectives: two-condition (AD vs. HC) and four-condition (AD vs. cMCI vs. MCI vs. HC) permutation distributions were used for statistical power assessment, in addition to one-condition analysis to conduct the estimation of type error I control with HC subjects.

A.2.1 Randomisation on multiclass distributions

For the analysis of statistical power with a multiclass dataset, the process is actually the same as in the other experiments, since a one-vs-one methodology [35] is applied. Following the proposed method in section 7.2.3, the test statistics were first obtained with values $\mathcal{T}^{KCV} = 0.5604$ and $\mathcal{T}^{RUB} = 0.6142$ in the four-condition analysis (AD vs. cMCI vs. MCI vs. HC). Regarding the upper bound, since $n = 400$, $d = 1$ and $\eta = 0.05$ and Equation (4.13) is applied, it gives $\mu = 0.0679$. Then, the permutation test was conducted. The results are summarised in Table A.1. Note that as a four-sample test, the mean of the permuted distribution tends to be 25% instead of 50%. The obtained p-values were 0.0040 and 0.0010 for CV and RUB, respectively. Therefore, the null hypothesis of independence between labels and samples was rejected with both methods.

		Accuracy	**Sensitivity**	***p*-value**
10-fold CV				
Original dataset	Training	0.5823 [0.0071]	0.5823 [0.0071]	-
	Test	0.4396 [0.0181]	0.4396 [0.0181]	-
Permuted dataset	Training	0.4066 [0.0250]	0.4066 [0.0250]	-
	Test	0.2487 [0.0687]	0.2487 [0.0687]	0.0040 [0.0020]
Resubstitution				
Original dataset		0.4537 [0.0162]	0.4537 [0.0162]	-
Permuted dataset		0.3375 [0.0184]	0.3375 [0.0184]	0.0010 [0.0010]
Upper-bounded resubstitution ($\mu = 0.0679$)				
Original dataset		0.3858 [0.0162]	0.3858 [0.0162]	-
Permuted dataset		0.2696 [0.0184]	0.2696 [0.0184]	0.0010 [0.0010]

Symbol "-" indicates that values were not computable and μ stands for upper bound.

Table A.1: Results related to the statistical power experiment using KAGGLE-AD original and permuted dataset (4 classes). Validation methods applied for the permuted dataset were 10-fold CV (100 iterations, medium computational cost, top), resubstitution (1000 iterations, medium computational cost, middle) and RUB (by applying the upper bound, low computational cost, bottom). Original dataset scores were obtained from 20 iterations (low computational cost). The significance level of the test was 0.05.

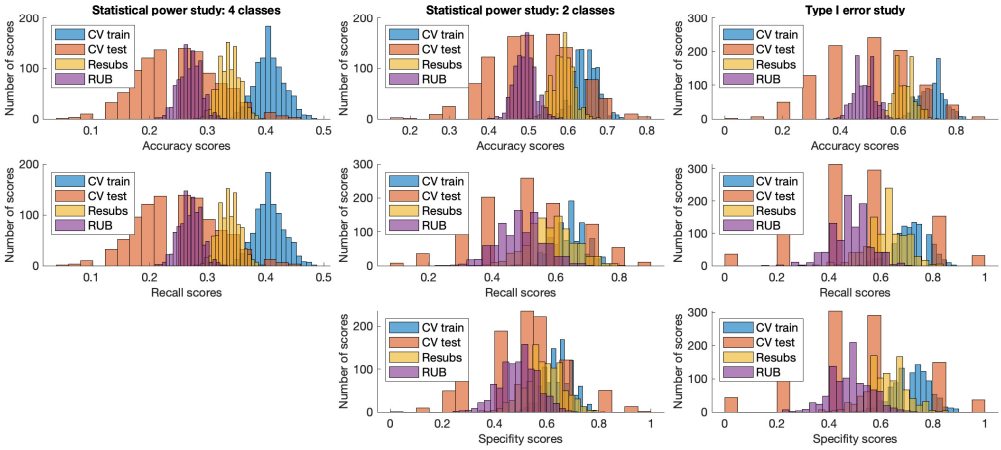


Figure A.1: Distribution of scores among the samples of the permuted KAGGLE-AD dataset (1000 iterations) in the statistical power study for four conditions (left), for two conditions (middle) and in type I error study (right). In the RUB study, the bounds applied are $\mu = 0.0679$, $\mu = 0.0960$ and $\mu = 0.1358$ in the four-condition, two-condition and one-condition experiments, respectively.

A.2.2 Randomisation on HC vs. AD

		Accuracy	Sensitivity	Specifity	p-value
10-fold CV					
Original dataset	Training	0.9141 [0.0158]	0.9288 [0.0227]	0.8994 [0.0224]	-
	Test	0.8458 [0.0775]	0.8510 [0.1084]	0.8405 [0.1165]	-
Permuted dataset	Training	0.6455 [0.0302]	0.6441 [0.0527]	0.6468 [0.0534]	-
	Test	0.5067 [0.1061]	0.5052 [0.1574]	0.5083 [0.1625]	0.0010 [0.0010]
Resubstitution					
Original dataset		0.8162 [0.0167]	0.8540 [0.0216]	0.7785 [0.0283]	-
Permuted dataset		0.5873 [0.0263]	0.5868 [0.0728]	0.5877 [0.0723]	0.0010 [0.0010]
Upper-bounded resubstitution ($\mu = 0.0960$)					
Original dataset		0.7202 [0.0167]	0.7580 [0.0216]	0.6825 [0.0283]	-
Permuted dataset		0.4913 [0.0263]	0.4908 [0.0728]	0.4917 [0.0723]	0.0010 [0.0010]

Symbol "-" indicates that values were not computable and μ stands for upper bound.

Table A.2: Results related to the statistical power experiment using KAGGLE-AD original and permuted AD-vs-HC subset. Validation methods applied for the permuted dataset were 10-fold CV (100 iterations, medium-low computational cost, top), resubstitution (1000 iterations, medium-low computational cost, middle) and RUB (by applying the upper bound, low computational cost, bottom). Original dataset scores were obtained from 20 iterations (low computational cost). The significance level of the test was 0.05.

Table A.2 summarises the results for the statistical power assessment in the two-condition distribution (AD vs. HC). In this case, the actual errors associated to the original dataset were $\mathcal{T}^{KCV} = 0.1542$ and $\mathcal{T}^{RUB} = 0.2798$ and the mean actual error of

the null distributions were 0.4933 using 10-fold CV and 0.5087 with RUB. The values used for computation of the upper bound were the same as those for the four-condition analysis, except the number of samples, which was 200 in this case. Thus, the value of the upper bound was 0.0960. Figure A.1 (middle) shows the distribution of the accuracies and other metrics obtained in this analysis. Figure A.1 (left) also illustrates the distribution related to the multiclass experiment. In this study, p-values were 0.0010 in both cases, CV and RUB. Thus, there was an effect in the samples given the labels.

A.2.3 Randomisation on controls

The HC subset was used for type I error control estimation. Results related to this permutation test are shown in Table A.3 whilst its distribution is illustrated in Figure A.1 (right). In this case the number of samples was 100, so the upper bound was equal to 0.0960. The mean actual errors obtained in this case were 0.5108 using 10-fold CV and 0.5171 using RUB. From these null distributions, the FWE rate obtained was 0.0480 applying CV, close to the significance level $\alpha = 0.05$. The value obtained by applying RUB was 0.0330.

		Accuracy	Sensitivity	Specifity	FWE rate
10-fold CV					
Permuted dataset	Training	0.7175 [0.0428]	0.7172 [0.0650]	0.7179 [0.0667]	-
	Test	0.4892 [0.1557]	0.4902 [0.2279]	0.4882 [0.2344]	0.0480 [0.0068]
Resubstitution					
Permuted dataset		0.6187 [0.0356]	0.6196 [0.0733]	0.6178 [0.0765]	0.0330 [0.0056]
Upper-bounded resubstitution (μ = 0.1358)					
Permuted dataset		0.4829 [0.0356]	0.4838 [0.0733]	0.4820 [0.0765]	0.0330 [0.0056]

Symbol "-" indicates that values were not computable and μ stands for upper bound.

Table A.3: Results related to the Type I error experiment using KAGGLE-AD permuted HC subset. Validation methods applied for the permuted dataset were 10-fold CV (100 iterations, medium-low computational cost, top), resubstitution (1000 iterations, medium-low computational cost, middle) and RUB (by applying the upper bound, low computational cost, bottom). The significance level of the test was 0.05.

A.3 Variability in feature extraction: extent of results

This section includes the accuracies related to the experiment where the permutation loop was modified. Its related p-values and FWE ratios are shown in Table 7.8. Table A.4 shows the results related to the statistical power experiments and Table A.5 indicates the accuracies obtained in the Type I error control assessments.

		ADNI-AD	KAGGLE-AD (multiclass)	KAGGLE-AD (binary)	DIAN-AD
10-fold CV					
Original dataset	Training	0.8611 [0.0063]	0.5904 [0.0150]	0.9109 [0.0129]	0.7157 [0.0166]
	Test	0.8341 [0.0535]	0.4581 [0.0721]	0.8540 [0.0741]	0.6148 [0.0948]
Permuted dataset	Training	0.4898 [0.0822]	0.4001 [0.0208]	0.6456 [0.0314]	0.6272 [0.0273]
	Test	0.4730 [0.0762]	0.2469 [0.0673]	0.5015 [0.1102]	0.4974 [0.0977]
Resubstitution					
Original dataset		0.8369 [0.0000]	0.4273 [0.0021]	0.8100 [0.0000]	0.6545 [0.0000]
Permuted dataset		0.4918 [0.0244]	0.2714 [0.0153]	0.5247 [0.0271]	0.5177 [0.0217]
RUB		μ = 0.0665	μ = 0.0679	μ = 0.0960	μ = 0.0866
Original dataset		0.7704 [0.0000]	0.3594 [0.0021]	0.7140 [0.0000]	0.5679 [0.0000]
Permuted dataset		0.4253 [0.0244]	0.2035 [0.0153]	0.4287 [0.0271]	0.4311 [0.0217]

Table A.4: Accuracies from the statistical power assessment using the alternative scheme. Validation methods applied for the permuted dataset were 10-fold CV (100 iterations, medium computational cost, top), resubstitution (1000 iterations, low computational cost, middle) and RUB (by applying the upper bound, μ, low computational cost, bottom). Original dataset scores were obtained from 20 iterations (low computational cost). The significance level, μ of the test was 0.05.

		ADNI	MCI Prediction	DIAN
10-fold CV				
Permuted dataset	Training	0.5064 [0.0924]	0.7178 [0.0419]	0.6893 [0.0376]
	Test	0.4267 [0.1177]	0.4949 [0.1517]	0.4939 [0.1422]
Resubstitution				
Permuted dataset		0.4882 [0.0328]	0.5324 [0.0353]	0.5299 [0.0343]
RUB		μ = 0.0897	μ = 0.1358	μ = 0.1225
Permuted dataset		0.3524 [0.0328]	0.4099 [0.0353]	0.4074 [0.0343]

Table A.5: Accuracies from the Type I error assessment using the alternative scheme. Validation methods applied for the permuted dataset were 10-fold CV (100 iterations, medium computational cost, top), resubstitution (1000 iterations, low computational cost, middle) and RUB (by applying the upper bound, μ, low computational cost, bottom). The significance level, μ, of the test was 0.05.

BIBLIOGRAPHY

[1] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "AI in health and medicine," *Nature Medicine*, vol. 28, pp. 31–38, jan 2022.

[2] S. O'Sullivan, F. Jeanquartier, C. Jean-Quartier, A. Holzinger, D. Shiebler, P. Moon, and C. Angione, "Developments in AI and machine learning for neuroimaging," in *Artificial Intelligence and Machine Learning for Digital Pathology*, pp. 307–320, Springer International Publishing, 2020.

[3] D. B. Dwyer, P. Falkai, and N. Koutsouleris, "Machine learning approaches for clinical psychology and psychiatry," *Annual review of clinical psychology*, vol. 14, pp. 91–118, 2018.

[4] M. Alber, A. Buganza Tepole, W. R. Cannon, S. De, S. Dura-Bernal, K. Garikipati, G. Karniadakis, W. W. Lytton, P. Perdikaris, L. Petzold, *et al.*, "Integrating machine learning and multiscale modeling—perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences," *NPJ digital medicine*, vol. 2, no. 1, p. 115, 2019.

[5] T. Carlson, E. Goddard, D. M. Kaplan, C. Klein, and J. B. Ritchie, "Ghosts in machine learning for cognitive neuroscience: Moving from data to theory," *NeuroImage*, vol. 180, pp. 88–100, oct 2018.

[6] A. N. Nielsen, D. M. Barch, S. E. Petersen, B. L. Schlaggar, and D. J. Greene, "Machine learning with neuroimaging: Evaluating its applications in psychiatry," *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, vol. 5, no. 8, pp. 791–798, 2020.

[7] L. Khedher, J. Ramírez, J. M. Górriz, A. Brahim, F. Segovia, and A. Disease Neuroimaging Initiative, "Early diagnosis of alzheimer's disease based on partial least squares, principal component analysis and support vector machine using segmented MRI images," *Neurocomputing*, vol. 151, pp. 139–150, mar 2015.

[8] S. Vieira, W. H. L. Pinaya, and A. Mechelli, "Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications," *Neuroscience & Biobehavioral Reviews*, vol. 74, pp. 58–75, mar 2017.

[9] D. Castillo-Barnes, L. Su, J. Ramírez, D. Salas-Gonzalez, F. J. Martinez-Murcia, I. A. Illan, F. Segovia, A. Ortiz, C. Cruchaga, M. R. Farlow, C. Xiong, N. R. Graff-Radford, P. R. Schofield, C. L. Masters, S. Salloway, M. Jucker, H. Mori, J. Levin, J. M. Gorriz, and D. I. A. N. (DIAN), "Autosomal dominantly inherited alzheimer disease: Analysis of genetic subgroups by machine learning," *Information Fusion*, vol. 58, pp. 153–167, jun 2020.

[10] H. Choi, "Deep learning in nuclear medicine and molecular imaging: Current perspectives and future directions," *Nuclear Medicine and Molecular Imaging*, vol. 52, pp. 109–118, nov 2017.

[11] A. M. Chekroud and N. Koutsouleris, "The perilous path from publication to practice," *Molecular Psychiatry*, vol. 23, pp. 24–25, nov 2017.

[12] D. Szucs and J. P. Ioannidis, "Sample size evolution in neuroimaging research: An evaluation of highly-cited studies (1990–2012) and of latest practices (2017–2018) in high-impact journals," *NeuroImage*, vol. 221, p. 117164, nov 2020.

[13] H. G. Schnack and R. S. Kahn, "Detecting neuroimaging biomarkers for psychiatric disorders: sample size matters," *Frontiers in psychiatry*, vol. 7, p. 50, 2016.

[14] C. S. Carter, T. A. Lesh, and D. M. Barch, "Thresholds, power, and sample sizes in clinical neuroimaging," *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, vol. 1, no. 2, pp. 99–100, 2016.

[15] G. Varoquaux, "Cross-validation failure: Small sample sizes lead to large error bars," *Neuroimage*, vol. 180, pp. 68–77, oct 2018.

[16] K. S. Button, J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. J. Robinson, and M. R. Munafò, "Power failure: why small sample size undermines the reliability of neuroscience," *Nature Reviews Neuroscience*, vol. 14, pp. 365–376, Apr. 2013.

[17] J. E. Desmond and G. H. Glover, "Estimating sample size in functional mri (fmri) neuroimaging studies: statistical power analyses," *Journal of neuroscience methods*, vol. 118, no. 2, pp. 115–128, 2002.

[18] R. R. Krishnaiah and L. N. Kanal, eds., *Handbook of Statistics*, ch. Dimensionality and sample size considerations in pattern recognition practice, pp. 825–855. North-Holland, 1982.

[19] Y.-D. Zhang, Z. Dong, S.-H. Wang, X. Yu, X. Yao, Q. Zhou, H. Hu, M. Li, C. Jiménez-Mesa, J. Ramirez, *et al.*, "Advances in multimodal data fusion in neuroimaging: overview, challenges, and novel orientation," *Information Fusion*, vol. 64, pp. 149–187, 2020.

[20] K. L. Moulder, B. J. Snider, S. L. Mills, V. D. Buckles, A. M. Santacruz, R. J. Bateman, and J. C. Morris, "Dominantly inherited alzheimer network: facilitating research and clinical trials," *Alzheimer's research & therapy*, vol. 5, pp. 1–7, 2013.

[21] P. A. Bandettini, *FMRI*, ch. Twenty-six controversies and challenges, pp. 163–214. MIT press, 2020.

[22] D. C. V. Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, and K. Ugurbil, "The WU-minn human connectome project: An overview," *NeuroImage*, vol. 80, pp. 62–79, oct 2013.

[23] K. L. Miller, F. Alfaro-Almagro, N. K. Bangerter, D. L. Thomas, E. Yacoub, J. Xu, A. J. Bartsch, S. Jbabdi, S. N. Sotiropoulos, J. L. R. Andersson, L. Griffanti, G. Douaud, T. W. Okell, P. Weale, I. Dragonu, S. Garratt, S. Hudson, R. Collins, M. Jenkinson, P. M. Matthews, and S. M. Smith, "Multimodal population brain imaging in the UK biobank prospective epidemiological study," *Nature Neuroscience*, vol. 19, pp. 1523–1536, sep 2016.

[24] J. Nalepa, M. Marcinkiewicz, and M. Kawulok, "Data augmentation for brain-tumor segmentation: A review," *Frontiers in Computational Neuroscience*, vol. 13, dec 2019.

[25] B. Mwangi, T. S. Tian, and J. C. Soares, "A review of feature reduction techniques in neuroimaging," *Neuroinformatics*, vol. 12, pp. 229–244, sep 2013.

[26] S. Rathore, M. Habes, M. A. Iftikhar, A. Shacklett, and C. Davatzikos, "A review on neuroimaging-based classification studies and associated feature extraction methods for alzheimer's disease and its prodromal stages," *NeuroImage*, vol. 155, pp. 530–548, jul 2017.

[27] X. Hao, Y. Bao, Y. Guo, M. Yu, D. Zhang, S. L. Risacher, A. J. Saykin, X. Yao, and L. Shen, "Multi-modal neuroimaging feature selection with consistent metric constraint for diagnosis of alzheimer's disease," *Medical Image Analysis*, vol. 60, p. 101625, feb 2020.

[28] F. J. Martinez-Murcia, A. Ortiz, J.-M. Gorriz, J. Ramirez, and D. Castillo-Barnes, "Studying the manifold structure of alzheimer's disease: A deep learning approach using convolutional autoencoders," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, pp. 17–26, jan 2020.

[29] M. Khodatars, A. Shoeibi, D. Sadeghi, N. Ghaasemi, M. Jafari, P. Moridian, A. Khadem, R. Alizadehsani, A. Zare, Y. Kong, A. Khosravi, S. Nahavandi, S. Hussain, U. R. Acharya, and M. Berk, "Deep learning for neuroimaging-based diagnosis and rehabilitation of autism spectrum disorder: A review," *Computers in Biology and Medicine*, vol. 139, p. 104949, dec 2021.

[30] P. T. Reiss, "Cross-validation and hypothesis testing in neuroimaging: An irenic comment on the exchange between friston and lindquist et al.," *NeuroImage*, vol. 116, pp. 248–254, aug 2015.

[31] J. M. Górriz, J. Ramirez, and J. Suckling, "On the computation of distribution-free performance bounds: Application to small sample sizes in neuroimaging," *Pattern Recognition*, vol. 93, pp. 1–13, sep 2019.

[32] K. Friston, "Sample size and the fallacies of classical inference," *NeuroImage*, vol. 81, pp. 503–504, nov 2013.

[33] M. A. Lindquist, B. Caffo, and C. Crainiceanu, "Ironing out the statistical wrinkles in "ten ironic rules"," *NeuroImage*, vol. 81, pp. 499–502, nov 2013.

[34] A. Eklund, T. E. Nichols, and H. Knutsson, "Cluster failure: Why fmri inferences for spatial extent have inflated false-positive rates," *Proceedings of the National Academy of Sciences*, vol. 113, pp. 7900–7905, jun 2016.

[35] C. Jimenez-Mesa, I. A. Illan, A. Martin-Martin, D. Castillo-Barnes, F. J. Martinez-Murcia, J. Ramirez, and J. M. Gorriz, "Optimized one vs one approach in multi-class classification for early Alzheimer's disease and Mild Cognitive Impairment diagnosis," *IEEE Access*, vol. 8, pp. 96981–96993, 2020.

[36] C. Jimenez-Mesa, J. Ramirez, J. Suckling, J. Vöglein, J. Levin, and J. M. Gorriz, "A non-parametric statistical inference framework for deep learning in current neuroimaging," *Information Fusion*, vol. 91, pp. 598–611, mar 2023.

[37] J. Gorriz, C. Jimenez-Mesa, R. Romero-Garcia, F. Segovia, J. Ramirez, D. Castillo-Barnes, F. Martinez-Murcia, A. Ortiz, D. Salas-Gonzalez, I. Illan, C. Puntonet, D. Lopez-Garcia, M. Gomez-Rio, and J. Suckling, "Statistical agnostic mapping: A framework in neuroimaging based on concentration inequalities," *Information Fusion*, vol. 66, pp. 198–212, feb 2021.

[38] C. Jiménez-Mesa, J. E. Arco, M. Valentí-Soler, B. Frades-Payo, M. A. Zea-Sevilla, A. Ortiz, M. Ávila-Villanueva, D. Castillo-Barnes, J. Ramírez, T. del Ser-Quijano, C. Carnero-Pardo, and J. M. Górriz, "Using explainable artificial intelligence in the clock drawing test to reveal the cognitive impairment pattern," *International Journal of Neural Systems*, jan 2023.

[39] C. Jimenez-Mesa, J. M. Peñalver, D. Lopez-Garcia, J. Ramirez, C. Gonzalez-Garcia, F. Segovia, J. Suckling, and J. M. Gorriz, "Standarization of agnostic learning techniques in Neuroimaging: a case study in EEG," in *2022 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, pp. 1–4, IEEE, (Conference proceedings not yet published), 2022.

[40] C. Jimenez-Mesa, D. Castillo-Barnes, J. E. Arco, F. Segovia, J. Ramirez, and J. M. Górriz, "Analyzing statistical inference maps using MRI images for Parkinson's

disease," in *Artificial Intelligence in Neuroscience: Affective Analysis and Health Applications*, pp. 166–175, Springer International Publishing, 2022.

[41] C. Jiménez-Mesa, J. E. Arco, M. Valentí-Soler, B. Frades-Payo, M. A. Zea-Sevilla, A. Ortiz, M. Ávila, D. Castillo-Barnes, J. Ramírez, T. del Ser-Quijano, C. Carnero-Pardo, and J. M. Górriz, "Automatic classification system for diagnosis of cognitive impairment based on the clock-drawing test," in *Artificial Intelligence in Neuroscience: Affective Analysis and Health Applications*, pp. 34–42, Springer International Publishing, 2022.

[42] A. Campero, P. Ajler, J. Emmerich, E. Goldschmidt, C. Martins, and A. Rhoton, "Brain sulci and gyri: A practical anatomical review," *Journal of Clinical Neuroscience*, vol. 21, pp. 2219–2225, dec 2014.

[43] D. Weishaupt, V. D. Köchli, B. Marincek, J. M. Froehlich, D. Nanz, and K. P. Pruessmann, *How does MRI work?: an introduction to the physics and function of magnetic resonance imaging*, vol. 2. Springer, 2006.

[44] J. R. Evans and A. Abarbanel, *Introduction to quantitative EEG and neurofeedback*. Elsevier, 1999.

[45] H. Begleiter and B. Porjesz, "Genetics of human brain oscillations," *International Journal of Psychophysiology*, vol. 60, no. 2, pp. 162–171, 2006.

[46] H. H. Jasper, "Report of the committee on methods of clinical examination in electroencephalography, electroencephalogr," *Clin. Neurophysiol*, vol. 10, pp. 370–375, 1958.

[47] P. Saxena, D. G. Pavel, J. C. Quintana, and B. Horwitz, "An automatic threshold-based scaling method for enhancing the usefulness of tc-hmpao spect in the diagnosis of alzheimer's disease," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 623–630, Springer, 1998.

[48] D. Salas-Gonzalez, J. M. Górriz, J. Ramírez, M. López, I. A. Illan, F. Segovia, C. G. Puntonet, and M. Gómez-Río, "Analysis of spect brain images for the diagnosis of alzheimer's disease using moments and support vector machines," *Neuroscience Letters*, vol. 461, no. 1, pp. 60–64, 2009.

[49] S. Arndt, T. Cizadlo, D. O'Leary, S. Gold, and N. C. Andreasen, "Normalizing counts and cerebral blood flow intensity in functional imaging studies of the human brain," *Neuroimage*, vol. 3, no. 3, pp. 175–184, 1996.

[50] D. Salas-Gonzalez *et al.*, "Linear intensity normalization of fp-cit spect brain images using the α-stable distribution," *NeuroImage*, vol. 65, pp. 449–455, jan 2013.

[51] S. J. Teipel, J. Kurth, B. Krause, and M. J. Grothe, "The relative importance of imaging markers for the prediction of Alzheimer's disease dementia in mild cognitive impairment — Beyond classical regression," *NeuroImage: Clinical*, vol. 8, pp. 583–593, Jan. 2015.

[52] E. H. Weissler, T. Naumann, T. Andersson, R. Ranganath, O. Elemento, Y. Luo, D. F. Freitag, J. Benoit, M. C. Hughes, F. Khan, *et al.*, "The role of machine learning in clinical research: transforming the future of evidence generation," *Trials*, vol. 22, no. 1, pp. 1–15, 2021.

[53] B. J. Mainland and K. I. Shulman, "Clock drawing test," in *Cognitive Screening Instruments*, pp. 67–108, Springer International Publishing, 2017.

[54] J. C. Morris, "Clinical dementia rating: A reliable and valid diagnostic and staging measure for dementia of the alzheimer type," *International Psychogeriatrics*, vol. 9, pp. 173–176, dec 1997.

[55] D. Palsetia, G. P. Rao, S. C. Tiwari, P. Lodha, and A. De Sousa, "The clock drawing test versus mini-mental status examination as a screening tool for dementia: a clinical comparison," *Indian journal of psychological medicine*, vol. 40, no. 1, pp. 1–10, 2018.

[56] A. So, D. Hooshyar, K. Park, and H. Lim, "Early diagnosis of dementia from clinical data by machine learning techniques," *Applied Sciences*, vol. 7, p. 651, jun 2017.

[57] S. Amini, L. Zhang, B. Hao, A. Gupta, M. Song, C. Karjadi, H. Lin, V. B. Kolachalama, R. Au, and I. C. Paschalidis, "An artificial intelligence-assisted method for dementia detection using images from the clock drawing test," *Journal of Alzheimer's Disease*, vol. 83, pp. 581–589, Sep 2021.

[58] Alzheimer's Association, "2023 Alzheimer's Disease Facts and Figures," *Alzheimer's & Dementia*, vol. 19, pp. 1598–1695, mar 2023.

[59] P. J. Modrego, "Predictors of conversion to dementia of probable Alzheimer type in patients with mild cognitive impairment," *Current Alzheimer Research*, vol. 3, pp. 161–170, Apr. 2006.

[60] C. Davatzikos, "Machine learning in neuroimaging: Progress and challenges," *NeuroImage*, vol. 197, pp. 652–656, Aug. 2019.

[61] F. d. Vos, T. M. Schouten, A. Hafkemeijer, E. G. P. Dopper, J. C. v. Swieten, M. d. Rooij, J. v. d. Grond, and S. A. R. B. Rombouts, "Combining multiple anatomical MRI measures improves Alzheimer's disease classification," *Human Brain Mapping*, vol. 37, no. 5, pp. 1920–1929, 2016.

[62] J. Claus, F. Van Harskamp, M. Breteler, E. Krenning, I. De Koning, T. Van der Cammen, A. Hofman, and D. Hasan, "The diagnostic value of spect with tc 99m hmpao in alzheimer's disease: A population-based study," *Neurology*, vol. 44, no. 3 Part 1, pp. 454–454, 1994.

[63] W. Poewe *et al.*, "Parkinson's disease," *Nature Reviews Disease Primers*, vol. 3, mar 2017.

[64] N. I. of Neurological Disorders and S. (US), *Parkinson's disease: Challenges, progress, and promise.* National Institute of Neurological Disorders and Stroke, National Institutes …, 2004.

[65] Y. Uno and J. T. Coyle, "Glutamate hypothesis in schizophrenia," *Psychiatry and clinical neurosciences*, vol. 73, no. 5, pp. 204–215, 2019.

[66] T. R. Insel, "Rethinking schizophrenia," *Nature*, vol. 468, no. 7321, pp. 187–193, 2010.

[67] P. C. Sallet, H. Elkis, T. M. Alves, J. R. Oliveira, E. Sassi, C. C. de Castro, G. F. Busatto, and W. F. Gattaz, "Reduced cortical folding in schizophrenia: an mri morphometric study," *American Journal of Psychiatry*, vol. 160, no. 9, pp. 1606–1613, 2003.

[68] J. Yan, Y. Cui, Q. Li, L. Tian, B. Liu, T. Jiang, D. Zhang, and H. Yan, "Cortical thinning and flattening in schizophrenia and their unaffected parents," *Neuropsychiatric Disease and Treatment*, pp. 935–946, 2019.

[69] E. Walton, D. P. Hibar, T. G. van Erp, S. G. Potkin, R. Roiz-Santiañez, B. Crespo-Facorro, P. Suarez-Pinilla, N. E. Van Haren, S. M. de Zwarte, R. S. Kahn, *et al.*, "Positive symptoms associate with cortical thinning in the superior temporal gyrus via the enigma schizophrenia consortium," *Acta Psychiatrica Scandinavica*, vol. 135, no. 5, pp. 439–447, 2017.

[70] E. Walton, D. P. Hibar, T. G. Van Erp, S. G. Potkin, R. Roiz-Santiañez, B. Crespo-Facorro, P. Suarez-Pinilla, N. E. van Haren, S. De Zwarte, R. S. Kahn, *et al.*, "Prefrontal cortical thinning links to negative symptoms in schizophrenia via the enigma consortium," *Psychological medicine*, vol. 48, no. 1, pp. 82–94, 2018.

[71] C. P. E. Rollins, J. R. Garrison, M. Arribas, A. Seyedsalehi, Z. Li, R. C. K. Chan, J. Yang, D. Wang, P. Liò, C. Yan, Z. hui Yi, A. Cachia, R. Upthegrove, B. Deakin, J. S. Simons, G. K. Murray, and J. Suckling, "Evidence in cortical folding patterns for prenatal predispositions to hallucinations in schizophrenia," *Translational Psychiatry*, vol. 10, nov 2020.

[72] W. J. Conover, *Practical nonparametric statistics*, vol. 350. john wiley & sons, 1999.

[73] D. B. Panagiotakos, "The value of p-value in biomedical research," *The open cardiovascular medicine journal*, vol. 2, p. 97, 2008.

[74] R. Kelter, "Analysis of type i and II error rates of bayesian and frequentist parametric and nonparametric two-sample hypothesis tests under preliminary assessment of normality," *Computational Statistics*, vol. 36, pp. 1263–1288, sep 2020.

[75] M. Hollander, D. A. Wolfe, and E. Chicken, *Nonparametric statistical methods*, vol. 751. John Wiley & Sons, 2013.

[76] S. S. SHAPIRO and M. B. WILK, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, pp. 591–611, dec 1965.

[77] B. L. Welch, "The generalization of 'student's'problem when several different population varlances are involved," *Biometrika*, vol. 34, no. 1-2, pp. 28–35, 1947.

[78] T. K. Kim, "T test as a parametric statistic," *Korean journal of anesthesiology*, vol. 68, no. 6, pp. 540–546, 2015.

[79] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The Annals of Mathematical Statistics*, vol. 18, pp. 50–60, mar 1947.

[80] M. P. Fay and M. A. Proschan, "Wilcoxon-mann-whitney or t-test? on assumptions for hypothesis tests and multiple interpretations of decision rules," *Statistics Surveys*, vol. 4, jan 2010.

[81] A. C. Aitken, "Iv.—on least squares and linear combination of observations," *Proceedings of the Royal Society of Edinburgh*, vol. 55, pp. 42–48, 1936.

[82] J.-B. Poline and M. Brett, "The general linear model and fmri: does love last forever?," *Neuroimage*, vol. 62, no. 2, pp. 871–880, 2012.

[83] K. J. Friston, A. P. Holmes, K. J. Worsley, J.-P. Poline, C. D. Frith, and R. S. J. Frackowiak, "Statistical parametric maps in functional imaging: A general linear approach," *Human Brain Mapping*, vol. 2, no. 4, pp. 189–210, 1994.

[84] J. Ashburner and K. J. Friston, "Voxel-based morphometry—the methods," *Neuroimage*, vol. 11, no. 6, pp. 805–821, 2000.

[85] W. D. Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols, *Statistical parametric mapping: the analysis of functional brain images*. Elsevier, 2011.

[86] T. E. Nichols and A. P. Holmes, "Nonparametric permutation tests for functional neuroimaging: a primer with examples," *Human brain mapping*, vol. 15, no. 1, pp. 1–25, 2002.

[87] M. Brett, W. Penny, and S. Kiebel, "Introduction to random field theory," *Human brain function*, vol. 2, pp. 867–879, 2003.

[88] T. Nichols and S. Hayasaka, "Controlling the familywise error rate in functional neuroimaging: a comparative review," *Statistical methods in medical research*, vol. 12, no. 5, pp. 419–446, 2003.

[89] Y. Hochberg, "Multiple comparison procedures," *Wiley Series in Probability and Statistics*, 1987.

[90] K. J. Worsley, A. C. Evans, S. Marrett, and P. Neelin, "A three-dimensional statistical analysis for cbf activation studies in human brain," *Journal of Cerebral Blood Flow & Metabolism*, vol. 12, no. 6, pp. 900–918, 1992.

[91] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[92] P. Golland, F. Liang, S. Mukherjee, and D. Panchenko, "Permutation tests for classification," in *Learning Theory*, pp. 501–515, Springer Berlin Heidelberg, 2005.

[93] P. Good, *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media, 2013.

[94] S. Hayasaka and T. E. Nichols, "Validating cluster size inference: random field and permutation methods," *Neuroimage*, vol. 20, no. 4, pp. 2343–2356, 2003.

[95] R. Heller, Y. Golland, R. Malach, and Y. Benjamini, "Conjunction group analysis: an alternative to mixed/random effect analysis," *Neuroimage*, vol. 37, no. 4, pp. 1178–1185, 2007.

[96] M. Ojala and G. C. Garriga, "Permutation tests for studying classifier performance," in *2009 Ninth IEEE International Conference on Data Mining*, IEEE, dec 2009.

[97] A. P. Holmes, R. Blair, J. Watson, and I. Ford, "Nonparametric analysis of statistic images from functional mapping experiments," *Journal of Cerebral Blood Flow & Metabolism*, vol. 16, no. 1, pp. 7–22, 1996.

[98] J. V. Haxby, "Multivariate pattern analysis of fmri: the early beginnings," *Neuroimage*, vol. 62, no. 2, pp. 852–855, 2012.

[99] J. V. Haxby, A. C. Connolly, J. S. Guntupalli, *et al.*, "Decoding neural representational spaces using multivariate pattern analysis," *Annual review of neuroscience*, vol. 37, no. 1, pp. 435–456, 2014.

[100] A. B. Graf and S. Borer, "Normalization in Support Vector Machines," in *Pattern Recognition* (B. Radig and S. Florczyk, eds.), Lecture Notes in Computer Science, pp. 277–282, Springer Berlin Heidelberg, 2001.

[101] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*, vol. 793. John Wiley & Sons, 2019.

[102] L. Yuan, Y. Wang, P. M. Thompson, V. A. Narayan, J. Ye, A. D. N. Initiative, *et al.*, "Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data," *NeuroImage*, vol. 61, no. 3, pp. 622–632, 2012.

[103] H. He and Y. Ma, *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons, 2013.

[104] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, vol. 4. Springer, 2006.

[105] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[106] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, pp. 389–422, 2002.

[107] M. H. Kutner, C. J. Nachtsheim, J. Neter, W. Li, *et al.*, *Applied linear statistical models*, vol. 5. McGraw-Hill Irwin Boston, 2005.

[108] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, no. 1, pp. 267–288, 1996.

[109] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, Oct. 2001.

[110] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, pp. 433–459, jun 2010.

[111] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016.

[112] S. Wold, A. Ruhe, H. Wold, and I. W. J. Dunn, "The collinearity problem in linear regression. the partial least squares (PLS) approach to generalized inverses," *SIAM Journal on Scientific and Statistical Computing*, vol. 5, pp. 735–743, sep 1984.

[113] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, pp. 504–507, jul 2006.

[114] D. Charte, F. Charte, S. García, M. J. del Jesus, and F. Herrera, "A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software and guidelines," *Information Fusion*, vol. 44, pp. 78–96, nov 2018.

[115] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[116] S. Singh and P. Gupta, "Comparative study id3, cart and c4. 5 decision tree algorithm: a survey," *International Journal of Advanced Information Science and Technology (IJAIST)*, vol. 27, no. 27, pp. 97–103, 2014.

[117] V. N. Vapnik, *Statistical Learning Theory*. John Wiley and Sons, Inc., New York, 1998.

[118] B. Schölkopf, A. J. Smola, F. Bach, *et al.*, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press, 2002.

[119] A. Ortiz, J. M. Górriz, J. Ramírez, F. J. Martínez-Murcia, A. D. N. Initiative, *et al.*, "LVQ-SVM based CAD tool applied to structural MRI for the diagnosis of the Alzheimer's disease," *Pattern Recognition Letters*, vol. 34, no. 14, pp. 1725–1733, 2013.

[120] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, may 2015.

[121] J. Seetha and S. S. Raja, "Brain tumor classification using convolutional neural networks," *Biomedical & Pharmacology Journal*, vol. 11, no. 3, p. 1457, 2018.

[122] M. Leming, J. M. Górriz, and J. Suckling, "Ensemble deep learning on large, mixed-site fMRI datasets in autism and other tasks," *International Journal of Neural Systems*, vol. 30, p. 2050012, apr 2020.

[123] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[124] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[125] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenet v2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

[126] R. Kohavi, "A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, (San Francisco, CA, USA), pp. 1137–1143, Morgan Kaufmann Publishers Inc., 1995. event-place: Montreal, Quebec, Canada.

[127] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer, 2009.

[128] I. Kim, A. Ramdas, A. Singh, and L. Wasserman, "Classification accuracy as a proxy for two-sample testing," *The Annals of Statistics*, vol. 49, no. 1, pp. 411–434, 2021.

[129] U. Braga-Neto, R. Hashimoto, E. R. Dougherty, D. V. Nguyen, and R. J. Carroll, "Is cross-validation better than resubstitution for ranking genes?," *Bioinformatics*, vol. 20, pp. 253–258, jan 2004.

[130] V. Vapnik, "Estimation of dependencies based on empirical data springer," *Information and Control*, 1982.

[131] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE transactions on electronic computers*, no. 3, pp. 326–334, 1965.

[132] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 1999.

[133] V. Vapnik, E. Levin, and Y. L. Cun, "Measuring the vc-dimension of a learning machine," *Neural computation*, vol. 6, no. 5, pp. 851–876, 1994.

[134] P. L. Bartlett, N. Harvey, C. Liaw, and A. Mehrabian, "Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks," *The Journal of Machine Learning Research*, vol. 20, no. 1, pp. 2285–2301, 2019.

[135] D. McAllester, "A pac-bayesian tutorial with a dropout bound," *arXiv preprint arXiv:1307.2118*, 2013.

[136] L. Devroye, L. Györfi, and G. Lugosi, *A probabilistic theory of pattern recognition*, vol. 31. Springer Science & Business Media, 2013.

[137] J. N. Mandrekar, "Receiver operating characteristic curve in diagnostic test assessment," *Journal of Thoracic Oncology*, vol. 5, no. 9, pp. 1315–1316, 2010.

[138] K. Hajian-Tilaki, "Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation," *Caspian journal of internal medicine*, vol. 4, no. 2, p. 627, 2013.

[139] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pp. 1135–1144, 2016.

[140] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), pp. 4765–4774, Curran Associates, Inc., 2017.

[141] L. S. Shapley *et al.*, "A value for n-person games," *Annals of Mathematical Studies*, vol. 28, p. 307–317, 1953.

[142] A. Alqaraawi, M. Schuessler, P. Weiß, E. Costanza, and N. Berthouze, "Evaluating saliency map explanations for convolutional neural networks: a user study," in *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pp. 275–285, 2020.

[143] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," *Advances in neural information processing systems*, vol. 31, 2018.

[144] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[145] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

[146] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE, oct 2017.

[147] J. Ashburner and K. J. Friston, "Unified segmentation," *NeuroImage*, vol. 26, pp. 839–851, jul 2005.

[148] I. Castiglioni, C. Salvatore, J. Ramírez, and J. M. Górriz, "Machine-learning neuroimaging challenge for automated diagnosis of mild cognitive impairment: Lessons learnt," *Journal of Neuroscience Methods*, vol. 302, pp. 10–13, May 2018.

[149] B. Fischl and A. M. Dale, "Measuring the thickness of the human cerebral cortex from magnetic resonance images," *Proceedings of the National Academy of Sciences*, vol. 97, pp. 11050–11055, Sept. 2000.

[150] B. Fischl, "FreeSurfer," *NeuroImage*, vol. 62, pp. 774–781, Aug. 2012.

[151] J. C. Morris, P. S. Aisen, R. J. Bateman, T. L. Benzinger, N. J. Cairns, A. M. Fagan, B. Ghetti, A. M. Goate, D. M. Holtzman, W. E. Klunk, E. McDade, D. S. Marcus, R. N. Martins, C. L. Masters, R. Mayeux, A. Oliver, K. Quaid, J. M. Ringman, M. N. Rossor, S. Salloway, P. R. Schofield, N. J. Selsor, R. A. Sperling, M. W. Weiner, C. Xiong, K. L. Moulder, and V. D. Buckles, "Developing an international network for alzheimer's research: the dominantly inherited alzheimer network," *Clinical Investigation*, vol. 2, pp. 975–984, oct 2012.

[152] A. M. Fagan, C. Xiong, M. S. Jasielec, R. J. Bateman, A. M. Goate, T. L. S. Benzinger, B. Ghetti, R. N. Martins, C. L. Masters, R. Mayeux, J. M. Ringman, M. N. Rossor, S. Salloway, P. R. Schofield, R. A. Sperling, D. Marcus, N. J. Cairns, V. D. Buckles,

J. H. Ladenson, J. C. Morris, and D. M. Holtzman, "Longitudinal change in CSF biomarkers in autosomal-dominant alzheimer's disease," *Science Translational Medicine*, vol. 6, pp. 226ra30–226ra30, mar 2014.

[153] M. S. Albert, S. T. DeKosky, D. Dickson, B. Dubois, H. H. Feldman, N. C. Fox, A. Gamst, D. M. Holtzman, W. J. Jagust, R. C. Petersen, *et al.*, "The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the national institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimer's & dementia*, vol. 7, no. 3, pp. 270–279, 2011.

[154] S. B. Guze, "Diagnostic and statistical manual of mental disorders, 4th ed. (DSM-IV)," *American Journal of Psychiatry*, vol. 152, pp. 1228–1228, aug 1995.

[155] P. R. Solomon, A. Hirschoff, B. Kelly, M. Relin, M. Brush, R. D. DeVeaux, and W. W. Pendlebury, "A 7 minute neurocognitive screening battery highly sensitive to Alzheimer's disease," *Archives of neurology*, vol. 55, no. 3, pp. 349–355, 1998.

[156] G. Grabner *et al.*, "Symmetric atlasing and model based segmentation: An application to the hippocampus in older adults," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2006*, pp. 58–66, Springer Berlin Heidelberg, 2006.

[157] D. Salas-Gonzalez *et al.*, "Building a FP-CIT SPECT brain template using a posterization approach," *Neuroinformatics*, vol. 13, pp. 391–402, mar 2015.

[158] J. M. Peñalver, D. López-García, C. González-García, B. Aguado-López, J. M. Górriz, and M. Ruz, "Top-down specific preparatory activations for selective attention and perceptual expectations," *NeuroImage*, vol. 271, p. 119960, may 2023.

[159] D. López-García, J. M. Peñalver, J. M. Górriz, and M. Ruz, "MVPAlab: A machine learning decoding toolbox for multidimensional electroencephalography data," *Computer Methods and Programs in Biomedicine*, vol. 214, p. 106549, feb 2022.

[160] D. S. Ma, J. Correll, and B. Wittenbrink, "The chicago face database: A free stimulus set of faces and norming data," *Behavior research methods*, vol. 47, no. 4, pp. 1122–1135, 2015.

[161] Z. Li, C. Yan, Q.-y. Lv, Z.-h. Yi, J.-y. Zhang, J.-h. Wang, S. S. Y. Lui, Y.-f. Xu, E. F. C. Cheung, R. E. Gur, R. C. Gur, and R. C. K. Chan, "Striatal dysfunction in patients with schizophrenia and their unaffected first-degree relatives," *Schizophrenia Research*, vol. 195, pp. 215–221, may 2018.

[162] A. Cachia, M.-L. Paillère-Martinot, A. Galinowski, D. Januel, R. de Beaurepaire, F. Bellivier, E. Artiges, J. Andoh, D. Bartrés-Faz, E. Duchesnay, D. Rivière, M. Plaze, J.-F. Mangin, and J.-L. Martinot, "Cortical folding abnormalities

in schizophrenia patients with resistant auditory hallucinations," *NeuroImage*, vol. 39, pp. 927–935, feb 2008.

[163] C. Mellerio, M.-N. Lapointe, P. Roca, S. Charron, L. Legrand, J.-F. Meder, C. Oppenheim, and A. Cachia, "Identification of reliable sulcal patterns of the human rolandic region," *Frontiers in Human Neuroscience*, vol. 10, aug 2016.

[164] D. Geffroy, D. Rivière, I. Denghien, N. Souedet, S. Laguitton, and Y. Cointepas, "Brainvisa: a complete software platform for neuroimaging," in *Python in Neuroscience workshop, Paris*, 2011.

[165] M. Perrot, D. Rivière, and J.-F. Mangin, "Cortical sulci recognition and spatial normalization," *Medical Image Analysis*, vol. 15, pp. 529–550, aug 2011.

[166] L. Borne, D. Rivière, M. Mancip, and J.-F. Mangin, "Automatic labeling of cortical sulci using patch- or CNN-based segmentation techniques combined with bottom-up geometric constraints," *Medical Image Analysis*, vol. 62, p. 101651, may 2020.

[167] J. Talairach, "Co-planar stereotaxic atlas of the human brain," *3-D proportional system: An approach to cerebral imaging*, 1988.

[168] D. L. Collins, P. Neelin, T. M. Peters, and A. C. Evans, "Automatic 3d intersubject registration of mr volumetric data in standardized talairach space.," *Journal of computer assisted tomography*, vol. 18, no. 2, pp. 192–205, 1994.

[169] F. Pizzagalli, G. Auzias, P. Kochunov, J. I. Faskowitz, P. M. Thompson, and N. Jahanshad, "The core genetic network underlying sulcal morphometry," in *SPIE Proceedings* (E. Romero, N. Lepore, J. Brieva, and I. Larrabide, eds.), SPIE, jan 2017.

[170] K. Jin, T. Zhang, M. Shaw, P. Sachdev, and N. Cherbuin, "Relationship between sulcal characteristics and brain aging," *Frontiers in Aging Neuroscience*, vol. 10, nov 2018.

[171] F. J. Martinez-Murcia, J. M. Górriz, J. Ramírez, A. Ortiz, and for the Alzheimer's Disease Neuroimaging Initiative, "A spherical brain mapping of mr images for the detection of alzheimer's disease," *Current Alzheimer Research*, vol. 13, no. 5, pp. 575–588, 2016.

[172] I. A. Illan, J. M. Górriz, J. Ramírez, and A. Meyer-Base, "Spatial component analysis of MRI data for Alzheimer's disease diagnosis: a Bayesian network approach," *Frontiers in Computational Neuroscience*, vol. 8, p. 156, 2014.

[173] F. Segovia, J. Górriz, J. Ramírez, D. Salas-Gonzalez, I. Álvarez, M. López, and R. Chaves, "A comparative study of feature extraction methods for the diagnosis of Alzheimer's disease using the ADNI database," *Neurocomputing*, vol. 75, no. 1, pp. 64–71, 2012.

[174] E. E. Bron, M. Smits, W. M. van der Flier, H. Vrenken, F. Barkhof, P. Scheltens, J. M. Papma, R. M. E. Steketee, C. Méndez Orellana, R. Meijboom, M. Pinto, J. R. Meireles, C. Garrett, A. J. Bastos-Leite, A. Abdulkadir, O. Ronneberger, N. Amoroso, R. Bellotti, D. Cárdenas-Peña, A. M. Álvarez Meza, C. V. Dolph, K. M. Iftekharuddin, S. F. Eskildsen, P. Coupé, V. S. Fonov, K. Franke, C. Gaser, C. Ledig, R. Guerrero, T. Tong, K. R. Gray, E. Moradi, J. Tohka, A. Routier, S. Durrleman, A. Sarica, G. Di Fatta, F. Sensi, A. Chincarini, G. M. Smith, Z. V. Stoyanov, L. Sørensen, M. Nielsen, S. Tangaro, P. Inglese, C. Wachinger, M. Reuter, J. C. van Swieten, W. J. Niessen, and S. Klein, "Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge," *NeuroImage*, vol. 111, pp. 562–579, May 2015.

[175] A. Sarica, A. Cerasa, A. Quattrone, and V. Calhoun, "Editorial on special issue: Machine learning on MCI," *Journal of Neuroscience Methods*, vol. 302, pp. 1–2, May 2018.

[176] S. I. Dimitriadis, D. Liparas, and M. N. Tsolaki, "Random forest feature selection, fusion and ensemble strategy: Combining multiple morphological MRI measures to discriminate among healhy elderly, MCI, cMCI and alzheimer's disease patients: From the alzheimer's disease neuroimaging initiative (ADNI) database," *Journal of Neuroscience Methods*, vol. 302, pp. 14–23, May 2018.

[177] R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehéricy, M.-O. Habert, M. Chupin, H. Benali, and O. Colliot, "Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database," *NeuroImage*, vol. 56, pp. 766–781, May 2011.

[178] K. R. Gray, P. Aljabar, R. A. Heckemann, A. Hammers, and D. Rueckert, "Random forest-based similarity measures for multi-modal classification of Alzheimer's disease," *NeuroImage*, vol. 65, pp. 167–175, Jan. 2013.

[179] A. Ortiz, J. Munilla, J. M. Górriz, and J. Ramírez, "Ensembles of deep learning architectures for the early diagnosis of the Alzheimer's disease," *International Journal of Neural Systems*, vol. 26, p. 1650025, aug 2016.

[180] J. Ramirez, J. M. Gorriz, R. Chaves, M. Lopez, D. Salas-Gonzalez, I. Alvarez, and F. Segovia, "Spect image classification using random forests," *Electronics Letters*, vol. 45, no. 12, pp. 604–605, 2009.

[181] Chih-Wei Hsu and Chih-Jen Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, pp. 415–425, Mar. 2002.

[182] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability Estimates for Multi-class Classification by Pairwise Coupling," *Journal of Machine Learning Research*, vol. 5, no. Aug, pp. 975–1005, 2004.

[183] A. Rojas, J. Górriz, J. Ramírez, I. Illán, F. Martínez-Murcia, A. Ortiz, M. G. Río], and M. Moreno-Caballero, "Application of empirical mode decomposition (emd) on datscan spect images to explore parkinson disease," *Expert Systems with Applications*, vol. 40, no. 7, pp. 2756 – 2766, 2013.

[184] J. Ramírez, J. Górriz, F. Segovia, R. Chaves, D. Salas-Gonzalez, M. López, I. Álvarez, and P. Padilla, "Computer aided diagnosis system for the Alzheimer's disease based on partial least squares and random forest SPECT image classification," *Neuroscience Letters*, vol. 472, pp. 99–103, Mar. 2010.

[185] F. Segovia, J. M. Górriz, J. Ramírez, D. Salas-González, and I. Álvarez, "Early diagnosis of Alzheimer's disease based on Partial Least Squares and Support Vector Machine," *Expert Systems with Applications*, vol. 40, pp. 677–683, Feb. 2013.

[186] V. Fritsch, G. Varoquaux, B. Thyreau, J.-B. Poline, and B. Thirion, "Detecting outliers in high-dimensional neuroimaging datasets with robust covariance estimators," *Medical Image Analysis*, vol. 16, pp. 1359–1370, Oct. 2012.

[187] J. M. Gorriz, J. Ramirez, J. Suckling, I. A. Illan, A. Ortiz, F. J. Martinez-Murcia, F. Segovia, D. Salas-Gonzalez, and S. Wang, "Case-based statistical learning: A non-parametric implementation with a conditional-error rate SVM," *IEEE Access*, vol. 5, pp. 11468–11478, 2017.

[188] T. G. Dietterich and G. Bakiri, "Solving Multiclass Learning Problems via Error-Correcting Output Codes," *Journal of Artificial Intelligence Research*, vol. 2, pp. 263–286, 1994.

[189] S. Escalera, O. Pujol, and P. Radeva, "On the Decoding Process in Ternary Error-Correcting Output Codes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 120–134, Jan. 2010.

[190] B. Dubois, H. H. Feldman, C. Jacova, S. T. DeKosky, P. Barberger-Gateau, J. Cummings, A. Delacourte, D. Galasko, S. Gauthier, G. Jicha, K. Meguro, J. O'Brien, F. Pasquier, P. Robert, M. Rossor, S. Salloway, Y. Stern, P. J. Visser, and P. Scheltens, "Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS–ADRDA criteria," *The Lancet Neurology*, vol. 6, pp. 734–746, Aug. 2007.

[191] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, pp. 123–140, Aug. 1996.

[192] Y. Freund and R. E. Schapire, *Experiments with a New Boosting Algorithm*. ICML'96: Proceedings of the Thirteenth International Conference on International Conference on Machine Learning, 1996.

[193] J. Ramírez, J. M. Górriz, A. Ortiz, F. J. Martínez-Murcia, F. Segovia, D. Salas-Gonzalez, D. Castillo-Barnes, I. A. Illán, and C. G. Puntonet, "Ensemble of

random forests One vs. Rest classifiers for MCI and AD prediction using ANOVA cortical and subcortical feature selection and partial least squares," *Journal of Neuroscience Methods*, vol. 302, pp. 47–57, May 2018.

[194] H. Li, M. Habes, D. A. Wolk, and Y. Fan, "A deep learning model for early prediction of Alzheimer's disease dementia based on hippocampal magnetic resonance imaging data," *Alzheimer's & Dementia*, vol. 15, pp. 1059–1070, Aug. 2019.

[195] K. Kawaguchi, L. P. Kaelbling, and Y. Bengio, "Generalization in deep learning," *arXiv preprint arXiv:1710.05468*, 2017.

[196] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[197] P. Thanapol, K. Lavangnananda, P. Bouvry, F. Pinel, and F. Leprevost, "Reducing overfitting and improving generalization in training convolutional neural network (CNN) under limited sample sizes in image recognition," in *2020 - 5th International Conference on Information Technology (InCIT)*, IEEE, oct 2020.

[198] Z. Tu, F. He, and D. Tao, "Understanding generalization in recurrent neural networks," in *International Conference on Learning Representations*, 2019.

[199] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro, "Exploring generalization in deep learning," *Advances in neural information processing systems*, vol. 30, pp. 5947–5956, 2017.

[200] J. D. Rosenblatt, Y. Benjamini, R. Gilron, R. Mukamel, and J. J. Goeman, "Better-than-chance classification for signal detection," *Biostatistics*, oct 2019.

[201] F. Pereira, T. Mitchell, and M. Botvinick, "Machine learning classifiers and fMRI: A tutorial overview," *NeuroImage*, vol. 45, pp. S199–S209, mar 2009.

[202] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[203] E. Olivetti, S. Greiner, and P. Avesani, "Induction in neuroscience with classification: Issues and solutions," in *Lecture Notes in Computer Science*, pp. 42–50, Springer Berlin Heidelberg, 2012.

[204] A. Isaksson, M. Wallman, H. Göransson, and M. Gustafsson, "Cross-validation and bootstrapping are unreliable in small sample classification," *Pattern Recognition Letters*, vol. 29, pp. 1960–1965, oct 2008.

[205] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Springer US, 1993.

[206] S. Basu, K. Wagstyl, A. Zandifar, L. Collins, A. Romero, and D. Precup, "Early prediction of alzheimer's disease progression using variational autoencoders," in *Lecture Notes in Computer Science*, pp. 205–213, Springer International Publishing, 2019.

[207] T. Jo, K. Nho, and A. J. Saykin, "Deep learning in alzheimer's disease: Diagnostic classification and prognostic prediction using neuroimaging data," *Frontiers in Aging Neuroscience*, vol. 11, p. 220, aug 2019.

[208] W. Feng, N. V. Halm-Lutterodt, H. Tang, A. Mecum, M. K. Mesregah, Y. Ma, H. Li, F. Zhang, Z. Wu, E. Yao, and X. Guo, "Automated MRI-based deep learning model for detection of alzheimer's disease process," *International Journal of Neural Systems*, vol. 30, p. 2050032, may 2020.

[209] M. López, J. Ramírez, J. Górriz, I. Álvarez, D. Salas-Gonzalez, F. Segovia, and R. Chaves, "SVM-based CAD system for early detection of the alzheimer's disease using kernel PCA and LDA," *Neuroscience Letters*, vol. 464, pp. 233–238, oct 2009.

[210] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 6232–6251, 2016.

[211] B. Neyshabur, R. Tomioka, and N. Srebro, "Norm-based capacity control in neural networks," in *Conference on Learning Theory*, pp. 1376–1401, 2015.

[212] G. K. Dziugaite and D. M. Roy, "Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data," in *In Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2017.

[213] R. Rosipal and N. Krämer, "Overview and Recent Advances in Partial Least Squares," in *Subspace, Latent Structure and Feature Selection* (C. Saunders, M. Grobelnik, S. Gunn, and J. Shawe-Taylor, eds.), Lecture Notes in Computer Science, pp. 34–51, Springer Berlin Heidelberg, 2006.

[214] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, pp. 448–456, PMLR, 2015.

[215] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot, "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain," *NeuroImage*, vol. 15, pp. 273–289, jan 2002.

[216] J. E. Arco, A. Ortiz, J. Ramirez, F. J. Martinez-Murcia, Y.-D. Zhang, J. Broncano, M. A. Berbis, J. R. del Val, A. Luna, and J. M. Gorriz, "Probabilistic combination

of non-linear eigenprojections for ensemble classification," *IEEE Transactions on Emerging Topics in Computational Intelligence*, pp. 1–11, 2022.

[217] M. N. Hebart and C. I. Baker, "Deconstructing multivariate decoding for the study of brain function," *NeuroImage*, vol. 180, pp. 4–18, oct 2018.

[218] Y. Jeong *et al.*, "18f-fdg pet findings in frontotemporal dementia: an spm analysis of 29 patients," *Journal of nuclear medicine : official publication, Society of Nuclear Medicine*, vol. 46, pp. 233–239, Feb. 2005.

[219] J. D. Rosenblatt, M. Vink, and Y. Benjamini, "Revisiting multi-subject random effects in fmri: Advocating prevalence estimation," *NeuroImage*, vol. 84, pp. 113–121, 2014.

[220] K. Friston, "Ten ironic rules for non-statistical reviewers," *Neuroimage*, vol. 61, no. 4, pp. 1300–1310, 2012.

[221] J. Gorriz *et al.*, "A connection between pattern classification by machine learning and statistical inference with the general linear model," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–1, 2021.

[222] C. Allefeld, K. Görgen, and J.-D. Haynes, "Valid population inference for information-based imaging: From the second-level t-test to prevalence inference," *Neuroimage*, vol. 141, pp. 378–392, 2016.

[223] G. Fung and J. Stoeckel, "SVM feature selection for classification of SPECT images of alzheimer's disease using spatial information," *Knowledge and Information Systems*, vol. 11, pp. 243–258, sep 2006.

[224] A. McIntosh, F. Bookstein, J. V. Haxby, and C. Grady, "Spatial pattern analysis of functional brain images using partial least squares," *Neuroimage*, vol. 3, no. 3, pp. 143–157, 1996.

[225] R. S. J. Frackowiak, J. T. Ashburner, W. D. Penny, and S. Zeki, *Human Brain Function, in: Introduction to Random Field Theory, Second Edition.* Academic Press, 2003.

[226] D. Castillo-Barnes, C. Jimenez-Mesa, F. J. Martinez-Murcia, D. Salas-Gonzalez, J. Ramírez, and J. M. Górriz, "Quantifying differences between affine and non-linear spatial normalization of FP-CIT spect images," *International Journal of Neural Systems*, vol. 32, mar 2022.

[227] S. J. Teipel, M. Grothe, S. Lista, N. Toschi, F. G. Garaci, and H. Hampel, "Relevance of magnetic resonance imaging for early detection and diagnosis of alzheimer disease," *Medical Clinics of North America*, vol. 97, pp. 399–424, may 2013.

[228] R. K. Lama, J. Gwak, J.-S. Park, and S.-W. Lee, "Diagnosis of Alzheimer's Disease Based on Structural MRI Images Using a Regularized Extreme Learning Machine and PCA Features," 2017.

[229] J. M. Górriz, J. Ramírez, F. Segovia, F. J. Martínez, M.-C. Lai, M. V. Lombardo, S. Baron-Cohen, M. A. Consortium, and J. Suckling, "A machine learning approach to reveal the neurophenotypes of autisms," *International journal of neural systems*, vol. 29, no. 07, p. 1850058, 2019.

[230] T. Bateman, "Advantages and disadvantages of PET and SPECT in a busy clinical practice," *Journal of Nuclear Cardiology*, vol. 19, pp. 3–11, jan 2012.

[231] O. Cigdem, I. Beheshti, and H. Demirel, "Effects of different covariates and contrasts on classification of parkinson's disease using structural MRI," *Computers in Biology and Medicine*, vol. 99, pp. 173–181, aug 2018.

[232] R. Martins *et al.*, "Automatic classification of idiopathic parkinson's disease and atypical parkinsonian syndromes combining [11c]raclopride PET uptake and MRI grey matter morphometry," *Journal of Neural Engineering*, vol. 18, p. 046037, apr 2021.

[233] B. Rana *et al.*, "Relevant 3d local binary pattern based features from fused feature descriptor for differential diagnosis of parkinson's disease using structural MRI," *Biomedical Signal Processing and Control*, vol. 34, pp. 134–143, apr 2017.

[234] P. Pan *et al.*, "Abnormalities of regional brain function in parkinson's disease: A meta-analysis of resting state functional magnetic resonance imaging studies," *Scientific Reports*, vol. 7, jan 2017.

[235] K. Ashton, B. D. Zinszer, R. M. Cichy, C. A. Nelson III, R. N. Aslin, and L. Bayet, "Time-resolved multivariate pattern analysis of infant eeg data: A practical tutorial," *Developmental cognitive neuroscience*, vol. 54, p. 101094, 2022.

[236] D. López-García, A. Sobrado, J. M. G. Peñalver, J. M. Górriz, and M. Ruz, "Multivariate pattern analysis techniques for electroencephalography data to study flanker interference effects," *International Journal of Neural Systems*, vol. 30, p. 2050024, jun 2020.

[237] T. Grootswagers, S. G. Wardle, and T. A. Carlson, "Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data," *Journal of cognitive neuroscience*, vol. 29, no. 4, pp. 677–697, 2017.

[238] J.-R. King and S. Dehaene, "Characterizing the dynamics of mental representations: the temporal generalization method," *Trends in cognitive sciences*, vol. 18, no. 4, pp. 203–210, 2014.

[239] M. N. Hebart, B. B. Bankson, A. Harel, C. I. Baker, and R. M. Cichy, "The representational dynamics of task and object processing in humans," *Elife*, vol. 7, p. e32816, 2018.

[240] M. van de Nieuwenhuijzen, A. Backus, A. Bahramisharif, C. Doeller, O. Jensen, and M. van Gerven, "MEG-based decoding of the spatiotemporal dynamics of visual category perception," *NeuroImage*, vol. 83, pp. 1063–1073, dec 2013.

[241] M. J. Mateos, A. Gastelum-Strozzi, F. A. Barrios, E. Bribiesca, S. Alcauter, and J. A. Marquez-Flores, "A novel voxel-based method to estimate cortical sulci width and its application to compare patients with alzheimer's disease to controls," *NeuroImage*, vol. 207, p. 116343, feb 2020.

[242] M. Plocharski and L. R. Østergaard, "Extraction of sulcal medial surface and classification of alzheimer's disease using sulcal features," *Computer Methods and Programs in Biomedicine*, vol. 133, pp. 35–44, sep 2016.

[243] J. Wang, H. You, J.-F. Liu, D.-F. Ni, Z.-X. Zhang, and J. Guan, "Association of olfactory bulb volume and olfactory sulcus depth with olfactory function in patients with parkinson disease," *American journal of neuroradiology*, vol. 32, no. 4, pp. 677–681, 2011.

[244] E. Collantoni, C. R. Madan, V. Meregalli, P. Meneguzzo, E. Marzola, M. Panero, F. D'Agata, G. Abbate-Daga, E. Tenconi, R. Manara, and A. Favaro, "Sulcal characteristics patterns and gyrification gradient at different stages of anorexia nervosa: A structural MRI evaluation," *Psychiatry Research: Neuroimaging*, vol. 316, p. 111350, oct 2021.

[245] A. Wagner, M. Ruf, D. F. Braus, and M. H. Schmidt, "Neuronal activity changes and body image distortion in anorexia nervosa," *Neuroreport*, vol. 14, no. 17, pp. 2193–2197, 2003.

[246] Y. Zhang, L. Lin, C.-P. Lin, Y. Zhou, K.-H. Chou, C.-Y. Lo, T.-P. Su, and T. Jiang, "Abnormal topological organization of structural brain networks in schizophrenia," *Schizophrenia Research*, vol. 141, pp. 109–118, nov 2012.

[247] N. Liu, Y. Xiao, W. Zhang, B. Tang, J. Zeng, N. Hu, S. Chandan, Q. Gong, and S. Lui, "Characteristics of gray matter alterations in never-treated and treated chronic schizophrenia patients," *Translational Psychiatry*, vol. 10, may 2020.

[248] L. Palaniyappan, O. Hodgson, V. Balain, S. Iwabuchi, P. Gowland, and P. Liddle, "Structural covariance and cortical reorganisation in schizophrenia: a MRI-based morphometric study," *Psychological Medicine*, vol. 49, pp. 412–420, may 2018.

[249] J. Janssen, C. Alloza, C. M. Díaz-Caneja, J. Santonja, L. Pina-Camacho, P. M. Gordaliza, A. Fernández-Pena, N. G. Lois, E. E. Buimer, N. E. van Haren, W. Cahn, E. Vieta, J. Castro-Fornieles, M. Bernardo, C. Arango, R. S. Kahn, H. E. H. Pol, and H. G. Schnack, "Longitudinal allometry of sulcal morphology in health and schizophrenia," *The Journal of Neuroscience*, vol. 42, pp. 3704–3715, mar 2022.

[250] J. G. Csernansky, S. K. Gillespie, D. L. Dierker, A. Anticevic, L. Wang, D. M. Barch, and D. C. V. Essen, "Symmetric abnormalities in sulcal patterning in schizophrenia," *NeuroImage*, vol. 43, pp. 440–446, nov 2008.

[251] J. P. A. Ioannidis, "Why most published research findings are false," *PLoS Medicine*, vol. 2, p. e124, aug 2005.

[252] B. H. Van der Velden, H. J. Kuijf, K. G. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (xai) in deep learning-based medical image analysis," *Medical Image Analysis*, p. 102470, 2022.

[253] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "XAI—explainable artificial intelligence," *Science Robotics*, vol. 4, dec 2019.

[254] N. C. Andreasen, G. Harris, T. Cizadlo, S. Arndt, D. S. O'Leary, V. Swayze, and M. Flaum, "Techniques for measuring sulcal/gyral patterns in the brain as visualized through magnetic resonance scanning: BRAINPLOT and BRAINMAP.," *Proceedings of the National Academy of Sciences*, vol. 91, pp. 93–97, jan 1994.

[255] K. J. Behnke, M. E. Rettmann, D. L. Pham, D. Shen, S. M. Resnick, C. Davatzikos, and J. L. Prince, "Automatic classification of sulcal regions of the human brain cortex using pattern recognition," in *SPIE Proceedings* (M. Sonka and J. M. Fitzpatrick, eds.), SPIE, may 2003.

[256] F. Yang and F. Kruggel, "Automatic segmentation of human brain sulci," *Medical Image Analysis*, vol. 12, pp. 442–451, aug 2008.

[257] S. Murphy, B. Mohr, Y. Fushimi, H. Yamagata, and I. Poole, "Fast, simple, accurate multi-atlas segmentation of the brain," in *Biomedical Image Registration: 6th International Workshop, WBIR 2014, London, UK, July 7-8, 2014. Proceedings 6*, pp. 1–10, Springer, 2014.

[258] G. Auzias, L. Brun, C. Deruelle, and O. Coulon, "Deep sulcal landmarks: Algorithmic and conceptual improvements in the definition and extraction of sulcal pits," *NeuroImage*, vol. 111, pp. 12–25, may 2015.

[259] C. R. Madan, "Robust estimation of sulcal morphology," *Brain Informatics*, vol. 6, jun 2019.

[260] S. Mikhael, C. Hoogendoorn, M. Valdes-Hernandez, and C. Pernet, "A critical analysis of neuroanatomical software protocols reveals clinically relevant differences in parcellation schemes," *NeuroImage*, vol. 170, pp. 348–364, apr 2018.

[261] J. Ramírez, J. Górriz, D. Salas-Gonzalez, A. Romero, M. López, I. Álvarez, and M. Gómez-Río, "Computer-aided diagnosis of alzheimer's type dementia combining support vector machines and discriminant set of features," *Information Sciences*, vol. 237, pp. 59–72, 2013.

[262] G. Orru, W. Pettersson-Yeo, A. F. Marquand, G. Sartori, and A. Mechelli, "Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review," *Neuroscience & Biobehavioral Reviews*, vol. 36, no. 4, pp. 1140–1152, 2012.

[263] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[264] X. Shen, T. Liu, D. Tao, Y. Fan, J. Zhang, S. Li, J. Jiang, W. Zhu, Y. Wang, Y. Wang, H. Brodaty, P. Sachdev, and W. Wen, "Variation in longitudinal trajectories of cortical sulci in normal elderly," *NeuroImage*, vol. 166, pp. 1–9, feb 2018.

[265] J. P. John, L. Wang, A. J. Moffitt, H. K. Singh, M. H. Gado, and J. G. Csernansky, "Inter-rater reliability of manual segmentation of the superior, inferior and middle frontal gyri," *Psychiatry Research: Neuroimaging*, vol. 148, pp. 151–163, dec 2006.

[266] J. Yang, D. Wang, C. Rollins, M. Leming, P. Liò, J. Suckling, G. Murray, J. Garrison, and A. Cachia, "Volumetric segmentation and characterisation of the paracingulate sulcus on mri scans," *bioRxiv*, p. 859496, 2019.

[267] M. Ribolsi, Z. J. Daskalakis, A. Siracusano, and G. Koch, "Abnormal asymmetry of brain connectivity in schizophrenia," *Frontiers in Human Neuroscience*, vol. 8, dec 2014.

[268] R. Nesvåg, M. Schaer, U. K. Haukvik, L. T. Westlye, L. M. Rimol, E. H. Lange, C. B. Hartberg, M.-C. Ottet, I. Melle, O. A. Andreassen, E. G. Jönsson, I. Agartz, and S. Eliez, "Reduced brain cortical folding in schizophrenia revealed in two independent samples," *Schizophrenia Research*, vol. 152, pp. 333–338, feb 2014.

[269] L. Palaniyappan, B. Park, V. Balain, R. Dangi, and P. Liddle, "Abnormalities in structural covariance of cortical gyrification in schizophrenia," *Brain Structure and Function*, vol. 220, pp. 2059–2071, apr 2014.

[270] J. R. Garrison, C. Fernyhough, S. McCarthy-Jones, M. Haggard, V. Carr, U. Schall, R. Scott, A. Jablensky, B. Mowry, P. Michie, S. Catts, F. Henskens, C. Pantelis, C. Loughland, and J. S. S. and, "Paracingulate sulcus morphology is associated with hallucinations in the human brain," *Nature Communications*, vol. 6, nov 2015.

[271] M. Yücel, G. W. Stuart, P. Maruff, S. J. Wood, G. R. Savage, D. J. Smith, S. F. Crowe, D. L. Copolov, D. Velakoulis, and C. Pantelis, "Paracingulate morphologic differences in males with established schizophrenia: a magnetic resonance imaging morphometric study," *Biological psychiatry*, vol. 52, no. 1, pp. 15–23, 2002.

[272] M. Freedman, L. Leach, E. Kaplan, G. Winocur, K. Shulman, and D. C. Delis, *Clock drawing: A neuropsychological analysis*. Oxford University Press, USA, 1994.

[273] K. I. Shulman, "Clock-drawing: is it the ideal cognitive screening test?," *International Journal of Geriatric Psychiatry*, vol. 15, no. 6, pp. 548–561, 2000.

[274] C. Carnero-Pardo, I. Rego-García, J. Barrios-López, S. Blanco-Madera, R. Calle-Calle, S. López-Alcalde, and R. Vílchez-Carrillo, "Assessment of the diagnostic accuracy and discriminative validity of the clock drawing and mini-cog tests in detecting cognitive impairment," *Neurología (English Edition)*, vol. 37, no. 1, pp. 13–20, 2022.

[275] S. Borson, J. M. Scanlan, P. Chen, and M. Ganguli, "The mini-cog as a screen for dementia: validation in a population-based sample," *Journal of the American Geriatrics Society*, vol. 51, no. 10, pp. 1451–1454, 2003.

[276] J. E. Arco, A. Ortiz, J. Ramírez, Y.-D. Zhang, and J. M. Górriz, "Tiled sparse coding in eigenspaces for image classification," *International Journal of Neural Systems*, vol. 32, no. 03, p. 2250007, 2022.

[277] P. Vuttipittayamongkol and E. Elyan, "Improved overlap-based undersampling for imbalanced dataset classification with application to epilepsy and Parkinson's disease," *International Journal of Neural Systems*, vol. 30, no. 08, p. 2050043, 2020.

[278] A. Bhattacharya, T. Baweja, and S. P. K. Karri, "Epileptic seizure prediction using deep transformer model," *International Journal of Neural Systems*, vol. 32, no. 02, p. 2150058, 2022.

[279] A. H. Ansari, P. J. Cherian, A. Caicedo, G. Naulaers, M. De Vos, and S. Van Huffel, "Neonatal seizure detection using deep convolutional neural networks," *International Journal of Neural Systems*, vol. 29, no. 04, p. 1850011, 2019.

[280] F. Cruciani, L. Brusini, M. Zucchelli, G. R. Pinheiro, F. Setti, I. B. Galazzo, R. Deriche, L. Rittner, M. Calabrese, and G. Menegaz, "Interpretable deep learning as a means for decrypting disease signature in multiple sclerosis," *Journal of Neural Engineering*, vol. 18, p. 0460a6, jul 2021.

[281] G. Mirzaei and H. Adeli, "Machine learning techniques for diagnosis of Alzheimer disease, mild cognitive disorder, and other types of dementia," *Biomedical Signal Processing and Control*, vol. 72, p. 103293, 2022.

[282] K. D. Tzimourta, V. Christou, A. T. Tzallas, N. Giannakeas, L. G. Astrakas, P. Angelidis, D. Tsalikakis, and M. G. Tsipouras, "Machine learning algorithms and statistical approaches for Alzheimer's disease analysis based on resting-state EEG recordings: A systematic review," *International journal of neural systems*, vol. 31, no. 05, p. 2130002, 2021.

[283] O. K. Cura, A. Akan, G. C. Yilmaz, and H. S. Ture, "Detection of Alzheimer's Dementia by using signal decomposition and machine learning methods.," *International Journal of Neural Systems*, pp. 2250042–2250042, 2022.

[284] G. Mirzaei, A. Adeli, and H. Adeli, "Imaging and machine learning techniques for diagnosis of Alzheimer's disease," *Reviews in the Neurosciences*, vol. 27, no. 8, pp. 857–870, 2016.

[285] L. Giancardo, A. Sánchez-Ferro, T. Arroyo Gallego, I. Butterworth, C. Mendoza, P. Montero-Escribano, M. Matarazzo, J. Obeso, M. Gray, and R. Estepar, "Computer keyboard interaction as an indicator of early Parkinson's disease," *Scientific Reports*, vol. 5, p. 34468, 10 2016.

[286] L. Chan, C. Simmons, S. Tillem, M. Conley, I. A. Brazil, and A. Baskin-Sommers, "Classifying conduct disorder using a biopsychosocial model and machine learning method," *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 2022.

[287] W. Souillard-Mandar, R. Davis, C. Rudin, R. Au, D. J. Libon, R. Swenson, C. C. Price, M. Lamar, and D. L. Penney, "Learning classification models of cognitive conditions from subtle behaviors in the digital clock drawing test," *Machine Learning*, vol. 102, pp. 393–441, oct 2015.

[288] Z. Harbi, Y. Hicks, and R. Setchi, "Clock drawing test interpretation system," *Procedia computer science*, vol. 112, pp. 1641–1650, 2017.

[289] A. Davoudi, C. Dion, S. Amini, P. J. Tighe, C. C. Price, D. J. Libon, and P. Rashidi, "Classifying non-dementia and Alzheimer's disease/vascular dementia patients using kinematic, time-based, and visuospatial parameters: The digital clock drawing test," *Journal of Alzheimer's Disease*, vol. 82, pp. 47–57, Jun 2021.

[290] S. Chen, D. Stromer, H. A. Alabdalrahim, S. Schwab, M. Weih, and A. Maier, "Automatic dementia screening and scoring by applying deep learning on clock-drawing tests," *Scientific Reports*, vol. 10, nov 2020.

[291] R. Binaco, N. Calzaretto, J. Epifano, S. McGuire, M. Umer, S. Emrani, V. Wasserman, D. J. Libon, and R. Polikar, "Machine learning analysis of Digital Clock Drawing test performance for differential classification of mild cognitive impairment subtypes versus Alzheimer's disease," *Journal of the International Neuropsychological Society*, vol. 26, pp. 690–700, mar 2020.

[292] S. Müller, O. Preische, P. Heymann, U. Elbing, and C. Laske, "Increased diagnostic accuracy of digital vs. conventional clock drawing test for discrimination of patients in the early course of Alzheimer's disease from Cognitively Healthy individuals," *Frontiers in Aging Neuroscience*, vol. 9, apr 2017.

[293] J. Y. C. Chan, B. K. K. Bat, A. Wong, T. K. Chan, Z. Huo, B. H. K. Yip, T. C. Y. Kowk, and K. K. F. Tsoi, "Evaluation of digital drawing tests and paper-and-pencil drawing tests for the screening of mild cognitive impairment and dementia: A systematic review and meta-analysis of diagnostic studies," *Neuropsychology Review*, pp. 1–11, oct 2021.

[294] Y. C. Youn, J.-M. Pyun, N. Ryu, M. J. Baek, J.-W. Jang, Y. H. Park, S.-W. Ahn, H.-W. Shin, K.-Y. Park, and S. Y. Kim, "Use of the clock drawing test and the Rey-Osterrieth complex figure test-copy with convolutional neural networks to predict cognitive impairment," *Alzheimer's Research & Therapy*, vol. 13, pp. 1–7, apr 2021.

[295] P. Soille, *Morphological image analysis: principles and applications*. Springer-Verlag, 2013.

[296] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[297] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 11, pp. 4793–4813, 2020.

[298] J. E. Arco, J. Ramírez, J. M. Górriz, and M. Ruz, "Data fusion based on searchlight analysis for the prediction of Alzheimer's disease," *Expert Systems with Applications*, vol. 185, p. 115549, 2021.

[299] I. J. Deary, J. Corley, A. J. Gow, S. E. Harris, L. M. Houlihan, R. E. Marioni, L. Penke, S. B. Rafnsson, and J. M. Starr, "Age-associated cognitive decline," *British medical bulletin*, vol. 92, no. 1, pp. 135–152, 2009.

[300] T. Salthouse, "Consequences of age-related cognitive declines," *Annual review of psychology*, vol. 63, p. 201, 2012.

[301] W. Souillard-Mandar, D. Penney, B. Schaible, A. Pascual-Leone, R. Au, and R. Davis, "DCT clock: Clinically-interpretable and automated artificial intelligence analysis of drawing behavior for capturing cognition," *Frontiers in Digital Health*, vol. 3, 2021.

[302] X. Feng, Q. Zou, Y. Zhang, Y. Tang, J. Ding, and X. Wang, "Clock drawing test evaluation via object detection for automatic cognitive impairment diagnosis," in *2020 IEEE 6th International Conference on Computer and Communications (ICCC)*, pp. 1229–1234, IEEE, 2020.

[303] G. Rippon, *The Gendered Brain: The new neuroscience that shatters the myth of the female brain*. Random House, 2019.

[304] K. A. Henderson, "Breaking with tradition—women & outdoor pursuits," *Journal of Physical Education, Recreation & Dance*, vol. 63, no. 2, pp. 49–51, 1992.

[305] L. J. Burton, "Underrepresentation of women in sport leadership: A review of research," *Sport management review*, vol. 18, no. 2, pp. 155–165, 2015.

[306] C. Sattler, P. Toro, P. Schönknecht, and J. Schröder, "Cognitive activity, education and socioeconomic status as preventive factors for mild cognitive impairment and Alzheimer's disease," *Psychiatry research*, vol. 196, no. 1, pp. 90–95, 2012.