Predicting Dementia Subtypes at Early Stages using Machine Learning and Structural Data

Hui Wei DAVIDWEI@NYU.EDU
Department of Population Health
NYU Langone Medical Center

Narges Razavian NARGES.RAZAVIAN@NYULANGONE.ORG
Center for Healthcare Innovation and Delivery Sciences
Departments of Radiology and Population Health
NYU Langone Medical Center

Editor: N/A

Abstract

Dementia like Alzheimer's disease and Lewy body dementia is currently incurable, which gives burden to not only patients but their family as well. It is hard for clinicians to diagnose accurately for dementia subtypes due to their similar symptoms, especially at early stages. Therefore, the necessity to tell exact differences between distinct dementia subtypes is urgent. Unlike previous works which only focused on classifying Alzheimer's disease and/or Lewy body dementia, this paper further refines subtypes and explores the possibility to distinguish Pure Alzheimer's disease, Pure Lewy body dementia, the Lewy body variant of Alzheimer's disease, and Other subtypes. Multiclass Logistic Regression and Multilayer Perceptron are applied to National Alzheimer's Coordinating Center dataset. To compare the best model and clinician performance, this paper utilizes neurologist-confirmed autopsy labels and top factors obtained from the weight matrix of the best model. It turns out that the best model outperformes clinicians by a large margin on the mixture subtype.

Keywords: Dementia Subtypes, NACC dataset, Multiclass Logistic Regression, Multilayer Perception

1. Introduction

Alzheimer's disease (AD) is the most common cause of dementia, which accounts for 60% to 80% of demented cases (Scheltens et al., 2016). Patients with AD always suffer from brain atrophy, which leads to memory loss, language problems and declining problem solving abilities. Until now, the cause of AD has not been fully understood, which makes it hard to be diagnosed and given effective treatments. However, several possible derivations were found. Pathologically, AD is closely related to the distribution of abnormal amyloid and tau proteins in the brain, which results in a severe neurodegeneration. Genetically, patients with APOEϵ gene can progress to more severe stages faster than those ones without it (Scheltens et al., 2016; Williams et al., 2013). Moreover, lifestyle factors like diabetes, obesity, smoking and low educational level can also contribute to AD.

Lewy body dementia (LBD) is the second most prevalent subtype (Walker et al., 2015; Mueller et al., 2017). Patients with LBD often experience visual hallucinations, anxiety, depression, and extrapyramidal effects when exposed to antipsychotics (Walker et al., 2015; Sfiimomura et al., 1998). Unlike AD, LBD does not draw sufficient clinical attentions (Walker et al., 2015). However, compared to patients with AD, LBD patients suffer from a more severe prognosis, which includes accelerated cognitive declines and shorter lifespans (Mueller et al., 2017). Due to overlapping biomarkers, genetics and symptoms with other dementias, LBD is more difficult to be identified, especially from AD (Walker et al., 2015; Mueller et al., 2017; Nelson et al., 2010).

Besides AD and LBD, a more severe disease combining them was also found, referred as the Lewy body variant of Alzheimer's disease (Mix AD + LBD) (Förstl, 1999). In this concomitant case, symptoms of AD and LBD can demonstrate at the same time, such as plaques and tangles, hallucinations, delusions and slow wave transients in electroencephalogram (EEG) (Weiner et al., 1996; Förstl, 1999; van der Zande et al., 2018). Therefore, available clinical diagnostic criteria for AD (McKhann, 2012) and LBD (McKeith et al., 2005) perform poorly when applied to this mixed dementia. Also, treatments should be used with special caution, since required dopaminergic treatments can lead to aggressive hallucination and delusion, and antipsychotics will result in Parkinsonian symptoms (Weiner et al., 1996). Thus, how to identify and treat this mixture is still an open research area.

Dementia subtypes diagnosis at early stages is significant since clinicians can take specific treatments based on the result. In this paper, we try to classify four dementia subtypes: Pure AD, Pure LBD, Mix AD+LBD and Other subtypes, at the first time patients show very mild or mild cognitive impairment. Their cognitive status are measured by global CDR score (CDRGLOB). We utilized demographics, medication, health history, physical, neuropsychological tests, cerebrospinal fluid (CSF) biomarker and genetic data in NACC dataset, which provide fundamental test information for a demented patient. (Nelson et al., 2010) and (Gaugler et al., 2013) showed that misdiagnosed rate is high for both AD and LBD in NACC dataest, so we used neuropathologic results instead of clinician judgements as true labels. Due to uncovered relationship between dementia subtypes and those features, two machine learning models: Multiclass Logistic Regression (LR) and Multilayer Perceptron (MLP), were examined. After selecting the best model, its performance was compared with clinician results, based on most important features of each disease. We also investigated reasons of clinician errors and best model improvements. Compared to prior works using other unstructural biomarkers like MRI or EEG, our work use only structural data to differentiate dementia subtypes. To our knowledge, this is the first paper using machine learning and structural data to distinguish Pure AD, Pure LBD, Mix AD+LBD, and other subtypes at the first time when patients show early stage dementia.

2. Related Works

Despite difficulties to distinguish AD and LBD, several tentative works have investigated the probability to combine machine leanring with biomarkers including imaging data and EEG to do the classification.

(Lebedev et al., 2013) uses multivariate sparse partial least squares classification of MRI cortical thickness measurements to differentiate AD and LBD. Although it can achieve a fairly good accuracy of 77.78%, the performance declines obviously when the model is tested on the cohort of a distinct protocol without protocol alignment. (Wada et al., 2019) investigates the probability of convolutional neural networks (Lecun et al., 1998) and structural MR connectomes. Their results cannot exceed previous studies which used CT or EEG. 3D local binary pattern texture features combined with a random forest (RF) classifier is investigated by (Oppedal et al., 2017). It shows separating AD and LBD is much harder than distinguishing between normal control (NC) and either of them. (Katako et al., 2018) demonstrates Positron Emission Tomography with fluorodeoxyglucose (FDG-PET) pattern is able to accurately classify AD and NC, but the specificity remains low when use it between dementia subtypes, especially AD and LBD.

Besides imaging data, EEG is also explored. (Lee et al., 2015) shows grand total EEG (GTE) cut-off score of 6.5 can be used for clinically distinguishing AD and LBD, with sensitivity of 79% and specificity of 76%. Combining RF and quantitative EEG, (Dauwan et al., 2016) argued that EEG can improve diagnosis accuracy for AD and LBD, and it is the most discriminative feature selected by RF classifier compared to clinical tests, MRI and CSF, and visual EEG. Furthermore, (Colloby et al., 2016) uses both EEG and MRI to do the differential diagnosis, whose result is better than EEG-only and MRI-only methods. (van der Zande et al., 2018) is the first work to investigate EEG to diagnose Pure LBD, Pure AD and Mix AD + LBD, and is the most related work to our task. Although they confirmed that EEG characteristics can separate Pure LBD and Pure AD, it cannot be applied for Pure LBD and Mix AD + LBD. Also, unlike our work, their model cannot distinguish Pure AD, Pure LBD and Mix AD+LBD at the same time.

It is worth noticing that all above works lack neuropathological diagnosis, so their clinical labels and conclusions can be suspicious.

3. Methods

3.1 Data Description and Analysis

The data used in this paper is from National Alzheimer's Coordinating Center (NACC) dataset. It consists of four directories, including Uniform dataset, CSF biomarker data, Genetic data and Neuropathology (NP) data. To be specific, they contain patients' demographics, health history, neuropsychological tests, clinician judgements, CSF values (for Amloyde beta, P-tau, T-tau), APOE genotypes and neuropathological findings. For all features used in this paper, please see the supplementary material.

NACC dataset records patient information when the patient visits individual Alzheimer's Disease Center (ADC). There are 40,858 patients in the acquired data with 3.49 visits per patient from 2005 to 2019. The minimum and maximum visit number is 1 and 14, respectively. Since this dataset is used for research, patients who have more visits may not have dementia.

CDRGLOB evaluates the cognitive impairment degree of patients, which is divided into five stages: 0.0 (no impairment), 0.5 (very mild), 1.0 (mild), 2.0 (moderate), 3.0 (severe). For all patients in our data, the average age is 71.78 and the average CDRGLOB score is 0.548 at their initial visit to ADCs. This fact indicates that dementia happens mostly in elderly people, and most patients are reluctant to admit their dementia until symptoms appear. Table 1 is the matrix of transitions between different CDRGLOB stages in two consecutive visits. Each number describes the proportion of patients converting from y_t to y_{t+1}. It turns out that between two visits, most patients remain the same cognitive status, while majority of the others will transfer to a more severe status. Two things need to be noticed: (1) demented patients can barely recover to a less severe stage, which indicates that dementia is almost incurable after the onset, (2) compared to earlier stages, the proportion of transition from mild ($y_t = 1$) or moderate ($y_t = 2$) to later stages ($y_t = 3$) is larger, implying cognitive decline accelerates once patients reach mild impairment stage. Therefore, to alleviate this situation, early diagnosis and treatment is urgently necessary.

Table 1: Transitions of global CDR score between two consecutive visits

CDRGLOB score	$y_{t+1} = 0$	$y_{t+1} = 0.5$	$y_{t+1} = 1$	$y_{t+1} = 2$	$y_{t+1} = 3$
$y_t = 0$	**0.8992**	0.0978	0.0025	0.0004	0.0002
$y_t = 0.5$	0.1149	**0.7104**	0.1508	0.0187	0.0051
$y_t = 1$	0.0013	0.0610	**0.5895**	0.2975	0.0508
$y_t = 2$	0.0000	0.0024	0.0455	**0.5909**	0.3611
$y_t = 3$	0.0000	0.0000	0.0029	0.0283	**0.9688**

3.2 Label Definition and Qualified Patients

As mentioned in the introduction section, clinician diagnosis is not reliable since it is very difficult for clinicians to make decisions due to similar symptoms shared by Pure AD, Pure LBD, and Mix AD+LBD, especially at early stages. To define output labels, we used neuropathological results, which are the gold standard for identifying dementia subtypes. Confirmed by a neurologist, the label is defined by NPADNC (NIA-AA Alzheimer's disease neuropathologic change score (ADNC)) and NACCLEWY (Lewy body pathology derived) in NACC dataset. The comprehensive definition is in Table 2.

Table 2: Autopsy-confirmed definition for each dementia subtype

Dementia Subtype	Pure AD	Pure LBD	Mix AD+LBD	Other Types

Definition	NPADNC=2 or 3
	and
	NACCLEWY=0
		NPADNC=0 or 1
		and
		NACCLEWY=3
			NPADNC=2 or 3
			and
			NACCLEWY=1,2, or 3
				others

NPADNC: 0=Not AD, 1=Low ADNC, 2=Intermediate ADNC, 3=High ADNC
NACCLEWY: 0=No Lewy body pathology, 1=Brainstem-predominant, 2=Limbic (transitional) or amygdala-predominant, 3=Neocortical(diffuse).

To compare the performance between our best model and clinicians, we also need clinician labels as reference. Like neuropathological labels, there is no direct result in NACC dataset illustrating dementia subtypes. However, two clinician diagnostic results: NACCALZD (Presumptive etiologic diagnosis of Alzhiemer's disease) and NACCLBDE (Presumptive etiologic diagnosis of Lewy body disease) are in UDS directory. Table 3 are derived definitions of clinician labels.

Table 3: Definition of clinician labels

Dementia Subtype	Pure AD	Pure LBD	Mix AD+LBD	Other Types

Definition	NACCALZD=1
	and
	NACCLBDE=0 or 8
		NACCALZD=0 or 8
		and
		NACCLBDE=1
			NACCALZD=1
			and
			NACCLBDE=1
				others

NACCALZD: 0=Cognitive impairment (dementia, MCI, or impaired, not MCI) and no AD. 1=Any cognitive impairment and AD etiologic diagnosis. 8=Normal cognition.

NACCLBDE: 0=Cognitive impairment and no LBD. 1=Any cognitive impairment and LBD etiologic diagnosis. 8=Normal cognition.

Although the number of patients in the acquired dataset is not small, not all of them satisfy criteria of qualification. To get neuropathological results for dementia subtypes, qualified candidates must be dead, having both NPADNC and NACCLEWY test results, and at least one demented visit (CDRGLOB$\geqslant$0.5). Moreover, those patients with moderate (CDRGLOB=2) or severe (CDRGLOB=3) impairment at their initial visit are eliminated, since symptoms can be so obvious for both the model and clinicians to diagnose. After these steps, all patients have at least one early stage visit (CDRGLOB=0.5 or 1) before more severe stages. For each eligible patient, the first early stage visit was picked up for experiments. The statistics for qualified patients is in Table 4.

Table 4: Statistics of all qualified patients

Dementia Subtype	Pure AD	Pure LBD	Mix AD+LBD	Other Types	Total
# of qualified patients	757	46	572	512	1887

After all qualified patients were selected, they were split randomly on the patient level into training, validation and test set, which means one patient data can only exist in one set to avoid data leakage between different sets. In order to keep the same distribution for all three sets, patients of each dementia subtype were divided by the ratio of 6:2:2. Then shuffled patients for each subtype were combined to training, validation and test set with the number of 1127, 380, 380, respectively.

3.3 Data Preprocessing

There are 721 features in the unpreprocessed data describing all information mentioned in section 3.1. However, to build reasonable input features to represent patients for machine learning models and avoid information leakage, we adopt following criteria to remove features from the original dataset: (1) all clinician notes and all features concluding text data, since our model only keeps focus on the structural data. (2) all administrative information, such as visited ADC and Packet code of investigation forms. This is to make our model more general, diminishing the administrative influence. (3) features which can provide hints for the model about true labels, which include whether patients are using anti-Alzheimer or anti-Parkinson medicine at the visit, whether patients have Parkinson's Disease in the health history, clinician judgement of symptoms, clinician diagnose, time information of diseases (since this will imply the patient has already had the disease), all features from NP dataset and death information. (4) features which the summary score can be derived from. Only keep CDRSUM for CDR, NACCGDS for GDS score, NACCMMSE for MMSE score, MOCATOTS for MOCA. After removing all those features, there are 191 left.

In clinical data, missing values are common. In NACC dataset, there are three kinds of missing values: (1) denoted by NaN (2) unknown data (3) not applicable or available data because of different ways to collect data. In preprocessing, all of them are treated in the same way and used the same symbol to represent. In addition, there are two data formats: continuous and categorical data. For continuous data, we assume that data of different visits has more similarity and closer relationship within the same patient than between different patients; thus, we fill the missing continuous data using the median of records from the same patient, which is more robust than mean values. After this step, there are still patients having continuous missing values, since feature values are missed entirely for all visits of that patient. In this case, those data are filled using the median of that feature among ALL patients. For the categorical features, we assume that they are not missed randomly and they can provide extra information for the model, so we just keep them as an additional category.

After filling the missing data, both categorical and continuous data are further processed by feature engineering, so that they can be directly fed into machine learning models. For continuous features in three sets, we normalize them using mean and standard deviation of the training set by $(feature\_value - training\_mean)/training\_std$. For categorical features, each category is encoded by a one-hot vector.

3.4 Model Description

We used Multiclass Logistic Regression (LR) and Multilayer perceptron (MLP) to predict dementia subtypes.

Multiclass LR has the form as

$$\hat{y} = \text{softmax}(Wx + b) \tag{1}$$

where $\hat{y}$ is the output vector, W is model wight matrix, x is an input feature vector, b is bias, and $\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_k e^{z_k}}$. The output $\hat{y}$ can be considered as a probabilistic distribution of outputs, and the category with the maximum probability is considered as the output label. In our experiments, the input feature dimension after feature engineering is 557 and the output dimension is 4, as the same number of subtypes we want to predict.

MLP (Fig.1) is a fully connected feedforward neural networks, which has one input layer and one hidden layer followed by an output layer. It can be considered as one of the simplest models in deep learning and can be trained using backpropagation.

To express more complex relations rather than only linear function between inputs and outputs, nonlinear activation functions are applied to the hidden layer. In our model, we use Leaky ReLU (Eq.2), where α is a positive hyperparameter. Unlike “vanilla” ReLU (Eq.3), which results in learning difficulty using backpropagation since the slope on the negative side is always 0, MLP with Leaky ReLU can learn even all inputs are negative.

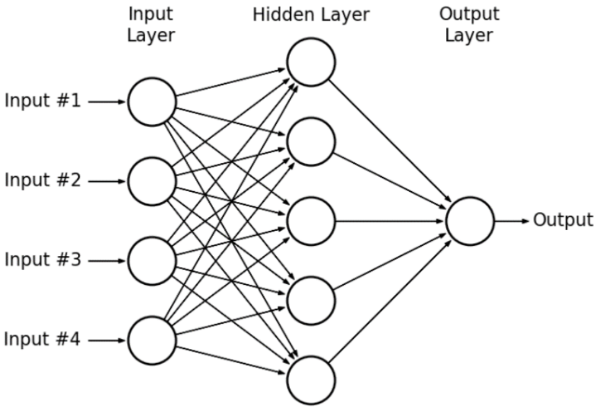

Figure 1: Multilayer perceptron

$$\text{LeakyReLU}(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha x & \text{otherwise} \end{cases} \tag{2}$$

$$\text{ReLU}(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

As other classification problems, we use Cross Entropy Loss to train the model. Within the training set, the ratio of patients of each disease is 453:26:342:306, which is not balanced, especially for Pure LBD. To compensate for this issue, we weight the loss function for each disease by (# of total training samples) / (# of training samples of this disease). To avoid overfitting, L1 regularization is added to the loss function. Compared to L2 regularization which uses the quadratic function as the norm function, L1 regularization makes weights in the model more sparse. The coefficient α weighing the importance of L1 regularization was searched within [0, 0.001, 0.01, 0.1]. The aggregated loss function is as follows:

$$\text{Loss Function} = -\frac{1}{N} \sum_{i \in [1,N]} \sum_{j \in [1,4]} p^{(i)} y_j^{(i)} \log \hat{y}_j^{(i)} + \alpha \sum_k |w_k| \tag{4}$$

where $p^{(i)}$ is the weight for i-th dementia subtype in the training set, y is the true label, $\hat{y}$ is the model output, w is a model weight.

3.5 Data Augmentation

Apart from regularization, data augmentation is useful to solve problems of overfitting. It adds training samples to provide models with more learning materials to better generalize on unseen data. In preprocessing, patients with no autopsy-confirmed labels at the first visit of early stages were removed. However, all visits of such patients have corresponding clinician diagnosed labels, so we propose four strategies to expand the training set using those abandoned visit data. The training data in these strategies are as follows: (1) all autopsy: all visits with autopsy-confirmed label instead of only the first visit of early stages

(2) all first early stages: first early stage visit of all patients in the original dataset no matter whether they have autopsy-confirmed labels or not (3) all early stages: all visits of early stages from the original dataset instead of only the first one (4) all stages: all visits of not only early stages, but later stages (CDRGLOB=2 or 3) as well. In all strategies, for those visits who have no neuropathological result, clinician diagnosed labels defined in section 3.2 are used as target labels for the model. In addition, since our task is for demented patients, for those visits with only clinician labels, we eliminate all visits whose clinician diagnosis are No Cognitive Impairment (NACCALZD=8 and NACCLBDE=8). For the number of training dataset of each data augmentation strategy, please see Table 5.

Table 5: The number of training dataset for each data augmentation strategy

Strategy	All autopsy	All first early stages	All early stages	All stages
# of samples	8,083	22,458	53,509	73,529

4. Experiments and Results

4.1 Model Training and Evaluation

For the convenience and flexibility to train and compare our models, we use Pytorch (Van Merriënboer et al., 2018) to implement both LR and MLP. Specifically, we train our model for 300 epochs, using batch size of 16 and Adam optimizer (Kingma and Ba, 2015) with (ϵ=1e-8, $\beta 1$=0.9, $\beta 2$=0.98). The initial learning rate is 0.001, then decreases by 0.5 when the training loss does not drop.

For each data augmentation strategy, both LR and MLP are trained. During training, models with the highest Macro F1 score and the highest Micro F1 score on the validation set are selected. Then we evaluate those models on the test set and compute F1 scores for each subtype to obtain the best overall model. The result is in Table 6.

In each cell, testing F1 scores of four subtypes are shown for best models with highest validation Macro and Micro F1 score. Regularization coefficient α of both models, and hidden dimension h of MLP are in the followed parenthesis. Please note that (1) even with same hyperparameters, F1 scores of each disease on the test set can be different, since the training epochs to get the best validation Macro and Micro F1 score are different, (2) if best validation Macro and Micro F1 score are obtained at the same epoch, there is only one set of F1 scores in the cell, such as Best LR on All early stages training set.

In table 6, those models with blue or red color is not overfitting. Furthermore, since the best model should not bias to any subtype, we selected the model with highest test Macro F1 score to compare with clinicians. The F1 scores and hyperparameters of this model are shown as red color.

4.2 Interpretability

In the computational healthcare, interpretability is important, we also examine which features the overall best model focuses on when it makes decisions for each disease. From the previous section, the best model we selected is LR, so we can rank the input feature impor-

Table 6: Testing F1 scores of Pure AD, Pure LBD, Mix AD+LBD and Other subtypes

Train set Models	Original	All autopsy	First early stages	All early stages	All stages
Best LR	Macro: [0.529, 0.171, 0.363, 0.563] (α=0.01) Micro:[0.541, 0, 0.348, 0.569] (α=0)	Macro: [0.428, 0, 0.444, 0.527] (α=0.001) Micro:[0.502, 0, 0.342, 0.538] (α=0.001)	Macro: [0.466, 0.2, 0.277, 0.473] (α=0.01) Micro:[0.482, 0.2, 0.374, 0.529] (α=0.001)	[0.562, 0.303, 0.268, 0.5] ($\alpha = 0.01$)	Macro: [0.439, 0.16, 0.340, 0.533] (α=0.001) Micro:[0.535, 0.146, 0.049, 0.531] (α=0.001)
Best MLP	Macro: [0.481, 0, 0.335, 0.562] (α=0, h=512) Micro:[0.528, 0, 0.309, 0.561] (α=0, h=256)	Macro: [0.482, 0.074, 0.315, 0.561] (α=0, h=256) Micro:[0.549, 0.16, 0.267, 0.571] (α=0.001, h=512)	Macro: [0.495, 0.098, 0.362, 0.517] (α=0, h=1024) Micro:[0.505, 0.185, 0.228, 0.523] (α=0.001, h=256)	Macro: [0.453, 0.233, 0.315, 0.477] (α=0.001, h=256) Micro:[0.544, 0.261, 0.245, 0.558] (α=0, h=1024)	[0.543, 0.261, 0.368, 0.553] (α=0, $h = 512$)

Blue and Red → not overfitting
Red → best model selected to compare with clinicians

tance for each disease based on weights in the linear layer. For each disease, top 10 features are shown in Table 7. If a feature name is followed by a number, it means that feature is categorical, and the number is the category the model attends on. For instance, SEX1 means Males, SEX2 means Females. The float number in the parenthesis is the weight of this particular feature in best LR. Please refer to supplement materials, and UDS[1], CSF[2] and Genetic[3] data of NACC data set to see the descriptions of features and each category.

From those top features, some conclusions align with medical findings, which proves the validity and credibility of our best model. For example, both Pure AD and Pure LBD have SEX, HALL, and NITE, but their categories are different based on the following numbers. In Pure AD, SEX2 means females, NITE0 means patients have no night behavior, and HALL0 means patients have no hallucination. In Pure LBD, having hallucinations (HALL1) and night behavior (NITE1) and male (SEX1) are more important when the model make decisions. They are consistent with following medical findings: (1) in AD, the number of males are larger, while in LBD, there are more females, (2) hallucination is a typical symptom of LBD, (3) patients with LBD always have sleep problems. Moreover, most top features of Mix AD + LBD are overlapping with Pure AD and Pure LBD, which indicates that patients with Mix AD + LBD have symptoms of both dementias.

[1] `https://www.alz.washington.edu/WEB/rdd_uds.pdf`
[2] `https://www.alz.washington.edu/WEB/csfded.pdf`
[3] `https://www.alz.washington.edu/WEB/rdd_gen.pdf`

Table 7: Top 10 features with weights of the best model for each disease

Dementia Subtype	Top Features
Pure AD	NACCAGE(2.054e-01), SEX2(1.984e-01), WAIS(1.222e-01), UDSVERLC(6.381e-02), CSFPTAU(4.581e-02), NITE0(4.47e-02), HALL0(4.403e-02), NACCADEP0(4.229e-02), NACCAPOE2(2.861e-02), NACCNE4S1(2.457e-02)
Pure LBD	HALL1(3.348e-01), NACCAANX(2.126e-01), BOSTON(1.847e-01), NACCMMSE(1744e-01), TRAILA(1.743e-01), NITE1(1.248e-01), SEX1(1.207e-01) INRELTO1(1.207e-01), TRAILB(9.504e-02), MEMUNITS(8.355e-02)
Mix AD + LBD	CDRSUM(2.373e-01), EDUC(1.148e-01), NACCAGE(1.006e-01), BOSTON(4.143e-02), NACCAGEB(2.667e-02), DIGIF(1.781e-02), HYPERCHO1(1.657e-02), TRAILA(1.451e-02), TRAILB(1.433e-02), HALL1(1.154e-02), HALLSEV1(1.046e-02)
None of these	WAIS(1.485e-01), NACCNE4S0(1.152e-01), MEMUNITS(7.549e-02), NACCBMI(7.176e-02), UDSBENTD(5.380e-02), REMDATES0(4.893e-02), NACCMMSE(4.695e-02), UDSBENTC(3.849e-02), QUITSMOK(2.742e-02), DISN1(2.230e-02)

4.3 Comparison with Clinicians

To see how well the best model performs for each subtype, we set clinician diagnosis on the test set as the baseline. In Table 8, bootstrap (sample size = 80% of the whole test set, 1000 iterations) is used to compute mean F1 score for each subtype, followed by 95% confidence intervals in the parenthesis.

Table 8: F1 scores of the best model and clinicians

Dementia Subtype	Best Model	Clinicians
Pure AD	0.563 (0.487, 0.627)	0.556 (0.489, 0.624)
Pure LBD	0.295 (0.080, 0.519)	0.283 (0.100, 0.526)
Mix AD + LBD	**0.265 (0.154, 0.372)**	0.062 (0.000, 0.133)
None of these	0.497 (0.406, 0.591)	0.584 (0.492, 0.667)

From Table 8, we can see that for all diseases except for Mix AD + LBD, mean F1 scores of the best model and clinicians fall within each other's 95% confidence interval. Thus, their ability to distinguish these subtypes is significantly the same. From confusion matrices in Figure 2, the best model greatly increased the number of true positive patients with Mix AD + LBD, which indicates that it outperforms clinicians by a large margin for this more severe subtype.

Compared with F1 scores, sensitivity and specificity are more meaningful in the medical field. Table 9 and Table 10 compare sensitivity and specificity of both the best model and clinicians for each dementia subtype. As F1 scores in Table 8, bootstrap (sample size = 80% of the whole test set, 1000 iterations) is also used to compute 95% confidence intervals shown in the parenthesis following the mean value.

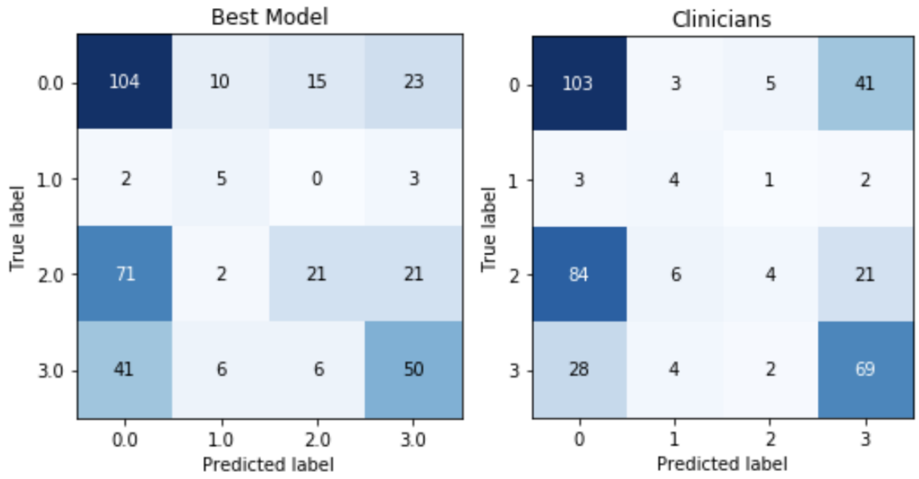

Figure 2: Confusion matrix of the best model and clinicians (0=Pure AD, 1=Pure LBD, 2=Mix AD+LBD, 3=None of these).

Table 9: Sensitivity of the best model and clinicians

Dementia Subtype	Best Model	Clinicians
Pure AD	0.684 (0.602, 0.769)	0.680 (0.597, 0.761)
Pure LBD	0.504 (0.142, 0.889)	0.406 (0.000, 0.800)
Mix AD + LBD	**0.185 (0.105, 0.271)**	0.033 (0.000, 0.075)
Other subtypes	0.485 (0.378, 0.595)	**0.671 (0.567, 0.764)**

Table 10: Specificity of the best model and clinicians

Dementia Subtype	Best Model	Clinicians
Pure AD	0.500 (0.429, 0.573)	0.495 (0.426, 0.563)
Pure LBD	0.951 (0.925, 0.973)	0.965 (0.942, 0.983)
Mix AD + LBD	0.921 (0.884, 0.955)	**0.970 (0.946, 0.990)**
Other subtypes	**0.831 (0.782, 0.880)**	0.770 (0.714, 0.822)

For Pure AD, specificity is relatively lower compared to sensitivity. The reason is that Pure AD is the most common dementia, and overlaps greatly with other subtypes, so when similar symptoms appear, both the model and clinicians tend to diagnose patients as Pure AD. For Pure LBD, low sensitivity and high specificity demonstrates that they are always misdiagnosed unless some extremely typical features appear. This is also confirmed by a large confidence interval range, which implies that some Pure LBD patients are easy to be identified, while others are not. Therefore, more research should be done on finding more explicit and differentiable features and criteria for this specific dementia. The same case is also for Mix AD + LBD, and it is more severe due to the much lower sensitivity compared to Pure LBD, which confirms that current criteria performs extremely poorly for this subtype.

As F1 scores, sensitivity and specificity for Pure AD and Pure LBD of the best model and clinicians fall within 95% confidence interval of each other, demonstrating that the best model and clinicians have the same clinical performance for these subtypes. The biggest difference lies in Mix AD + LBD and Other subtypes. For Mix AD + LBD, the best model has higher sensitivity, where our model performs better. For Other subtypes, clinicians achieve higher sensitivity, probably because they can refer to more explicit symptoms and criteria of each of those subtypes. However, for our model, other dementias are grouped together, which is harder to diagnose. Additionally, moderate clinician specificity (0.770) of Other subtypes implies that symptoms between AD, LBD and other dementias are also overlapping.

After comparing the performance of the best model and clinicians from macro perspective, we are further interested in (1) in what population clinicians make more errors (2) where our best model improves. Since Others subtypes is an aggregated category, so reasons cannot be explored explicitly for each subtype. Based on it, we only analyze Pure AD, PURE LBD, and Mix AD+LBD in this part. In the best model, magnitudes of linear layer weights are from -10^{-1} to 10^{-1}. For Pure AD and Pure LBD, We explore top features with the maximum magnitude, i.e. 10^{-1}. For Mix AD + LBD, since the difference between clinicians and the model is relatively large, top features with magnitudes of 10^{-1} and 10^{-2} are analyzed. For each subtype, we first divide the population truly with this subtype into several groups according to a specific feature, then compute the probability of errors for each particular group $p(wrong|group)$. If the probability of a group is highest, then if clinicians only know patients belong to this specific group, they will most likely make errors. Table 11 and Table 12 show the probabilities of clinician errors for Pure AD and Pure LBD. Combining them with Table 13, we can see in what population clinicians are easier to misdiagnose. For exploring the second question, we listed error probabilities for both clinicians and the best model, and compare them to see why the best model performs better. The number in parenthesis is the proportion of patients belonged to this group are misdiagnosed, and each number in brackets is the number of misdiagnosed patients who are falsely assigned to each subtype.

Table 11: Clinician wrong probabilities given each group for Pure AD and Number of falsely predicted labels for each subtype. The highlighted number is the highest wrong probability within each group.

Features	Groups	Wrong probabilities	# of falsely predictive labels
NACCAGE	$30 \sim 40$	0.000 (0/1)	[0, 0, 0, 0]
	$50 \sim 60$	0.267 (4/15)	[0, 0, 0, 4]
	$60 \sim 70$	0.292 (7/24)	[0, 0, 1, 6]
	$70 \sim 80$	0.345 (19/55)	[0, 1, 2, 16]
	$80 \sim 90$	**0.354 (17/48)**	[0, 2, 2, 13]
	$90 \sim 100$	0.222 (2/9)	[0, 0, 0, 2]
SEX	1	**0.352(25/71)**	[0, 2, 3, 20]
	2	0.296 (24/81)	[0, 1, 2, 21]
WAIS	$0 \sim 10$	0.333 (2/6)	[0, 0, 0, 2]
	$10 \sim 20$	0.300 (6/20)	[0, 0, 0, 6]
	$20 \sim 30$	0.361 (13/36)	[0, 1, 3, 9]
	$30 \sim 40$	0.333 (9/27)	[0, 1, 1, 7]
	$40 \sim 50$	0.354 (8/24)	[0, 0, 0, 8]
	$50 \sim 60$	**1.000 (4/4)**	[0, 0, 0, 4]
	$70 \sim 80$	**1.000 (1/1)**	[0, 0, 0, 1]

Table 12: Clinician wrong probabilities given each group for Pure LBD and Number of falsely predicted labels for each subtype. The highlight number is the highest wrong probability for each group.

Features	Groups	Wrong probabilities	# of falsely predictive labels
HALL	0	**0.625 (5/8)**	[2, 0, 1, 2]
	1	0.500 (1/2)	[1, 0, 0, 0]
NACCAANX	0	**0.833 (5/6)**	[2, 0, 1, 2]
	1	0.250 (1/4)	[1, 0, 0, 0]
BOSTON	20 ∼ 25	**1.000 (3/3)**	[3, 0, 0, 0]
	25 ∼ 30	0.429 (3/7)	[0, 0, 1, 2]
TRAILA	0 ∼ 29	**1.000 (1/1)**	[0, 0, 0, 1]
	29 ∼ 78	0.556 (5/9)	[3, 0, 1, 1]
NITE	0	**1.000 (1/1)**	[0, 0, 0, 1]
	1	0.429 (3/7)	[1, 0, 1, 1]
SEX	1	0.571 (4/7)	[1, 0, 1, 2]
	2	**0.667 (2/3)**	[2, 0, 0, 0]
INRELTO	1	0.571 (4/7)	[1, 0, 1, 2]
	2	**1.000 (1/1)**	[1, 0, 0, 0]
	3	0.500 (1/2)	[1, 0, 0, 0]

Highlighted groups are harder for clinicians to make decisions. For example, in Pure LBD, NITE is whether patients have night behaviors or not. According to Table 12, clinicians are easier to make errors for those patients who don't have the night behavior (NITE=0) compared to those having it (NITE=1). For patients who don't have night behaviors and have wrong diagnosis, they are falsely assigned to Pure AD and Other subtypes. In addition, compared to NITE=0 patients, NITE=1 patients is easier to be assigned as Mix AD +LBD, which confirms that the night behavior is also one of the symptoms of the mixed subtype.

Table 13: Clinician and Best Model wrong probabilities given each group for Mix AD+LBD and Number of falsely predicted labels for each subtype. The highlighted number is lower wrong probability for each group.

Features	Groups	Clinicians	Best Model
CDRSUM	0 ∼ 6	0.968 (90/93) [64, 6, 0, 20]	**0.892 (83/93)** [60, 2 , 0, 21]
	1 ∼ 12	0.955 (21/22) [20, 0, 0, 1]	**0.500 (11/22)** [11, 0, 0, 0]
EDUC	0 ∼ 12	1.000 (5/5) [4, 0, 0, 1]	1.000 (5/5) [4, 0, 0, 1]
	12 ∼ 16	0.976 (40/41) [33, 3, 0, 4]	**0.829 (34/41)** [25, 1, 0, 8]
	16 ∼ 18	0.969 (31/32) [23, 1, 0, 7]	**0.813 (26/32)** [23, 0, 0, 3]
	18 ∼ 20	0.950 (19/20) [12, 2, 0, 5]	**0.850 (17/20)** [12, 0, 0, 5]
	20 ∼ 36	0.941 (16/17) [12, 0, 0, 4]	**0.706 (12/17)** [7, 1, 0, 4]
NACCAGE	40 ∼ 50	1.000 (2/2) [2, 0, 0, 0]	1.000 (2/2) [0, 0, 0, 2]
	50 ∼ 60	1.000 (15/15) [13, 1, 0, 1]	**0.333 (5/15)** [2, 0, 0, 3]
	60 ∼ 70	0.909 (10/11) [7, 0, 0, 3]	0.909 (10/11) [9, 0, 0, 1]
	70 ∼ 80	0.944 (51/54) [35, 4, 0, 12]	**0.833 (45/54)** [33, 2, 0, 10]
	80 ∼ 90	1.000 (29/29) [24, 1, 0, 4]	**0.966 (28/29)** [24, 0, 0, 4]

	90 ∼ 100	1.000 (4/4) [3, 0, 0, 1]	1.000 (4/4) [3, 0, 0, 1]
BOSTON	0 ∼ 5	1.000 (4/4) [4, 0, 0, 0]	**0.500 (2/4)** [2, 0, 0, 0]
	5 ∼ 10	1.000 (2/2) [2, 0, 0, 0]	**0.500 (1/2)** [1, 0, 0, 0]
	10 ∼ 15	1.000 (7/7) [5, 0, 0, 2]	**0.857 (6/7)** [5, 0, 0, 1]
	15 ∼ 20	0.950 (19/20) [17, 0, 0, 2]	**0.900 (18/20)** [13, 0, 0, 5]
	20 ∼ 25	1.000 (18/18) [15, 1, 0, 2]	**0.833 (15/18)** [13, 0, 0, 2]
	25 ∼ 30	1.000 (49/52) [31, 5, 0, 13]	**0.827 (43/52)** [31, 1, 0, 11]
NACCAGEB	40 ∼ 50	1.000 (2/2) [2, 0, 0, 0]	1.000 (2/2) [0, 0, 0, 2]
	50 ∼ 60	1.000 (15/15) [13, 1, 0, 1]	**0.333 (5/15)** [2, 0, 0, 3]
	60 ∼ 70	0.909 (10/11) [7, 0, 0, 3]	0.909 (10/11) [9, 0, 0, 1]
	70 ∼ 80	0.944 (55/58) [36, 5, 0, 14]	**0.845 (49/58)** [34, 2, 0, 13]
	80 ∼ 90	1.000 (25/25) [23, 0, 0, 2]	**0.960 (24/25)** [23, 0, 0, 1]
	90 ∼ 100	1.000 (4/4) [3, 0, 0, 1]	1.000 (4/4) [3, 0, 0, 1]
DIGIF	0 ∼ 1	1,000 (1/1) [1, 0, 0, 0]	**0.000 (0/1)** [0, 0, 0, 0]
	2 ∼ 3	1.000 (1/1) [1, 0, 0, 0]	1.000 (1/1) [0, 0, 0, 1]
	3 ∼ 4	1.000 (2/2) [2, 0, 0, 0]	**0.500 (1/2)** [1, 0, 0, 0]
	4 ∼ 5	1.000 (9/9) [7, 0, 0, 2]	**0.889 (8/9)** [7, 0, 0, 1]
	5 ∼ 6	1.000 (6/6) [4, 1, 0, 1]	**0.333 (2/6)** [0, 0, 0, 2]
	6 ∼ 7	0.958 (23/24) [20, 0, 0, 3]	**0.792 (19/24)** [13, 0, 0, 6]
	7 ∼ 8	1.000 (17/17) [13, 2, 0, 2]	**0.824 (14/17)** [12, 0, 0, 2]
	8 ∼ 9	1.000 (14/14) [11, 0, 0, 3]	**0.786 (11/14)** [9, 0, 0, 2]
	9 ∼ 10	1.000 (14/14) [7, 2, 0, 5]	1.000 (14/14) [10, 1, 0, 2]
	10 ∼ 11	**0.714 (5/7)** [3, 1, 0, 1]	0.857 (6/7) [6, 0, 0, 0]
	11 ∼ 12	**0.889 (8/9)** [5, 0, 0, 3]	1.000 (9/9) [8, 0, 0, 1]
HYPERCHO	0	0.967 (58/60) [9, 0, 0, 2]	**0.750 (45/60)** [32, 1, 0, 12]
	1	0.961 (49/51) [48, 6, 0, 16]	**0.902 (46/51)** [37, 1, 0, 8]
	2	1.000 (3/3) [16, 0, 0, 1]	**0.667 (2/3)** [2, 0, 0, 0]
TRAILA	0 ∼ 29	0.917 (11/12) [9, 0, 0, 2]	0.917 (11/12) [8, 0, 0, 3]
	29 ∼ 78	0.972 (70/72) [48, 6, 0, 16]	**0.917 (66/72)** [51, 1, 0, 14]
	78 ∼ 150	1.000 (17/17) [16, 0, 0, 1]	**0.412 (7/17)** [5, 1, 0, 1]
TRAILB	0 ∼ 75	1.000 (7/7) [2, 1, 0, 4]	1.000 (7/7) [3, 0, 0, 4]
	75 ∼ 273	**0.946 (53/56)** [36, 3, 0, 14]	0.964 (54/56) [45, 0, 0, 9]
	273 ∼ 300	1.000 (29/29) [27, 2, 0, 0]	**0.655 (19/29)** [13, 2, 0, 4]
HALL	0	0.972 (103/106) [79, 5, 0, 19]	**0.830 (88/106)** [69, 1, 0, 18]
	1	0.750 (3/4) [2, 1, 0, 0]	**0.250 (1/4)** [0, 1, 0, 0]
HALLSEV	1	0.667 (2/3) [2, 0, 0, 0]	**0.000 (0/3)** [0, 0, 0, 0]
	3	1.000 (1/1) [0, 1, 0, 0]	1.000 (1/1) [0, 1, 0, 0]
	8	0.972 (103/106) [79, 5, 0, 19]	**0.830 (88/106)** [69, 1, 0, 18]

For Mix AD + LBD in Table 13, we analyze hallucination (HALL) as an example. No matter whether patients show hallucination or not, the best model always performs better than clinicians. Additionally, we can draw the following conclusions: (1) clinicians most likely diagnose Mix AD + LBD as Pure AD or Pure LBD for either patients with or without hallucination. (2) for non-hallucinated patients, the best model reduces the number of wrongly assigned Pure AD and Pure LBD. (3) for hallucinated patients, the best model greatly decreases the wrong probability since it has a stronger ability to differentiate Pure AD and Mix AD + LBD, compared to clinicians.

5. Discussion

5.1 Limitations

Although our best model can outperform clinicians, there are still some limitations of this work. Firstly, in NACC dataset, there is no explicit clinician diagnostic label for Pure AD, Pure LBD, Mix AD + LBD and Other subtypes. Thus, we only defined clinician diagnosis based on separate diagnosis for AD and LBD, which is more rough. However, this will improve clinician performance compared to the case if real clinicians are requested to diagnose for the refined subtypes discussed in the work. The reason is that once AD symptoms appears, the clinician label is 1, so when clinician assign positive for AD, patients could have Pure AD or Mix AD+LBD. Pure LBD has the same case. Therefore, our model can still beat clinicians. Secondly, for our particular task, after selecting qualified patients, the number of samples in the dataset become relatively small. Because of this and individual variations for dementia symptoms, our best model may be overfitting to this specific dataset. Therefore, to test its generalization, a larger dataset for this task needs to be collected and evaluated on.

5.2 Future Works

To improve our model performance and also simplify the patient examination process, some possible ideas will be explored in the future: (1) only keep those input features which can be measured at home so that patients can predict dementia subtypes on their own. (2) change the way to define labels for data augmentation: using the clinician diagnosis of each patient's last visit as the label for all previous visits of this patient. The reason is that later diagnosis is more accurate since symptoms will be more obvious so that it is easier for clinicians to distinguish subtypes. (3) Based on aggregated test scores like CDRSUM, the model may not be able to differentiate subtypes confidently. So expanding test scores which compose those aggregated ones may help if that will not lead to overfitting. (4) change the way to train the model using transfer learning: train the model on samples who only have clinician labels, then fine-tune it using samples with neuropathological results.

6. Conclusion

In conclusion, in this work we examined Multiclass Logistic Regression and Multilayer Perceptron to diagnose Pure Alzheimes's Disease, Pure Lewy body Dementia, Lewy body variant of Alzheimer's Disease and Other subtypes at the first time patients demonstrate very mild or mild coginitive impairment. This is challenging due to their extremely overlapping symptoms. Multiclass Logitstic Regression is the best model on the test set. It can achieve clinician performance for the first two subtypes, and exceed it on the mixed one, with regard to F1 scores and sensitivity. We also investigated top features of each subtype for clinical interpretability and also calculated misdiagnosed probability for each group. To our best knowledge, this is the first work to use machine learning and only structural data to predict such refined dementia subtypes at the first early stage visit.

References

Sean J. Colloby, Ruth A. Cromarty, Luis R. Peraza, Kristinn Johnsen, Gísli Jóhannesson, Laura Bonanni, Marco Onofrj, Robert Barber, John T. O'Brien, and John Paul Taylor. Multimodal EEG-MRI in the differential diagnosis of Alzheimer's disease and dementia with Lewy bodies. *Journal of Psychiatric Research*, 78:48–55, 2016. ISSN 18791379. doi: 10.1016/j.jpsychires.2016.03.010.

Meenakshi Dauwan, Jessica J van der Zande, Edwin van Dellen, Iris E C Sommer, Philip Scheltens, Afina W Lemstra, and Cornelis J Stam. Random forest to differentiate dementia with Lewy bodies from Alzheimer's disease. *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, 4:99–106, 2016. ISSN 23528729. doi: 10.1016/j.dadm.2016.07.003. URL `http://dx.doi.org/10.1016/j.dadm.2016.07.003`.

Hans Förstl. The Lewy body variant of Alzheimer's disease: Clinical, pathophysiological and conceptual issues. *European Archives of Psychiatry and Clinical Neuroscience*, 249 (SUPPL. 3):64–67, 1999. ISSN 09401334. doi: 10.1007/pl00014176.

Joseph E Gaugler, Haya Ascher-Svanum, David L Roth, Tolulope Fafowora, Andrew Siderowf, and Thomas G Beach. Characteristics of patients misdiagnosed with Alzheimer's disease and their medication use: An analysis of the NACC-UDS database. *BMC Geriatrics*, 13(1), 2013. ISSN 14712318. doi: 10.1186/1471-2318-13-137.

Audrey Katako, Paul Shelton, Andrew L. Goertzen, Daniel Levin, Bohdan Bybel, Maram Aljuaid, Hyun Jin Yoon, Do Young Kang, Seok Min Kim, Chong Sik Lee, and Ji Hyun Ko. Machine learning identified an Alzheimer's disease-related FDG-PET pattern which is also expressed in Lewy body dementia and Parkinson's disease dementia. *Scientific Reports*, 8(1):1–13, 2018. ISSN 20452322. doi: 10.1038/s41598-018-31653-6. URL `http://dx.doi.org/10.1038/s41598-018-31653-6`.

Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pages 1–15, 2015.

Alexander V Lebedev, E Westman, M K Beyer, M G Kramberger, C Aguilar, Z Pirtosek, and D Aarsland. Multivariate classification of patients with Alzheimer's and dementia with Lewy bodies using high-dimensional cortical thickness measurements: An MRI surface-based morphometric study. *Journal of Neurology*, 260(4):1104–1115, 2013. ISSN 03405354. doi: 10.1007/s00415-012-6768-z.

Yann Lecun, Leon Bottou, Yoshua Bengio, and Patrick Ha. Gradient-Based Learnining Applied to Document Recognition. *Proceedings of the IEEE*, (November):1–46, 1998. ISSN 00189219. doi: 10.1109/5.726791.

Hennie Lee, Geert J.F. Brekelmans, and Gerwin Roks. The EEG as a diagnostic tool in distinguishing between dementia with Lewy bodies and Alzheimer's disease. *Clinical Neurophysiology*, 126(9):1735–1739, 2015. ISSN 18728952. doi: 10.1016/j.clinph.2014.11.021. URL http://dx.doi.org/10.1016/j.clinph.2014.11.021.

I G McKeith, D W Dickson, J Lowe, M Emre, J T O\textquoterightBrien, H Feldman, J Cummings, J E Duda, C Lippa, E K Perry, D Aarsland, H Arai, C G Ballard, B Boeve, D J Burn, D Costa, T Del Ser, B Dubois, D Galasko, S Gauthier, C G Goetz, E Gomez-Tortosa, G Halliday, L A Hansen, J Hardy, T Iwatsubo, R N Kalaria, D Kaufer, R A Kenny, A Korczyn, K Kosaka, V M Y Lee, A Lees, I Litvan, E Londos, O L Lopez, S Minoshima, Y Mizuno, J A Molina, E B Mukaetova-Ladinska, F Pasquier, R H Perry, J B Schulz, J Q Trojanowski, and M Yamada. Diagnosis and management of dementia with Lewy bodies. *Neurology*, 65(12):1863–1872, 2005. ISSN 0028-3878. doi: 10.1212/01.wnl.0000187889.17253.b1. URL https://n.neurology.org/content/65/12/1863.

GM McKhann. The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging- Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement*, 7(3):263–269, 2012. doi: 10.1016/j.jalz.2011.03.005.The.

Christoph Mueller, Clive Ballard, Anne Corbett, and Dag Aarsland. The prognosis of dementia with Lewy bodies. *The Lancet Neurology*, 16(5):390–398, 2017. ISSN 14744465. doi: 10.1016/S1474-4422(17)30074-1. URL http://dx.doi.org/10.1016/S1474-4422(17)30074-1.

Peter T. Nelson, Gregory A. Jicha, Richard J. Kryscio, Erin L. Abner, Frederick A. Schmitt, Gregory Cooper, Li O. Xu, Charles D. Smith, and William R. Markesbery. Low sensitivity in clinical diagnoses of dementia with Lewy bodies. *Journal of Neurology*, 257(3):359–366, 2010. ISSN 03405354. doi: 10.1007/s00415-009-5324-y.

Ketil Oppedal, Kjersti Engan, Trygve Eftestøl, Mona Beyer, and Dag Aarsland. Classifying Alzheimer's disease, Lewy body dementia, and normal controls using 3D texture analysis in magnetic resonance images. *Biomedical Signal Processing and Control*, 33:19–29, 2017. ISSN 17468108. doi: 10.1016/j.bspc.2016.10.007. URL http://dx.doi.org/10.1016/j.bspc.2016.10.007.

Philip Scheltens, Kaj Blennow, Monique M.B. Breteler, Bart de Strooper, Giovanni B. Frisoni, Stephen Salloway, and Wiesje Maria Van der Flier. Alzheimer's disease. *The Lancet*, 388(10043):505–517, 2016. ISSN 1474547X. doi: 10.1016/S0140-6736(15)01124-1. URL http://dx.doi.org/10.1016/S0140-6736(15)01124-1.

Tcitstio Sfiimomura, Efsuro Mon, Hikari Yamashitatont Inuimtira, Nolwfsiigtt Hirono, Mflmoni Hashimoto, Satoshi Tanimukai, Hiroajd Kaziti, and Toleiji Hcmihara. Cognitive loss in dementia with Lewy bodies and Alzheimer disease. *Archives of Neurology*, 55(12):1547–1552, 1998. ISSN 00039942.

Jessica J. van der Zande, Alida A. Gouw, Inger van Steenoven, Philip Scheltens, Cornelis Jan Stam, and Afina W. Lemstra. EEG characteristics of dementia with Lewy

Bodies, Alzheimer's Disease and mixed pathology. *Frontiers in Aging Neuroscience*, 10 (JUL):1–10, 2018. ISSN 16634365. doi: 10.3389/fnagi.2018.00190.

Bart Van Merriënboer, Olivier Breuleux, Pascal Lamblin, and Arnaud Bergeron. Automatic differentiation in ML: Where we are and where we should be going. *Advances in Neural Information Processing Systems*, 2018-Decem(NeurIPS):8757–8767, 2018. ISSN 10495258.

Akihiko Wada, Kohei Tsuruta, Ryusuke Irie, Koji Kamagata, Tomoko Maekawa, Shohei Fujita, Saori Koshino, Kanako Kumamaru, Michimasa Suzuki, Atsushi Nakanishi, Masaaki Hori, and Shigeki Aoki. Differentiating Alzheimer's disease from dementia with lewy bodies using a deep learning technique based on structural brain connectivity. *Magnetic Resonance in Medical Sciences*, 18(3):219–224, 2019. ISSN 18802206. doi: 10.2463/mrms.mp.2018-0091.

Zuzana Walker, Katherine L. Possin, Bradley F. Boeve, and Dag Aarsland. Lewy body dementias. *The Lancet*, 386(10004):1683–1697, 2015. ISSN 1474547X. doi: 10.1016/S0140-6736(15)00462-6. URL `http://dx.doi.org/10.1016/S0140-6736(15)00462-6`.

Myron Weiner, Richard Risser, Munro Cullum, Lawrence Honig, Charles White III, Samuel Speciale, and Roger Rosenberg. Alzheimer's Disease and Its Lewy Body Variant: A Clinical Analysis of Postmortem Verified Cases. *American Journal of Psychiatry*, 153:10 (October):1269–1273, 1996.

Monique M. Williams, Martha Storandt, Catherine M. Roe, and John C. Morris. Progression of Alzheimer's disease as measured by Clinical Dementia Rating Sum of Boxes scores. *Alzheimer's and Dementia*, 9(1 SUPPL.):S39–S44, 2013. ISSN 15525260. doi: 10.1016/j.jalz.2012.01.005. URL `http://dx.doi.org/10.1016/j.jalz.2012.01.005`.

Supplementary Material

Table 14: Clinical features used in this paper from NACC dataset

NACC Form	Feature Name	Descriptions
A1 Subject Demographics	NACCID	Subject ID number
	DATE	Visit Date, derived from VISITMO, VISITDAY, VISITYR
	SEX	Subject's sex
	HISPANIC	Hispanic/Latino ethnicity
	HIPOR	Hispanic origins
	PRIMLANG	Primary language
	EDUC	Years of education
	NACCAGEB	Subject's age at initial visit
	NACCNIHR	Subject's race
	NACCAGE	Subject's age at visit
A2 Co-participant Demographics	INSEX	Co-participant's sex Original
	NACCNINR	Co-participant's race
	INEDUC	Co-participant's years of education
	INRELTO	Co-participant's relationship to subject
A3 Subject Family History	NACCFAM	Indicator of first-degree family member with cognitive impairment
A4 Subject Medications	ANYMEDS	Subject taking any medications
	NACCAAAS	Reported current use of an antiadrenergic agent
	NACCAANX	Reported current use of an anxiolytic, sedative, or hypnotic agent
	NACCAC	Reported current use of an anticoagulant or antiplatelet agent
	NACCACEI	Reported current use of an angiotensin converting enzyme (ACE) inhibitor
	NACCADEP	Reported current use of an antidepressant
	NACCAHTN	Reported current use of any type of an antihypertensive or blood pressure medication
	NACCAMD	Total number of medications reported at each visit
	NACCANGI	Reported current use of an angiotensin II inhibitor
	NACCAPSY	Reported current use of an antipsychotic agent
	NACCBETA	Reported current use of a betaadrenergic blocking agent (Beta-Blocker)

A4 Subject Medications	NACCCCBS	Reported current use of a calcium channel blocking agent
	NACCDBMD	Reported current use of a diabetes medication
	NACCDIUR	Reported current use of a diuretic
	NACCEMD	Reported current use of estrogen hormone therapy
	NACCEPMD	Reported current use of estrogen + progestin hormone therapy
	NACCHTNC	Reported current use of an antihypertensive combination therapy
	NACCLIPL	Reported current use of lipid lowering medication
	NACCNSD	Reported current use of nonsteroidal anti-inflammatory medication
	NACCVASD	Reported current use of a vasodilator
A5 Subject Health History	TOBAC30	Smoked cigarettes in last 30 days O
	TOBAC100	Smoked more than 100 cigarettes in life
	SMOKYRS	Total years smoked cigarettes
	PACKSPER	Average number of packs smoked per day
	QUITSMOK	If the subject quit smoking, age at which he/she last smoked (i.e., quit)
	ALCOCCAS	In the past three months, has the subject consumed any alcohol?
	ALCFREQ	During the past three months, how often did the subject have at least one drink of any alcoholic beverage such as wine, beer, malt liquor, or spirits?
	CVHATT	Heart attack/cardiac arrest
	HATTMULT	More than one heart attack/cardiac arrest?
	CVAFIB	Atrial fibrillation
	CVANGIO	Angioplasty/endarterectomy/stent
	CVBYPASS	Cardiac bypass procedure
	CVPACDEF	Pacemaker and/or defibrillator
	CVPACE	Pacemaker
	CVCHF	Congestive heart failure
	CVANGINA	Angina
	CVHVALVE	Heart valve replacement or repair
	CVOTHR	Other cardiovascular disease
	CBSTROKE	Stroke
	STROKMUL	More than one stroke reported as of the Initial Visit
	CBTIA	Transient ischemic attack (TIA)

	TIAMULT	More than one TIA reported as of the Initial Visit
	SEIZURES	Seizures
	NACCTBI	History of traumatic brain injury (TBI)
	TBI	Traumatic brain injury (TBI)
	TBIBRIEF	Traumatic brain injury (TBI) with brief loss of consciousness
	TRAUMBRF	Brain trauma — brief unconsciousness
	TBIEXTEN	TBI with extended loss of consciousness — 5 minutes of longer
	TRAUMEXT	Brain trauma — extended unconsciousness
	TBIWOLOS	TBI without loss of consciousness — as might result from military detonations or sports injury
	TRAUMCHR	Brain trauma — chronic deficit
	NCOTHR	Other neurological condition
	DIABETES	Diabetes
	DIABTYPE	If Recent/active or Remote/ inactive diabetes, which type?
	HYPERTEN	Hypertension
	HYPERCHO	Hypercholesterolemia
	B12DEF	Vitamin B12 deficiency
	THYROID	Thyroid disease
	ARTHRIT	Arthritis
	ARTHTYPE	Type of arthritis
	ARTHUPEX	Arthritis, region affected — upper extremity
	ARTHLOEX	Arthritis, region affected — lower extremity
	ARTHSPIN	Arthritis, region affected — spine
	ARTHUNK	Region affected — unknown
	INCONTU	Incontinence — urinary
	INCONTF	Incontinence — bowel
	APNEA	Sleep apnea history reported at Initial Visit
	RBD	REM sleep behavior disorder (RBD) history reported at Initial Visit
	INSOMN	Hyposomnia/insomnia history reported at Initial Visit
	OTHSLEEP	Other sleep disorder history reported at Initial Visit
	ALCOHOL	Alcohol abuse — clinically significant occurring over a 12-month period manifested in one of the following areas: work, driving, legal, or social

	ABUSOTHR	Other abused substances — clinically significant impairment occurring over a 12-month period manifested in one of the following areas: work, driving, legal, or social
	PTSD	Post-traumatic stress disorder (PTSD) O
	BIPOLAR	Bipolar disorder
	SCHIZ	Schizophrenia
	DEP2YRS	Active depression in the last two years
	DEPOTHR	Depression episodes more than two years ago
	ANXIETY	Anxiety
	OCD	Obsessive-compulsive disorder (OCD)
	NPSYDEV	Developmental neuropsychiatric disorders (e.g., autism spectrum disorder [ASD], attention-deficit hyperactivity disorder [ADHD], dyslexia)
	PSYCDIS	Other psychiatric disorder
B1 Physical	NACCBMI	NACCBMI
	BPSYS	Subject blood pressure (sitting), systolic
	BPDIAS	Subject blood pressure (sitting), diastolic
B4 CDR score	CDRSUM	CDR sum of boxes
B5 Neuropsychiatric Inventory Questionnaire (NPI-Q)	NPIQINF	NPI-Q co-participant
	DEL	Delusions in the last month
	DELSEV	Delusions severity
	HALL	Hallucinations in the last month
	HALLSEV	Hallucinations severity
	AGIT	Agitation or aggression in the last monthv
	AGITSEV	Agitation or aggression severity
	DEPD	Depression or dysphoria in the last month
	DEPDSEV	Depression or dysphoria severity
	ANX	Anxiety in the last month
	ANXSEV	Anxiety severity
	ELAT	Elation or euphoria in the last month
	ELATSEV	Elation or euphoria severity
	APA	Apathy or indifference in the last month
	APASEV	Apathy or indifference severity
	DISN	Disinhibition in the last month
	DISNSEV	Disinhibition severity
	IRR	Irritability or lability in the last month
	IRRSEV	Irritability or lability severity
	MOT	Motor disturbance in the last month
	MOTSEV	Motor disturbance severity
	NITE	Nighttime behaviors in the last month
	NITESEV	Nighttime behaviors severity

	APP	Appetite and eating problems in the last month
	APPSEV	Appetite and eating severity
B6 Geriatric Depression Scale (GDS)	NACCGDS	Total GDS Score
B7 Functional Activities Questionnaire (FAQ)	BILLS	In the past four weeks, did the subject have any difficulty or need help with: Writing checks, paying bills, or balancing a checkbook
	TAXES	In the past four weeks, did the subject have any difficulty or need help with: Assembling tax records, business affairs, or other paper
	SHOPPING	In the past four weeks, did the subject have any difficulty or need help with: Shopping alone for clothes, household necessities, or groceries
	GAMES	In the past four weeks, did the subject have any difficulty or need help with: Playing a game of skill such as bridge or chess, working on a hobby
	STOVE	In the past four weeks, did the subject have any difficulty or need help with: Heating water, making a cup of coffee, turning off the stove
	MEALPREP	In the past four weeks, did the subject have any difficulty or need help with: Preparing a balanced meal
	EVENTS	In the past four weeks, did the subject have any difficulty or need help with: Keeping track of current events
	PAYATTN	In the past four weeks, did the subject have any difficulty or need help with: Paying attention to and understanding a TV program, book, or magazine

	REMDATES	In the past four weeks, did the subject have any difficulty or need help with: Remembering appointments, family occasions, holidays, medications
	TRAVEL	In the past four weeks, did the subject have any difficulty or need help with: Traveling out of the neighborhood, driving, or arranging to take public transportation
MMSE Score	NACCMMSE	Total MMSE score (using D-L-R-O-W)
C1 Neuropsychological Battery Summary Scores	LOGIMEM	Total number of story units recalled from this current test administration
	MEMUNITS	Logical Memory IIA — Delayed — Total number of story units recalled
	MEMTIME	Logical Memory IIA — Delayed — Time elapsed since Logical Memory IA — Immediate
	UDSBENTC	Total score for copy of Benson figure
	UDSBENTD	Total score for 10- to 15-minute delayed drawing of Benson figure
	UDSBENRS	Recognized original stimulus from among four options
	DIGIF	Digit span forward trials correct
	DIGIFLEN	Digit span forward length
	DIGIB	Digit span backward trials correct
	DIGIBLEN	Digit span backward length
	ANIMALS	Animals — Total number of animals named in 60 seconds
	VEG	Vegetable — Total number of vegetables named in 60 seconds
	TRAILA	Trail Making Test Part A — Total number of seconds to complete
	TRAILARR	Part A — Number of commission errors
	TRAILALI	Part A — Number of correct lines
	TRAILB	Trail Making Test Part B — Total number of seconds to complete
	TRAILBRR	Part B — Number of commission errors
	TRAILBLI	Part B — Number of correct lines
	WAIS	WAIS-R Digit Symbol
	BOSTON	Boston Naming Test (30) — Total score
	UDSVERFC	Number of correct F-words generated in 1 minute

	UDSVERFN	Number of F-words repeated in 1 minute
	UDSVERNF	Number of non-F-words and rule violation errors in 1 minute
	UDSVERLC	Number of correct L-words generated in 1 minute
	UDSVERLR	Number of L-words repeated in 1 minute
	UDSVERLN	Number of non-L-words and rule violation errors in 1 minute
	UDSVERTN	Total number of correct F-words and L-words
	UDSVERTE	Total number of F-word and L-word repetition errors
	UDSVERTI	Total number of non-F/L-words and rule violation errors
MoCA score	MOCATOTS	MoCA Total Raw Score — uncorrected
C2 Neuropsychological Batery Scores	CRAFTVRS	Craft Story 21 Recall (Immediate) — Total story units recalled, verbatim scoring
	CRAFTURS	Craft Story 21 Recall (Immediate) — Total story units recalled, paraphrase scoring
	DIGFORCT	Number Span Test: Forward — Number of correct trials
	DIGFORSL	Number Span Test: Forward — Longest span forward
	DIGBACCT	Number Span Test: Backward — Number of correct trials
	DIGBACLS	Number Span Test: Backward — Longest span backward
	CRAFTDVR	Craft Story 21 Recall (Delayed) — Total story units recalled, verbatim scoring
	CRAFTDRE	Craft Story 21 Recall (Delayed) — Total story units recalled, paraphrase scoring
	CRAFTDTI	Craft Story 21 Recall (Delayed) — Delay time
	CRAFTCUE	Craft Story 21 Recall (Delayed) — Cue (boy) needed
	MINTTOTS	Multilingual Naming Test (MINT) — Total score
	MINTTOTW	Multilingual Naming Test (MINT) — Total correct without semantic cue
	MINTSCNG	Multilingual Naming Test (MINT) — Semantic cues: Number given

	MINTSCNC	Multilingual Naming Test (MINT) — Semantic cues: Number correct with cue
	MINTPCNG	Multilingual Naming Test (MINT) — Phonemic cues: Number given
	MINTPCNC	Multilingual Naming Test (MINT) — Phonemic cues: Number correct with cue
Genetic Data (RDD-Gen)	NACCNE4S	Number of APOE e4 alleles
	NACCAPOE	APOE genotype
CSF Biomarker Data	CSFABETA	$A\beta_{1-42}$ reported value/concentration (pg/mL)
	CSFPTAU	$P - tau_{181P}$ reported value/concentration (pg/mL)
	CSFTTAU	T-tau reported value/concentration (pg/mL)
	CSFABMD	$A\beta_{1-42}$ assay method
	CSFPTMD	$P - tau_{181P}$ assay method
	CSFTTMD	T-tau assay method