Deep sMRI features for ordinal classification of Alzheimer's Disease staging

Miguel Á. Contreras-Córdoba[1 [0009−0006−2977−903X]] (✉), Juan E. Arco[2,3 [0000−0002−5896−4975]], Francisco Bérchez-Moreno[1 [0000−0003−1648−2780]], Víctor M. Vargas[4 [0000−0002−0700−275X]], Juan M. Górriz[2,3 [0000−0001−7069−1714]], José A. Gámez[5,6 [0000−0003−1188−1117]], and Pedro Antonio Gutiérrez[1 [0000−0002−2657−776X]]

[1] Departamento de Ciencia de la Computación e Inteligencia Artificial, Universidad de Córdoba, Córdoba, España
[2] SiPBA at Data Science and Computational Intelligence Institute, University of Granada, Granada, Spain
[3] Data Science and Computational Intelligence Institute (DaSCII), University of Granada, Granada, Spain
[4] Departamente de Teoría de la Señal y Comunicaciones, Universidad de Alcalá, Alcalá de Henares, Spain
[5] Departamento de Sistemas Informáticos, Universidad de Castilla-La Mancha, Albacete, Spain
[6] Instituto de Investigación en Informática de Albacete (I3A), Universidad de Castilla-La Mancha, Albacete, Spain
macontreras@uco.es; jearco@ugr.es; fberchez@uco.es; victor.vargas@uah.es; gorriz@ugr.es; jose.gamez@uclm.es; pagutierrez@uco.es

Abstract. The progression of Alzheimer's Disease (AD) follows a clear clinical sequence, moving from a cognitively NORmal state (NOR) through Mild Cognitive Impairment (MCI) to clinical AD. To leverage this natural order, this study evaluates nominal and ordinal classification methodologies using structural Magnetic Resonance Imaging (sMRI) data from the ADNI database. Instead of training the classifiers directly on the high-dimensional raw voxel space, we employ a fine-tuned 3D ResNet18 to extract low-dimensional semantic features. We then apply nominal models (Logistic Regression, LR, and Support Vector Machine, SVM) and ordinal methods (two ordinal variants of LR: LR with All-Thresholds, LogAT, and LR with Immediate-Thresholds, LogIT; and an ordinal version of the SVM: Support Vector Ordinal Regression with EXplicit constraints, SVOREX) to this new deep representation space. The experiments cover binary, 3-class, and 4-class scenarios under a stratified 5-fold cross-validation, using accuracy, Mean Absolute Error (MAE), and Average Mean Absolute Error (AMAE). Additionally, we compare this hybrid methodology against a pure end-to-end ResNet18 in the 4-class staging task. Our results indicate that the hybrid pipeline substantially outperforms the end-to-end Deep Learning (DL) approach for the most difficult experiment. Interestingly, while the LR performs better on raw voxels, ordinal models (LogAT and SVOREX) greatly benefit from the deep feature space, achieving the lowest overall ordinal errors. This confirms

that combining CNN-based feature extraction with ordinal classification is highly effective for reducing severe diagnostic mistakes in AD staging.

Keywords: Alzheimer's disease · ordinal classification · sMRI · SVM · logistic regression · SVOREX · ADNI · ResNet18 · Deep Learning

1 Introduction

Alzheimer's Disease (AD) is the leading cause of dementia, characterised by a silent preclinical phase where brain changes accumulate long before symptoms appear [1]. Structural Magnetic Resonance Imaging (sMRI) is a fundamental tool for non-invasive tracking of this macroscopic brain atrophy. Over the years, numerous Machine Learning (ML) and Deep Learning (DL) models have been proposed to automate AD diagnosis [8,17,3], driven by recent improvements in computational methods and DL [12,13]. However, most of these approaches treat the task as a standard nominal classification problem. Consequently, they apply the same penalty to any diagnostic error, failing to consider the clinical distance between the predicted and the actual severity stages.

To solve this, a methodology based on ordinal regression, also referred to as Ordinal Classification (OC), can be considered. OC, unlike nominal classification, explicitly models datasets with more than two classes that presents a natural order ($C_1 \prec C_2 \prec \cdots \prec C_Q$) within its categories [14]. For AD, this progression goes from cognitively NORmal (NOR) to stable Mild Cognitive Impairment (MCI), progressive MCI (Mild Cognitive Impairment converter (MCIc)), and finally AD. This work compares ordinal and nominal classification models using the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [22].

Our main contributions include: 1) the use of an alternative to the raw voxels inputs based on Convolutional Neural Networks (CNNs) to extract image features. This process radically reduces the input dimensionality and transforms the data distribution before classification. For this, a fine-tuning is applied to a 3-Dimensional Convolutional Neural Network (3DCNN) [16], as an extractor of the most critical features of the images. 2) Under a nested cross-validation procedure, we evaluate how different linear models perform on this low-dimensional deep space compared to their performance on raw voxels. 3) We compare this hybrid feature-extraction pipeline against a standard end-to-end ResNet18 classification model.

The remainder of this paper is structured as follows. Section 2 includes a description of the dataset used and the main demographic details for the different groups. Section 3 describes the methodology applied and the different methods considered. Section 4 explains the experimentation done on the study. Section 5 shows the results obtained after the experimentation and our interpretation of them, followed by the conclusions in 6.

2 Dataset

Data used in the preparation of this article were obtained from the ADNI database [22]. ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD.

The dataset of this work consists of 818 sMRI scans of different patients, with a size of $121 \times 145 \times 121$ per patient. The data is labelled into four stages of severity. While the temporal progression of the disease naturally advances from a cognitively normal state to full dementia, we assign the labels following the inverse direction, and ordering the classes goes from highest to lowest clinical severity. Throughout this paper, the symbol $\prec$ denotes the order within this assigned label space (i.e., lower to higher label value) rather than the temporal direction of disease progression, explicitly establishing the classification sequence as AD $\prec$ MCIc $\prec$ MCI $\prec$ NOR. The main demographic details and clinical scores for these groups are outlined in Table 1, where MMSE is the Mini-Mental State Examination score.

Table 1: Demographic and clinical characteristics of the dataset (Mean_{Std}).

Characteristic	**AD (label 0)**	**MCIc (label 1)**	**MCI (label 2)**	**NOR (label 3)**
Number of images	188	216	185	229
Age	$75.3_{7.5}$	$74.5_{7.0}$	$75.0_{7.7}$	$75.9_{5.0}$
Sex (M/F)	99/89	136/80	122/63	119/110
MMSE	$23.2_{2.2}$	$26.8_{1.7}$	$27.2_{2.5}$	$29.1_{1.0}$

All scans, the 3-dimensional images obtain with the MRI, were preprocessed via CAT12 [10] and SPM12 [9] to obtain segmented Grey Matter (GM) volumetric maps. Data preparation varies depending on the target representation space. For the DL method, the raw 3D images undergo intensity normalisation to a $[0, 1]$ range and are spatially resized to $128 \times 128 \times 128$ voxels. After fine-tuning the ResNet18, we use it as a deep feature extractor to project each 3D scan into a compact 512-dimensional vector.

Conversely, the data pipeline for the classical linear models remains strictly consistent, whether they ingest these new deep features or the massive raw GM voxels. Every input matrix $X \in \mathbb{R}^{N \times D}$ (N being subjects and D being features/voxels) is processed in two stages: 1) variance filtering, discarding uninformative columns with zero variance; and 2) standardisation, scaling the remaining variables to a zero mean and unit variance. To prevent data leakage, these steps are strictly wrapped inside a training pipeline, ensuring they are only fitted on the training split during cross-validation

3 Proposed methodology

In this section, the hybrid approach we considered for classification is described. As previously discussed, instead of applying standard classification, we use a specific OC framework. The primary goal of a classification task in machine learning is to assign the appropriate label to an input pattern. Distinct from standard nominal classification, which uses unordered categories, OC [15] addresses problems where the classes follow a natural sequence determined by their real-world context. This order is typically denoted as $C_1 \prec C_2 \prec \cdots \prec C_Q$. Here, the symbol $\prec$ indicates the progression of the categories, acknowledging that the actual mathematical distance between them is usually unknown and left unspecified.

To facilitate the application of ordinal evaluation metrics, every class within the dataset is assigned a numerical value along this scale. This mapping offers a direct mechanism to measure the magnitude of misclassifications, penalizing prediction errors according to the established natural progression of the categories.

The feature extraction phase of a fine-tuned ResNet18 is used to obtain the most important characteristics. Those features are then fed to different classic classifiers, both nominal (Logistic Regression (LR), Support Vector Machines (SVM)) and ordinal models (Logistic All-Threshold (LogAT), Logistic Immediate-Threshold (LogIT), Support Vector Ordinal Regression with EXplicit constraints (SVOREX)), that perform the final prediction.

We compared our approach with the same classifiers using the raw voxels of the images as input and with the full ResNet18 as a classifier. The feature extraction and the raw voxel baseline approaches were evaluated across three different configurations:

- Binary (2-class). The six possible pairwise combinations are evaluated: AD vs. MCIc, AD vs. MCI, AD vs. NOR, MCIc vs. MCI, MCIc vs. NOR, and MCI vs. NOR.
- Three-class. A scenario merging MCIc and MCI into a single Combined MCI (MCIcomb) class: AD $\prec$ MCIcomb $\prec$ NOR.
- Four-class. The complete clinical staging sequence: AD $\prec$ MCIc $\prec$ MCI $\prec$ NOR.

The pure end-to-end evaluation of the ResNet18 was conducted exclusively on the most challenging scenario (the four-class problem). To properly analyse how feature extraction influences the algorithms, we fine-tuned two ResNet18 models, one using the classic Cross-Entropy Loss (CELoss) and other trained with an ordinal loss, the Weighted Kappa Loss (WKLoss). The CELoss is given by:

$$\mathcal{L}_{CE} = -\sum_{k=1}^{N}\sum_{i=1}^{Q} p_{y_k,i} \log\left(\mathrm{P}(\mathrm{y} = i \mid \mathbf{x}_k)\right), \tag{1}$$

where $p_{y_k,i} \in \{0, 1\}$ represents the one hot encoding of the true class for the k-th pattern ($p_{y_k,i} = 1$ when $y_k = \mathcal{C}_i$, and 0 otherwise), and $\mathrm{P}(\mathrm{y} = i \mid \mathbf{x}_k)$ is the

predicted probability for class i and pattern k. The WKLoss [21] is given by:

$$\mathcal{L}_{\mathrm{WK}}(\mathbf{X}, \mathbf{y}) = \frac{\sum_{i=1}^{J}\sum_{j=1}^{J}\omega_{i,j}\sum_{k=1}^{N}\mathrm{P}\left(\mathrm{y}=\mathcal{C}_i \mid \mathbf{x}_k\right)\ p_{y_k,j}}{\frac{1}{N}\sum_{i=1}^{J}\sum_{j=1}^{J}\omega_{i,j}\left(\sum_{k=1}^{N}\mathrm{P}\left(\mathrm{y}=\mathcal{C}_i \mid \mathbf{x}_k\right)\right)\left(\sum_{k=1}^{N}p_{y_k,j}\right)}, \tag{2}$$

where ω is the penalisation matrix, defined either linearly or quadratically. Its elements are given by $\omega_{i,j} = \frac{|i-j|}{J-1}$ for the linear case and by $\omega_{i,j} = \frac{(i-j)^2}{(J-1)^2}$ for the quadratic case.

4 Experimental settings

In this section we explain the experimentation done and the performance metrics used to validate it. We start by introducing the classification models followed by an explanation on how the training and testing were carried out.

4.1 Classification Models

To clearly observe how dimensionality and error penalisation affect performance, we tested the following linear models on both the raw GM voxels and the deep feature vectors. This set includes a trivial baseline (Dummy), two nominal methodologies, and three ordinal methods. Note that the direct end-to-end evaluation of the ResNet18 will be presented separately.

Table 2: Brief description of the classifiers used in the work

Classifier	Description	Ordinal
Dummy	This classifier assigns the majority class as the prediction.	No
LR [18]	Standard Logistic Regression.	No
LogAT / LogIT [19]	Two ordinal variations of LR.	Yes
SVM [7]	Linear SVM acting as a second nominal baseline.	No
SVOREX [6]	An ordinal counterpart of SVM.	Yes

4.2 Models training

To guarantee unbiased generalisation estimates and avoid data leakage, we implemented a nested stratified setting consisting of an outer 5-fold and an inner validation loop, with a fixed random seed to ensure reproducibility. The outer folds are used to test the models on unseen subjects. Also, the ResNet18 is initialised from pre-trained weights of ImageNet [5] and fine-tuned, with two different losses (nominal and ordinal), for each outer fold, ensuring the network acts

as a "blind" feature extractor that has never processed the test images during its optimisation. This results in two different feature extractor models.

Inside every outer training fold, hyperparameters are tuned by grid search, with an inner validation strategy that depends on the algorithm. For the CNN fine-tuning, each configuration is scored on a stratified 20% hold-out, while a separate internal split tracks the validation loss for early stopping. For the CNN we rely on this single hold-out rather than an inner 5-fold because each fit requires training a full 3D ResNet-18, and a k-fold would multiply this already high cost by 5. Since the inner split only determines decisions on the training partition, it leaves the outer folds, and the reported estimates, unaffected. Meanwhile, for the nominal and ordinal classifiers, the grid search relies on an inner stratified 5-fold validation.

The regularisation (C) parameter is explored for all linear algorithms considering the values $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$.The optimisation target is Average Mean Absolute Error (AMAE) for multi-class setups and Balanced Accuracy (BACC) for binary ones (see Section 4.3). Overall performance is given as the mean and standard deviation of the outer test folds.

4.3 Evaluation Metrics

Different evaluation metrics have been considered for comparison of the classifiers, including both nominal and ordinal alternatives. The metrics considered in this study are:

- Correct Classification Rate (CCR) [4]. Standard accuracy, defined as the proportion of correctly classified samples.
- Balanced Accuracy (BACC) [20]. Calculated as the macro-average of per-class recall to account for class imbalance.
- Mean Absolute Error (MAE) [2]. Average absolute distance (in number of categories) between predicted and true ordinal labels.
- Average Mean Absolute Error (AMAE) [11]. Class-averaged MAE, introduced to handle class imbalance in ordinal problems.

For interpretation, higher CCR and BACC values imply a superior accuracy, whereas lower MAE and AMAE scores indicate fewer ordinal mistakes.

5 Results

The experimental results are summarised in Tables 3 to 5. In Table 5, we also include the results for the full end-to-end training of the ResNet model. Note that, for binary classification (Table 3), ordinal metrics and loss are not considered, as they do not make sense.

The resulting confusion matrices for both the three-stage and four-stage classification tasks are shown in Table 6 and 7. These matrices are obtained as the sum of the five ones derived from the test set of each iteration of the 5-fold validation. The tables contrast the predictive behaviour of the models when trained

Table 3: Results for binary classification experiments across different feature spaces (Mean_{Std}). Best results per scenario are highlighted in bold. Second best are presented in italics.

Combination	Classifier	Feature Space	CCR ↑	BACC ↑
AD vs MCI	Dummy	Baseline	$0.4772_{0.0633}$	$0.4770_{0.0628}$
	LR	Raw Voxels	*$0.7320_{0.0240}$*	$0.7319_{0.0241}$
		ResNet-CE	$0.7319_{0.0260}$	*$0.7323_{0.0260}$*
	SVM	Raw Voxels	**$0.7345_{0.0354}$**	**$0.7344_{0.0354}$**
		ResNet-CE	$0.5838_{0.1026}$	$0.5840_{0.1027}$
AD vs MCIc	Dummy	Baseline	$0.5172_{0.0764}$	$0.5127_{0.0770}$
	LR	Raw Voxels	*$0.5791_{0.0221}$*	$0.5719_{0.0246}$
		ResNet-CE	**$0.5844_{0.0575}$**	**$0.5821_{0.0572}$**
	SVM	Raw Voxels	$0.5717_{0.0263}$	*$0.5727_{0.0283}$*
		ResNet-CE	$0.5324_{0.0602}$	$0.5235_{0.0539}$
AD vs NOR	Dummy	Baseline	$0.5133_{0.0601}$	$0.5046_{0.0617}$
	LR	Raw Voxels	**$0.8442_{0.0164}$**	**$0.8428_{0.0166}$**
		ResNet-CE	$0.8247_{0.0496}$	$0.8199_{0.0483}$
	SVM	Raw Voxels	*$0.8299_{0.0432}$*	*$0.8317_{0.0426}$*
		ResNet-CE	$0.6918_{0.1603}$	$0.7036_{0.1465}$
MCI vs NOR	Dummy	Baseline	$0.5026_{0.0678}$	$0.4923_{0.0700}$
	LR	Raw Voxels	*$0.6112_{0.0389}$*	*$0.6070_{0.0447}$*
		ResNet-CE	**$0.6374_{0.0521}$**	**$0.6329_{0.0529}$**
	SVM	Raw Voxels	$0.5725_{0.0154}$	$0.5721_{0.0113}$
		ResNet-CE	$0.5434_{0.1239}$	$0.5451_{0.1225}$
MCIc vs MCI	Dummy	Baseline	$0.4788_{0.0625}$	$0.4727_{0.0604}$
	LR	Raw Voxels	*$0.5988_{0.0709}$*	*$0.5977_{0.0733}$*
		ResNet-CE	**$0.6284_{0.0278}$**	**$0.6300_{0.0298}$**
	SVM	Raw Voxels	$0.5963_{0.0583}$	$0.5969_{0.0593}$
		ResNet-CE	$0.4990_{0.0691}$	$0.4974_{0.0795}$
MCIc vs NOR	Dummy	Baseline	$0.5191_{0.0698}$	$0.5184_{0.0707}$
	LR	Raw Voxels	*$0.7528_{0.0396}$*	*$0.7513_{0.0386}$*
		ResNet-CE	**$0.7573_{0.0614}$**	**$0.7571_{0.0613}$**
	SVM	Raw Voxels	$0.7101_{0.0149}$	$0.7091_{0.0154}$
		ResNet-CE	$0.5438_{0.1580}$	$0.5455_{0.1548}$

on raw voxel inputs versus when utilising the low-dimensional ResNet-CE representations. The actual clinical diagnoses correspond to the rows of each matrix, while the columns represent the predictions of the models.

The results demonstrate a profound interaction between feature space geometry and the chosen classification algorithm, going against the trend of using fully end-to-end DL pipelines for medical imaging. While standard architectures like ResNet18 are highly capable of anatomical compression, they tend to overfit small datasets, resulting in very poor accuracy (e.g., CCR ≈ 0.30 in the full four-class scenario). Separating the nominal semantic extraction from the classi-

Table 4: Results for three-class (AD vs. MCIcomb vs. NOR) experiments comparing raw voxels and deep feature extraction (Mean_{Std}). Best results per scenario are highlighted in bold. Second best are presented in italics.

Classifier	Feature Space	CCR ↑	BACC ↑	MAE ↓	AMAE ↓
Dummy	Baseline	$0.3484_{0.0212}$	$0.3032_{0.0290}$	$0.8008_{0.0261}$	$0.8938_{0.0372}$
LR	Raw Voxels	$0.5232_{0.0176}$	$0.5555_{0.0209}$	$0.5171_{0.0165}$	$\textit{0.4971}_{\textit{0.0179}}$
	ResNet-CE	$\mathbf{0.5574_{0.0495}}$	$0.5153_{0.0501}$	$\textit{0.4696}_{\textit{0.0753}}$	$0.5208_{0.0827}$
	ResNet-WK	$0.5306_{0.0179}$	$0.4860_{0.306}$	$0.4927_{0.0319}$	$0.5443_{0.0462}$
LogAT	Raw Voxels	$0.4780_{0.0300}$	$0.5216_{0.0268}$	$0.5635_{0.0347}$	$0.5304_{0.0342}$
	ResNet-CE	$0.5415_{0.0491}$	$0.5458_{0.0582}$	$0.4879_{0.0645}$	$\mathbf{0.4937_{0.0790}}$
	ResNet-WK	$0.5366_{0.0236}$	$0.5387_{0.0284}$	$0.4915_{0.0383}$	$0.4984_{0.0490}$
LogIT	Raw Voxels	$0.4755_{0.0245}$	$0.5164_{0.0253}$	$0.5624_{0.0254}$	$0.5309_{0.0278}$
	ResNet-CE	$0.5060_{0.0410}$	$\textit{0.5579}_{\textit{0.0515}}$	$0.5404_{0.0541}$	$0.5045_{0.0667}$
	ResNet-WK	$0.5073_{0.0307}$	$\mathbf{0.5626_{0.0328}}$	$0.5429_{0.0608}$	$0.5047_{0.0737}$
SVM	Raw Voxels	$0.4853_{0.0357}$	$0.4913_{0.0411}$	$0.5538_{0.0446}$	$0.5608_{0.0590}$
	ResNet-CE	$0.3690_{0.0778}$	$0.4492_{0.1040}$	$0.7546_{0.1534}$	$0.7055_{0.1784}$
	ResNet-WK	$0.4108_{0.0564}$	$0.4663_{0.0477}$	$0.6784_{0.1114}$	$0.6441_{0.1176}$
SVOREX	Raw Voxels	$0.5245_{0.0209}$	$0.4523_{0.0196}$	$0.4914_{0.0272}$	$0.5687_{0.0264}$
	ResNet-CE	$\textit{0.5452}_{\textit{0.0467}}$	$0.4888_{0.0492}$	$\mathbf{0.4658_{0.0470}}$	$0.5265_{0.0495}$
	ResNet-WK	$\textit{0.5452}_{\textit{0.0229}}$	$0.4871_{0.0351}$	$0.4805_{0.0261}$	$0.5466_{0.0438}$

Table 5: Results for four-class (AD vs. MCIc vs. MCI vs. NOR) experiments comparing raw voxels and deep feature extraction (Mean_{Std}). Best results per scenario are highlighted in bold. Second best are presented in italics.

Classifier	Feature Space	CCR ↑	BACC ↑	MAE ↓	AMAE ↓
Dummy	Baseline	$0.2384_{0.0245}$	$0.2363_{0.0240}$	$1.2848_{0.0529}$	$1.2851_{0.0553}$
LR	Raw Voxels	$0.4107_{0.0304}$	$0.4005_{0.0291}$	$0.8080_{0.0321}$	$0.8176_{0.0275}$
	ResNet-CE	$\mathbf{0.4511_{0.0190}}$	$\mathbf{0.4386_{0.0136}}$	$0.7764_{0.0739}$	$0.7886_{0.0649}$
	ResNet-WK	$0.4364_{0.0283}$	$0.4224_{0.0280}$	$0.8450_{0.1306}$	$0.8576_{0.1316}$
LogAT	Raw Voxels	$0.3924_{0.0234}$	$0.4013_{0.0176}$	$0.8740_{0.0241}$	$0.8603_{0.0298}$
	ResNet-CE	$0.4034_{0.0337}$	$0.4077_{0.0309}$	$\mathbf{0.7287_{0.0685}}$	$\mathbf{0.7254_{0.0658}}$
	ResNet-WK	$0.3679_{0.0470}$	$0.3688_{0.0426}$	$0.7679_{0.0880}$	$0.7675_{0.0847}$
LogIT	Raw Voxels	$0.3925_{0.0141}$	$0.3995_{0.0150}$	$0.8837_{0.0577}$	$0.8712_{0.0589}$
	ResNet-CE	$\textit{0.4412}_{\textit{0.0637}}$	$\textit{0.4377}_{\textit{0.0607}}$	$\textit{0.7569}_{\textit{0.0939}}$	$0.7627_{0.0887}$
	ResNet-WK	$0.4278_{0.0250}$	$0.4227_{0.0232}$	$0.8034_{0.1087}$	$0.8104_{0.1067}$
SVM	Raw Voxels	$0.4181_{0.0173}$	$0.4127_{0.0183}$	$0.7763_{0.0298}$	$0.7815_{0.0269}$
	ResNet-CE	$0.2994_{0.0403}$	$0.3160_{0.0359}$	$1.0815_{0.2775}$	$1.0566_{0.2628}$
	ResNet-WK	$0.3275_{0.0569}$	$0.3357_{0.0503}$	$1.0652_{0.2645}$	$1.0428_{0.2510}$
SVOREX	Raw Voxels	$0.3790_{0.0419}$	$0.3744_{0.0322}$	$0.7592_{0.0250}$	$0.7598_{0.0244}$
	ResNet-CE	$0.4193_{0.0261}$	$0.4075_{0.0238}$	$\mathbf{0.7287_{0.0624}}$	$\textit{0.7381}_{\textit{0.0579}}$
	ResNet-WK	$0.4377_{0.0372}$	$0.4233_{0.0392}$	$0.7616_{0.0537}$	$0.7708_{0.0537}$
ResNet18	End-to-End CE	$0.3019_{0.0385}$	$0.2978_{0.0344}$	$1.2462_{0.2052}$	$1.2455_{0.1947}$
	End-to-End WK	$0.2849_{0.0534}$	$0.2787_{0.0448}$	$1.2320_{0.2504}$	$1.2432_{0.2347}$

fication layer proves to be a more robust strategy, even though it is not universally beneficial across all classical models. Margin-based classifiers, particularly

Table 6: Confusion matrices for the three-class experiment. The rows represent the True classes and the columns represent the Predicted classes (A: AD, M: MCIcomb (MCIcb), N: NOR). The correct predictions (main diagonal) are highlighted in bold. Bellow each matrix, the corresponding metrics (Mean_{Std}) are included, with the best value per metric highlighted in bold. Second best are presented in italics.

Features	**True**	**LR**			**LogAT**			**LogIT**			**SVM**			**SVOREX**		
		A	M	N	A	M	N	A	M	N	A	M	N	A	M	N
Raw Voxels	AD	**109**	64	15	**153**	27	8	**150**	31	7	**85**	85	18	**50**	131	7
	MCIcb	114	**163**	124	175	**154**	72	173	**156**	72	114	**182**	105	38	**301**	62
	NOR	18	55	**156**	26	119	**84**	24	122	**83**	14	85	**130**	6	145	**78**
	BACC ↑	*0.5555*$_{0.0209}$			0.5216$_{0.0268}$			0.5164$_{0.0253}$			0.4913$_{0.0411}$			0.4523$_{0.0196}$		
	AMAE ↓	*0.4971*$_{0.0179}$			0.5304$_{0.0342}$			0.5309$_{0.0278}$			0.5608$_{0.0590}$			0.5687$_{0.0264}$		
ResNet-CE	AD	**64**	112	12	**98**	76	14	**120**	46	22	**148**	18	22	**54**	127	7
	MCIcb	56	**270**	75	102	**208**	91	134	**132**	135	261	**59**	81	53	**284**	64
	NOR	10	97	**122**	10	82	**137**	16	51	**162**	79	55	**95**	2	119	**108**
	BACC ↑	0.5153$_{0.0501}$			0.5458$_{0.0582}$			**0.5579$_{0.0515}$**			0.4492$_{0.1040}$			0.4888$_{0.0492}$		
	AMAE ↓	0.5208$_{0.0827}$			**0.4937$_{0.0790}$**			0.5045$_{0.0667}$			0.7055$_{0.1784}$			0.5265$_{0.0495}$		

Table 7: Confusion matrices for the four-class experiment. The rows represent the True classes and the columns represent the Predicted classes (A: AD, Mc: MCIc, M: MCI, N: NOR). The correct predictions (main diagonal) are highlighted in bold. Bellow each matrix there are its metrics (Mean_{Std}), with the best value per metric highlighted in bold. Second best are presented in italics.

Features	**True**	**LR**				**LogAT**				**LogIT**				**SVM**				**SVOREX**			
		A	Mc	M	N	A	Mc	M	N	A	Mc	M	N	A	Mc	M	N	A	Mc	M	N
Raw Voxels	AD	**94**	61	16	17	**157**	18	6	7	**157**	17	7	7	**89**	63	23	13	**49**	116	16	7
	MCIc	66	**62**	33	55	120	**47**	30	19	117	**47**	26	26	55	**80**	44	37	28	**115**	51	22
	MCI	28	51	**30**	76	69	34	**40**	42	68	35	**31**	51	26	47	**51**	61	9	74	**66**	36
	NOR	12	22	45	**150**	35	40	77	**77**	38	35	70	**86**	12	23	72	**122**	3	46	100	**80**
	BACC ↑	0.4005$_{0.0291}$				0.4013$_{0.0176}$				0.3995$_{0.0150}$				0.4127$_{0.0183}$				0.3744$_{0.0322}$			
	AMAE ↓	0.8176$_{0.0275}$				0.8603$_{0.0298}$				0.8712$_{0.0589}$				0.7815$_{0.0269}$				0.7598$_{0.0244}$			
ResNet-CE	AD	**81**	65	21	21	**73**	74	35	6	**108**	37	28	15	**147**	20	16	5	**54**	103	20	11
	MCIc	63	**77**	29	47	51	**80**	73	12	84	**44**	60	28	97	**23**	45	20	43	**115**	39	19
	MCI	22	39	**45**	79	12	40	**96**	37	30	20	**59**	76	97	23	**45**	20	8	71	**50**	56
	NOR	17	20	26	**166**	6	25	117	**81**	13	20	46	**150**	75	27	97	**30**	3	46	56	**124**
	BACC ↑	**0.4386$_{0.0136}$**				0.4077$_{0.0309}$				*0.4377*$_{0.0607}$				0.3160$_{0.0359}$				0.4075$_{0.0238}$			
	AMAE ↓	0.7886$_{0.0649}$				**0.7254$_{0.0658}$**				0.7627$_{0.0887}$				1.0566$_{0.2628}$				*0.7381*$_{0.0579}$			

the linear SVM, have a drop in performance when transitioned from the massive dimensionality of raw brain voxels to the continuous 512-dimensional embeddings generated by the CELoss. In contrast, probabilistic algorithms like LR and ordinal methods (LogAT, SVOREX) thrive in this compact feature space. This aligns with previous observations in ordinal regression literature [15], which demonstrate that supervised dimensionality reduction and the preservation of the intrinsic data manifold greatly facilitate the optimisation of parallel decision thresholds by minimising within-class variance.

Additionally, we can also see a lower performance of the ordinal loss for the fine-tuning of the ordinal extractor. Even though ordinal methods perform better in classification than their nominals counterparts.

From a clinical perspective, identifying extreme structural atrophy (such as differentiating AD from NOR) is an easier task across all evaluated representation spaces. The true difficulty lies in resolving the transitional cognitive phases, particularly within the MCI spectrum (MCIc vs. MCI). In these highly overlapping boundaries, traditional nominal models often produce a wider variance in their error distribution, leading to critical medical mistakes, such as predicting a cognitively normal state for a patient with confirmed AD ($|y_i - \hat{y}_i| > 1$). By applying explicit mathematical constraints directly on the threshold parameters [6], ordinal methodologies tend to concentrate predictions on the main diagonal and its immediate adjacent entries. Also, even when algorithms like LogAT misclassify a sample, the error is typically within the closest clinical severity level, achieving the lowest overall AMAE scores. This confirms that while deep neural networks are unmatched as semantic extractors, the clinical progression rules should be mathematically encoded *a posteriori* by ordinal classifiers to guarantee a safe and reliable computer-aided diagnosis.

6 Conclusion

In this work, we introduced a hybrid approach for the classification of sMRI. We used a ResNet18 feature extractor phase to extract the most important characteristics of the images and then feed those feature to classic classifiers to obtain the final predictions. We compared our approach with the classification of the images using the raw voxels as inputs and with the complete ResNet18 classifier.

The fundamental challenge in automated AD staging lies not only in maximising overall accuracy but in ensuring clinical safety. This study shows that, even though deep convolutional networks are extremely good at compressing complex anatomical changes from sMRI, leaving the final diagnostic decision to classical ordinal algorithms provides a crucial layer of robustness.

The results suggest that the reduced space generated by the CNN is ideal for the classical classifiers in which they can effectively apply rules of disease progression to the final prediction. Also, CNNs have proven to be excellent at extracting features but they are weak when used as an end-to-end model for classifying medical images that come from a small dataset. Lastly, the reduced

misclassification on distant stages provides clinical safety when used in real medical environments, ensuring that the errors committed by the models are not clinically severe.

Taken together, the findings presented in this work suggest that clinical progression rules must be specifically considered as constraints during model training, rather than implicitly learned, in order to guarantee safe and reliable computer-aided diagnosis. Future research will focus on integrating different CNN architectures as feature extractors, enriching the input space and possibly improving classification performance.

Acknowledgements. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). The present study has been supported by the "Agencia Estatal de Investigación (España)" (grant ref.: PID2023-150663NB-C22 / AEI / 10.13039 / 501100011033), and by the University of Córdoba and Junta de Andalucía (grant ref.: PP2F_L1_15). Francisco Bérchez-Moreno has been supported by "Plan Propio de Investigación Submodalidad 2.2 Contratos predoctorales" of the Universidad de Córdoba. Víctor Manuel Vargas has been supported by the Ministerio de Ciencia, Innovación y Universidades, the Agencia Estatal de Investigación and the European Social Fund Plus (grant ref.: MICIU/AEI/10.13039/501100011033, JDC2024-054787-I).

References

1. Alzheimer Europe: Alzheimer's dementia. https://www.alzheimer-europe.org/dementia/alzheimers-dementia, accessed: 2026-02-21
2. Baccianella, S., Esuli, A., Sebastiani, F.: Evaluation measures for ordinal regression. In: 9th International Conference on Intelligent Systems Design and Applications. pp. 283–287 (2009). https://doi.org/10.1109/ISDA.2009.230
3. Barbero-Gómez, J., Cruz, R., Cardoso, J., Gutiérrez, P., Hervás-Martínez, C.: Cnn explanation methods for ordinal regression tasks. Neurocomputing (2025). https://doi.org/10.1016/j.neucom.2024.128878
4. Bérchez-Moreno, F., Fernández, J.C., Hervás-Martínez, C., Gutiérrez, P.A.: Fusion of standard and ordinal dropout techniques to regularise deep models. Information Fusion **106**, 102299 (2024). https://doi.org/https://doi.org/10.1016/j.inffus.2024.102299, https://www.sciencedirect.com/science/article/pii/S1566253524000770
5. Cardoso, M.J., et. al: Monai: An open-source framework for deep learning in healthcare (2022). https://doi.org/10.48550/ARXIV.2211.02701, https://arxiv.org/abs/2211.02701
6. Chu, W., Keerthi, S.: New approaches to support vector ordinal regression. In: Proceedings of the 22nd ICML. pp. 145–152 (2005). https://doi.org/10.1145/1102351.1102370
7. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**(3), 273–297 (09 1995)

8. Cuingnet, R., Gerardin, E., Tessieras, J., et al.: Automatic classification of patients with alzheimer's disease from structural mri: a comparison of ten methods using the adni database. NeuroImage **56**(2), 766–781 (2011). https://doi.org/10.1016/j.neuroimage.2010.06.013
9. Friston, K., et al.: Statistical Parametric Mapping: The Analysis of Functional Brain Images. Academic Press, London (2007)
10. Gaser, C., Dahnke, R.: Cat12: a computational anatomy toolbox for the analysis of structural mri data. GigaScience **13**, giae049 (2024). https://doi.org/10.1093/gigascience/giae049
11. Gómez-Orellana, A., Guijo-Rubio, D., Gutiérrez, P., Hervás-Martínez, C., Vargas, V.: ORFEO: Ordinal classifier and regressor fusion for estimating an ordinal categorical target. Eng. Applications of Artificial Intelligence **133**, 108462 (2024)
12. Gorriz, J.M., et al.: Artificial intelligence within the interplay between natural and artificial computation: Advances in data science, trends and applications. Neurocomputing **410**, 237–270 (2020)
13. Gorriz, J.M., et al.: Computational approaches to explainable artificial intelligence: Advances in theory, applications and trends. Information Fusion **100**, 101945 (2023)
14. Gutiérrez, P., García, S.: Current prospects on ordinal and monotonic classification. Progress in Artificial Intelligence pp. 171–179 (2016). https://doi.org/10.1007/s13748-016-0088-y
15. Gutiérrez, P.A., Pérez-Ortiz, M., Sánchez-Monedero, J., Fernández-Navarro, F., Hervás-Martínez, C.: Ordinal regression methods: Survey and experimental study. IEEE Transactions on Knowledge and Data Engineering **28**(1), 127–146 (2016). https://doi.org/10.1109/TKDE.2015.2457911
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
17. Jain, R., Jain, N., Aggarwal, A., Hemanth, D.: Convolutional neural network based alzheimer's disease classification from magnetic resonance brain images. Cognitive Systems Research pp. 147–159 (2019). https://doi.org/10.1016/j.cogsys.2018.12.015
18. Lemon, S.C., Roy, J., Clark, M.A., Friedmann, P.D., Rakowski, W.: Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. Annals of Behavioral Medicine **26**(3), 172–181 (12 2003)
19. Rennie, J., Srebro, N.: Loss functions for preference levels: Regression with discrete ordered labels. Proceedings of the IJCAI Multidisciplinary Workshop on Advances in Preference Handling (01 2005)
20. Vargas, V.M., Gómez-Orellana, A.M., Gutiérrez, P.A., Hervás-Martínez, C., Guijo-Rubio, D.: Ebano: A novel ensemble based on unimodal ordinal classifiers for the prediction of significant wave height. Knowledge-Based Systems **300**, 112223 (2024)
21. Vargas, V.M., Gutiérrez, P.A., Hervás-Martínez, C.: Cumulative link models for deep ordinal classification. Neurocomputing **401**, 48–58 (2020). https://doi.org/https://doi.org/10.1016/j.neucom.2020.03.034, https://www.sciencedirect.com/science/article/pii/S0925231220303805
22. Weiner, M.: Adni. https://adni.loni.usc.edu/ (2003)