Improving the interpretability of computer-assisted analysis tools in neuroimaging

Ninon Burgos, CNRS Researcher
Aramis Lab, Paris Brain Institute, France
Neuroimaging

Unmatched description of the brain’s structure and physiology

Magnetic resonance imaging (MRI)

- T1-weighted MRI
- T2-weighted MRI

Positron emission tomography (PET)

- FDG PET
- Amyloid PET
Clinical pattern recognition

Basic elements of a machine learning classification pipeline

• Training data set
• Feature extraction from raw data and dimensionality reduction
• Model training and optimisation
• Application to test data

Klöppel et al., NeuroImage, 2012
Clinical pattern recognition

Basic elements of a deep learning classification pipeline

• Training data set
• Feature extraction from raw data and dimensionality reduction
• Model training and optimisation
• Application to test data
Use case: Alzheimer’s disease (AD)

What is Alzheimer’s disease?

- Most common cause of dementia
- Disorder caused by abnormal brain changes
- Trigger decline in cognitive abilities, severe enough to impair daily life
- Affect behaviour, feelings and relationships
- Progressive disease
Use case: Alzheimer’s disease (AD)

AD-related markers

• Clinical/cognitive tests
  • Neuropsychological testing of cognitive functions (memory, language, etc.)

• Structural MRI
  • Atrophy

• FDG PET
  • Hypometabolism

• Amyloid PET
  • Accumulation of amyloid-β proteins

• CSF Aβ42, CSF tau, tau PET, diffusion MRI, etc.
Clinical pattern recognition

Use case: Alzheimer’s disease (AD)

• Classification
  • Controls vs AD patients
  • Stable vs progressive mild cognitive impairment (MCI)

• Regression
  • Time of onset
  • Future clinical score

ML/DL for Alzheimer’s diagnosis & prognosis

A very active field of research

Number of articles


Machine learning
Deep learning
## Literature review

- Alzheimer classification from structural MRI using deep learning

### Average accuracy (AD vs CN): 86.4%

<table>
<thead>
<tr>
<th>Study</th>
<th>Performance</th>
<th>Approach</th>
<th>Data leakage</th>
<th>Number of citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Aderghal et al., 2017)</td>
<td>ACC=0.84</td>
<td></td>
<td></td>
<td>16</td>
</tr>
<tr>
<td>(Aderghal et al., 2018)</td>
<td>BA=0.90</td>
<td></td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>(Bäckström et al., 2018)†</td>
<td>ACC=0.90</td>
<td></td>
<td></td>
<td>20</td>
</tr>
<tr>
<td>(Cheng et al., 2017)</td>
<td>ACC=0.87</td>
<td></td>
<td></td>
<td>12</td>
</tr>
<tr>
<td>(Cheng and Liu, 2017)</td>
<td>ACC=0.85</td>
<td></td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>(Islam and Zhang, 2017)</td>
<td>ACC=0.90</td>
<td></td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>(Korolev et al., 2017)</td>
<td>ACC=0.97</td>
<td></td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>(Li et al., 2017)</td>
<td>ACC=0.90</td>
<td></td>
<td></td>
<td>12</td>
</tr>
<tr>
<td>(Lian et al., 2018)</td>
<td>ACC=0.90</td>
<td></td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>(Mingxia Liu et al., 2018a)</td>
<td>ACC=0.91</td>
<td></td>
<td></td>
<td>30</td>
</tr>
<tr>
<td>(Mingxia Liu et al., 2018c)</td>
<td>ACC=0.91</td>
<td></td>
<td></td>
<td>59</td>
</tr>
<tr>
<td>(Qu et al., 2018)</td>
<td>ACC=0.91</td>
<td></td>
<td></td>
<td>26</td>
</tr>
<tr>
<td>(Senanyake et al., 2018)</td>
<td>ACC=0.76</td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>(Shinulev et al., 2018)</td>
<td>ACC=0.62</td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>(Valliani and Soni, 2017)</td>
<td>ACC=0.81</td>
<td></td>
<td></td>
<td>8</td>
</tr>
</tbody>
</table>

### Average accuracy (AD vs CN): 93.8%

<table>
<thead>
<tr>
<th>Study</th>
<th>Performance</th>
<th>Approach</th>
<th>Data leakage</th>
<th>Number of citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Aderghal et al., 2017a)</td>
<td>ACC=0.91</td>
<td></td>
<td></td>
<td>13</td>
</tr>
<tr>
<td>(Basa et al., 2019)</td>
<td>BA=0.99</td>
<td></td>
<td></td>
<td>25</td>
</tr>
<tr>
<td>(Hon and Khan, 2017)</td>
<td>ACC=0.96</td>
<td></td>
<td></td>
<td>32</td>
</tr>
<tr>
<td>(Hosseini Ad et al., 2018)</td>
<td>ACC=0.99</td>
<td></td>
<td></td>
<td>107</td>
</tr>
<tr>
<td>(Islam and Zhang, 2017)</td>
<td></td>
<td></td>
<td></td>
<td>23</td>
</tr>
<tr>
<td>(Korolev et al., 2017)</td>
<td></td>
<td></td>
<td></td>
<td>72</td>
</tr>
<tr>
<td>(Li et al., 2018)</td>
<td></td>
<td></td>
<td></td>
<td>12</td>
</tr>
<tr>
<td>(Wang et al., 2017)</td>
<td></td>
<td></td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>(Wang et al., 2018)</td>
<td></td>
<td></td>
<td></td>
<td>36</td>
</tr>
<tr>
<td>(Wang et al., 2019)</td>
<td></td>
<td></td>
<td></td>
<td>17</td>
</tr>
</tbody>
</table>

---

Wen, Thibeau-Sutre et al., Medical Image Analysis, 2020
Main causes of data leakage in DL scenarios

- No independent test set
- Late split
- Biased within-subject split
- Biased transfer learning

Wen, Thibeau-Sutre et al., Medical Image Analysis, 2020
Framework for the objective & reproducible classification of Alzheimer's disease using deep learning
Framework for the objective & reproducible classification of Alzheimer’s disease using deep learning

Inputs
- Preprocessing types:
  - Minimal
  - Extensive

Classifiers
- 3D subject-level CNN
- 3D patch-level CNN
- 2D slice-level CNN

CV
- K-fold

3D subject-level CNN
3D patch-level CNN
2D slice-level CNN

Wen, Thibeau-Sutre et al., Medical Image Analysis, 2020; https://github.com/aramislab/AD-DL
Influence of the type of network architecture

<table>
<thead>
<tr>
<th></th>
<th>3D subject-level</th>
<th>3D patch-level</th>
<th>2D slice-level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Controls vs AD patients</td>
<td>85%</td>
<td>86%</td>
<td>74%</td>
</tr>
<tr>
<td>Stable vs progressive mild cognitive impairment</td>
<td>73%</td>
<td>70%</td>
<td>-</td>
</tr>
</tbody>
</table>

Balanced accuracy on the test set — Values are presented as mean ± SD.

- 3D subject-level and 3D patch-level approaches: similar balanced accuracies
- 2D-slice approach: consistently lower balanced accuracy
Deep learning interpretability

Alzheimer’s disease
Deep learning interpretability

State-of-the-art methods

- **Gradients visualization** or "saliency maps"  
  \[\text{Simonyan et al, 2013}\]

- **Occlusion heatmaps**  
  \[\text{Zeiler and Fergus, 2016}\]

→ Not so interpretable...

→ Zones are studied independently

→ Depends on the occlusion size
Deep learning interpretability

Learnable mask based on saliency maps

The image $X'_m$ masked by $m$ at voxel $u$ is defined as:

$$X'_m(u) = m(u)X(u) + (1 - m(u))X_{perturbed}(u)$$

Masking possibilities: blurring, replacing by a constant value, adding noise, etc.

[Meaningful perturbation]

[Fong and Vedaldi, 2017]
Learnable mask based on saliency maps

Goal: mask $m$ minimizing the score of the CNN $f$ on a set of occluded images $X'_m$ by covering a minimal amount of pixels in connected parts.

$$m^* = \arg\min_m f(X'_m) + \lambda_1 \|1 - m\|_{\beta_1} + \lambda_2 \|
abla m\|_{\beta_2}$$

Thibeau Sutre et al., SPIE Medical Imaging 2020
Deep learning interpretability

Application to Alzheimer’s disease: quantitative maps

Mask with a constant value
- 0 to simulate grey matter atrophy (control → patient)
- 1 to simulate grey matter restoration (patient → control)

Goal: mask the minimal part of the image necessary to transform a patient into a control for the CNN.

Thibreau Sutre et al., SPIE Medical Imaging 2020
Deep learning interpretability

Application to Alzheimer’s disease: group level masking

Mask computed on a set of images

→ General pattern that characterizes the disease for the CNN

Thibeau Sutre et al., SPIE Medical Imaging 2020
Deep learning interpretability

Application to Alzheimer’s disease: individual level masking

Mask computed on one single image

→ Possible study of individual differences

Thibeau Sutre et al., SPIE Medical Imaging 2020
Deep learning interpretability

Robustness of the masking method

- Experiments performed:
  1. Comparison of inter- and intra-subject similarity
  2. Robustness to mask hyperparameters
  3. Robustness to datasets

→ Our method is robust and can be used as a tool to study CNN training robustness.

Thibreau Sutre et al., SPIE Medical Imaging 2020
Robustness of CNN trainings

- Inter-run variability larger than inter-subject variability

→ CNN training does not robustly identify the relevant regions

Thibeau Sutre et al., SPIE Medical Imaging 2020
Interpretability of computer-assisted diagnosis

Two strategies

- Classify Alzheimer’s disease
- Interpret

- Detect anomalies
- Classify Alzheimer’s disease
Subject-specific abnormality maps

Objective

• Develop a new framework for the individual analysis of PET data to better identify patterns of abnormality

Method

• Subject-specific model of healthy PET uptake
• Subject-specific abnormality map
Subject-specific abnormality maps

- Subject-specific model of healthy PET uptake

Burgos et al., MICCAI 2015, CMMI 2017
Subject-specific abnormality maps

- Subject-specific abnormality map

MRI-PET control dataset

Target subject’s MRI

Registered control dataset

Subject-specific PET model

Target subject’s real PET

SIM$_1$

SIM$_2$

SIM$_n$

w$_1$

w$_2$

w$_n$

Burgos et al., MICCAI 2015, CMMI 2017
Application to Alzheimer’s disease

- FDG PET: marker of synaptic dysfunction

Cognitively normal  Early MCI  Late MCI  Alzheimer’s disease

Disease progression
Computer-aided diagnosis using abnormality maps

Classification algorithm based on linear support vector machines (SVM)

Features:

- Subject’s PET images
- Subject-specific abnormality maps

Classification performance (balanced accuracy [%]):

<table>
<thead>
<tr>
<th>Task</th>
<th>PET</th>
<th>Abn. map</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitively normal vs Alzheimer’s disease</td>
<td>89.3 ± 4.3</td>
<td>92.4 ± 3.7</td>
</tr>
<tr>
<td>Cognitively normal vs Late mild cognitive impairment</td>
<td>73.4 ± 6.1</td>
<td>76.7 ± 5.3</td>
</tr>
<tr>
<td>Cognitively normal vs Early mild cognitive impairment</td>
<td>58.3 ± 6.7</td>
<td>64.7 ± 5.8</td>
</tr>
</tbody>
</table>
Application to primary progressive aphasia

• FDG PET: marker of synaptic dysfunction
Subject-specific abnormality maps

Visualisation tool for clinicians

<table>
<thead>
<tr>
<th>Early MCI</th>
<th>Alzheimer’s disease</th>
<th>Semantic dementia</th>
</tr>
</thead>
</table>

Improved interpretability of subsequent analyses

Alzheimer’s disease
Deep autoencoding models for unsupervised anomaly segmentation in brain MR images

Fig. 1. The proposed anomaly detection concept at a glance. A simple subtraction of the reconstructed image from the input reveals lesions in the brain.
Pseudo healthy image synthesis with pathology disentanglement and adversarial learning

Xia et al., Medical Image Analysis, 2020
Pseudo healthy image synthesis with DL

Pseudo-healthy synthesis with pathology disentanglement and adversarial learning

• ISLES dataset (ischemic lesions)

• BraTS dataset (brain tumours)

Xia et al., Medical Image Analysis, 2020
Improving the interpretability of computer-assisted analysis tools in neuroimaging

**Classification**

Alzheimer’s disease

**Interpretation**

\[ m^* = \arg\min_m f(X_m') + \lambda_1 \|1 - m\|^2_{\theta_1} + \lambda_2 \|\nabla m\|^2_{\theta_2} \]

**Anomaly detection**

**Classification**

Alzheimer’s disease

---

Thibeu Sutre et al., SPIE Medical Imaging 2020 - Burgos et al., MICCAI 2015, CMMI 2017