

Seleção Simplificada de Atributos para Auxílio Ao Diagnóstico da Doença de Alzheimer Utilizando Aprendizado de Máquina

Raimundo G. Saraiva Jr ^(1,2), Joao L. A. Carvalho ⁽¹⁾; Eduardo Peixoto ⁽¹⁾, e Francisco A. O. Nascimento ⁽¹⁾

⁽¹⁾ *Departamento de Engenharia Elétrica, Universidade de Brasília (UnB), Brasília, Brasil*

⁽²⁾ *Controle e Processos Industriais, Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE), Limoeiro do Norte, Brasil*

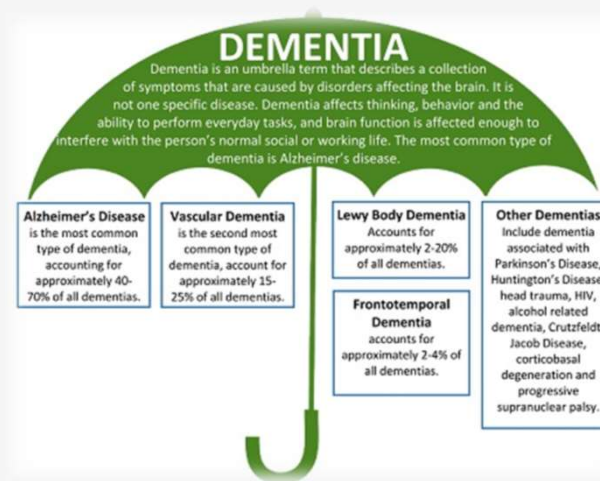
Alzheimer's Disease

- 1 Characterized as a progressive kind of dementia that starts with minor cognitive impairment and progresses to the most severe stage, which robs the patient of their potential for independence and function.
- 2 Dementia is derived from the Latin word "dementia," which means "loss of reason".
- 3 Alois Alzheimer, a German doctor, first identified Alzheimer's Disease in a 55-year-old patient with dementia in 1906 after performing a brain biopsy and correlating the patient's symptoms and medical records with the brain findings.

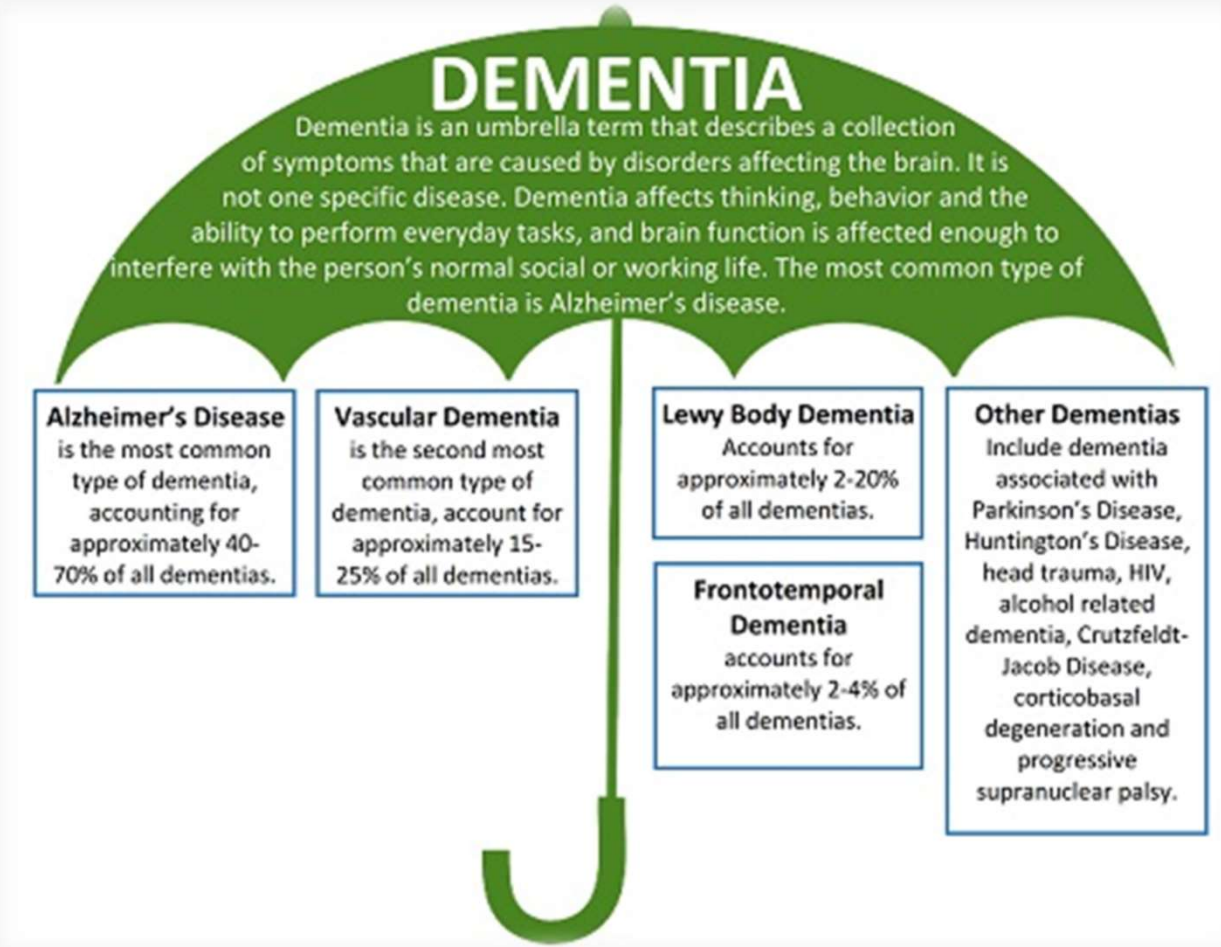


Dementia classifications

- 1 There are over 100 different types of dementia known, but due to the prevalence of Alzheimer's disease, they are commonly referred to as Alzheimer's disease.
- 2 Alzheimer's disease (AD) is the most common type of dementia among all known dementias, accounting for up to 70% of all diagnosed patients.

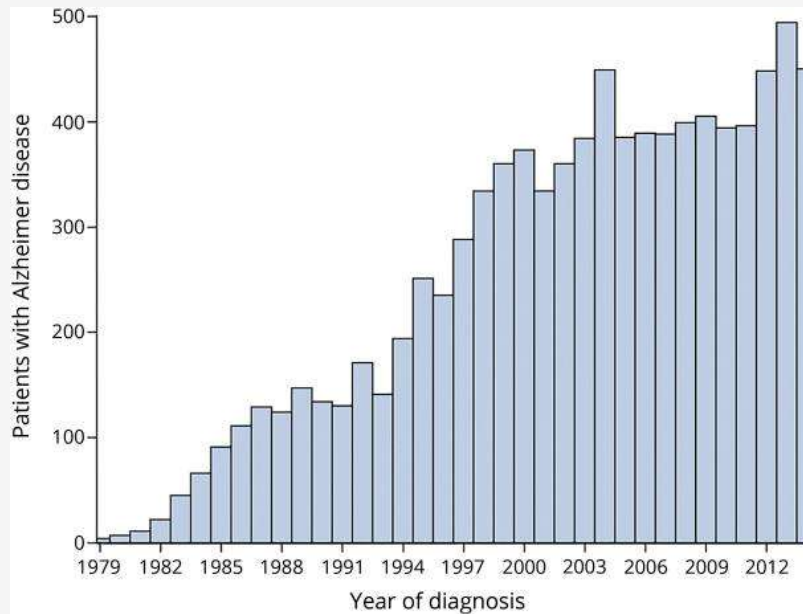


Dementia Types



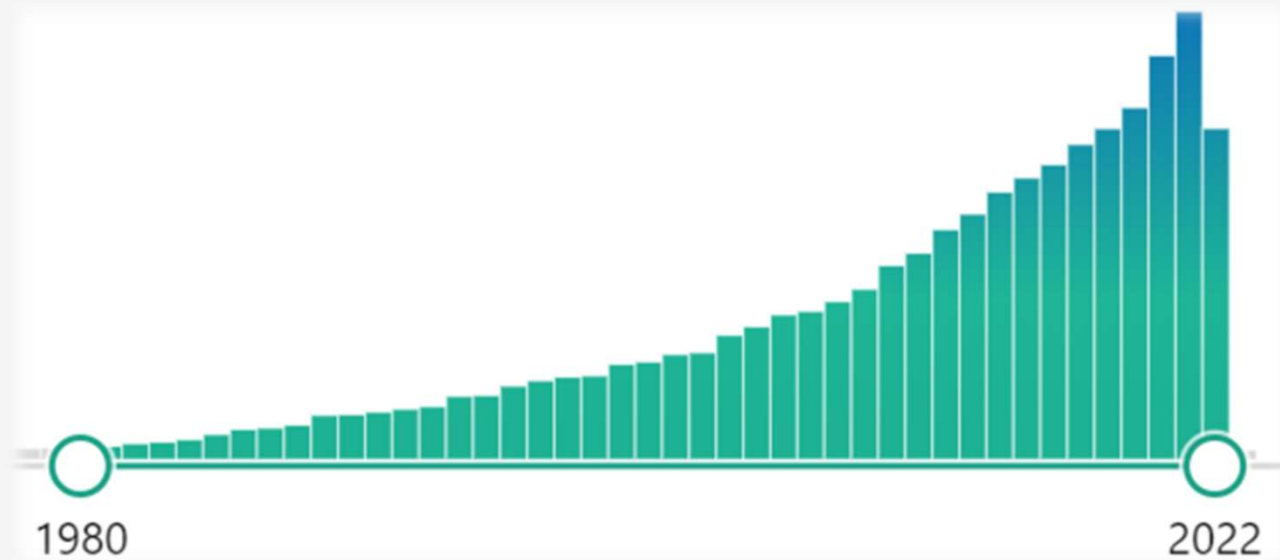
Alzheimer's disease diagnosis

- 1 Despite being known for AD since the early twentieth century, the disease represents civilization's distant past, laden with prejudice and lack of understanding, and has become a social stigma.
- 2 It was only in the second half of the 1970s that AD received an International Classification of Disease (ICD) code and thus started to be diagnosed.



Unique patient diagnoses of Alzheimer disease in persons in Olmsted County, Minnesota, from 1980 to 2014.

Alzheimer's Disease Research



PubMed portal search results with the word "Alzheimer's" from 1980 to september 2022.

Stages of Alzheimer's Disease - Clinical Dementia Rating (CDR) Scale

Stage	Description	Expected Duration of Stage
CDR-0	No Dementia	N/A
CDR-0.5	Very Mild Dementia Memory problems are slight, but consistent. <ul style="list-style-type: none"> - Some difficulty with time and problem solving - Daily life is slightly impaired - Individuals can perform personal care activities 	Average duration is a few years up to 7 years.
CDR-1	Mild Dementia Memory loss is moderate, especially for recent events, and interferes with daily activities. <ul style="list-style-type: none"> - Moderate difficulty with solving problems - Cannot function independently at community affairs - Difficulty with daily activities and hobbies, especially complex ones 	Average duration is 2 years.
CDR-2	Moderate Dementia More profound memory loss, only retaining highly learned material. <ul style="list-style-type: none"> - Disorientation with respect to time and place - Impaired judgment, with difficulty handling problems - Little to no independent function at home - Can only do simple chores - Fewer interests 	Average duration is just under 2 years to 4 years.
CDR-3	Severe Dementia Severe memory loss. <ul style="list-style-type: none"> - Disorientation with respect to time or place - No judgment or problem solving abilities - Cannot participate in community affairs outside the home - Requires help with all tasks of daily living - Requires help with most personal care - Frequent incontinence 	Average duration is 1 year to 2.5 years.

Image and Data Archive (IDA) - Alzheimer's Disease Neuroimaging Initiative (ADNI)

The screenshot displays the IDA website interface. At the top, there is a navigation bar with the IDA logo, a 'Select Study' dropdown menu set to 'ALL', and links for 'ALL@LONI', 'Download', and 'Search'. On the right side of the navigation bar, there are links for 'IDA Home', 'Support', and a user profile for 'junior.saraiva' with the email '@ifce.edu.br'. Below the navigation bar, there are tabs for 'Featured Studies', 'About', 'Utilization', and 'Quick Start'. The main content area features a 'Welcome to the IDA' section with a description: 'The Image and Data Archive (IDA) is a **secure** online resource for **archiving, exploring** and **sharing** neuroscience data.' To the right of this text, there are statistics: '141 studies', '142.735 users', '83.521 subjects', and '161 countries'. Below the welcome section is a 'Featured Studies' section with the text 'Select criteria below to search studies' and a 'Reset Search' button. The search filters are represented by donut charts for 'SEX' (Male, Female, Any), 'MODALITY' (SPECT, PET, EEG, MRI, CT), and 'COLLECTED DATA' (Genetic, Biospecimen, Exam & Assessment, Demographics). A red circle highlights a sidebar on the right that shows 'Featured : 23 studies • 23.897 subjects' and 'Studies you have access to' with icons for ADNI and AIBL. Below this, there is a section for 'Studies you can apply to' with icons for 4RTNI, A4, ABVIB, and ADNIDOD.

Selected studies



The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a longitudinal multicenter study designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of Alzheimer's disease (AD). Since its launch more than a decade ago, the landmark public-private partnership has made major contributions to AD research, enabling the sharing of data between researchers around the world.

<https://adni.loni.usc.edu/>



The Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL) is a study to discover which biomarkers, cognitive characteristics, and health and lifestyle factors determine subsequent development of symptomatic Alzheimer's Disease (AD).

<https://aibl.csiro.au/>

AIBL: Features description

Feature	Description
HMT3	Red Blood Cell Count
HMT7	White Blood Cell Count
HMT13	Platelets
HMT40	Hemoglobin
HMT100	Mean Cell Hemoglobin
HMT102	Mean Cell Hemoglobin Concentrate
AXT117	Thyroid Stimulate Hormone
BAT126	Vitamin 12
RCT6	Urea Nitrogen
RCT11	Serum Glucose
RCT20	Cholestrol
RCT392	Creatinine
LIMMTOTAL	Total number of story units recalled-Logical Memory Immediate Recall
LDELTOTAL	Total number of story units recalled-Partial Score of Logical Memory test
MMSCORE	Mini Mental State Examination Score
CDGLOBAL	Clinical Dementia Rating Global
APGEN1	Apoe gene 1
APGEN2	Apoe gene 2
DXCURREN	Diagnosis status

ADNI: Features description

Feature	Description
Ventricles	Ventricles Volume
Hippocampus	Hippocampus Volume
CDRSB	Clinical Dementia Rating Scale Box
MMSE	Mini-Mental State Examination
WholeBrain	WholeBrain Volume
Entorhinal	Entorhinal Volume
Fusiform	Fusiform Volume
MidTemp	Med Temp Volume
ICV	Intracranial Volume
ADAS11	Alzheimers Disease Assesement Cognition Scale 11
ADAS13	Alzheimers Disease Assesement Cognition Scale 13
Age	Age at baseline
PTEDUCAT	Education
RAVLT_learning	Rey Auditory Verbal Learning Test
RAVLT_immediate	Rey Auditory Verbal Learning Test (5 sum)
RAVLT_Forgetting	Rey Auditory Verbal Learning Test Forgetting
RAVLT_Perc_Forgetting_bl	Rey Auditory Verbal Learning Test Percentile Forgetting
AV45	Average AV45 SUVR of frontal, AC, precuneus, and parietal cortex relative to the cerebellum
APOE4	Apolipoprotein E4
FAQ	Functional Activities Questionnaire
FDG	Average FDG-PET of angular, temporal, and posterior cingulate
DX	Diagnosis Status

Dataset composition

AIBL-18: dataset composition.

Quantidade	Atributos
18	CDGLOBAL AXT117 BAT126 HMT3 HMT7 HMT13 HMT40 HMT100 HMT102 RCT6 RCT11 RCT20 RCT392 MMSCORE LIMMTOTAL LDELTOTAL APGEN1 APGEN2

ADNI-21: dataset composition.

Quantidade	Atributos
21	Ventricles Hippocampus WholeBrain Entorhinal Fusiform MidTemp ICV CDRSB MMSE RAVLT_learning RAVLT_immediate RAVLT_forgetting RAVLT_perc_forgetting_bl ADAS11 ADAS13 FAQ FDG AV45 APOE4 AGE PTEDUCAT

Preprocessing

AIBL-18 resulted in:

- 785 instances,
- 18 features and
- 3 classes

AIBL-18: instance distribution by class.

Class	Number of instances
CN	547
MCI	125
AD	113

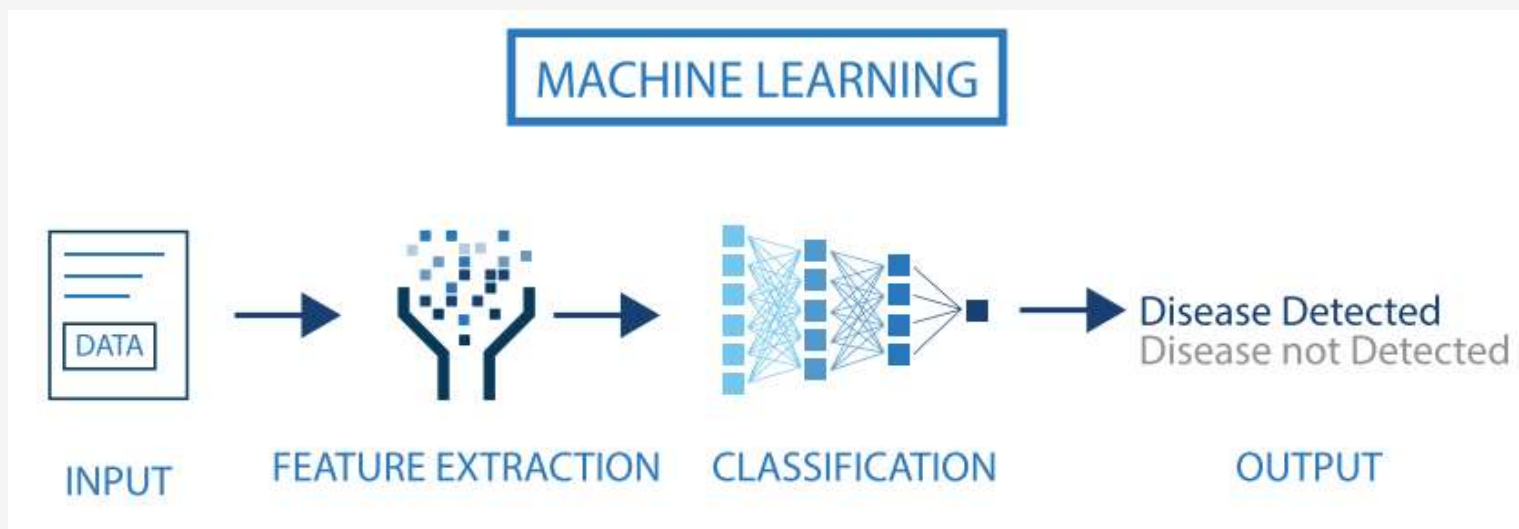
ADNI-21 resulted in

- 1033 instances,
- 21 attributes and
- 3 classes

ADNI-21: instance distribution by class.

Class	Number of instances
CN	505
MCI	367
AD	161

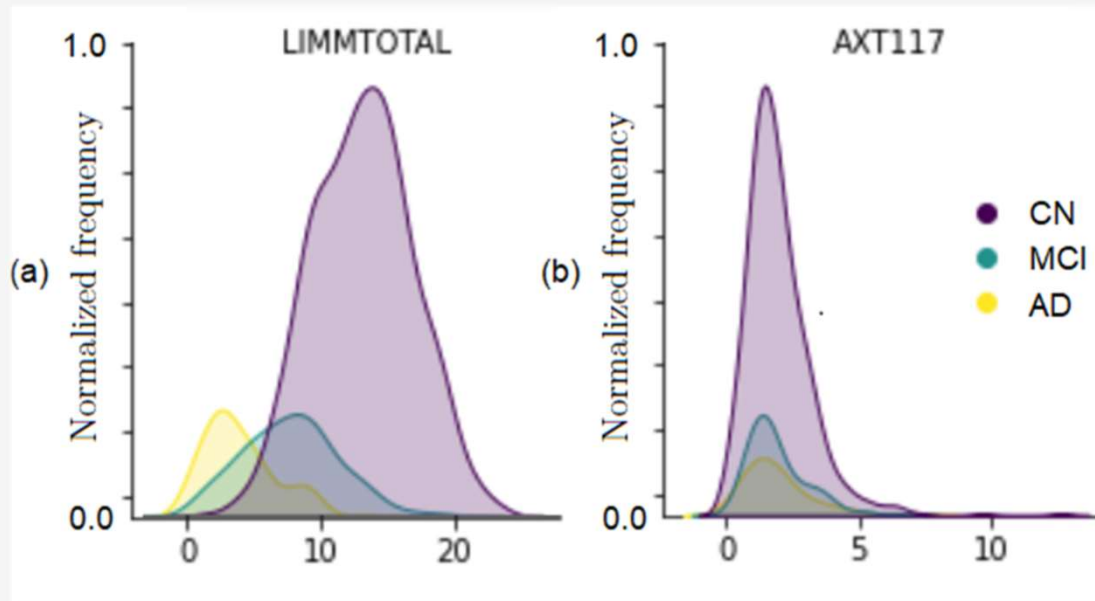
Machine Learning Schema



Propose a new feature selection mechanism

FMFS - first moment feature selection

Data Analysis



Histogram by classes of features LIMMTOTAL(a) and AXT117(b), from AIBL dataset.

FMFS description

Nomenclature definition.

$$\mathbf{x} = \mathbf{M} \text{ instances} \left\{ \begin{array}{l} \overbrace{\left[\begin{array}{cccc} \cdot & & & \\ \cdot & & & \\ \cdot & \cdot & x_{i,j} & \cdot \\ \cdot & & & \\ \cdot & & & \end{array} \right]}^{\text{N features}} \\ i \in [0 \dots \mathbf{N}-1] \\ j \in [0 \dots \mathbf{M}-1] \\ \mathbf{x} \in R^{\mathbf{N} \times \mathbf{M}} \end{array} \right. \quad \mathbf{y} = \begin{bmatrix} \cdot \\ \cdot \\ y_{j,c_k} \\ \cdot \\ \cdot \end{bmatrix} \begin{array}{l} c_k \in [c_1 \dots c_p] \\ j \in [0 \dots \mathbf{M}-1] \\ \mathbf{y} \in R^{\mathbf{N} \times 1} \end{array}$$

1. calculation of the overall data set average

$$\mu_g = \frac{1}{\mathbf{N}} \frac{1}{\mathbf{M}} \sum_{i=0}^{\mathbf{M}-1} \sum_{j=0}^{\mathbf{N}-1} x_{i,j}$$

FMFS description

2. average calculation of each feature

$$\mu_i = \frac{1}{N} \sum_{j=0}^{N-1} x_{i,j}$$

3. scaling all samples

$$z_{ij} = x_{ij}(\mu_g/\mu_i)$$

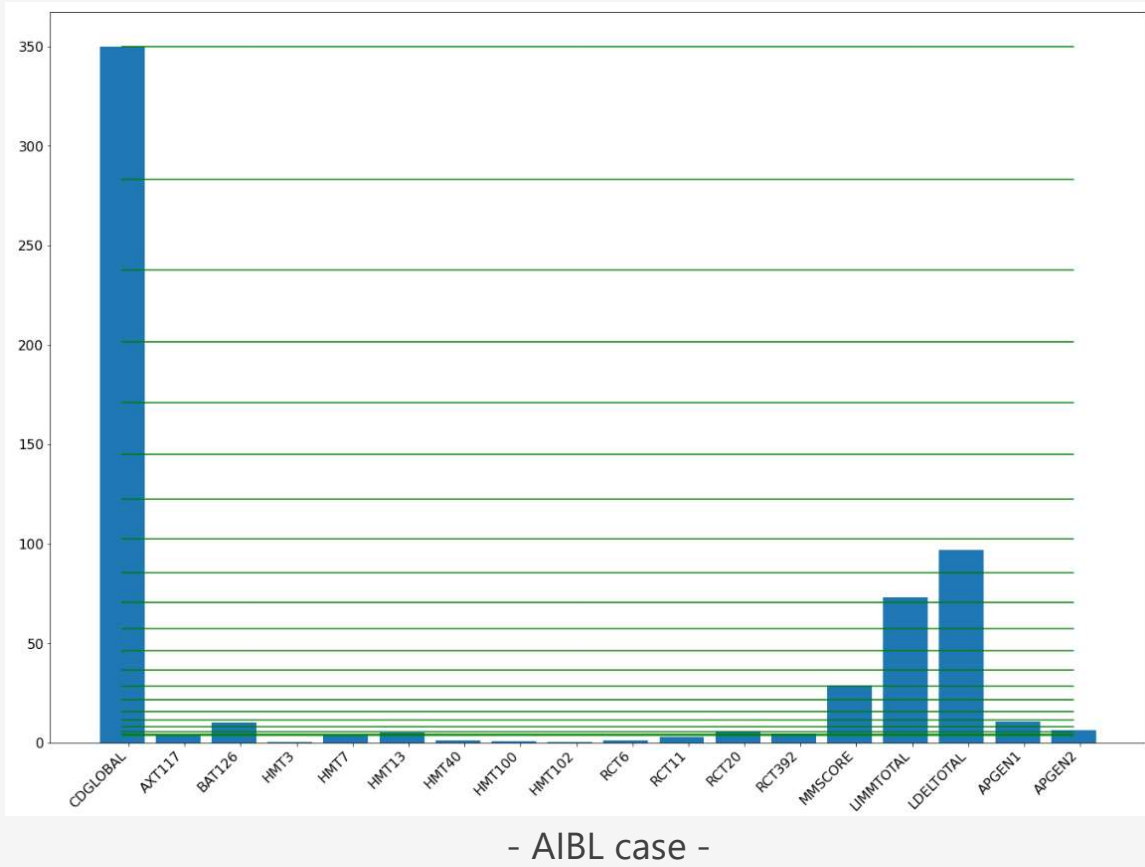
4. calculating the average of each class for each of the feature

$$\mu_{i,c_k} = \frac{1}{N_{i,c_k}} \sum_{j=0}^{N_{j,c_k}-1} z_{i,j} \text{ onde } (y_j = c_k)$$

5. distance between all class means for each attribute

$$dca_i = \sqrt{\sum_{\substack{0 \leq k \leq p \\ 0 \leq q \leq p \\ k \neq q}} (\mu_{i,c_k} - \mu_{i,c_q})^2}$$

FMFS description



1. linear scan

$$lim_x = \max(dca) - x(\max(dca) - \min(dca))/w$$

2. logarithmic scan

$$lim_x = \max(dca) - x \log\left(\frac{x+1}{w} - 1\right) \frac{(\max(dca) - \min(dca))}{w}$$

FMFS results

Reduced AIBL dataset: selected attributes.

Model	Dataset	Features
KNN	AIBL-6	CDGLOBAL BAT126 MMSCORE LIMMTOTAL LDELTOTAL APGEN1
SVM	AIBL-4	CDGLOBAL MMSCORE LIMMTOTAL LDELTOTAL

Reduced ADNI dataset: selected attributes.

Model	Dataset	Features
KNN	ADNI-3	CDRSB ADAS11 FAQ
SVM	ADNI-11	Ventricles Entorhinal CDRSB MMSE RAVLT_learning RAVLT_immediate RAVLT_perc_forgetting_bl ADAS11 ADAS13 FAQ APOE4

Performance evaluation

AIBL dataset results.

Model	Dataset	$B_{ca} \pm IC$	A_{cy}	P_{rn}	S_{ny}	S_{py}	F1	MAUC
KNN	AIBL-18	74,15 \pm 2,08	82,28	73,85	63,61	84,69	66,50	0,912
	AIBL-6	90,44* \pm 3,89	91,32	87,37	86,08	94,81	85,81	0,950
SVM	AIBL-18	91,96 \pm 2,25	91,72	86,97	88,32	95,61	87,09	0,967
	AIBL-4	92,52 \pm 2,50	92,22	88,25	89,06	95,98	87,74	0,959

IC - confidence interval for 95% confidence level

(*) Statistically significant result ($p < 0,05$)

Except for MAUC, every numerical result is expressed as a percentage.

ADNI - Resultados.

Model	Dataset	$B_{ca} \pm IC$	A_{cy}	P_{rn}	S_{ny}	S_{py}	F1	MAUC
KNN	ADNI-21	81,66 \pm 3,39	76,29	79,34	76,89	87,23	77,41	0,878
	ADNI-3	90,06* \pm 2,26	89,45	90,32	86,52	93,60	87,97	0,960
SVM	ADNI-21	92,60 \pm 2,06	91,97	92,52	90,09	95,11	91,07	0,969
	ADNI-11	92,61 \pm 2,29	91,77	92,34	90,20	95,02	90,90	0,969

IC - confidence interval for 95% confidence level

(*) Statistically significant result ($p < 0,05$)

Except for MAUC, every numerical result is expressed as a percentage.

B_{ca} - Balanced Classification Accuracy

A_{cy} - Accuracy

P_{rn} - Precision

S_{ny} - Sensitivity

S_{py} - Specificity

F1 - f1 score

MAUC - Multi-class AUC

Comparison with previous study

Comparison with previous study (AIBL dataset).

Reference	ML/DL	Model	BCA	MAUC
Present study	ML	KNN	90,44	0,950
		SVM	92,52	0,959
(NIYAS; P, 2021b)	ML	KNN	82,50	NA
		SVM	92,50	NA

NA - not available

Except for MAUC, all numerical results are expressed as percentages.

Comparison with previous study

Comparison with previous studies (ADNI dataset).

Reference	ML/DL	Model	BCA	MAUC
Present study	ML	KNN	90,06	0,960
		SVM	92,61	0,969
(NIYAS; P, 2021a)	ML	RF	87	NA
(NIYAS; P, 2021b)	ML	KNN	90	0,96
		SVM	88	0,95
(NGUYEN M., 2020)	DL	RNN	88,7	0,944
(LIU J L I, 2020)	DL	RNN	87,93	0.9439
(ALBRIGHT, 2019)	DL	RNN	NA	0,866
(IDDI <i>et al.</i> , 2019)	ML	RF	86	NA
(MOORE <i>et al.</i> , 2019)	ML	RF	73	0,82
(GHAZI <i>et al.</i> , 2019)	ML	LDA	NA	0,75
(NGUYEN <i>et al.</i> , 2018)	ML	SVM	79	0,86

NA - not available. LDA-linear discriminant analysis. RNN-recurrent neural network. RF-random forest

Except for MAUC, all numerical results are expressed as percentages.

Any questions about this work?

Presenter's Name: Raimundo Guimarães Saraiva Junior

Affiliation: *Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)*

E-mail: junior.saraiva@ifce.edu.br

Thank you!

