

Supplementary Content

Machine learning-based prediction of future dementia using routine clinical MRI brain scans and healthcare data

Table of Contents

eMethod 1: Cases and controls identification using EMR datasets	2
eMethod 2: Identifying valid brain scans.....	3
eMethod 3: Preprocessing valid brain scans.....	4
eMethod 4: Feature reduction.....	5
eMethod 5: Hyper-parameter details.....	6
eMethod 6: Patient cohort identified using EMR datasets	7
eTable 1: Selection of ICD-9 and ICD-10 codes used to select different dementia types...	8
eFigure 1: Example of time window showing the timeline of events used for selecting valid scans.....	9
eFigure 2: Snapshot showing GUI of bespoke annotation tool developed in RShiny and used to manually inspect and annotate brain scans.....	10
eFigure 3: Flowchart showing the volume calculation to filter partial brain scans.	11
eFigure 4: Summary of sequence parameters for T1w MRI images. Where TR: repetition time, TE: echo time.....	12
eFigure 5: Number of voxels selected during feature reduction in 5-fold nested cross-validation.....	13
eFigure 6: Heatmap showing grid search performance across hyperparameters	14
References	15

eMethod 1: Cases and controls identification using EMR datasets

Dementia Cases

Cases were selected from GoDARTS and SHARE studies using the following four linked datasets:

1. **Dispensed prescribing:** This dataset includes the detailed longitudinal prescription records of all medications that are dispensed from the pharmacies to the individuals. The drugs specific to dementia i.e. Chapter 4.11 of the British National Formulary (BNF) were identified.
2. **Death:** The Scottish General Register Office (GRO) records information of death certification. The ICD-9 and ICD-10 codes related to dementia were identified in these certifications (See Appendix Table 1).
3. **General hospitalisation admissions (SMR01):** The general hospital inpatient and day case discharge information include the individual episode level data. The ICD-9 and ICD-10 codes related to dementia were identified in these records (See Appendix Table 1).
4. **Psychiatric hospital discharge (SMR04):** The psychiatric hospital inpatient and day case discharge information include the individual episode level data. The ICD-9 and ICD-10 related to dementia were identified from these records (See Appendix Table 1).

For each patient with dementia, the incident date of dementia was evaluated based on the earliest date of occurrence in any of these datasets. As EMR records are independently coded episodes over time, patients could receive codes from different dementia subtypes due to the intrinsic inaccuracy in the clinical coding. A probabilistic model was used¹ to assign a dementia subtype to such patients.

Non-Dementia Controls

These patients were selected from GoDARTS and SHARE studies and have:

1. No dementia specific drugs (Chapter 4.11 of the BNF) in the dispensed prescribing records, and
2. No ICD-9 and ICD-10 codes related to dementia in death, general hospital admissions, and psychiatric hospital discharge datasets.

The algorithm for identifying the dementia cases and non-dementia controls was implemented in R² using DBI³, dplyr⁴, string⁵ and odbc⁶ packages.

eMethod 2: Identifying valid brain scans

A valid scan for the purpose of this study was defined as a T1 weighted volume of the entire brain, without significant artefact or pathological structural distortion (e.g. large stroke or large tumour). Firstly, the DICOM metadata was searched for text strings 'T1' AND 'MR' AND 'Head' sequence type scans. A time window approach is then used to select the valid scans. Valid case scans were identified as scans acquired one to six years before the earliest date of diagnosis of dementia in their medical record. Similarly, valid control scans were identified as scans acquired 6 years before the last date of follow-up or date of death. The valid images are identified using the following time window (See Appendix Figure 1):

For valid dementia patient scans:

$$(edd - ds) \geq 1\text{yr AND } (edd - ds) \leq 6\text{ yrs}$$

For valid control patient scans:

$$(ldf - ds) \geq 6\text{ yrs}$$

Where, *ds* is date of scan, *edd* is the earliest date of diagnosis for dementia patient and *ldf* is the latest date of follow-up for non-dementia control patients.

Next, a bespoke 'annotation tool' was developed and used to manually inspect the anatomical view in all three planes (axial, sagittal, and coronal) of these remaining scans along with their DICOM metadata. This manual review was useful to exclude localizer survey scans (with <10 2D slices), partial brains (identified using bespoke volume calculation, See Appendix Note 3 and Appendix Figure 4), brains with large lesions/tumours and other incorrectly tagged body part scans.

In order to filter partial brain scans, each valid scan (having number of slices > 10) is read from the memory to capture the orthographic view (See Appendix Figure 3). Finally, the value of volume *V* is calculated (in cm³) as follows:

$$V(MRAquisitionType_{(0018,0023)}) = \begin{cases} (SpacingBetweenSlices_{(0018,0088)} * noofSOPInstances * noofRows * noofColumns * PixelSpacing_{(0028,0030)}^2) / 1000 & MRAquisitionType_{(0018,0023)} = 2D \\ (SliceThickness_{(0018,0050)} * noofSOPInstances * noofRows * noofColumns * PixelSpacing_{(0028,0030)}^2) / 1000 & MRAquisitionType_{(0018,0023)} = 3D \end{cases}$$

The orthographic view and the corresponding volume is displayed in the RShiny-based annotation tool.

The final list of valid scans was then saved. To reduce gender and age confounds potentially arising through the IDCP process, the number of males and females as well as their age were balanced between the control and the patient groups, such that both groups had equal proportions of aged-matched males and females. Subsequently, the 'annotation tool' was used to manually review the matched control scans for exclusion of partial brain views, large lesions/tumours and other anatomical abnormalities. In case of any partial brains/lesions present in control scan, an intermediate list was saved, and the age sex matching performed again to obtain another valid control scan from the pool. Finally, the updated matched cohort of 'normal' (without abnormalities) whole brain scans were converted from dicom to nifti format for SPM⁷ software usage. The IDCP, annotation tool and volume calculation were implemented in R Language² using DT⁸, plyr⁹, oro.dicom¹⁰, oro.nifti¹¹, neurobase¹², future.apply¹³ and shiny¹⁴ packages.

eMethod 3: Preprocessing valid brain scans

All T1-weighted magnetic resonance scans from IDCP were pre-processed using SPM12⁷, running under MATLABTM R2021a¹⁵. MRI scans were bias-field corrected, spatially normalised to the MNI template and segmented creating spatially normalized images of different brain ‘tissue’ types: grey matter (GM), white matter (WM), and cerebrospinal fluid (CSF).

Following tissue segmentation, a nonlinear within-subject registration of the segmented GM and WM tissues was performed using the SPM12 Geodesic Shooting Toolbox¹⁶. This involves iteratively refining a template created from the averaged of all subjects’ segmented GM and WM images, computing deformations between the template and segmented subjects’ images, applying inverse deformations, and averaging the results to update the template. This continues until the template accurately represents common features across all the input images.

The *scalar momentum*¹⁷ features were extracted that encodes the deformation fields associated with the segmented GM and WM tissues, and also the residual between these tissue types and the template, as defined by equation 1-3 below:

$$|D\phi|(\mu_1 - c_1(\phi)) \dots (1)$$

$$|D\phi|(\mu_2 - c_2(\phi)) \dots (2)$$

$$|D\phi|(c_1(\phi) + c_2(\phi) - \mu_1 - \mu_2) \dots (3)$$

Where:

- ϕ represents the forward transformation field
- $|D\phi|$ denotes the Jacobian determinants of ϕ
- μ_1 is the GM template
- c_1 represents GM
- μ_2 is the WM template
- c_2 represents WM

This captures anatomical variability, tissue density differences, morphological changes and local geometric features, providing a comprehensive representation of the brain's structure,¹⁸ which has found to be empirically useful for maximising ML performance¹⁸.

eMethod 4: Feature reduction

Feature reduction to avoid over fitting due to the 'curse of dimensionality', it is important to select a subset of informative voxels and to discard noisy (random variation) and redundant voxels¹⁹. Successful feature reduction improves the performance of the classification and reduces computational cost^{19,20}. The present study employed a voxel-wise two sample t-test embedded within a nested cross-validation SVM framework²¹. This framework is very flexible and not specific to particular illnesses, having been used for depression²², non-brain structure fMRI data²³, and non-neuroimaging cognitive data²⁴.

The t-test is computationally fast, easy to implement and scales well to high dimensional data¹⁹. To avoid double-dipping and overfitting the model, the univariate t-map was computed using only the training set data²⁵. The t-map represents the t-statistic value for each voxel, which indicates the statistical significance of the difference between the two groups. The t-map was then used to create a t-mask, which is a binary mask that selects only the voxels with a p-value below an automatically selected threshold. This p-value threshold indicates that these voxels are statistically significant for training purposes. this threshold being optimized at the same time as the classifier's hyperparameters using nested cross-validation. The t-masks were applied to the testing set to evaluate the classification performance at different p-value thresholds to improve model generalizability.

eMethod 5: Hyper-parameter details

In this study, a radial basis function (RBF)-SVM, as implemented in MATLAB, was employed²⁶. SVM is one of the most widely used supervised ML methods in Multi-Variate Pattern Analysis (MVPA) for neuroimaging studies, due to its effectiveness in dealing with high-dimensional datasets and its flexibility in decoding diverse sources of brain data²⁷. RBF-SVM offer better performance on small datasets in comparison to deep learning models such as Convolutional Neural Networks²⁸.

A nested cross-validation approach was applied to optimise two parameters: the p-value threshold of the t-test feature reduction, and the box-constraint “C”, which determines the extent to which wrong predictions are penalised. The predefined range of values for each hyperparameter was as follows: p-value [0.000001, 0.00001], kernel width ‘ λ ’ [0.001, 200], and C [0.001, 200], were tested using evenly spaced values within these ranges. These parameters were tested using SLURM on cloud-based TRE²⁹.

Our feature reduction t-test was applied to the training subset of the inner folds, generated multiple binary t-masks for each p-value threshold, to filter voxels for inclusion in the subsequent ML training and testing.

Feature reduction and RBF-SVM used a combination of hyperparameters, with model performance evaluated on the left-out inner fold. This procedure was repeated such that each inner fold served as the validation set once, and the combination of hyperparameters that yielding the highest average classification accuracy across the inner folds was selected. This optimal set of hyperparameters was then used for training and testing of the RBF-SVM on the outer fold. This entire procedure was then repeated for each outer fold, ensuring that each fold was used as a test set once. The overall accuracy was calculated by averaging the classification accuracies across all outer folds.

eMethod 6: Patient cohort identified using EMR datasets

Dementia Patients

Following count of patients were selected from GoDARTS and SHARE studies using the following four linked datasets:

1. **Dispensed prescribing:** Patients identified [n=1932].
2. **Death:** Patients identified: Alzheimer's Dementia [n=513], Vascular Dementia [n=572], Unspecified Dementia [n=435], and Other Dementia [n=42]. Total records [1562] and total common patients [1463].
3. **General hospitalisation admissions (SMR01):** Patients identified: Alzheimer's Dementia [n=791], Vascular Dementia [n=759], Unspecified Dementia [n=869], and Other Dementia [n=353]. Total records [n=2772] and total common patients [n=2219].
4. **Psychiatric hospital discharge (SMR04):** Patients identified: Alzheimer's Dementia [n=208], Vascular Dementia [n=131], Unspecified Dementia [n=101], and Other Dementia [n=30]. Total records [n=470] and total common patients [n=411].

Overall, 3477 dementia patients were identified which included Alzheimer's Dementia [n=1858], Vascular Dementia [n=800], Unspecified Dementia [n=617], and Other Dementia [n=202].

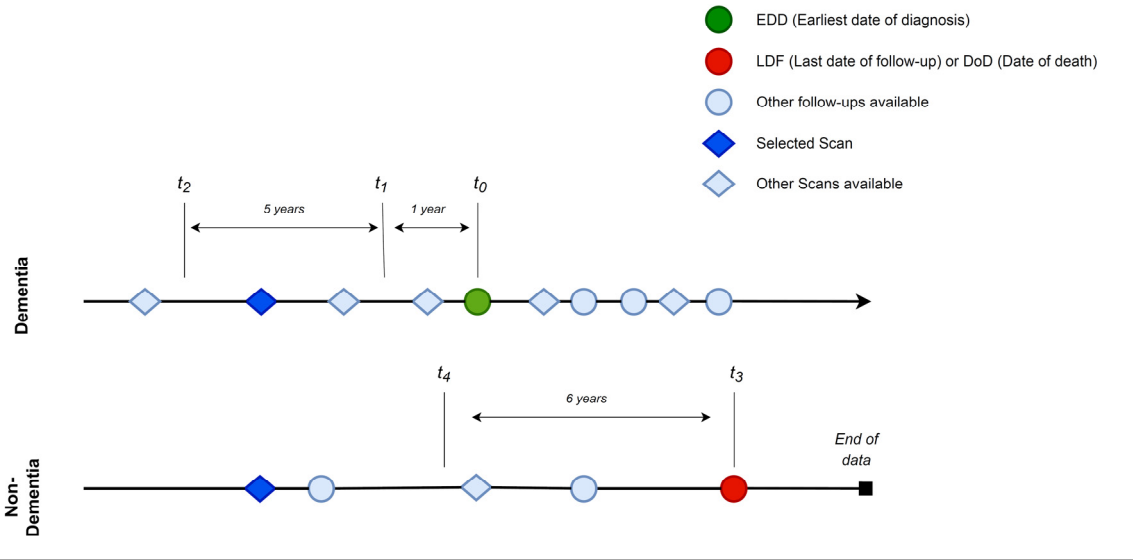
Non-Dementia Control Patients

In total, 99105 non-dementia patients were identified from the GoDARTS and SHARE studies which neither have any dementia related ICD-9 and ICD-10 codes nor prescription BNF codes.

eTable 1: Selection of ICD-9 and ICD-10 codes used to select different dementia types

Alzheimer's Disease (AD)	Vascular Dementia (VD)	Unspecified Dementia (UD)	Other Dementia (OD)
^3310	^2904	^2900	^2912
F00	F01	^2901	^2982
G30		^2902	^2941
		^2903	^3311
		^2942	^3312
		^3312	^33119
		F03	^33111
		G311	^33182
			^29282
			F02
			A810
			F051
			G310
			G318
			^0461
			^797
			F1027
			F1097
			G3110
			G3109
			G3183

eFigure 1: Example of time window showing the timeline of events used for selecting valid scans.



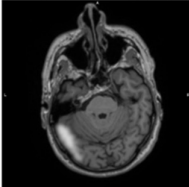
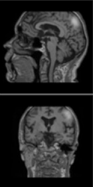
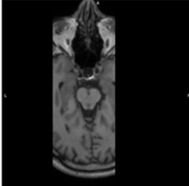
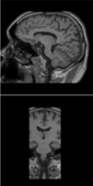
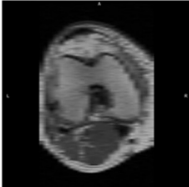
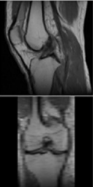
eFigure 2: Snapshot showing GUI of bespoke annotation tool developed in RShiny and used to manually inspect and annotate brain scans.

Imaging Data Curation Tool

Showing 1 to 3 of 10 entries Select all Deselect all Copy CSV

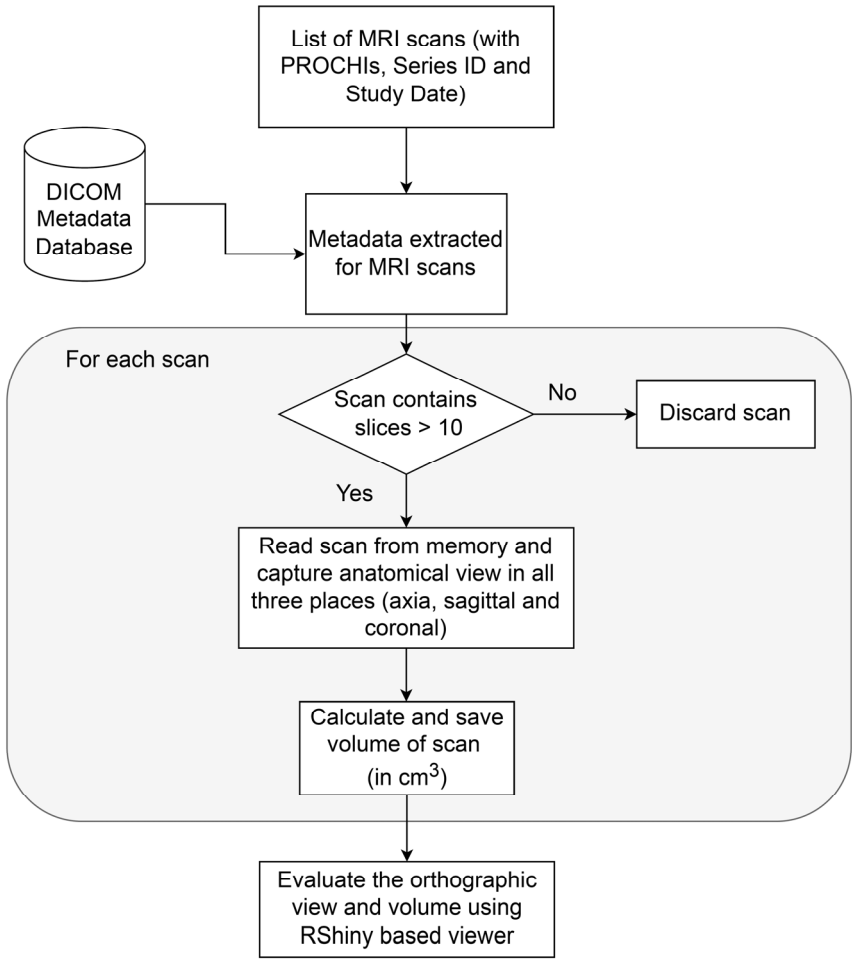
Search:

Previous 1 2 3 4 Next

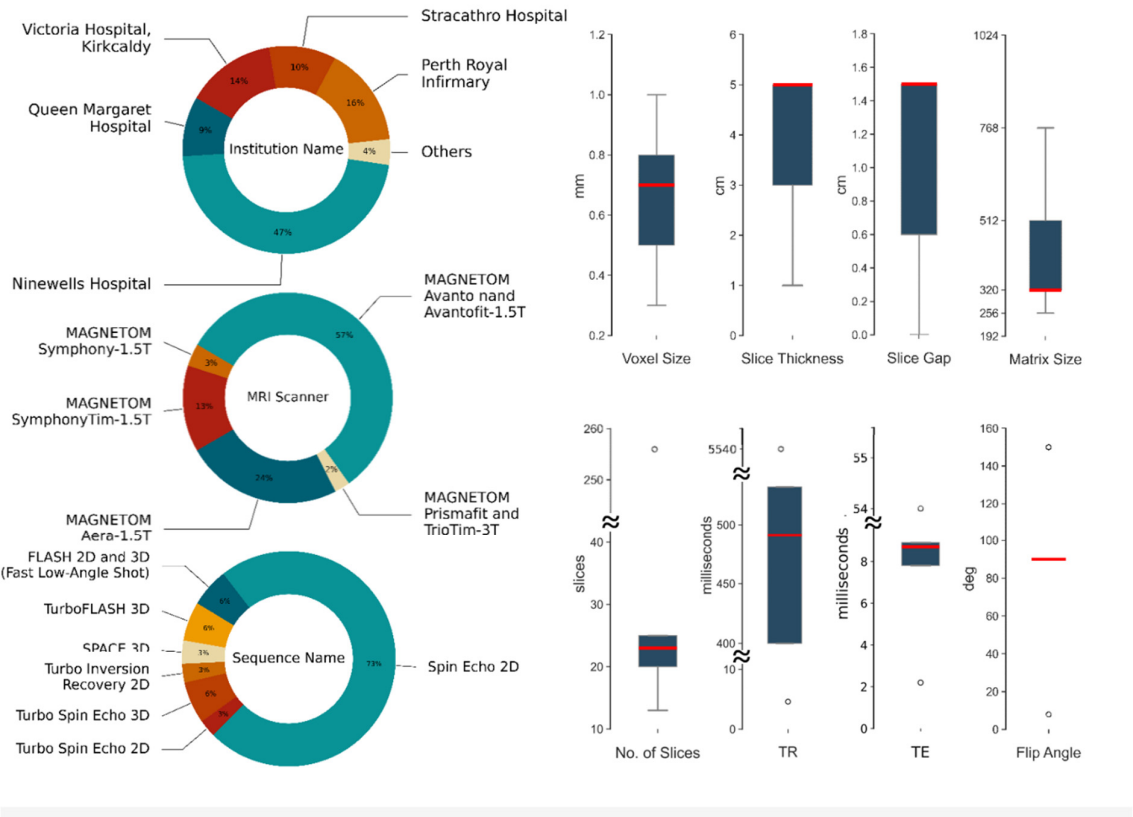
ortho_view	BodyPart	Completeness	MRModality	OtherComments	HiResPlane	Include	volcm3	dim	error	anonSeriesInstanceID	Final_Diagnosis
 	Brain	Whole	T1	Tumor ?? Check with Clinician	All	Yes	8326	256 X 256 X 150		Image5	AZ
 	Brain	Partial	T1	Check with Clinician	All	No	3967	256 X 256 X 150		Image7	OD
 	Knee	Whole	T1		3	No	3024	23 X 512 X 512		Image10	OD

Previous 1 2 3 4 Next

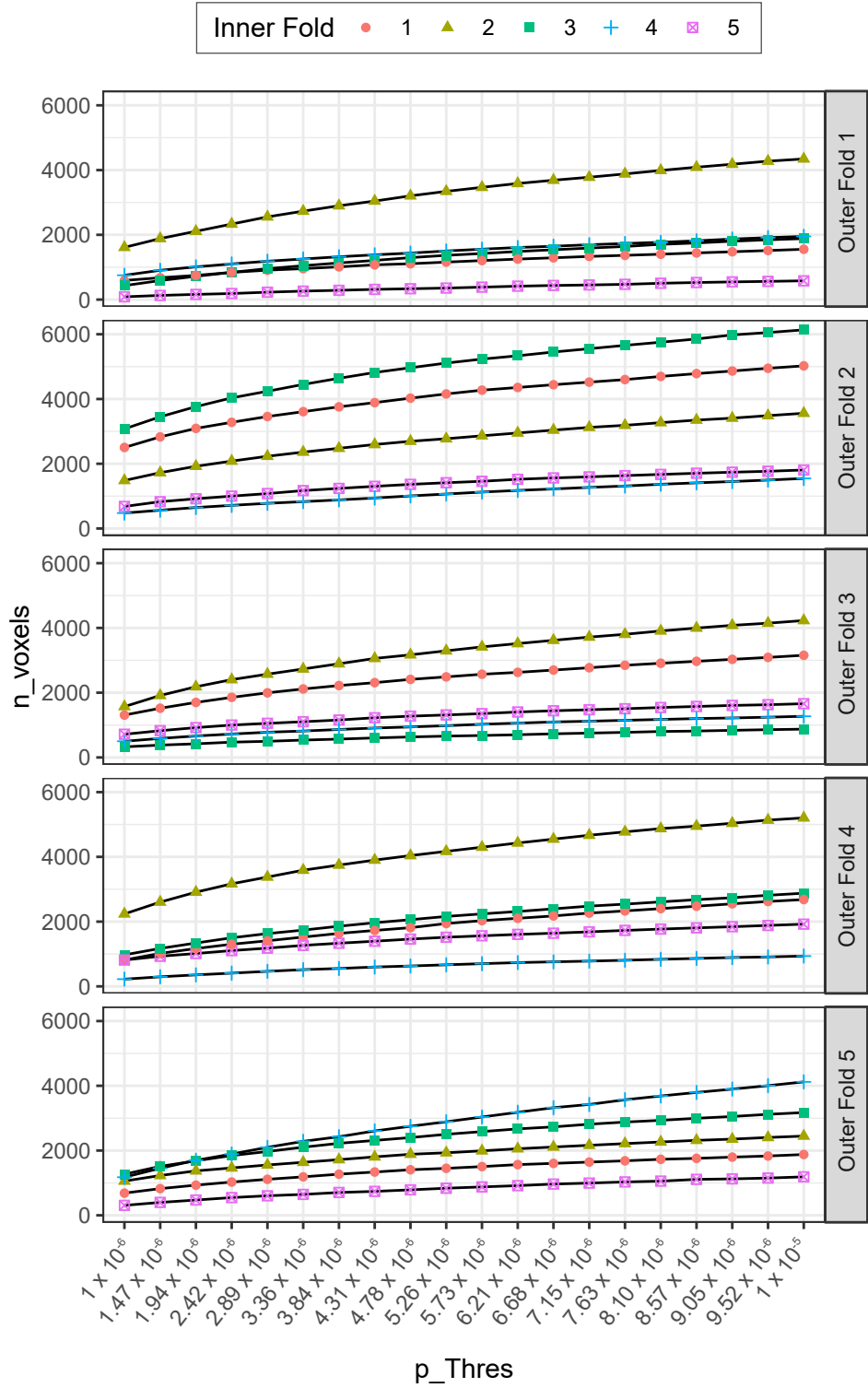
eFigure 3: Flowchart showing the volume calculation to filter partial brain scans.



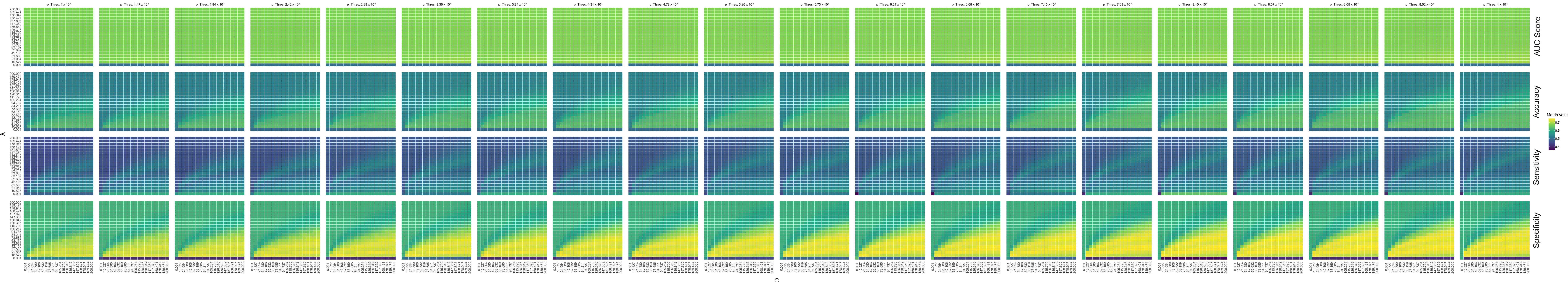
eFigure 4: Summary of sequence parameters for T1w MRI images. Where TR: repetition time, TE: echo time.



eFigure 5: Number of voxels selected during feature reduction in 5-fold nested cross-validation



eFigure 6: Heatmap showing grid search performance across hyperparameters



References

1. Doney ASF, Bonney W, Jefferson E, et al. Investigating the Relationship Between Type 2 Diabetes and Dementia Using Electronic Medical Records in the GoDARTS Bioresource. *Diabetes Care*. 2019;42(10):1973-1980. doi:10.2337/dc19-0380
2. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2013. <http://www.R-project.org/>
3. Databases (R-SIG-DB) RSIG on, Wickham H, Müller K, Consortium R. DBI: R Database Interface. Published online June 18, 2022. Accessed December 11, 2022. <https://CRAN.R-project.org/package=DBI>
4. Wickham H, François R, Henry L, Müller K, RStudio. dplyr: A Grammar of Data Manipulation. Published online September 1, 2022. Accessed December 11, 2022. <https://CRAN.R-project.org/package=dplyr>
5. Wickham H, RStudio. stringr: Simple, Consistent Wrappers for Common String Operations. Published online December 2, 2022. Accessed December 11, 2022. <https://CRAN.R-project.org/package=stringr>
6. Hester J, Wickham H, Gjoneski (detule) O, library) lexicalunit (nanodbc, library) GI (cctz, RStudio. odbc: Connect to ODBC Compatible Databases (using the DBI Interface). Published online November 30, 2021. Accessed December 11, 2022. <https://CRAN.R-project.org/package=odbc>
7. SPM - Statistical Parametric Mapping. Accessed October 12, 2024. <https://www.fil.ion.ucl.ac.uk/spm/>
8. Xie Y, Cheng J, Tan X, et al. DT: A Wrapper of the JavaScript Library "DataTables." Published online October 19, 2022. Accessed January 16, 2023. <https://CRAN.R-project.org/package=DT>
9. Wickham H. plyr: Tools for Splitting, Applying and Combining Data. Published online November 11, 2022. Accessed January 16, 2023. <https://CRAN.R-project.org/package=plyr>
10. Whitcher B. oro.dicom: Rigorous - DICOM Input / Output. Published online October 28, 2019. Accessed January 16, 2023. <https://CRAN.R-project.org/package=oro.dicom>
11. Whitcher B, Schmid V, Thornton A, Tabelow K, Clayden J, Muschelli J. oro.nifti: Rigorous - "NIFTI" + "ANALYZE" + 'AFNI' : Input / Output. Published online August 10, 2022. Accessed January 16, 2023. <https://CRAN.R-project.org/package=oro.nifti>
12. Muschelli J. neurobase: "Neuroconductor" Base Package with Helper Functions for "nifti" Objects. Published online October 23, 2022. Accessed January 16, 2023. <https://CRAN.R-project.org/package=neurobase>
13. Bengtsson H, R Core Team. future.apply: Apply Function to Elements in Parallel using Futures. Published online November 5, 2022. Accessed January 16, 2023. <https://CRAN.R-project.org/package=future.apply>
14. Chang W, Cheng J, Allaire JJ, et al. shiny: Web Application Framework for R. Published online December 15, 2022. Accessed January 16, 2023. <https://CRAN.R-project.org/package=shiny>
15. MATLAB. Accessed October 12, 2024. <https://www.mathworks.com/products/matlab.html>
16. Ashburner J, Friston KJ. Diffeomorphic registration using geodesic shooting and Gauss–Newton optimisation. *NeuroImage*. 2011;55(3):954-967. doi:10.1016/j.neuroimage.2010.12.049

17. Singh N, Fletcher PT, Preston JS, et al. Multivariate Statistical Analysis of Deformation Momenta Relating Anatomical Shape to Neuropsychological Measures. In: Jiang T, Navab N, Pluim JPW, Viergever MA, eds. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2010*. Springer; 2010:529-537. doi:10.1007/978-3-642-15711-0_66
18. Monté-Rubio GC, Falcón C, Pomarol-Clotet E, Ashburner J. A comparison of various MRI feature types for characterizing whole brain anatomical differences using linear pattern recognition methods. *NeuroImage*. 2018;178:753-768. doi:10.1016/j.neuroimage.2018.05.065
19. Mwangi B, Tian TS, Soares JC. A review of feature reduction techniques in neuroimaging. *Neuroinformatics*. 2014;12(2):229-244. doi:10.1007/s12021-013-9204-3
20. Demirhan A. The effect of feature selection on multivariate pattern analysis of structural brain MR images. *Physica Medica: European Journal of Medical Physics*. 2018;47:103-111. doi:10.1016/j.ejmp.2018.03.002
21. Johnston BA, Mwangi B, Matthews K, Coghill D, Steele JD. Predictive classification of individual magnetic resonance imaging scans from children and adolescents. *Eur Child Adolesc Psychiatry*. 2013;22(12):733-744. doi:10.1007/s00787-012-0319-0
22. Mwangi B, Ebmeier KP, Matthews K, Douglas Steele J. Multi-centre diagnostic classification of individual structural neuroimaging scans from patients with major depressive disorder. *Brain*. 2012;135(5):1508-1521. doi:10.1093/brain/awv084
23. Johnston BA, Tolomeo S, Gradin V, Christmas D, Matthews K, Douglas Steele J. Failure of hippocampal deactivation during loss events in treatment-resistant depression. *Brain*. 2015;138(9):2766-2776. doi:10.1093/brain/awv177
24. Johnston BA, Coghill D, Matthews K, Steele JD. Predicting methylphenidate response in attention deficit hyperactivity disorder: A preliminary study. *J Psychopharmacol*. 2015;29(1):24-30. doi:10.1177/0269881114548438
25. Ball TM, Squeglia LM, Tapert SF, Paulus MP. Double Dipping in Machine Learning: Problems and Solutions. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*. 2020;5(3):261-263. doi:10.1016/j.bpsc.2019.09.003
26. Statistics and Machine Learning Toolbox. Accessed October 12, 2024. <https://www.mathworks.com/products/statistics.html>
27. Sabuncu MR, Konukoglu E, for the Alzheimer's Disease Neuroimaging Initiative. Clinical Prediction from Structural Brain MRI Scans: A Large-Scale Empirical Study. *Neuroinform*. 2015;13(1):31-46. doi:10.1007/s12021-014-9238-1
28. Wang P, Fan E, Wang P. Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. *Pattern Recognition Letters*. 2021;141:61-67. doi:10.1016/j.patrec.2020.07.042
29. Reel PS, Aaron J, Krueger S, et al. Designing Machine Learning Experiments using SLURM within a Cloud Trusted Research Environment. Presented at: September 6, 2023. doi:10.5281/zenodo.13921331