UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Computational Intelligence Laboratory

DIPARTIMENTO DI INFORMATICA

Visual and Textual Explainability in Alzheimer's Disease Detection in MRI

Supervisors
Prof. Giovanna Castellano
Dr. Gennaro Vessio

Katya Trufanova
Computer Vision
MSc Computer Science

Funded by the European Union NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

Future Artificial Intelligence Research

Introduction

Alzheimer's Disease

A growing challenge in healthcare due to its increasing prevalence and impact on cognitive function

AI in AD Detection

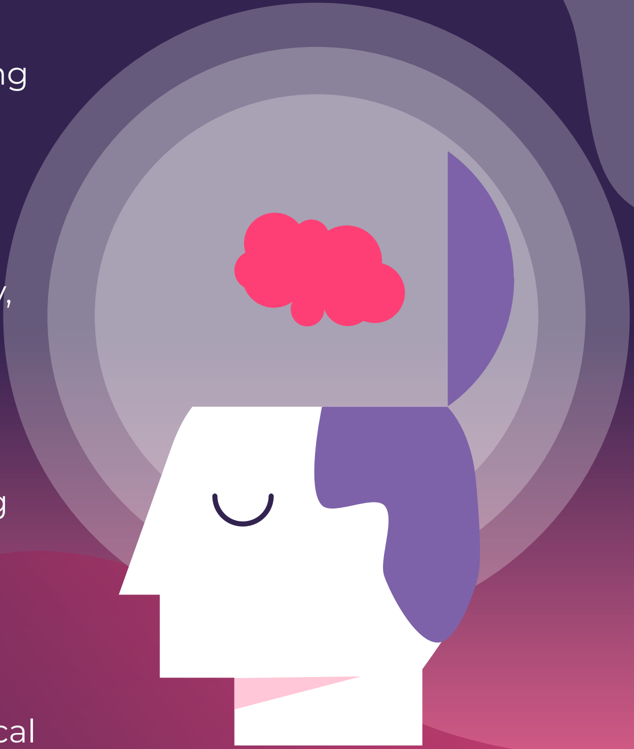
AI can offer accurate diagnosis but lacks transparency, hindering clinical adoption

Proposed Framework

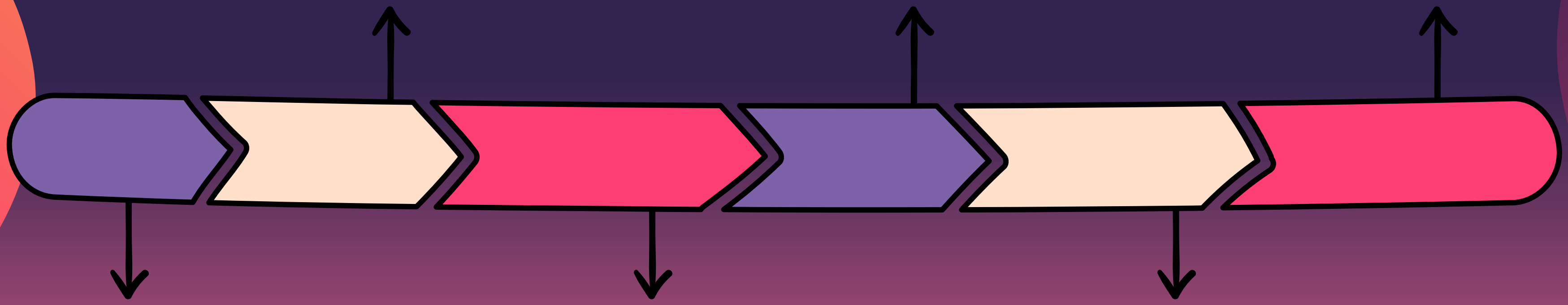
An explainable AI system for AD detection combining accurate classification with visual and textual explanations

Significance

Enhances trust and interpretability in AI-based medical diagnosis, promoting clinical integration



Methodology Overview



Dataset

- Sources: **OASIS-3** and **OASIS-4**
- Filtered for **T1w** and **T2w** MRI modalities
- Selected subjects with **consistent diagnoses**
- **Harmonized features** from both datasets: **demographic** and **clinical**

Preprocessing

- **Orientation** standardization
- **Intensity** scaling
- **Background** removal
- Image **resizing** and **padding**
- **Intensity** normalization
- **Data augmentation**: flipping, rotation, scaling, shifting

Experimental Setting

Experiments

Different data combinations: T1w and T2w channels, demographic and clinical features

Data Split

80% training (20% for validation), 20% testing; stratified by subject to prevent data leakage

Model

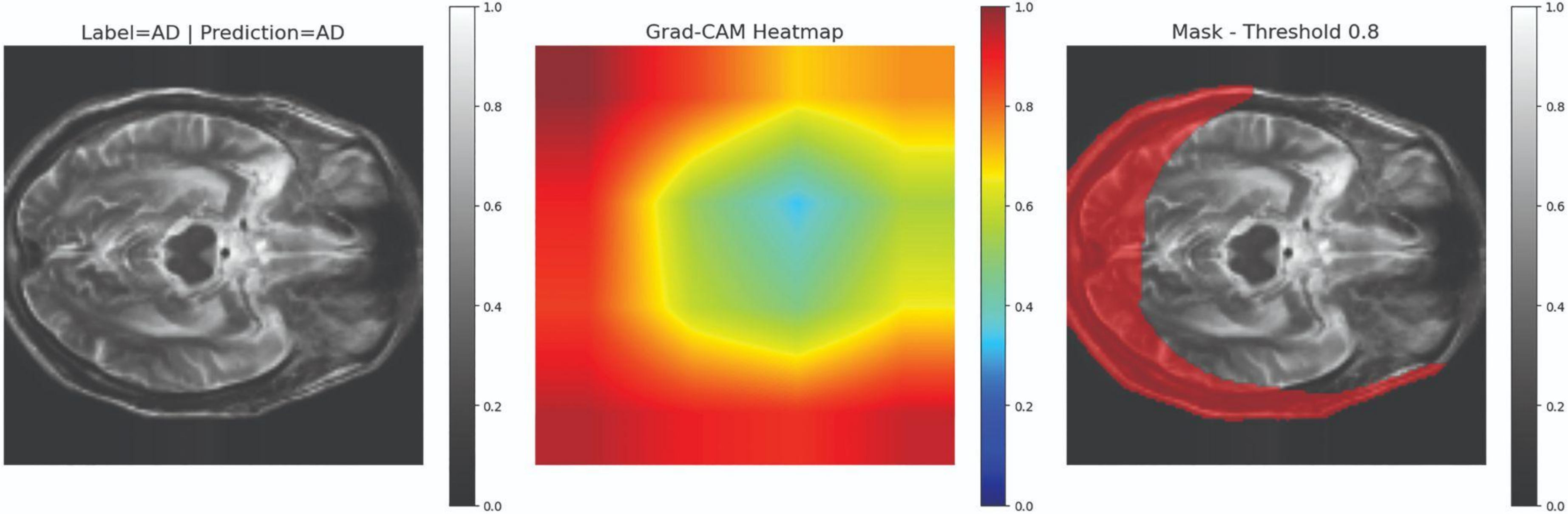
Multimodal adaptation of DenseNet, with separate streams for imaging and non-imaging data

Classification

Binary (AD vs. Cognitively Normal) and ternary (Cognitively Normal, Early-stage AD, AD)

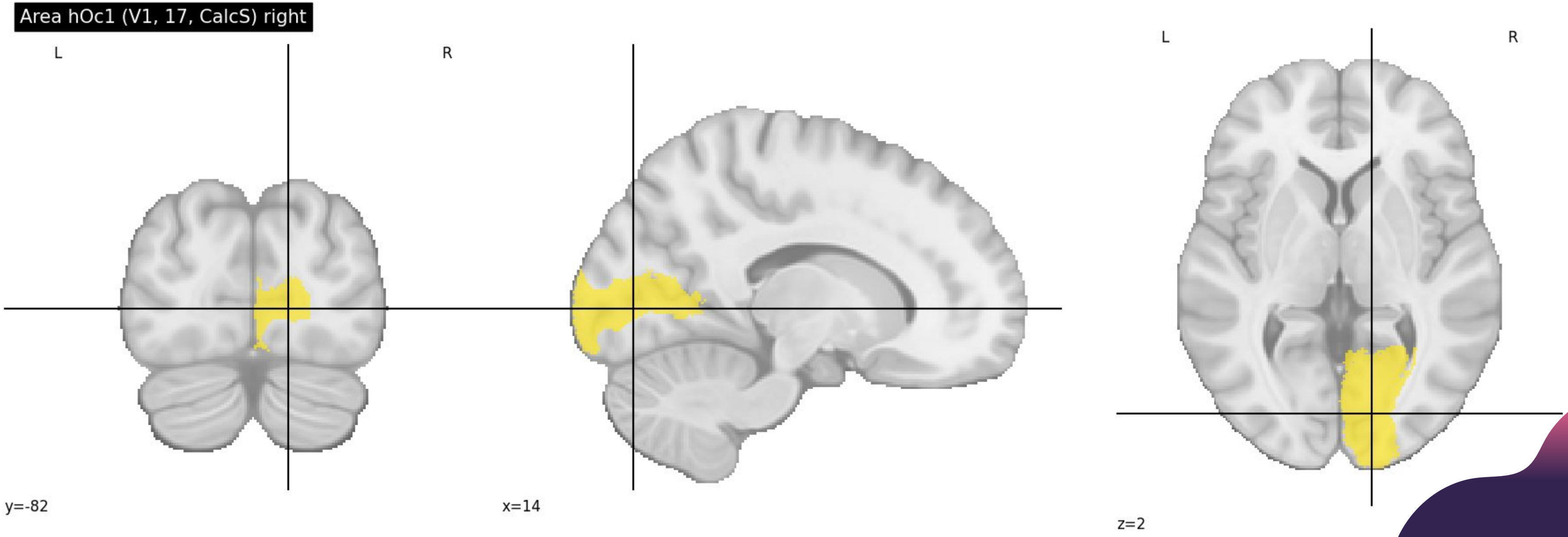
Visual Explainability

Gradient-weighted Class Activation Mapping (Grad-CAM) was used to highlight regions of input MRI scans influencing model predictions. The 3D segmentation mask was computed using a threshold of 0.8



Brain Atlas Mapping

Grad-CAM heatmaps were mapped to specific brain regions using the Julich Brain Atlas. The top 5 brain regions with percentages of heatmap coverage and region volume involvement were identified.



```json
{
  "MRI_Scan": {
    "Disease_Analysis": {
      "Classification_Result": "CLASSIFICATION_RESULT",
      "Features_Used": {
        "age": "Age of the patient",
        "ethnicity": "Ethnicity of the patient"
      },
      "Confidence_Score": {
        "accuracy": "ACCURACY",
        "precision": "PRECISION",
        "recall": "RECALL",
        "F1_score": "F1_SCORE",
        "ROC_AUC": "ROC_AUC"
      },
      "Models_Used": {
        "Classification_Model": "DenseNet",
        "Visual_Explainability_Model": "Grad-CAM"
      },
      "Visual_Heatmap": [
        {
          "Region": "Temporal-to-Parietal (GapMap) left",
          "Percentage_of_Heatmap": 19.22,
          "Percentage_of_Region_Affected": 6.07
        }
      ]
    }
  }
}
```

JSON File Construction

01 Classification Result

02 Feature Description

03 Classification Performance Metrics

04 Employed Models

05 Brain Atlas Mapping Data

Prompt Example

Using the provided JSON data from a brain MRI study, please generate a comprehensive report detailing the visual heatmap analysis for Alzheimer's Disease detection. The regions highlighted by the heatmap are not areas affected by the disease, but rather areas that the classifier model focused on to make its classification decision. Describe these regions according to the Julich-Brain Atlas, noting the percentages of the heatmap within each region and the impact on the region.
Provide explanations in a clear manner that can be easily understood by medical professionals. The goal is to enhance the explainability of the AI disease detection model to support clinical decision-making. While discussing the confidence of the classification based on the performance metrics and the models used, keep the explanations high-level and avoid overly technical details.

As an example, your report might follow this structure:

1. Summary of Heatmap Analysis
    - For each region highlighted in the heatmap:
        - Describe the percentage of the heatmap that covers the region.
        - Discuss the potential reasons why the model might have focused on this region, given its known roles and functions in the brain.
        - Explain the clinical implications of the model's focus on this region, such as how it might encourage further investigation into certain signs of Alzheimer's Disease.
2. Conclusion
    - Summarize the insights gained from the heatmap analysis.
    - Discuss how these insights could enhance clinical decision-making and potentially reveal new aspects of Alzheimer's Disease pathology and diagnosis."

Please remember that the heatmap does NOT describe areas that are affected by the disease, but simply the areas that were the most relevant for the ML model to reach its decision (classification). The specific model used to perform visual explainability should not be revealed to the medical professional in the report.

Classification Results

Channel	Size	Accuracy	Precision	Recall	F1	ROC AUC	Average
T1w + T2w	128	0.7955	0.7279	0.9554	0.8263	0.7925	0.8195
T1w	128	0.7682	0.7402	0.8393	0.7866	0.7669	0.7802
T2w	128	0.8364	0.7923	0.9196	0.8512	0.8348	**0.8469**
T2w	256	0.8227	0.7552	0.9643	0.8471	0.8201	0.8419
T2w + features	128	0.8136	0.7448	0.9643	0.8405	0.8108	0.8348

Channel	Metric	Cognitively normal	Early Stage AD	AD	Average
T2w	Average	0.6723	0.6225	0.2455	0.5134
T2w + features	Average	0.7585	0.5206	0.6088	**0.6293**

Generated Report (LLaMA 3)

Summary of Heatmap Analysis

The heatmap analysis highlights the regions of the brain that the classifier model focused on to make its classification decision for Alzheimer's Disease detection. The regions highlighted are not areas affected by the disease, but rather areas that the model used to make its prediction.

- **Area hOc2 (V2, 18) right:** This region accounts for 22.72% of the heatmap. The Julich-Brain Atlas describes this region as a visual processing area, specifically involved in processing complex visual information. The model's focus on this region may indicate that it is using visual processing features to distinguish Alzheimer's Disease from other conditions. Clinically, this could encourage further investigation into visual processing deficits as a potential sign of Alzheimer's Disease.
- **Area hOc1 (V1, 17, CalcS) right:** This region accounts for 22.2% of the heatmap. This region is also a visual processing area, involved in processing simple visual information. The model's focus on this region may indicate that it is using visual processing features to distinguish Alzheimer's Disease from other conditions. Clinically, this could encourage further investigation into visual processing deficits as a potential sign of Alzheimer's Disease.
- **Temporal-to-Parietal (GapMap) left:** This region accounts for 20.31% of the heatmap. The Julich-Brain Atlas describes this region as involved in processing auditory and visual information, as well as attention and memory. The model's focus on this region may indicate that it is using features related to attention and memory to distinguish Alzheimer's Disease from other conditions. Clinically, this could encourage further investigation into attention and memory deficits as potential signs of Alzheimer's Disease.

Generated Report (LLaMA 3)

- **Area 6mp (SMA, mesial SFG) right:** This region accounts for 9.92% of the heatmap. The Julich-Brain Atlas describes this region as involved in motor control, particularly in the planning and execution of movements. The model's focus on this region may indicate that it is using features related to motor control to distinguish Alzheimer's Disease from other conditions. Clinically, this could encourage further investigation into motor control deficits as potential signs of Alzheimer's Disease.
- **Area hOc3d (Cuneus) right:** This region accounts for 7.57% of the heatmap. The Julich-Brain Atlas describes this region as involved in processing visual information, particularly in the processing of complex shapes and patterns. The model's focus on this region may indicate that it is using visual processing features to distinguish Alzheimer's Disease from other conditions. Clinically, this could encourage further investigation into visual processing deficits as a potential sign of Alzheimer's Disease.

Conclusion

The heatmap analysis provides insights into the regions of the brain that the classifier model used to make its classification decision for Alzheimer's Disease detection. The highlighted regions are not areas affected by the disease, but rather areas that the model used to make its prediction. The model's focus on visual processing areas, attention and memory processing areas, and motor control areas may indicate that these features are important for distinguishing Alzheimer's Disease from other conditions. Clinically, these findings could encourage further investigation into these areas as potential signs of Alzheimer's Disease. The confidence of the classification decision is supported by the performance metrics, including an accuracy of 81.36% and an F1-score of 84.05%. These findings could enhance clinical decision-making and potentially reveal new aspects of Alzheimer's Disease pathology and diagnosis.

Generated Report (LLaMAntino 3)

Sintesi dell'analisi della heatmap

La nostra analisi della heatmap ha rilevato che il modello di classificazione si è concentrato su cinque aree cerebrali specifiche. Queste aree non sono colpite dalla malattia di Alzheimer, ma sono state le più rilevanti per il modello per prendere la sua decisione di classificazione.

- **Area hOc2 (V2, 18) right**: Questa area occupa il 22,72% della heatmap e solo il 10,74% della sua superficie è coperta dalla regione cerebrale in questione. L'area hOc2 (V2, 18) è coinvolta nella elaborazione visiva e nella rappresentazione spaziale. Il fatto che il modello si sia concentrato su questa area potrebbe suggerire che i cambiamenti nella funzionalità visiva o nella rappresentazione spaziale possano essere indicatori preclini di Alzheimer.
- **Area hOc1 (V1, 17, CalcS) right**: Questa area occupa il 22,2% della heatmap e solo il 6,48% della sua superficie è coperta dalla regione cerebrale in questione. L'area hOc1 (V1, 17, CalcS) è coinvolta nella elaborazione visiva e nella percezione dei dati sensoriali. La focalizzazione del modello su questa area potrebbe suggerire che i problemi nella percezione visiva o nella elaborazione dei dati sensoriali possano essere indicatori preclini di Alzheimer.
- **Temporal-to-Parietal (GapMap) left**: Questa area occupa il 20,31% della heatmap e solo il 1,31% della sua superficie è coperta dalla regione cerebrale in questione. L'area Temporal-to-Parietal (GapMap) è coinvolta nella elaborazione dei dati sensoriali, in particolare nell'elaborazione degli stimoli auditivi e tattili. La focalizzazione del modello su questa area potrebbe suggerire che i problemi nella integrazione dei dati sensoriali possano essere indicatori preclini di Alzheimer.

Generated Report (LLaMAntino 3)

- **Area 6mp (SMA, mesial SFG) right:** Questa area occupa il 9,92% della heatmap e il 16,75% della sua superficie è coperta dalla regione cerebrale in questione. L'area 6mp (SMA, mesial SFG) è coinvolta nella pianificazione motoria e nella gestione dei movimenti. La focalizzazione del modello su questa area potrebbe suggerire che i problemi nella pianificazione motoria o nella gestione dei movimenti possano essere indicatori preclini di Alzheimer.
- **Area hOc3d (Cuneus) right:** This region accounts for 7.57% of the heatmap. The Julich-Brain Atlas describes this region as involved in processing visual information, particularly in the processing of complex shapes and patterns. The model's focus on this region may indicate that it is using visual processing features to distinguish Alzheimer's Disease from other conditions. Clinically, this could encourage further investigation into visual processing deficits as a potential sign of Alzheimer's Disease.

Conclusione

La nostra analisi della heatmap ha rilevato che il modello di classificazione si è concentrato su aree cerebrali specifiche, che non sono colpite dalla malattia di Alzheimer, ma sono state le più rilevanti per il modello per prendere la sua decisione di classificazione. Queste aree potrebbero essere indicatori preclini della malattia, suggerendo che i cambiamenti nella funzionalità visiva, nella percezione, nella integrazione dei dati sensoriali, nella pianificazione motoria e nella gestione dei movimenti possano essere indicatori preclini di Alzheimer.

Queste intuizioni potrebbero migliorare il processo decisionale clinico, facendo emergere nuovi aspetti della patologia e della diagnosi della malattia di Alzheimer.

Textual Explainability Metrics

- **Type-Token Ratio:** Measure of **lexical diversity**.
- **Maas' Index:** Adaptation of the **Type-Token Ratio** designed to align the measure with a **logarithmic scale**. A lower score signifies greater diversity in vocabulary.
- **Flesch Reading Ease (English Readability):** Evaluates the ease of reading a given text based on factors such as **word length** and **sentence complexity**.
- **Gulpease Index (Italian Readability):** Equivalent of Flesch Reading Ease for the Italian language.
- **Coherence Score (CohS):** Quantifies the coherence and flow of the generated textual explanations, assessing the logical structure and transitions between ideas within the output.
- **Embedding-based Coverage Score (ECS):** Measures the coverage of information in the LLM's output by comparing the **embeddings** of the **sentences** in the **prompt** and the corresponding **output**.
- **Token-based Coverage Score (TCS):** Quantifies the **overlap** between the **tokens** in the **prompt** and the **output**. It aims to evaluate how well the LLM incorporated **relevant information** from the prompt into its generated output.

Textual Explainability Evaluation

	TTR	Maas'	Readability	CohS	ECS	TCS	Time
BioMistral	**0.86**	**0.02**	43.32	0.09	0.33	0.04	**24.34**
LLaMA 3	0.3	0.03	42.22	0.37	0.44	0.21	515.55
Mistral	0.32	0.03	46.95	0.3	0.33	0.18	483.88
LLaMAntino	0.62	0.02	**70.3**	0.35	0.4	0.07	297.53
LLaMAntino 3	0.38	0.03	43.6	**0.4**	**0.48**	0.18	516.08
Minerva	0.14	0.05	17.18	0.27	0.33	**0.28**	392.44

Heuristic Evaluation

English (prompts 1 and 2)	LLaMA	Mistral	BioMistral
Query understanding	✓	✓	✗
Context understanding	✓	✓	✗
Response Length	✓	✓	✗

Italian (prompt 2)	LLaMA	Mistral	LLaMAntino	Minerva
Query understanding	✓	✗	✓	✗
Context understanding	✓	✗	✓	✗
Response Length	✓	✓	✓	✗

Conclusion

Key Results

- Accurate **AD detection** from **3D MRI** scans
- Interpretable **visual** and **textual explanations**
- Integration of **deep learning**, **Grad-CAM**, and **LLMs**
- Comprehensive **evaluation** framework

Significance

- Enhances **transparency** and **trust** in AI-assisted diagnosis
- Facilitates informed **decision-making** in healthcare
- Contributes to wider adoption of **AI in medicine**

Future Work

- **Multimodal** data integration (PET, CSF, genetic data)
- Exploration of other **XAI techniques**
- **Clinical validation** studies
- Expansion to **other medical domains**
- Automatic **dataset annotation**