Artificial Intelligence In Medicine 140 (2023) 102543

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Research paper

A unique color-coded visualization system with multimodal information fusion and deep learning in a longitudinal study of Alzheimer's disease☆

Mohammad Eslami [a,b,*], Solale Tabarestani [b], Malek Adjouadi [b]

[a] Harvard Ophthalmology AI lab, Schepens Eye Research Institute of Massachusetts Eye and Ear, Harvard Medical School, Boston, MA, USA
[b] Center for Advanced Technology and Education, Florida International University, Miami, FL, United States

ARTICLE INFO

Keywords:
Alzheimer's disease
Diagnosis
Prognosis
Deep learning
Trustfulness visualization

ABSTRACT

Purpose: Automated diagnosis and prognosis of Alzheimer's Disease remain a challenging problem that machine learning (ML) techniques have attempted to resolve in the last decade. This study introduces a first-of-its-kind color-coded visualization mechanism driven by an integrated ML model to predict disease trajectory in a 2-year longitudinal study. The main aim of this study is to help capture visually in 2D and 3D renderings the diagnosis and prognosis of AD, therefore augmenting our understanding of the processes of multiclass classification and regression analysis.

Method: The proposed method, Machine Learning for Visualizing AD (ML4VisAD), is designed to predict disease progression through a visual output. This newly developed model takes baseline measurements as input to generate a color-coded visual image that reflects disease progression at different time points. The architecture of the network relies on convolutional neural networks. With 1123 subjects selected from the ADNI QT-PAD dataset, we use a 10-fold cross-validation process to evaluate the method. Multimodal inputs* include neuroimaging data (MRI, PET), neuropsychological test scores (excluding MMSE, CDR-SB, and ADAS to avoid bias), cerebrospinal fluid (CSF) biomarkers with measures of amyloid beta (ABETA), phosphorylated tau protein (PTAU), total tau protein (TAU), and risk factors that include age, gender, years of education, and ApoE4 gene.

Findings/results: Based on subjective scores reached by three raters, the results showed an accuracy of 0.82 ± 0.03 for a 3-way classification and 0.68 ± 0.05 for a 5-way classification. The visual renderings were generated in 0.08 msec for a 23×23 output image and in 0.17 ms for a 45×45 output image. Through visualization, this study (1) demonstrates that the ML visual output augments the prospects for a more accurate diagnosis and (2) highlights why multiclass classification and regression analysis are incredibly challenging. An online survey was conducted to gauge this visualization platform's merits and obtain valuable feedback from users. All implementation codes are shared online on GitHub.

Conclusion: This approach makes it possible to visualize the many nuances that lead to a specific classification or prediction in the disease trajectory, all in context to multimodal measurements taken at baseline. This ML model can serve as a multiclass classification and prediction model while reinforcing the diagnosis and prognosis capabilities by including a visualization platform.

1. Introduction

The challenges of understanding AD and its prodromal stages are associated with the meaningful interpretation of the interplay between the different biomarkers for diagnosis, multiclass classification, and regression analysis, especially as it relates to the pathogenesis of the disease [1–5] and its early detection [6,7]. There is also wide-ranging deliberation on the nature of cognitive reserve [8,9], potentially biasing the neuropsychological examinations and, ultimately, the diagnosis. Additionally, there is the issue of chronology in the manifestation of amyloid-beta plaques and tau tangles [10–12] and their synergistic effects on AD pathology. We also need to consider APOE genotypes

☆ The data used for this study can be found in the "QT-PAD Project Data" from the Alzheimer's Disease Modelling Challenge [http://www.pi4cs.org/qt-pad-challenge].

* Corresponding author at: Harvard Ophthalmology AI lab, Schepens Eye Research Institute of Massachusetts Eye and Ear, Harvard Medical School, Boston, MA, USA.
E-mail addresses: Mohammad_eslami@meei.harvard.edu (M. Eslami), Staba006@fiu.edu (S. Tabarestani), adjouadi@fiu.edu (M. Adjouadi).

https://doi.org/10.1016/j.artmed.2023.102543
Received 6 September 2022; Received in revised form 28 March 2023; Accepted 2 April 2023
Available online 7 April 2023
0933-3657/© 2023 Elsevier B.V. All rights reserved.

Downloaded for Anonymous User (n/a) at Florida International University from ClinicalKey.com by Elsevier on August 04, 2023. For personal use only. No other uses without permission. Copyright ©2023. Elsevier Inc. All rights reserved.

[13–15] for their association with a cognitive reserve and cortical thinning, as well as with its potential link to both amyloid-beta, tau aggregation, and the cerebrospinal fluid (CSF) biomarkers [16,17]. The central aim of all these studies is to identify the earliest manifestations of AD to take preventive measures and provide early treatment/therapeutic interventions [18,19].

Implementing machine learning is an effective way to approach the complex challenge of multimodal data [20–22]. However, ML models are not always easily interpretable. Visualization of the ML results can enhance our understanding of the inner workings of the algorithmic process in context to what it has learned from the baseline measurements. The assertation here is that visualization will enhance the means to assess the importance of features and the interpretability of results [23–25].

Using the "QT-PAD Project Data" from the Alzheimer's Disease Modelling Challenge [http://www.pi4cs.org/qt-pad-challenge], the proposed machine learning, named Machine Learning for Visualizing AD (ML4VisAD), construct aims to produce a color-coded visualization scheme with a unique tensorization method to yield images that express disease state and progression through the different time points in a longitudinal study. Although the goals of high accuracy in multiclass classification and prediction of disease trajectory using only baseline features is essential, the information provided visually by the ML4VisAD model brings forth subtle nuances of the machine learning decision-making process, which is especially crucial when dealing with converter cases. Ultimately, the proposed visualization method exemplifies the challenges faced in multimodal and multiclass classification and the decision-making process. Visualization may also shed some light on the "black box" problem associated with machine learning. Moreover, ML4VisAD will also augment the deliberation process through a visual opportunity to reassess ambiguous cases, like the converter cases, to determine whether a misclassification happened or that the ML visual outcome is the one projecting a correct classification, although different from the target image. In such a case, clinicians could deliberate on the visual output in context to the available measurements.

Along this line of research, the studies reported in [20,21] suggest that most machine/deep learning methods rely more often on data-related issues, proposed methodologies, and the different clinical aspects under study but ignore visualization. Similarly, most studies emphasize the relevant clinical features and the computational methods, which are more likely to produce high classification and prediction results [20,26,27]. Machine learning can also help develop medical imaging methods that address the challenging task of segmentation and noise removal [28–30]. Also, in [31], efforts are made at data reduction and using different data visualization techniques to embed complex information in 2-D images to reflect gene expression and clinical data for diagnosis.

Several other studies focused their classification and prediction algorithms on visualizing data in a dimensionally reduced decisional space. The dimensionality reduction methods typically involve the use of principal component analysis (PCA), locally linear embedding (LLE), latent profile analysis (LPA), 3D scattering transforms, and the concept of histones [32–45]. Traditionally, standard methods used to aid in the visualization and diagnosis of AD typically involve heat maps, brain connectivity maps, and specific AD signatures, such as Standard Uptake Value Ratios (SUVRs) of disease-prone brain regions [46–53]. All methods that address the challenge of high-dimensional data also use visualization methods that produce optimal decisional spaces helpful to the classification process but not necessarily geared towards facilitating a visual interpretation of a diagnosis and prognosis of the disease which ML4VisAD seeks to address.

The manuscript's structure is as follows: Section 2 provides the details of the data used in this study and the methods implemented, including the color-coding mechanism, the machine learning architecture, and its computational capability. Section 3 reports the results of varying disease states and disease progression cases. Section 4 provides a discussion reflecting on the different findings and merits of the proposed ML4VisAD model. Finally, Section 5 concludes with a retrospective on the contributions made, highlighting the complexity faced when using machine learning for multiclass classification and prediction in AD.

2. Methods

2.1. Study design

Clinical data used in the preparation of this study is from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). Only subjects that have a baseline (T0) scan (in at least one modality) and showed up for follow-up visits at T6 (6th month), T12 (12th month), and T24 (24th month) have been considered in this study, leading to a total of 1123 subjects as shown in Table 1. ADNI categorizes these subjects into the three classes of CN, MCI, and AD at baseline and for each referral session.

The input features used for each modality and the number of observations made at the different time points are obtained from the "QT-PAD Project Data" AD Modelling Challenge [http://www.pi4cs.org/qt-pad-challenge] as given in Table 2. Hence, inputs to the ML model contain features from the baseline, including MRI and PET sequences, demographic information, and specific cognitive measurements. Automatically generated outputs of the ML network are images containing colorful strips expressing disease progression at different time points. It is important to emphasize that in designing this color-coded visualization scheme, and to avoid any bias, we exclude the Mini-Mental State Examination (MMSE) and the Clinical Dementia Rating Sum of Boxes (CDR-SB) scores from the input feature space in the training and testing phases since both are used for the labeling of subjects. Furthermore, we also remove from consideration the Alzheimer's Disease Assessment Scores (ADAS11, ADAS13) as they correlate well with MMSE and CDR-SB. Each feature f of the input feature vector, e.g. FDG, is normalized by mean normalization over all its non-missing values (set F) i.e. $f\_normalized = (f - mean(F))/(max(F) - min(F))$.

After normalization, we ensure the missing values do not affect network training. It is worth mentioning that the QT dataset implicitly reports values of some features as ABETA>1700, for example. For this reason, during preprocessing of the data, ABETA of those samples higher than 1700 or smaller than 200 have been replaced by 1700 and 200, respectively. Similarly, PTAU values greater than 120 and smaller than 8 have been replaced by 120 and 8, respectively. Also, TAU values greater than 1300 and less than 80 are replaced by 1300 and 80, respectively.

2.2. Color coding

The adage "a picture is worth a thousand words," together with the challenge imposed by both the variability and interrelatedness of the multimodal features, served as an incentive to create the ML4VisAD model. The (23×23×3) target images shown in Fig. 1 are color-coded and include a region of uncertainty (RU) represented by the black bar entry. We use the three (R, G, B) channels to represent the state of the disease with different colors, AD as red, Mild Cognitive impairment-MCI: as blue, and Cognitively Normal-CN as green. In this color-coded scheme, subjects that are stable over time would display a single color as in cases (a) through (c), and subjects who convert at specific time points to other states would display two or more colors as in cases (d) through (g).

Cognitive status through a 24-month timeline (including baseline T0 and three referral sessions T6 (6th month), T12 (12th month), and T24 (24th month) define trajectories of the disease state. To assess the degree of uncertainty that the machine learning model may inject into the process, we add a black bar after the bar representing the T24 time point. This black bar could be situated anywhere in this display and is there solely to estimate the degree of uncertainty the ML model injects into the

Downloaded for Anonymous User (n/a) at Florida International University from ClinicalKey.com by Elsevier on August 04, 2023. For personal use only. No other uses without permission. Copyright ©2023. Elsevier Inc. All rights reserved.

Table 1
Study population and subgroups.

Categories based on diagnosis				
	Number of samples #			Total
	CN	AD	MCI	
Baseline	331	163	629	1123
6th month	331	195	597	1123
12th month	332	243	548	1123
24th month	334	342	447	1123

Categories based on conversion			
		#	Description
CN		329	Stable Normal
Others		8	Others (e.g. MCI to CN)
Impaired	AD	163	Stable Dementia
	MCI	442	Stable MCI
	MCIc	181	MCI converter to AD
		786	Total Impaired
		1123	**Total**

Table 2
ADNI (QT-pad challenge) dataset with the features extracted from each modality/source at baseline.

Number of subjects: 1123					
Modality	Feature	Minimum Value	Average Value	Maximum Value	Number of missed values at baseline
MRI	Ventricular volume	5650.0	39,420.220	145,115.0	39
	Hippocampus volume	3091.0	6798.67	10,769.0	158
	Whole Brain volume	738,813.0	1,022,118.21	1,443,990.50	18
	Entorhinal Cortical thickness	1426.0	3507.23	5896.0	160
	Fusiform	8991.0	17,354.76	26,280.0	160
	Middle temporal gyrus	9375.0	19,545.76	29,435.0	160
	Intracranial volume (ICV)	1,116,279.11	1,536,383.48	2,072,473.30	8
PET	'FDG'	0.69	1.24	1.707168	321
	Pittsburgh Compound-B (PIB)	1.18	1.53	1.89	1116
	'AV45'	0.83	1.19	2.02	614
Cognitive Test	RAVLT immediate	7.0	35.59	71.0	3
	RAVLT learning	−2.0	4.29	11.0	3
	RAVLT forgetting	−5	4.35	13.0	3
	RAVLT percforgetting	−100.0	57.37	100.0	4
	Functional Activities Questionnaires (FAQ)	0.0	3.73	30.0	4
	Montreal Cognitive Assessment (MoCA)	10.0	23.78	30.0	616
	Everyday Cognition (Ecog): 'EcogPtMem'	1.0	2.12	4.0	613
	Ecog: 'EcogPtLang'	1.0	1.73	4.0	612
	Ecog: 'EcogPtVisspat'	1.0	1.37	4.0	614
	Ecog: 'EcogPtPlan'	1.0	1.40	4.0	612
	Ecog: 'EcogPtOrgan'	1.0	1.48	4.0	624
	Ecog: 'EcogPtDivatt'	1.0	1.79	4.0	615
	Ecog: 'EcogPtTotal'	1.0	1.67	3.82	612
	Ecog: 'EcogSPMem'	1.0	2.01	4.0	615
	Ecog: 'EcogSPLang'	1.0	1.56	4.0	614
	Ecog: 'EcogSPVisspat'	1.0	1.38	4.0	622
	Ecog: 'EcogSPPlan'	1.0	1.50	4.0	616
	Ecog: 'EcogSPOrgan'	1.0	1.57	4.0	638
	Ecog: 'EcogSPDivatt'	1.0	1.78	4.0	621
	Ecog: 'EcogSPTotal'	1.0		3.89	614
CSF	Amyloid Beta (ABETA)	200.0	984.94	1700.0*	335
	phosphorylated tau protein (PTAU)	8	27.45	94.86	335
	Total tau protein (TAU)	80	284.98	816.9	335
Risk factors	Age	55.0	73.93	91.4	0
	years of education	6.0	15.92	20	0
	APOE4	0	0.56	2	0
	Gender				0

visual output through its many inner computations. The assertion here is that a perfect ML model should leave the black bar unchanged (i.e., zero effect), meaning that the ML model is stable and has performed its task reliably. The size 23 × 23 of the RGB image could have been of any NxN dimension. In the discussion section, we explain that a target image with a higher resolution (e.g., 45 × 45) would provide an output image that is more detailed and with smoother transition phases. However, the ML model will need more convolutional layers with a higher N; hence, the need for more training/processing time, as explained in Section 2.4.

2.3. Machine learning architecture

In the machine learning architecture shown in Fig. 2, the overall objective was to model a network estimator E in which $I_{target} = E(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5)$ is a colorful image similar to those shown in Fig. 1, and where the input space is the multimodal features of $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5\}$ in which each vector $\boldsymbol{x}_m$ comprises the extracted measurements from modality m at baseline. Features extracted from MRI, PET, CSF, cognitive tasks, and the risk factors, as shown earlier in Table 2, serve as input to the ML4VisAD model. The network is designed to have two parts ($p1, p2$) such that the initial layers address the intra/inter-modality feature extraction via fully connected layers ($p1, $) and the second part involves tensorization, extra feature extraction and image production ($p2$). Thus, $I_{predicted} = p2(\, p1(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5)\,)$ and the difference between $I_{predicted}$ and I_{target} are optimized for each observation/patient.

With the fully connected layers, the network converts the basic features for each modality into a primary representation space. Layers L0, L1, and L2 are to transform the features extracted from MRI, PET, CSF, neurocognitive measurements, and risk factors into an initial feature space representation specific to each modality. The size of the input node in layer L0 for each modality m is the length of the input

Downloaded for Anonymous User (n/a) at Florida International University from ClinicalKey.com by Elsevier on August 04, 2023. For personal use only. No other uses without permission. Copyright ©2023. Elsevier Inc. All rights reserved.

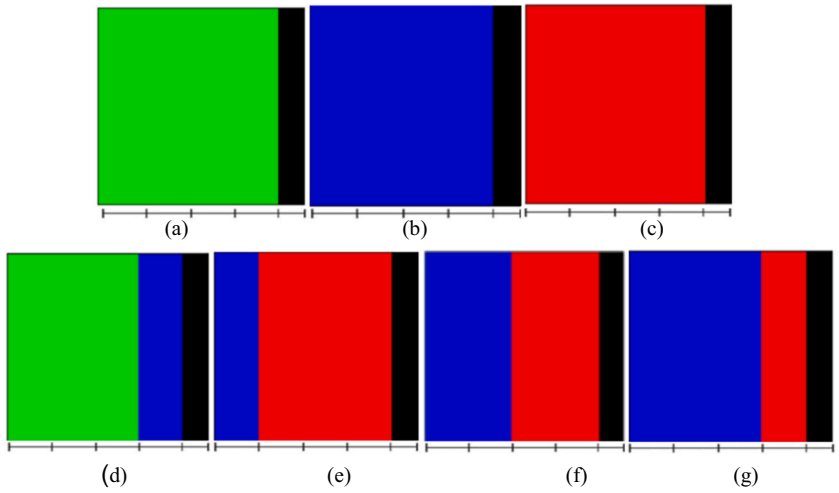


Fig. 1. Designed target images showing: (a) stable CN, (b) stable MCI, (c) stable AD, (d) CN converting to MCI at T24 (24th month), e, f, and g are MCI that progressed to AD at time points T6 (6th month), T12 (12th month), and T24 (24th month), respectively.

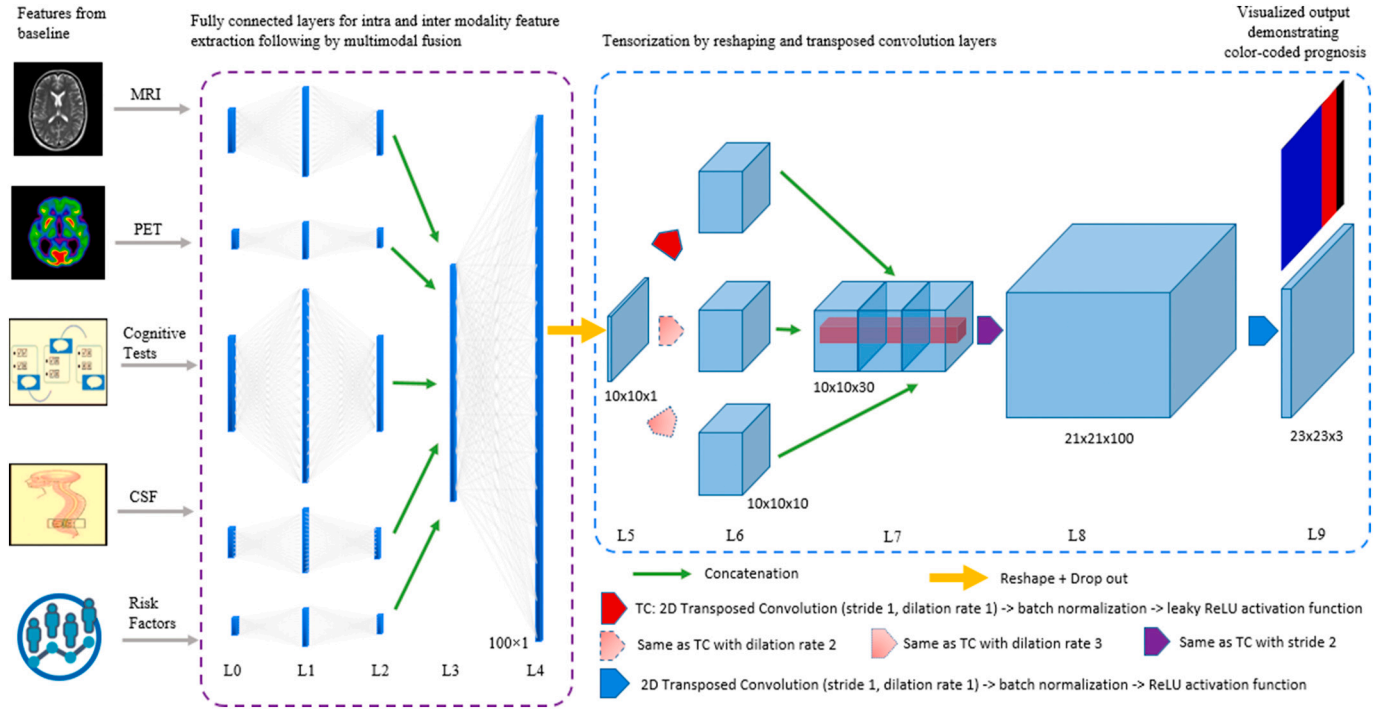


Fig. 2. Designed architecture of the network with a color-coded visual output describing disease trajectory.

feature vector $n_m = length(x_m)$, which then goes through two more fully connected layers, L1 and L2, with $2 \times n_m$ and n_m nodes, respectively, followed by linear activation layers. The previous fully connected layers of L2 are integrated into L3 by concatenating the outputs of the L2 layer to initiate the inter-modality feature space and create a new modality representation. We accomplish feature fusion and feature extraction in the inter-modality phase using concatenation (L3) and a fully connected layer (L4).

Layers L5 to L9 are for tensorization and two reasons were in mind: (1) since the input data format to our network consists of vectors from different modalities, and the target output is a colorful image, we needed to reshape the vectors to matrices to generate colorful 2-D images. 2) Layers L0 to L4 were necessary to use information from the different modalities and model progression of the disease. However, combining the features from different modalities in a standard network may not consider the heterogeneity of the data. Using a non-linear mapping function to transform the feature space into a higher-dimensional receptive field can help the network identify more significant relationships.

Our network architecture utilizes reshaping and convolutional neural layers for tensorization and extracting higher-order features from multimodal features. A tensor with dimensions of $10 \times 10 \times 30$ is generated using the following steps through layers L5, L6, and L7. Layer L5 reshapes the 100-node output vector of layer L4 to create a 2D tensor with dimensions of 10×10. Layer L6 conducts 2D transpose convolutional filtering with three different dilation rates 1, 2, and 3. For each dilation rate, we have ten kernels with 3×3 kernel size, the stride of 1, and padding of type *same*. Layer L7 is a concatenation of the three output tensors from layer L6. Layer L8 is also a 2D transpose convolution but with 100 kernels of size 3×3 and a stride of 2. Lastly, the L9 produces

Downloaded for Anonymous User (n/a) at Florida International University from ClinicalKey.com by Elsevier on August 04, 2023. For personal use only. No other uses without permission. Copyright ©2023. Elsevier Inc. All rights reserved.

the output image $I_{predicted}$ by 2D transpose convolution with three kernels of size 3 × 3 and a stride of 1. Padding in layers L8 and L9 are of type *valid*, which does not include zero padding.

Drop-out and batch normalization are also applied in layers L6, L8, and L9 to prevent overfitting. Design details and tensor dimensions for the different layers are shown in Fig. 2 as well. The GitHub repository (https://github.com/mohaEs/ML4VisAD) provides the implementation codes.

2.4. Training and evaluation

The loss function is the Mean Absolute Error (MAE) between the target image and the produced output (i.e., $loss = MAE(I_{target} - I_{predicted})$). We use the 10-fold cross-validation over subjects, and in each training session, we use 10 % of the training set as a validation set (i.e., ten times of training data split to 81/9/10 % as train/validation/test). We use 4000 epochs with a batch size of 500 to train the network. To produce a larger 45 × 45 image size is like the network shown in Fig. 2, but with the L8 layer replicated. The network makes use of the Keras TensorFlow deep learning frameworks. Using the GPU NVIDIA Geforce RTX 2080, Table 3 provides the processing time it took from feeding the input to the ML model to obtaining the visual outcome as a function of the image size.

3. Results

To demonstrate the merits of the visualization platform, we consider different scenarios, as shown in Fig. 3, that include stable subjects over time and subjects that transition from one state to another at different time points. All the results and supplementary materials are also available in the GitHub repository. These varied examples highlight the practical merits this color-coded visualization could have in facilitating diagnosis and prognosis. For each subject in the testing phase (not seen in the training phase), color-coded patterns are generated based solely on observed features at baseline.

Fig. 3 provides several examples that reflect different target images and the respective visual outputs that the ML model produces in the test phase. The target image is on the left, and the ML visual output is on the right for each displayed case. To include different scenarios of all the 1123 subjects considered, we show 4 cases each for the stable cognitively normal (CN) group with the green-colored target at all four-time points in cases (a)-(d), stable mild cognitive impaired (MCI) with a blue-colored target in cases (e)–(h), eight different transition cases that include examples of subjects who transitioned from CN to MCI or from MCI to AD with one case from MCI to CN at different time points as illustrated in cases (i)-(p), followed by four examples of stable AD subjects in cases (q)-(t), and lastly, we show few selected cases (all stable cases) that the ML model misclassified as belonging to a different stable disease state as shown in (u)-(w).

For a more meaningful assessment of disease trajectory, as we consider all these different cases, context is provided in Fig. 3 for augmented interpretability of the challenging cases. For this added context, we provide MMSE, CDR-SB, and RAVLT scores for all four-time points (T0, T6, T12, and T24), age, sex, years of education, the APOE, AV45, FDG, TAU, ABETA, number of missing features at baseline, and graphs of the SUVR measurements at T0 (baseline) and T24 (24th month), where the x-axis reflects the different brain regions for the SUVRs as annotated in Table 5. The scores/values used for MMSE and CDR-SB conform to the standards defined by ADNI. The APOE value of 0,1 or 2 specifies a carrier of zero, one, or two APOE e4 alleles. For all these displayed cases, the intent here is to use such context to deliberate on what may have led to the differences between target images and the ML visual outcomes. The Discussion *sec*tion provides more details. Notice that the legend of Fig. 3 includes patients' Record ID (RID) for other researchers interested in validating these results or who would like to perform further analysis given the nuances of the ML visual outputs that differ from their target images.

Table 3
Processing time of machine learning model.

Image size (pixels)	Trainable parameters	Train time (s)	Test time per subject (s)
23 × 23	36,143	4000 epochs: 275.67	0.008
45 × 45	126,443	4000 epochs: 987.94	0.017

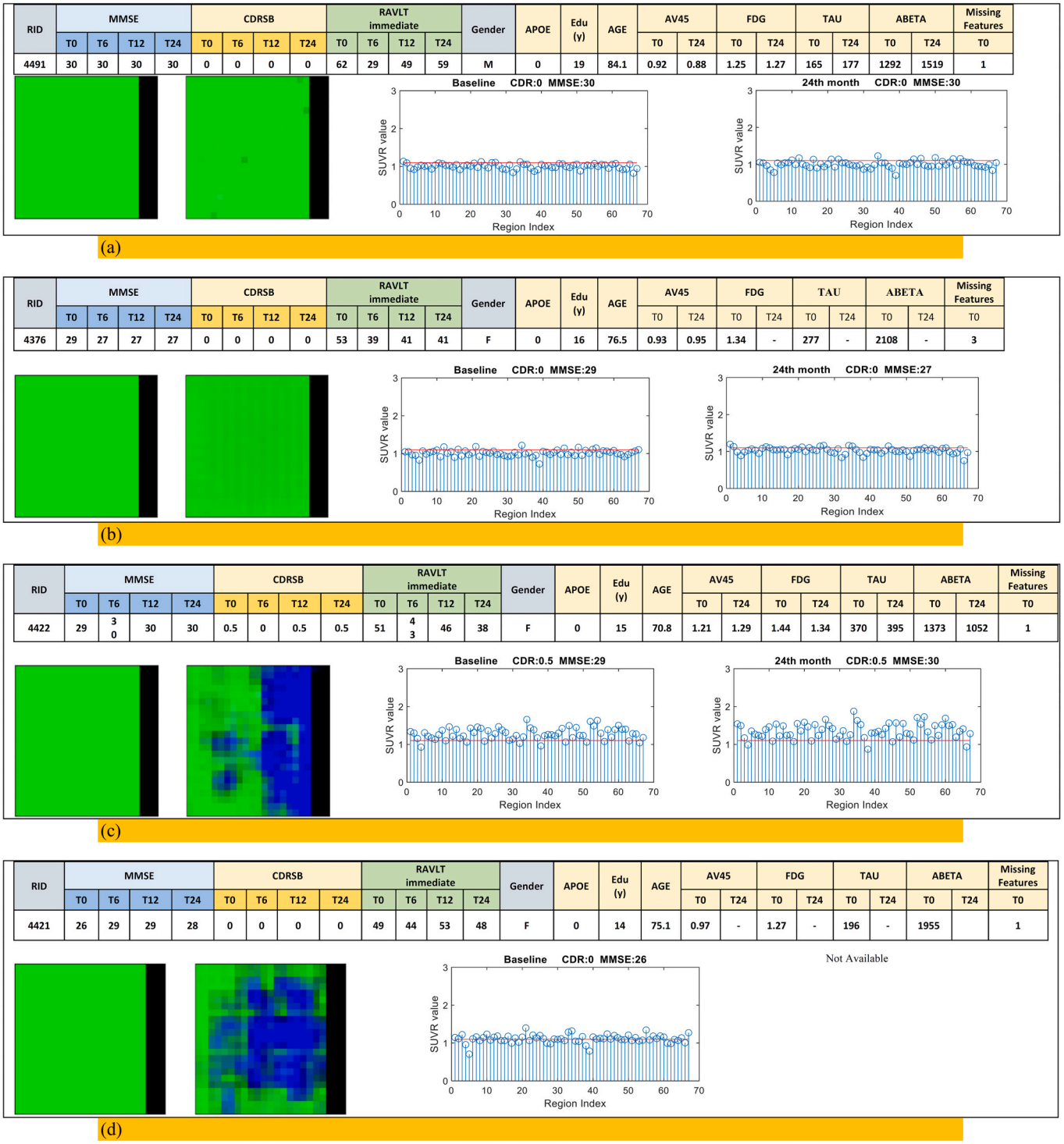
Furthermore, to acquire feedback from the research community on the practicality of this visualization platform, an online survey provided in the Appendix was conducted using the Qualtrics platform and shared via Facebook and LinkedIn. More than 100 persons participated in this survey globally, confirming the importance of the proposed method in its ease of use and in facilitating the decision-making process. This survey shows that 83.49 % of participants agree that the visual representation is easy to remember and interpret, with 79.55 % stating that they would prefer to receive the results in a graphic format. With an overwhelmingly favorable rating of 82.35 % in terms of ease of memorizing/remembering the output through visualization and 73.79 % agreeing that the visualized form speeds up the decision-making process. As for the level of uncertainty (i.e., trustfulness of the output), 81.65 % stated that different levels of trustfulness are visible in the visualized format. These are very encouraging results, and the feedback received would allow us to continue improving the platform.

In addition to these survey results, three raters (M.E., S.T., and M.S.) independently reviewed all ML-generated visual outcomes for both types of classification: 3-way (C.N., impaired, others) and 5-way (CN, MCI, MCIc, AD, others) using a developed MATLAB-based user interface (demo: https://youtube/yQWFo33RYiQ). Each rater is to view each ML visual output and classify it. "*Others*" include those that converted back to CN from MCI or to MCI from AD. The results in Table 4 show that when using a 3-way classification, the ML model was relatively accurate with an 82 % ± 3 % accuracy, and for a 5-way classification, the accuracy dropped to 68 % ± 5 %. The achieved accuracy is consistent with state-of-the-art literature.

We observe that most stable cases were classified correctly and that the misclassified cases often were those that experienced a transition phase of the disease. From the examples shown in Fig. 3, cases (a), (b), (k), (n), and (q) clearly show that the ML visual outcome agrees with the target image. Even in cases like (c), (e), (o), and (r), although the ML outcome is slightly different from the target, they are still mostly similar, and the three raters had no problem classifying them correctly. However, for these last three cases, although the changes were minor, this could still lead to a misclassification with a strict rater or when relying solely on machine learning without the benefit of visual output. The intent here is to initiate a conversation contrasting the visual outcome of the ML model in context to all the quantitative measures known during the different time points of this longitudinal study. Through these many nuanced visual versions of the ML model in contrast to the target image, we could appreciate the difficulties typically faced in reaching high classification results, especially in multiclass classification and longitudinal studies.

It is fascinating to note from the results shown in Fig. 3 that although we exclude the neuropsychological test scores MMSE, CDR-SB, and ADAS from the training and testing phases of the ML model, these cognitive scores still show significant consistency with the outcome of the machine learning. For example, in case (c), the stable CN is shown to transition to MCI in T12 and T24 just as the CDRSB scores changed from 0 to 0.5, which indicates questionable impairment in the staging category [59], even when the MMSE score is stable at 30, which is the maximum score one can get. Case (h) is another interesting outcome of the ML model, as it shows a transition to AD in T24 due perhaps to the change of the CDR-SB score to 3 and 2.5, respectively, with a score of 3

Downloaded for Anonymous User (n/a) at Florida International University from ClinicalKey.com by Elsevier on August 04, 2023. For personal use only. No other uses without permission. Copyright ©2023. Elsevier Inc. All rights reserved.



RID	MMSE				CDRSB				RAVLT immediate				Gender	APOE	Edu (y)	AGE	AV45		FDG		TAU		ABETA		Missing Features
	T0	T6	T12	T24	T0	T6	T12	T24	T0	T6	T12	T24					T0	T24	T0	T24	T0	T24	T0	T24	T0
4491	30	30	30	30	0	0	0	0	62	29	49	59	M	0	19	84.1	0.92	0.88	1.25	1.27	165	177	1292	1519	1

(a)

RID	MMSE				CDRSB				RAVLT immediate				Gender	APOE	Edu (y)	AGE	AV45		FDG		TAU		ABETA		Missing Features
	T0	T6	T12	T24	T0	T6	T12	T24	T0	T6	T12	T24					T0	T24	T0	T24	T0	T24	T0	T24	T0
4376	29	27	27	27	0	0	0	0	53	39	41	41	F	0	16	76.5	0.93	0.95	1.34	-	277	-	2108	-	3

(b)

RID	MMSE				CDRSB				RAVLT immediate				Gender	APOE	Edu (y)	AGE	AV45		FDG		TAU		ABETA		Missing Features
	T0	T6	T12	T24	T0	T6	T12	T24	T0	T6	T12	T24					T0	T24	T0	T24	T0	T24	T0	T24	T0
4422	29	30	30	30	0.5	0	0.5	0.5	51	43	46	38	F	0	15	70.8	1.21	1.29	1.44	1.34	370	395	1373	1052	1

(c)

RID	MMSE				CDRSB				RAVLT immediate				Gender	APOE	Edu (y)	AGE	AV45		FDG		TAU		ABETA		Missing Features
	T0	T6	T12	T24	T0	T6	T12	T24	T0	T6	T12	T24					T0	T24	T0	T24	T0	T24	T0	T24	T0
4421	26	29	29	28	0	0	0	0	49	44	53	48	F	0	14	75.1	0.97	-	1.27	-	196	-	1955		1

(d)

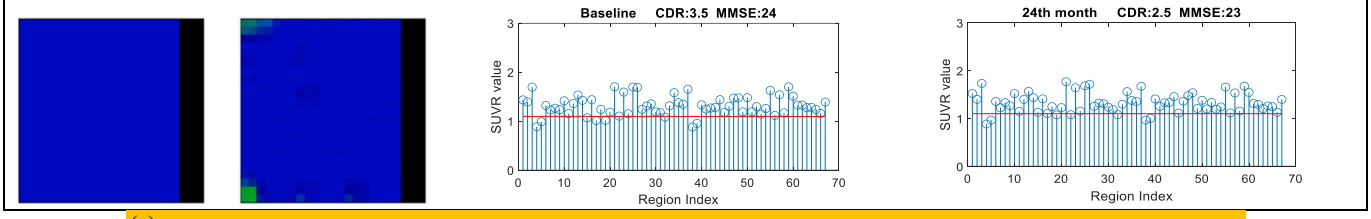
Fig. 3. Visualization of AD: The left and right images in each sub-figure are target and ML visual output for test subjects, respectively. (a) through (d) show 4 different cases of stable CN subjects; (e) through (h) 4 different cases of stable MCI subjects; cases (i) through (p) show subjects who have transitioned either from CN to MCI or from MCI to AD at different time points; cases (q) through (t) show 4 different cases of stable AD subjects. Cases (u), (v) and (w) in the last row are challenging stable cases where the ML outcome is completely different than the target.

* The patient/record (RIDs) of the shown cases of ADNI dataset are as follow: a) 4491, b) 4376, c) 4422, d) 4421, e) 4531, f) 2068, g) 4871, h) 4346, i) 4277, j) 4813, k) 2047, l) 4426, m) 4595, n) 4167, o) 4542, p) 4189, q) 4252, r) 4338, s) 4494, t) 4001, u) 4226, v) 4339, and w) 4676.

indicating very mild dementia in the staging category. Note also for this case that the MMSE dropped from 30 to 27, with mild AD defined in the 21–26 range. Another case that is hard to explain is (j), which we define as "other" in the classification categories. In this case, it seems that the MCI patient reverted to CN at T24, yet the ML model determined that this is a case of a stable CN. In such cases, where the MMSE scores, as well as the CDR-SB, are ambiguous from the diagnosis standpoint at baseline, such cases should be reviewed in context to all other inputs to

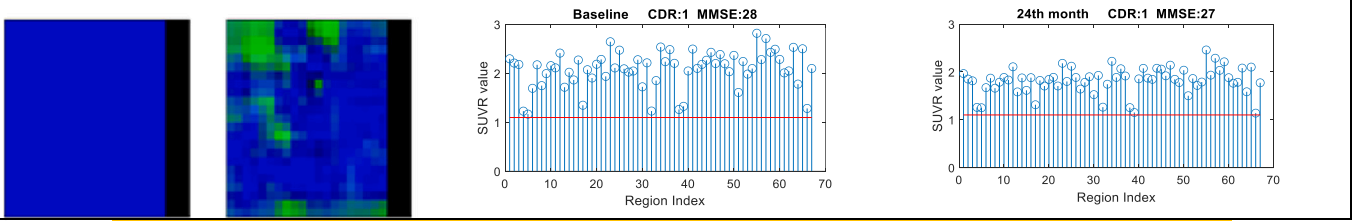
Downloaded for Anonymous User (n/a) at Florida International University from ClinicalKey.com by Elsevier on August 04, 2023. For personal use only. No other uses without permission. Copyright ©2023. Elsevier Inc. All rights reserved.

RID	MMSE				CDRSB				RAVLT immediate				Gender	APOE	Edu (y)	AGE	AV45		FDG		TAU		ABETA		Missing Features
	T0	T6	T12	T24	T0	T6	T12	T24	T0	T6	T12	T24					T0	T24	T0	T24	T0	T24	T0	T24	T0
4531	24	26	26	23	3.5	3	2	2.5	15	16	22	17	M	1	20	73.7	1.32	1.26	1.13	1.14	185	-	464		4



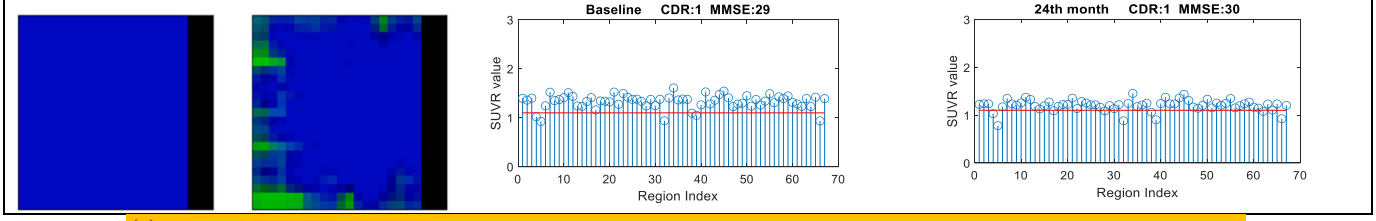
(e)

RID	MMSE				CDRSB				RAVLT immediate				Gender	APOE	Edu (y)	AGE	AV45		FDG		TAU		ABETA		Missing Features
	T0	T6	T12	T24	T0	T6	T12	T24	T0	T6	T12	T24					T0	T24	T0	T24	T0	T24	T0	T24	T0
2068	28	23	26	27	1	1	1	1	32	20	29	33	M	1	20	83.1	1.57	1.46	1.19	1.11	322	351	722	656	1



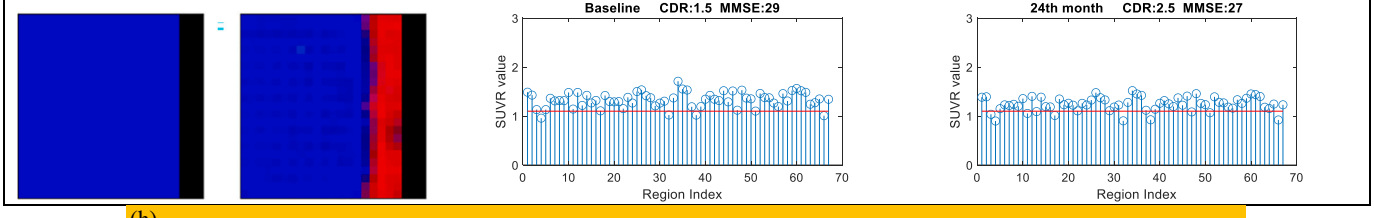
(f)

RID	MMSE				CDRSB				RAVLT immediate				Gender	APOE	Edu (y)	AGE	AV45		FDG		TAU		ABETA		Missing Features
	T0	T6	T12	T24	T0	T6	T12	T24	T0	T6	T12	T24					T0	T24	T0	T24	T0	T24	T0	T24	T0
4871	29	30	28	30	1	1.5	1.5	1	52	43	52	47	F	0	14	66.4	1.07	1.01	1.34	-	162	-	1244	-	1



(g)

RID	MMSE				CDRSB				RAVLT immediate				Gender	APOE	Edu (y)	AGE	AV45		FDG		TAU		ABETA		Missing Features
	T0	T6	T12	T24	T0	T6	T12	T24	T0	T6	T12	T24					T0	T24	T0	T24	T0	T24	T0	T24	T0
4346	29	30	29	27	1.5	2.5	3	2.5	44	39	33	37	M	2	18	71.3	1.26	1.21	1.23	1.23	288	293	664	624	1



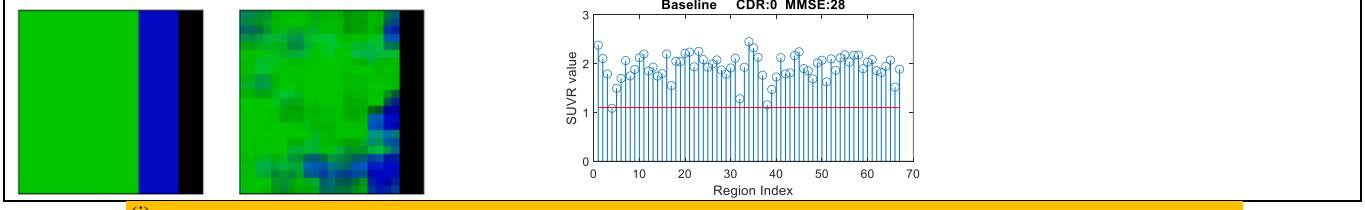
(h)

Fig. 3. (*continued*).

the ML model to look into the neuroimaging data and other cognitive scores to determine what led to this transition in the diagnosis at baseline. Case (i) is also interesting, where a stable CDR-SB of 0 scores (which means no impairment) and high MMSE scores from 28 to 30, the ML model is attempting to render visually a stable CN instead of the clear transition to MCI seen in the target image. The more complex cases of (m) and (p) may reveal that the ML model does struggle at times when the MMSE scores and CDR-SB scores vary in ways that are difficult to decipher from one phase in time to another with the target image reflecting the diagnosis at baseline may be the correct one. Cases (s) and (t) are misclassified, especially given the low MMSE scores and the high CDR-SB scores; note, however, the high number of missing values for case (t). With these examples discussed, we highlight the merits of such a visualization process where these types of contextual deliberations would not otherwise be possible if we relied solely on the ML classification algorithm without recourse to a visualized output.

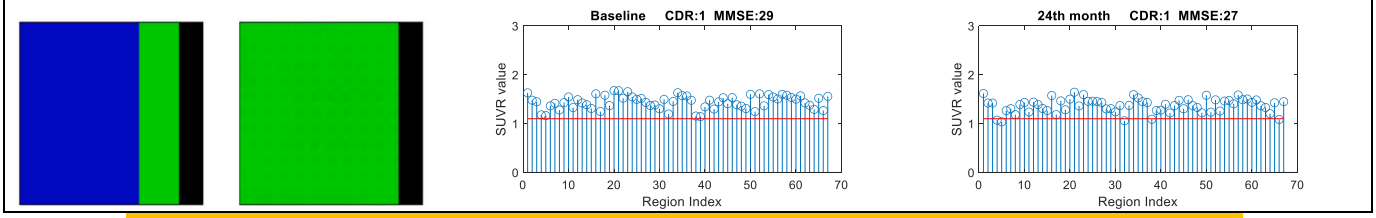
Downloaded for Anonymous User (n/a) at Florida International University from ClinicalKey.com by Elsevier on August 04, 2023. For personal use only. No other uses without permission. Copyright ©2023. Elsevier Inc. All rights reserved.

RID	MMSE				CDRSB				RAVLT immediate				Gender	APOE	Edu (y)	AGE	AV45		FDG		TAU		ABETA		Missing Features
	T0	T6	T12	T24	T0	T6	T12	T24	T0	T6	T12	T24					T0	T24	T0	T24	T0	T24	T0	T24	T0
4277	28	30	28	30	0	0	0	0	52	48	51	63	F	0	18	71.4	1.37	-	1.50	-	232	-	2717	-	2



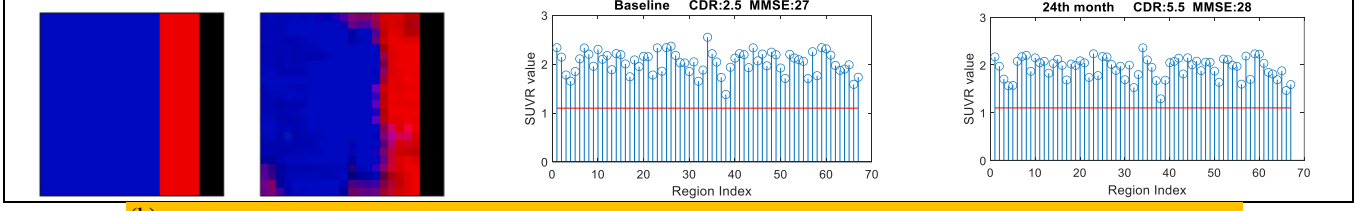
(i)

RID	MMSE				CDRSB				RAVLT immediate				Gender	APOE	Edu (y)	AGE	AV45		FDG		TAU		ABETA		Missing Features
	T0	T6	T12	T24	T0	T6	T12	T24	T0	T6	T12	T24					T0	T24	T0	T24	T0	T24	T0	T24	T0
4813	29	29	27	27	1	1	1	1	61	56	45	45	M	0	16	67.6	1.22	1.20	1.28	-	220	214	2198	2330	1



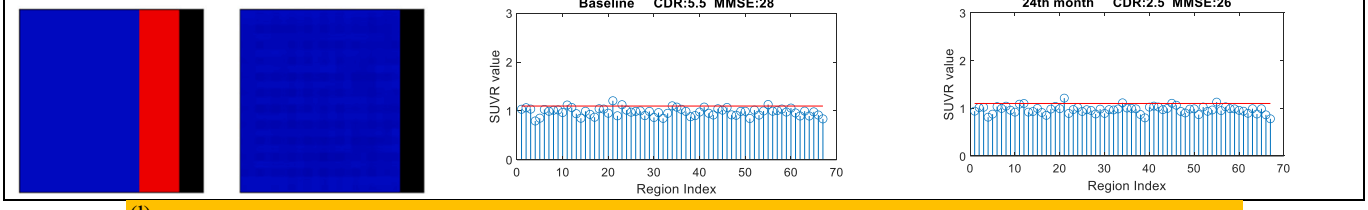
(j)

RID	MMSE				CDRSB				RAVLT immediate				Gender	APOE	Edu (y)	AGE	AV45		FDG		TAU		ABETA		Missing Features
	T0	T6	T12	T24	T0	T6	T12	T24	T0	T6	T12	T24					T0	T24	T0	T24	T0	T24	T0	T24	T0
2047	27	27	26	28	2.5	4.5	6	5.5	26	33	33	25	M	1	18	77.4	1.70	1.67	1.44	1.45	404	424	709	723	1



(k)

RID	MMSE				CDRSB				RAVLT immediate				Gender	APOE	Edu (y)	AGE	AV45		FDG		TAU		ABETA		Missing Features
	T0	T6	T12	T24	T0	T6	T12	T24	T0	T6	T12	T24					T0	T24	T0	T24	T0	T24	T0	T24	T0
4426	28	27	29	26	5.5	2	2.5	2.5	29	20	29	22	M	0	16	77.2	0.92	0.89	1.37	1.31	188	207	819	840	1



(l)

Fig. 3. (*continued*).

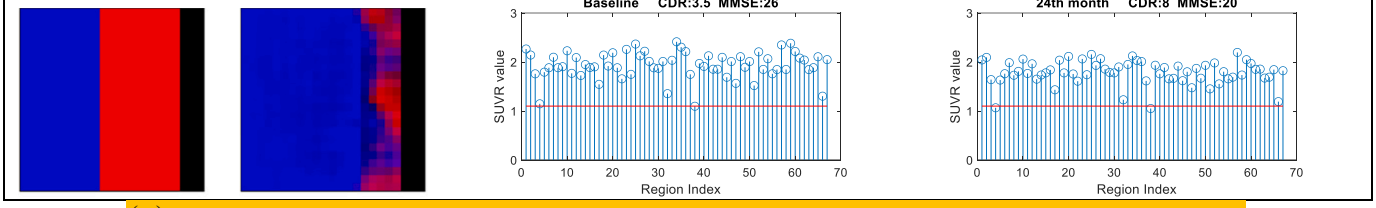
By having recourse to a visual outcome, we could reassess challenging cases to determine what could have led to such an ML outcome and whether there is more reason to assert a misclassification or instead accept the ML outcome as the more appropriate diagnosis. When reviewing these challenging cases, as illustrated in Fig. 3, recall that the target image is on the left, and the ML visual outcome is on the right. Furthermore, when deliberating on which outcome could be more telling or more convincing, review the provided MMSE and CDR-SB scores as well as all other features provided in the figure as context. Remember that MMSE, CDR-SB, and ADAS were excluded from consideration in the training and testing phases when we designed this ML model.

3.1. Comparison to other methods

Since the classification results of our proposed method rely on an agreement reached between the three raters looking at the visual outcomes of the machine learning independently, it is not straightforward

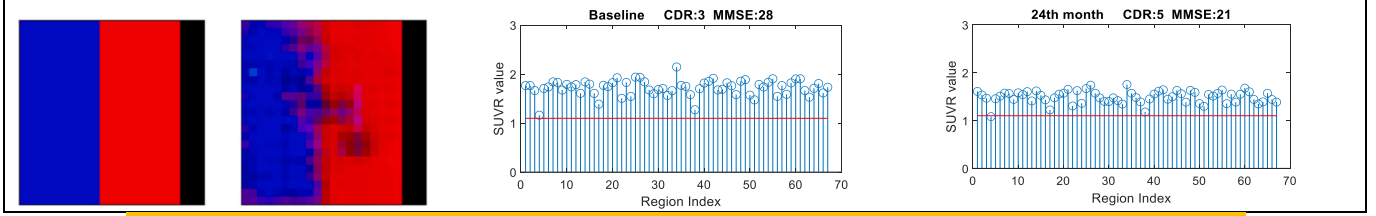
Downloaded for Anonymous User (n/a) at Florida International University from ClinicalKey.com by Elsevier on August 04, 2023. For personal use only. No other uses without permission. Copyright ©2023. Elsevier Inc. All rights reserved.

RID	MMSE				CDRSB				RAVLT immediate				Gender	APOE	Edu (y)	AGE	AV45		FDG		TAU		ABETA		Missing Features
	T0	T6	T12	T24	T0	T6	T12	T24	T0	T6	T12	T24					T0	T24	T0	T24	T0	T24	T0	T24	T0
4595	26	27	24	20	3.5	2.5	4.5	8	26	14	28	10	M	0	16	76.6	1.69	1.59	0.94	1.02	480	621	957	781	1



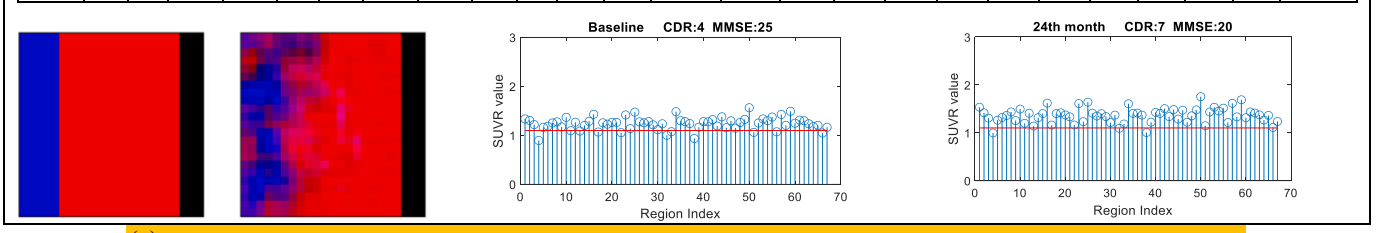
(m)

RID	MMSE				CDRSB				RAVLT immediate				Gender	APOE	Edu (y)	AGE	AV45		FDG		TAU		ABETA		Missing Features
	T0	T6	T12	T24	T0	T6	T12	T24	T0	T6	T12	T24					T0	T24	T0	T24	T0	T24	T0	T24	T0
4167	28	25	23	21	3	3	3	5	24	23	23	24	F	1	16	69.5	1.54	1.35	0.93	0.84	301	281	645	519	2



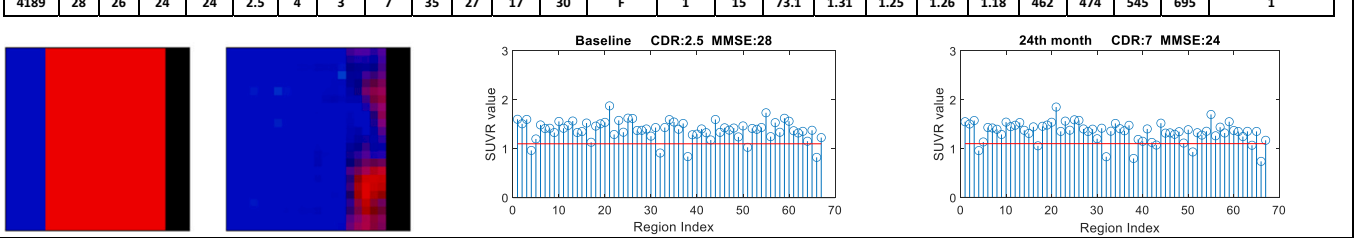
(n)

RID	MMSE				CDRSB				RAVLT immediate				Gender	APOE	Edu (y)	AGE	AV45		FDG		TAU		ABETA		Missing Features
	T0	T6	T12	T24	T0	T6	T12	T24	T0	T6	T12	T24					T0	T24	T0	T24	T0	T24	T0	T24	T0
4542	25	23	21	20	4	4.5	5	7	25	11	21	17	F	2	16	79.3	1.24	1.33	1.15	1.00	380	409	479	256	1



(o)

RID	MMSE				CDRSB				RAVLT immediate				Gender	APOE	Edu (y)	AGE	AV45		FDG		TAU		ABETA		Missing Features
	T0	T6	T12	T24	T0	T6	T12	T24	T0	T6	T12	T24					T0	T24	T0	T24	T0	T24	T0	T24	T0
4189	28	26	24	24	2.5	4	3	7	35	27	17	30	F	1	15	73.1	1.31	1.25	1.26	1.18	462	474	545	695	1



(p)

Fig. 3. (*continued*).

to compare our results with other methods quantitively. But for a fair comparison with the proposed method, we review below only the results of other studies that relied on multiclass classification methods that involved at least 200 subjects from the ADNI data set. Liu et al. (2018) in [54] considered the baseline ADNI-1 dataset, which contained 181 AD subjects, 226 control normal (CN), 165 progressive (or converter) MCI (pMCI), and 225 stable MCI or non-converters (sMCI) subjects; and in the baseline ADNI-2 dataset, there were 143 AD, 185 NC, 37 pMCI, and 234 sMCI subjects. By using a CNN model for joint regression and classification tasks, they refer to as a deep multitask multichannel learning (DM2L) framework; they reached an accuracy of 51.8 % in a four-way (CN, sMCI, pMCI, AD) classification process (this 4-way classification is similar to our 5-way results due to they removed the *others* cases). Another study by Zhu et al. (2016a) [55] considered 202 subjects using baseline MRI and PET images, which included 51 AD subjects, 52 Normal Control (NC) subjects, and 99 MCI subjects. Of the 99 MCI, 43

Downloaded for Anonymous User (n/a) at Florida International University from ClinicalKey.com by Elsevier on August 04, 2023. For personal use only. No other uses without permission. Copyright ©2023. Elsevier Inc. All rights reserved.

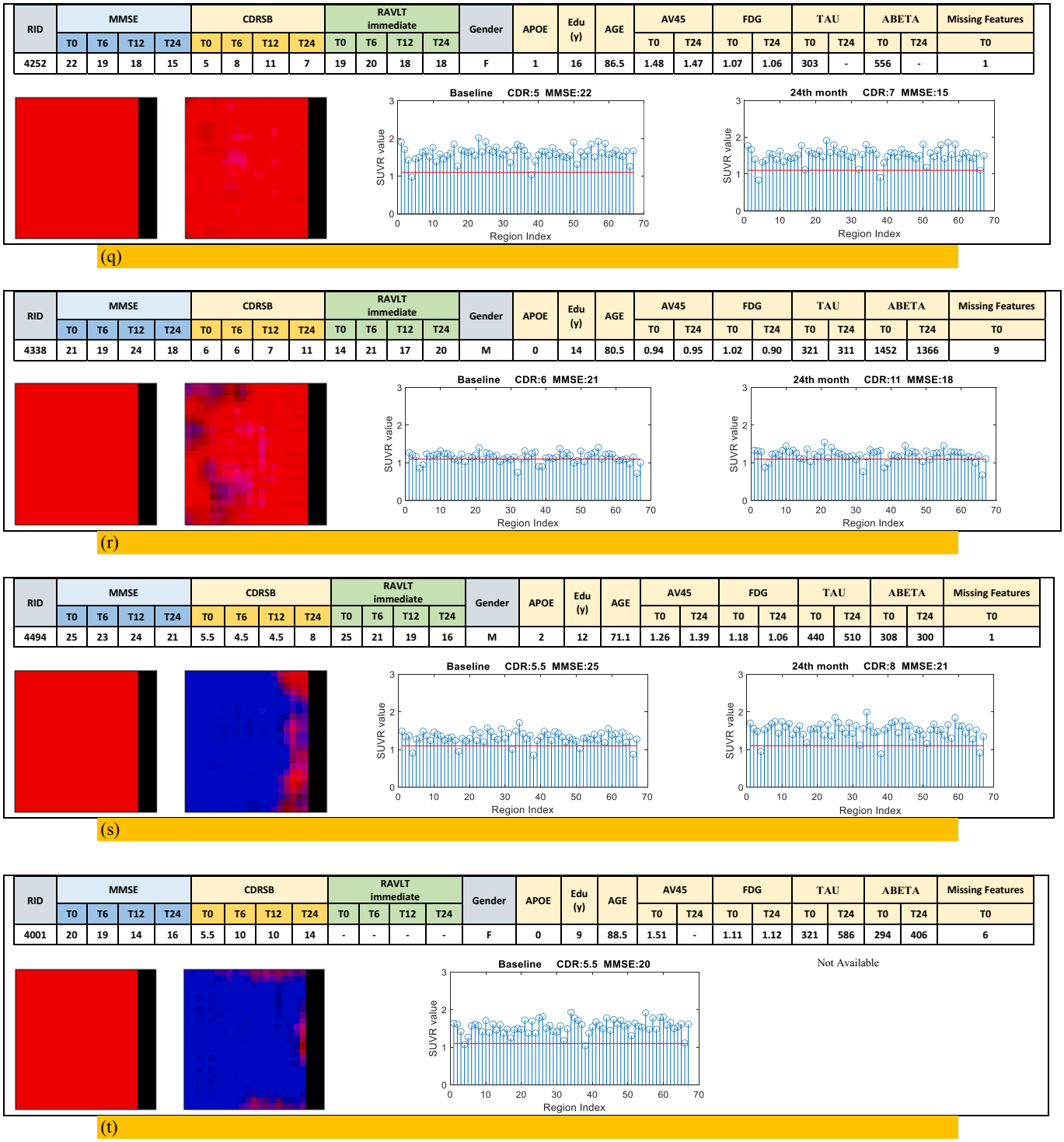


Fig. 3. (*continued*).

were converters (MCI-C), and 56 were non-converters (MCI-NC). Their 4-way (CN, MCI-C, MCI-NC, AD) yielded just over 61 % accuracy. Another study by Shi et al. (2018) [56] developed a method to perform both tasks of binary and multiclass classification on the same 202 subjects used in Zhu et al. (52 CN, 43 MCI-C, 56 MCI-NC, and 51 AD), where they implement a two-stage stacked deep polynomial network, obtaining an accuracy of 53.65 % in multiclass classification with higher accuracies obtained as expected for binary classification. Lin et al. in [57] performed a multiclass classification on 746 subjects (200 NC, 441 MCI, and 105 AD subjects), with 110 of the 441 MCI subjects converting to AD at future time points in the three-year longitudinal study. These subjects of the 1800 subjects had all the measures the authors considered (MRI, PET, cerebrospinal fluid (CSF), and some genetic features). Their multiclass results based on a linear discriminant analysis (LDA) scoring method to fuse the multimodal data yielded an accuracy of 66.7 % for a three-way (CN, MCI, and AD) classification and a lower 57.3 % for a four-way classification with the MCI converters separated from the stable MCIs. Moreover, in earlier studies by our research group, a study by Fang et al. [58] considered 906 subjects (251 CN, 297 EMCI, 196 late MCI (LMCI), and 162 AD) subjects from the ADNI dataset, using the

Downloaded for Anonymous User (n/a) at Florida International University from ClinicalKey.com by Elsevier on August 04, 2023. For personal use only. No other uses without permission. Copyright ©2023. Elsevier Inc. All rights reserved.

Fig. 3. (*continued*).

Table 4
Classification outcomes as assessed by three raters.

Classification type	Correctly classified	Misclassified outcomes	Inconclusive outcomes
3-Way (CN, impaired, others)	0.82 ± 0.03	0.15 ± 0.004	0.023 ± 0.002
5-way (CN, MCI, MCIc, AD, others)	0.68 ± 0.05	0.29 ± 0.01	0.023 ± 0.002

neuroimaging modalities of MRI and PET. A 4-way (CN, EMCI, LMCI, AD) multiclass classification, using a Gaussian discriminative component analysis in a supervised dimensionality reduction algorithm, resulted in an accuracy of 67.69 %. In another study by Tabarestani et al. [26], 1117 subjects were considered (328CN, 441 MCI-NC, 191 MCI-C, 157 AD), using kernelized and tensorized multitask network (KTMnet) for both prediction and multiclass classification. Combining features from PET, MRI, CSF, cognitive scores, and other risk factors that included age, gender, education, and the APOE gene, a 4-way (CN, MCI-C, MCI-NC, AD) resulted in a classification accuracy of 66.85 %.

3.2. Extending the 2D visualization platform to 3D

The design of the proposed ML model can display all these results in 3D as well, as shown in Fig. 4. For 3D visualization, the L component of L-*a*-*b* format, a 3D variation of the CIE Chromaticity diagram, can be used to display in 3D the RGB format without changing the contextual meaning of the outcomes reflected in the examples considered in Figs. 4 and 5. In this L-*a*-*b* format, L refers to lightness normalized from zero to 1, and *a* and *b* reflect the colors from green to red for *a* and from blue to yellow for *b*. Fig. 4a and b show the target and ML output images, Fig. 4c and d illustrate the blue and red channels, respectively, and Fig. 4e through h provide the 3D displays of (a) through (d). Note the gradual change in the ML-generated visual outcomes. Observe that at T24 (24th month), the ML visual outcome in 4f stabilizes at the highest levels near the normalized value of 1. Moreover, observe that as the blue channel reflecting the MCI state declines rapidly between T12 and T24, the red channel in 4 h reflecting the AD state increases between T12 through T24 to stabilize at the maximum value of 1. Note how easy it is to ascertain the effect of the ML model has on the region of uncertainty in displays (f), (g), and (h). For the visual appreciation of this 3D display model, we provide four different cases (a), (b), (h), and (u) of Fig. 3 displayed in 3D in Fig. 5.

4. Discussion

The results of ML4VisAD's implementation show the need for deep reflection when assessing multiclass classification or prediction results using machine learning, especially when observing all the subtle nuances of the visual outcome. There were a few cases where the ML4VisAD visual output seemed to make more sense than what the target

Downloaded for Anonymous User (n/a) at Florida International University from ClinicalKey.com by Elsevier on August 04, 2023. For personal use only. No other uses without permission. Copyright ©2023. Elsevier Inc. All rights reserved.

Table 5-
Brain regions for the SUVRs shown in Fig. 3.

SUVR regions considered		
1) LH_CAUDALANTERIORCINGULATE	24) LH_PRECENTRAL	47) RH_LINGUAL
2) LH_CAUDALMIDDLEFRONTAL	25) LH_PRECUNEUS	48) RH_MEDIALORBITOFRONTAL
3) LH_CUNEUS	26) LH_ROSTRALANTERIORCINGULATE	49) RH_MIDDLETEMPORAL
4) LH_ENTORHINAL	27) LH_ROSTRALMIDDLEFRONTAL	50) RH_PARACENTRAL
5) LH_FRONTALPOLE	28) LH_SUPERIORFRONTAL	51) RH_PARAHIPPOCAMPAL
6) LH_FUSIFORM	29) LH_SUPERIORPARIETAL	52) RH_PARSOPERCULARIS
7) LH_INFERIORPARIETAL	30) LH_SUPERIORTEMPORAL	53) RH_PARSORBITALIS
8) LH_INFERIORTEMPORAL	31) LH_SUPRAMARGINAL	54) RH_PARSTRIANGULARIS
9) LH_INSULA	32) LH_TEMPORALPOLE	55) RH_PERICALCARINE
10) LH_ISTHMUSCINGULATE	33) LH_TRANSVERSETEMPORAL	56) RH_POSTCENTRAL
11) LH_LATERALOCCIPITAL	34) RH_BANKSSTS	57) RH_POSTERIORCINGULATE
12) LH_LATERALORBITOFRONTAL	35) RH_CAUDALANTERIORCINGULATE	58) RH_PRECENTRAL
13) LH_LINGUAL	36) RH_CAUDALMIDDLEFRONTAL	59) RH_PRECUNEUS
14) LH_MEDIALORBITOFRONTAL	37) RH_CUNEUS	60) RH_ROSTRALANTERIORCINGULATE
15) LH_MIDDLETEMPORAL	38) RH_ENTORHINAL	61) RH_ROSTRALMIDDLEFRONTAL
16) LH_PARACENTRAL	39) RH_FRONTALPOLE	62) RH_SUPERIORFRONTAL
17) LH_PARAHIPPOCAMPAL	40) RH_FUSIFORM	63) RH_SUPERIORPARIETAL
18) LH_PARSOPERCULARIS	41) RH_INFERIORPARIETAL	64) RH_SUPERIORTEMPORAL
19) LH_PARSORBITALIS	42) RH_INFERIORTEMPORAL	65) RH_SUPRAMARGINAL
20) LH_PARSTRIANGULARIS	43) RH_INSULA	66) RH_TEMPORALPOLE
21) LH_PERICALCARINE	44) RH_ISTHMUSCINGULATE	67) RH_TRANSVERSETEMPORAL
22) LH_POSTCENTRAL	45) RH_LATERALOCCIPITAL	
23) LH_POSTERIORCINGULATE	46) RH_LATERALORBITOFRONTAL	

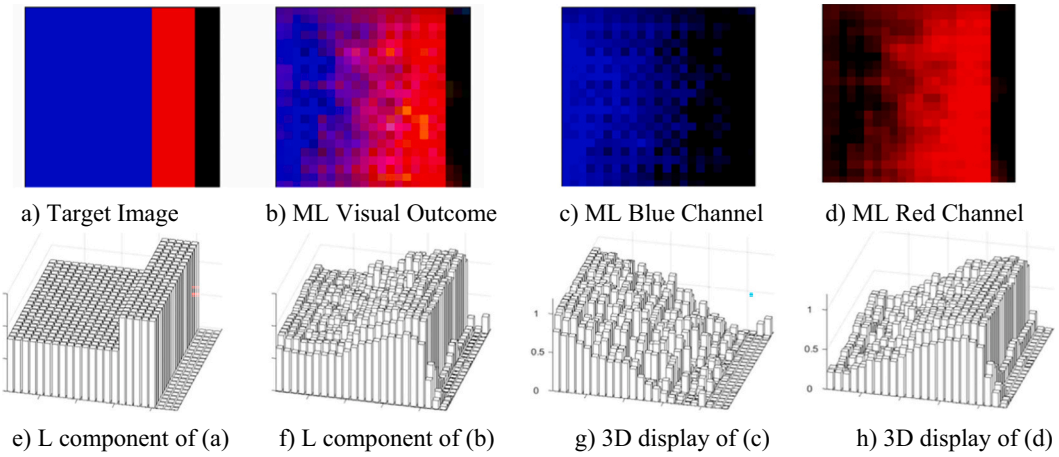


Fig. 4. 3D Display of the RGB channels of an MCI case that transitioned to AD at T24(24th month). Note the gradual change in the ML generated displays. Also note how minimally the ML model affected the region of uncertainty (RU) in the 3D displays in f, g and h.

images portrayed, especially concerning the available measurements at the different time points. Case (j) shows a subject that transitioned back to normal control (CN) from an MCI diagnosis in the previous three time points. The ML model did not see it the same way and had the subject as stable CN through all four time points, and most measurements support this classification. Another example is case (l), where the target shows a transition from MCI to AD in T24, while the ML4VisAD visual output displays a stable MCI through all time points. Here again, the measurements are somewhat ambiguous but more in favor of the ML model in that the MMSE did drop but only by one point in T24 compared to T6, and the CDR-SB scores are otherwise consistent through T6 to T24 with the SUVR also consistent in T0 and T24. Another interesting case is (v), where the target image shows a stable MCI, while the ML4VisAD visual output places this subject as stable CN. In this case, from the high MMSE score, the low SUVR values, and an APOE of 0, although the CDR is 0.5, the ML visual outcome of a stable CN seems more reasonable. But other cognitive tests (ADAS, RAVLT) may have influenced the diagnosis, and these scores were not used in the ML4VisAD model to avoid bias. In many of these cases, there may be some merits in generating a composite score, as proposed in [46]. Moreover, for cases (u), (v), and (w), all stable cases misclassified as another type of stable cases, there seems to be an influence of the APOE value on the ML4VisAD outcome (0 influences the CN state, 2 switched CN to MCI, and 1 reverted AD to MCI). See also findings reported in [53].

These ML visual outcomes clearly show why clinicians face difficulty each time they deliberate on a patient's disease state. For example, it is hard to understand why the subject in case (u) in Fig. 3 had an MMSE score of 29 for T0, T12, and T24 but an MMSE score of 24 at time T6. Also, for the same patient in (u) the CDR score was 1 at T0 and reverted to 0 for all subsequent time points. Although the diagnosis is that of a stable CN for (u), the machine learning visual outcome places this subject as stable MCI when considering all other features. Recall that the APOE for (u) is 2 at baseline and that the SUVRs are relatively high. Also, the high number of years of education for this subject (17) may have led to the high MMSE scores of 29 for T0, T12, and T24, although stumbling in the test given at T6.

The subtle nuances encountered with the ML4VisAD visual outcomes could reduce the misclassifications with added scrutiny on the visual output in context to specific measurements clinicians may be interested in. Consequently, the first point is that multiclass classification, whether it is automated or made through a rating process, does not allow for a more thorough deliberation process if these nuances and subtle

Downloaded for Anonymous User (n/a) at Florida International University from ClinicalKey.com by Elsevier on August 04, 2023. For personal use only. No other uses without permission. Copyright ©2023. Elsevier Inc. All rights reserved.

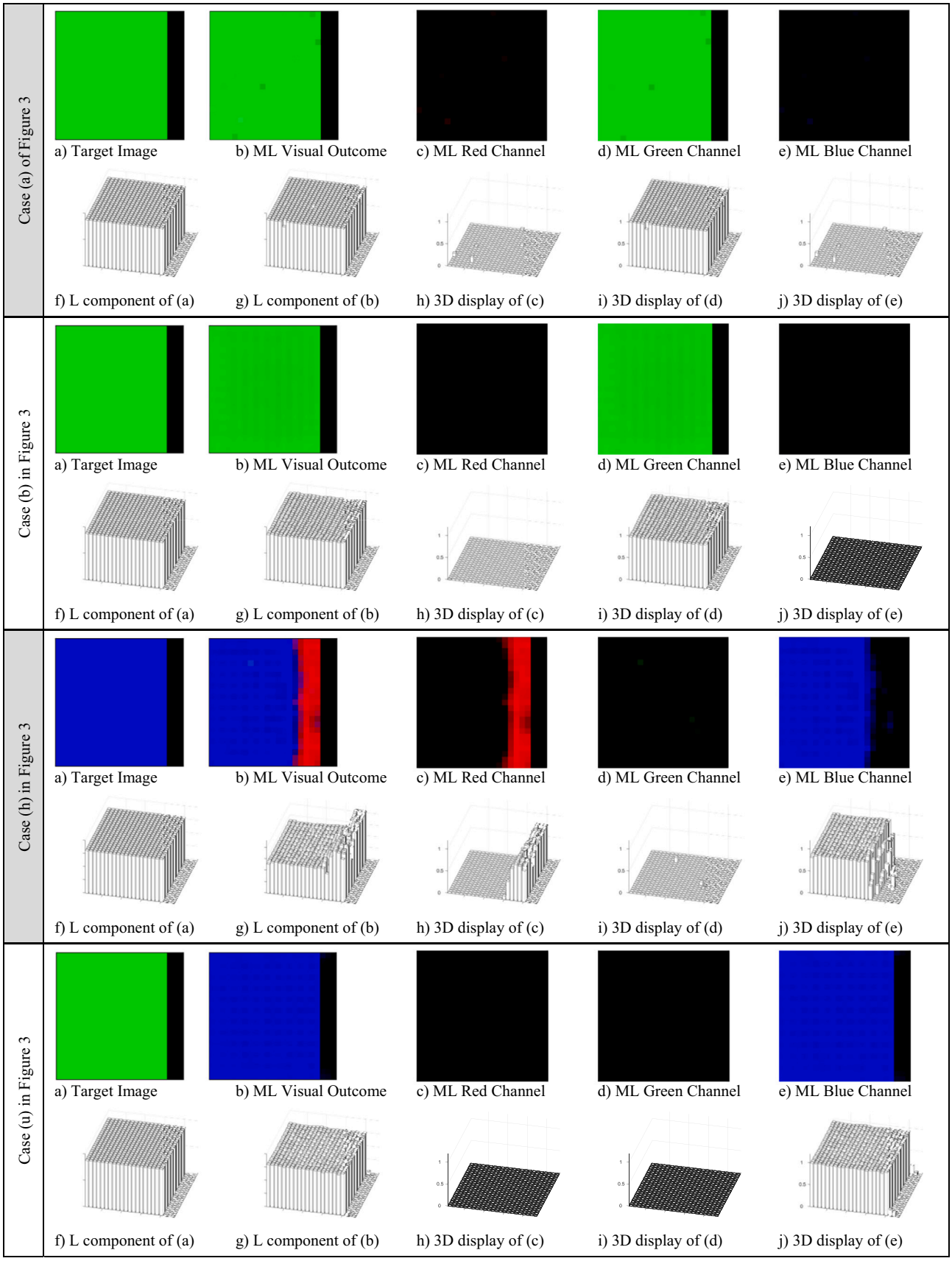


Fig. 5. 3D Display of the RGB channels of cases (a), (b), (h) and (u) from Fig. 3. Also note the minimal effects on the region of uncertainty (RU).

Downloaded for Anonymous User (n/a) at Florida International University from ClinicalKey.com by Elsevier on August 04, 2023. For personal use only. No other uses without permission. Copyright ©2023. Elsevier Inc. All rights reserved.

differences cannot be visually observed and would be so hard to decipher otherwise through tabulated data or decisional spaces showing different overlapped regions among the considered classes. Therefore, it is no revelation that the more classes considered in a multiclass classification algorithm, the less accurate will the classification results be.

4.1. Future work

The proposed visual outcome enhances the processes of multiclass classification and disease projection with the ability to visualize the inner workings of the machine learning and observe what the differences between the ML visual outcome and target image could mean. In other words, the difference between them does not necessarily mean an outright misclassification but emphasizes the nuances between them and implies that a review is necessary of what may have led to such change, especially if the region of uncertainty (RU) in the visual display remains unaffected.

It is thus essential to recognize that the interrelatedness in features, along with the many variations of such multimodal features, some being temporal, others structural, functional, metabolic, genetic, demographic, or cognitive are extremely difficult to disentangle, especially when combined with subjective thresholds or ranges of scores such as with SUVRs, MMSE, and CDR-SB. When considering ADNI data, there is an overlap in MMSE scores between CN, MCI, and even AD groups, and the CDR-SB values may resolve this overlap. Still, for an ML model, more datasets are required to learn more of the interplay between such cognitive features, especially when used for baseline diagnosis.

We contend that it is possible to define some objective criteria to quantify the uncertainty of the machine's estimation per case/patient, which is one of the significant open problems for utilizing AI in medicine. As a good first step, we included in our visual template a black bar that evaluates the amount of uncertainty infused by the machine learning model into the classification/prediction results. But we believe further investigation is needed to better understand this effect. For instance, we could investigate the checkerboard effects observed in cases (b) and (j) to determine what led to their presence. Are these effects due to the convolutions and other calculations performed by the ML model, or are they an indication of some subtle changes in the feature space of that specific patient that were not seen in the training phase?

As for the number of classes considered in the study, the proposed method relied on the three primary RGB colors for the three groups (CN, MCI, AD) available in the dataset. However, suppose additional classes such as EMCI, LMCI, or aMCI are also available. In that case, we could always augment the primary color with the secondary colors of yellow, cyan, and magenta (Y, C, M) on the visual template.

As it stands, from the availability of data, there were nearly four times more MCIs than AD and twice as many MCIs than CNs. Since the input features fed into the ML model were those acquired at baseline, a balance of samples between CN, MCI, and AD groups would be ideal in future implementations. Moreover, although ML4VisAD utilized 1123 subjects, its efficacy could be enhanced by the availability of a much larger and more balanced dataset if the ML model in the training phase is to capture all the nuances that distinguish the different subgroups. ADRC centers and ADNI, who continue to build a much larger population of the various subgroups for research with balanced data regarding ethnicity and disease state, are crucial to future experimentation.

5. Conclusion

The genesis of this study was to create a new approach for the visualization-based estimation of disease trajectory to augment the diagnosis and prognosis of AD. A new deep learning (DL) architecture based on Convolutional Neural Networks generates a visual image that portrays AD trajectory in a 2-year longitudinal study using baseline features only. From these baseline features, to avoid bias, we remove all cognitive scores MMSE, CDR-SB, and ADAS from consideration in the design of the ML model as the input features. Target images using different colors define each stage of the disease at the four observation time points (T0, T6, T12, and T24), with T0 being the baseline time point. A unique characteristic of this model is that it is trained with known target images with color-coded diagnoses at all four time points to generate a visual output that predicts disease trajectory based on baseline features only. Since we use only baseline features as input, this design is amenable to cross-sectional and longitudinal studies based on similar datasets. This research could also lead to new insights as to the gradual changes between transition phases as a function of the input feature space considered.

Funding

This research is supported by the National Science Foundation (NSF) under grants CNS-1920182, CNS-1532061, CNS-2018611, and National Institutes of Health under grants NIA/NIH 5R01AG061106-02, NIA/NIH 5R01AG047649-05, and the NIA/NIH 1P30AG066506-01 with the 1Florida ADRC.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Malek Adjouadi, Mohammad Eslami, Solale Tabarestani reports financial support was provided by National Institutes of Health. Malek Adjouadi, Mohammad Eslami, Solale Tabarestani reports financial support was provided by National Science Foundation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.artmed.2023.102543.

References

[1] Lynch M, Alzheimer's Association. New Alzheimer's Association report reveals sharp increases in Alzheimer's prevalence, deaths and cost of care. report. Alzheimer's Association; 2018. https://www.alz.org/news/2018/new_alzheimer_s_association_report_reveals_sharp_i.
[2] Einav L, Finkelstein A, Mullainathan S, Obermeyer Z. Predictive modeling of U.S. health care spending in late life. Science June 2018;360(6396):1462–5.
[3] Scheltens P, Blennow K, Breteler MMB, de Strooper B, Frisoni GB, Salloway S, Van der Flier WM. Alzheimer's disease. Lancet July 2016;388(10043):505–17.
[4] Young PNE, Estarellas M, Coomans E. Imaging biomarkers in neurodegeneration: current and future practices. Alzheimers Res Ther April 27, 2020;12(1):49.
[5] Jack CR, CR, Knopman DS, Jagust WJ, Petersen RC, Weiner MW, Aisen PS, Shaw LM, Vemuri P. Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. Lancet Neurol Feb 2013;12(2):207–16.
[6] Loewenstein DA, Curiel RE, DeKosky S, Bauer RM, Rosselli M, Guinjoan SM, Adjouadi M, Peñate MA, Barker WW, Goenaga S, Golde T, Greig-Custo MT, Hanson KS, Li C, Lizarraga G, Marsiske M, Duara R. Utilizing semantic intrusions to identify amyloid positivity in mild cognitive impairment. Neurology September 2018;27(11):E976–84.
[7] Selkoe DJ. Early network dysfunction in Alzheimer's disease. Science August 2019;365(6453):540–1.
[8] Jessen F, Amariglio RE, Buckley RF, et al. The characterisation of subjective cognitive decline. Lancet Neurol March 2020;19(3):271–8.
[9] van Loenhoud AC, van der Flier WM, Wink AM, Dicks E, Groot C, Twisk J, Barkhof F, Scheltens P, Ossenkoppele R. For the Alzheimer's disease neuroimaging initiative, "Cognitive reserve and clinical progression in Alzheimer disease: a paradoxical relationship". Neurology 2019 Jul 23;93(4):e334–46.
[10] Nortley R, Korte N, Izquierdo P, Hirunpattarasilp C, Mishra A, Jaunmuktane Z, Kyrargyri V, et al. Amyloid beta oligomers constrict human capillaries in Alzheimer's disease via signaling to pericytes. Science July 2019;365(6450). 11 pages.
[11] Scholl M, Maass A. Does early cognitive decline require the presence of both tau and amyloid-beta? Brain January 2020;143(Part: 1):10–3.
[12] Lockhart SN, Scholl M, Baker SL, et al. Amyloid and tau PET demonstrate region-specific associations in normal older people. Neuroimage April 2017;150:191–9.
[13] Montagne A, Nation DA, Sagare AP, Barisano G, Sweeney MD, Chakhoyan A, Pachicano M, Joe E, Nelson AR, D'Orazio LM. APOE4 leads to blood-brain barrier dysfunction predicting cognitive decline. Nature April 2020;581:71–6.

Downloaded for Anonymous User (n/a) at Florida International University from ClinicalKey.com by Elsevier on August 04, 2023. For personal use only. No other uses without permission. Copyright ©2023. Elsevier Inc. All rights reserved.

[14] Li T, Wang B, Gao Y, Wang X, Yan T, Xiang J. APOE epsilon 4 and cognitive reserve effects on the functional network in the Alzheimer's disease spectrum. Brain Imaging Behav 2020. https://doi.org/10.1007/s11682-020-00283-w. Early Access.
[15] Therriault J, Joseph, Benedet AL, Pascoal TA, Mathotaarachchi S, Chamoun M, Savard M, Tharick A. Association of apolipoprotein E epsilon 4 with medial temporal tau independent of amyloid-beta. JAMA Neurol. April 2020;77(4):470–9.
[16] Olsson B, Lautner R, Andreasson U. CSF and blood biomarkers for the diagnosis of Alzheimer's disease: a systematic review and meta-analysis. Lancet Neurol. June 2016;15:673–84.
[17] Lewczuk P, Matzen A, Blennow K, Parnetti L, Molinuevo JL, Eusebi P, et al. Cerebrospinal fluid A beta(42/40) corresponds better than A beta(42) to amyloid PET in Alzheimer's disease. J Alzheimers Dis 2017;55(2):813–22.
[18] Sabbagh M, Sadowsky C, Tousi B, Agronin ME, Alva G, Armon C, Bernick C, Keegan AP, Karantzoulis S, Baror E, et al. Effects of a combined transcranial magnetic stimulation (TMS) and cognitive training intervention in patients with Alzheimer's disease. Alzheimers Dement April 2020;16(4):641–50.
[19] Romanella M, Roe D, Paciorek R, Cappon D, Ruffini G, Menardi A, Rossi A, Rossi S, Santarnecchi E. Sleep, noninvasive brain stimulation, and the aging brain: challenges and opportunities. Ageing Res Rev August 2020;61:101067.
[20] Tăuțan AM, Ionescu B, Santarnecchi E. Artificial intelligence in neurodegenerative diseases: a review of available tools with a focus on machine learning techniques. Artif Intell Med 2021;117:102081.
[21] Khojaste-Sarakhsi M, Haghighi SS, Ghomi SF, Marchiori E. Deep learning for Alzheimer's disease diagnosis: a survey. Artif Intell Med 2022;130:102332.
[22] Pellegrini E, Ballerini L, Valdes Hernandez MC, Chappell FM, González-Castro V, Anblagan D, et al. Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: a systematic review. Alzheimers Dement 2018;10:519–35.
[23] Ezzati A, Zammit AR, Harvey DJ, et al. Optimizing machine learning methods to improve predictive models of Alzheimer's disease. J Alzheimers Dis 2019;71(3): 1027–36.
[24] Tabarestani S, Aghili M, Eslami M, Cabrerizo M, Barreto A, Rishe N, Curiel RE, Loewenstein D, Duara R, Adjouadi M. A distributed multitask multimodal approach for the prediction of Alzheimer's disease in a longitudinal study. Neuroimage February 2020;206:116317.
[25] Donini M, Monteiro JM, Pontil M, Hahn T, Fallgatter AJ, Shawe-Taylor J, Mourão-Miranda J. Combining heterogeneous data sources for neuroimaging based diagnosis: re-weighting and selecting what is important. Neuroimage July 2019; 195:215–31.
[26] Tabarestani S, Eslami M, Cabrerizo M, Barreto A, Rishe N, Curiel RE, Barreto A, Rishe N, Vaillancourt D, DeKosky ST, Loewenstein DA, Duara R, Adjouadi M. A tensorized multitask deep learning network for progression prediction of Alzheimer's Disease. Front Aging Neurosci May 2022;14:810873. https://doi.org/10.3389/fnagi.2022.810873. PMID: 35601611.
[27] Shojaie M, Tabarestani S, Cabrerizo M, DeKosky ST, Vaillancourt DE, Loewenstein D, Duara R, Adjouadi M. PET imaging of tau pathology and amyloid-β, and MRI for Alzheimer's disease feature fusion and multimodal classification. J Alzheimers Dis 2021;84(4):1497–514. https://doi.org/10.3233/JAD-210064. PMID 34719488.
[28] Ranjbarzadeh R, Caputo A, Tirkolaee EB, Ghoushchi SJ, Bendechache M. Brain tumor segmentation of MRI images: a comprehensive review on the application of artificial intelligence tools. Comput Biol Med January 2023;152:106405.
[29] Aghamohammadi A, Ranjbarzadeh R, Naiemi F, Mogharrebi M, Dorosti S, Bendechache M. TPCNN: two-path convolutional neural network for tumor and liver segmentation in CT images using a novel encoding approach. Expert Syst Appl November 2021;183(30):115406.
[30] Kazeminia S, Baur C, Kuijper A, van Ginneken B, Navab N, Albarqouni S, Mukhopadhyay A. GANs for medical image analysis. Artif Intell Med 2020. https://doi.org/10.1016/j.artmed.2020.101938, Aug.
[31] Bruno P, Calimeri F, Kitanidis AS, Momi EDe. Data reduction and data visualization for automatic diagnosis using gene expression and clinical data. Artif Intell Med JUL 2020;107:101884. https://doi.org/10.1016/j.artmed.2020.101884.
[32] Lizarraga G, Li C, Cabrerizo M, Barker W, Loewenstein DA, Duara R, Adjouadi M. A neuroimaging web services interface as a cyber physical system for medical imaging and data management in brain research: design study. JMIR Med Inform Apr-Jun 2018;6(2):228–44.
[33] Yuan J, Sartor EA, Au R, Kolachalama VB. Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. Brain 2020;143(6):1920–33. https://doi.org/10.1093/brain/awaa137. June.
[34] Li Qi. Overview of data visualization. Embodying Data June 2020:17–47. https://doi.org/10.1007/978-981-15-5069-0_2. Published online 2020 Jun 20.
[35] Seo K, Pan R, Lee D, Thiyyagura P, Chen K. Visualizing Alzheimer's disease progression in low dimensional manifolds August 2019;5(8):e02216.
[36] Blanken AE, Jang JY, Ho JK, Edmonds EC, Han SD, Bangen KJ, Nation DA. Distilling heterogeneity of mild cognitive impairment in the National Alzheimer Coordinating Center Database Using Latent Profile Analysis. JAMA Netw Open March 2020;3(3).
[37] Liu X, Tosun D, Weiner MW, Schuff N. Locally linear embedding (LLE) for MRI based Alzheimer's disease classification. Neuroimage December 2013;83:148–57.
[38] Gerber S, Tasdizen T, Fletcher PT, Joshi S, Whitaker R. Manifold modeling for brain population analysis. Med Image Anal October 2010;14(5):643–53.
[39] Berron D, van Westen D, Ossenkoppele R, Strandberg O, Hansson O. Medial temporal lobe connectivity and its associations with cognition in early Alzheimer's disease. Brain April 2020;143(4):1233–48.
[40] Montez T, Simon-Shlomo Poil BF, Jones I, Manshanden JPA, van Verbunt BW, Dijk AB, van Brussaard A, Ooyen CJ, Stam P, Scheltens KLinkenkaer-Hansen. Altered temporal correlations in parietal alpha and prefrontal theta oscillations in early-stage alzheimer disease. Proc Natl Acad Sci U S A February 2009;106(5): 1614–9.
[41] Buckley RF, Schultz AP, Hedden T, Papp KV, Hanseeuw BJ, Marshall G, Sepulcre J, Smith EE, et al. Functional network integrity presages cognitive decline in preclinical alzheimer disease. Neurology July 2017;89(1):29–37.
[42] Wisch JK, Roe CM, Babulal GM, Schindler SuE, Fagan AM, Benzinger TL, Morris JC, Ances BM. Resting state functional connectivity signature differentiates cognitively Normal from individuals who convert to symptomatic alzheimer disease. J Alzheimers Dis 2020;74(4):1085–95.
[43] Toddenroth D, Ganslandt T, Castellanos I, Prokosch HU, Barkle T. Employing heat maps to mine associations in structured routine care data. Artif Intell Med February 2014;60(2). 2 Pages: 79-88.
[44] Klemm P, Lawonn K, Glasser S, Niemann U, Hegenscheid K, Volzke H, Preim B. 3D regression heat map analysis of population study data. IEEE Trans Vis Comput Graph January 2016;22(1):81–90.
[45] Qiu S, Joshi PS, Miller MI, Xue CH. Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. Brain June 2020; 143(6):1920–33.
[46] Jelistratova I, Teipel SJ, Grothe MJ, Michel J. Longitudinal validity of PET-based staging of regional amyloid deposition. Hum Brain Mapp 2020. https://doi.org/10.1002/hbm.25121. Early Access: July.
[47] Ossenkoppele R, Jansen WJ, Rabinovici GD, Knol DL, van der Flier WM, van Berckel BNM, et al. Prevalence of amyloid PET positivity in dementia syndromes a meta-analysis. JAMA May 2015;313(19):1939–49.
[48] Landau SM, Breault C, Joshi AD, Pontecorvo M, Mathis CA, Jagust WJ, Mintun MA. Amyloid-β imaging with Pittsburgh compound B and florbetapir: comparing radiotracers and quantification methods. J Nucl Med January 2013;54(1):70–7.
[49] Grothe MJ, Barthel H, Sepulcre J, Dyrba M, Sabri O, Teipel SJ. In vivo staging of regional amyloid deposition. Neurology November 2017;89(20):2031–8.
[50] Parbo P, Ismail R, Hansen KV, Amidi A, Mårup FH, Gottrup H, Brændgaard H, Eriksson BO, Eskildsen SF, Lund TE, et al. Brain inflammation accompanies amyloid in the majority of mild cognitive impairment cases due to Alzheimer's disease. Brain May 2017;140(7):2002–11.
[51] Loewenstein DA, Lizarraga G, Adjouadi M, Barker WW, Greig-Custo MT, Penate A, Shea YF, Behar R, Ollarves A, Robayo C, Hanson K, Marsiske M, Burke S, Ertekin-Taner N, Vaillancourt D, Santi SDe, Golde T, Duara RR, Rosselli MM. Effect of age, ethnicity, sex, cognitive status and APOE genotype on amyloid load and the threshold for amyloid positivity. UNSP Neuroimage Clin 2019;22:101800. PMID: 30991618.
[52] Qiu S, Joshi PS, Miller MI, Xue CH. Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. Brain June 2020; 143(6):1920–33.
[53] Crane PK, Carle A, Gibbons LE, Insel P, Mackin RS, Gross A, Jones RN, Mukherjee S, Curtis SM, Harvey D, Weiner M, Mungas, for the Alzheimer's Disease Neuroimaging Initiative. Development and assessment of a composite score for memory in the Alzheimer's Disease Neuroimaging Initiative (ADNI). Brain Imaging Behav 2012 Dec;6(4):502–16.
[54] Liu M, Zhang J, Adeli E, Shen D. Joint classification and regression via deep multi-task multi-channel learning for Alzheimer's disease diagnosis. IEEE Trans Biomed Eng 2018;66:1195–206. https://doi.org/10.1109/TBME.2018.2869989.
[55] Zhu X, Suk H-I, Lee S-W, Shen D. Canonical feature selection for joint regression and multi-class identification in Alzheimer's disease diagnosis. Brain Imaging Behav 2016;10:818–28. https://doi.org/10.1007/s11682-015-9430-9434.
[56] Shi J, Zheng X, Li Y, Zhang Q, Ying S. Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease. IEEE J Biomed Health Inform 2018;22(1):173–83. https://doi.org/10.1109/JBHI.2017.2655720.
[57] Lin WM, Gao QQ, Tong T. Multiclass diagnosis of stages of Alzheimer's disease using linear discriminant analysis scoring for multimodal data. Comput Biol Med Jul 2021;134:104478.
[58] Fang C, Li C, Forouzannezhad P, Cabrerizo M, Curiel RE, Loewenstein D, Duara R, Adjouadi M. Gaussian discriminative component analysis for early detection of Alzheimer's disease: a supervised dimensionality reduction algorithm. J Neurosci Methods October 2020;344:108856. PMID: 32663548.
[59] 'Bryant SE, Waring SC, Cullum CM. Staging dementia using clinical dementia rating scale sum of boxes scores: a Texas Alzheimer's Research Consortium Study. Arch Neurol 2008;65(8):1091–5. https://doi.org/10.1001/archneur.65.8.1091.

Downloaded for Anonymous User (n/a) at Florida International University from ClinicalKey.com by Elsevier on August 04, 2023. For personal use only. No other uses without permission. Copyright ©2023. Elsevier Inc. All rights reserved.