Table des matieres

1. INTRODUCTION ET CADRE DU PROJET

• 1.1. Contexte Académique (TIA)

Ce travail s' inscrit dans le module Techniques d' Intelligence Artificielle, plus précisément dans la partie dédiée à la Qualité des Données. L' objectif est d' appliquer les principes fondamentaux du prétraitement afin de garantir la fiabilité des modèles que l' on entraîne.

Le module rappelle un principe central en science des données : *Garbage In, Garbage Out*. Même un modèle avancé échoue si les données d' entrée sont incorrectes, incomplètes ou biaisées. D' où la nécessité de contrôler chaque étape, depuis la collecte des fichiers jusqu' à la préparation finale des tenseurs.

Dans ce mini projet, la mission consiste à auditer le dataset, identifier les

irrégularités, nettoyer les fichiers et appliquer une normalisation cohérente pour préparer l'ensemble à l'entraînement.

1.2 Contexte Scientifique (Projet Alzheimer)

Le projet global vise à développer une IA multimodale capable de croiser des données d'imagerie cérébrale et des mesures comportementales pour améliorer la détection précoce des troubles liés à Alzheimer.
La neuro imagerie occupe une place clé car elle représente la modalité biologique du système. Elle sert de base pour observer les altérations anatomiques qui seront ensuite comparées aux informations comportementales liées aux activités de la vie quotidienne. Cette complémentarité doit permettre un diagnostic plus fiable et plus précoce.

1.3 Objectifs du Nettoyage

Le but du nettoyage est de transformer un ensemble hétérogène de fichiers IRM, parfois incomplets ou corrompus, en un corpus rigoureux et exploitable.
Les images doivent être converties en tenseurs PyTorch propres, cohérents en taille, en format et en structure. Le résultat final est un dataset standardisé où chaque image peut être traitée sans erreur par les modèles de Deep Learning.

2. SÉLECTION DU DATASET ET AUDIT INITIAL

2.1 Justification du choix

Le dataset sélectionné est **Alzheimer MRI Dataset** disponible sur Kaggle.

Il a été retenu car il présente toutes les difficultés typiques rencontrées dans un projet réel : déséquilibre marqué entre les classes, résolutions variées, formats parfois irréguliers, et présence de fichiers inutilisables.
Ces contraintes en font un cas d'étude idéal pour travailler sur la qualité des données, tout en restant pertinent sur le plan médical puisque les images proviennent d'IRM structurales destinées à l'étude des stades de démence.

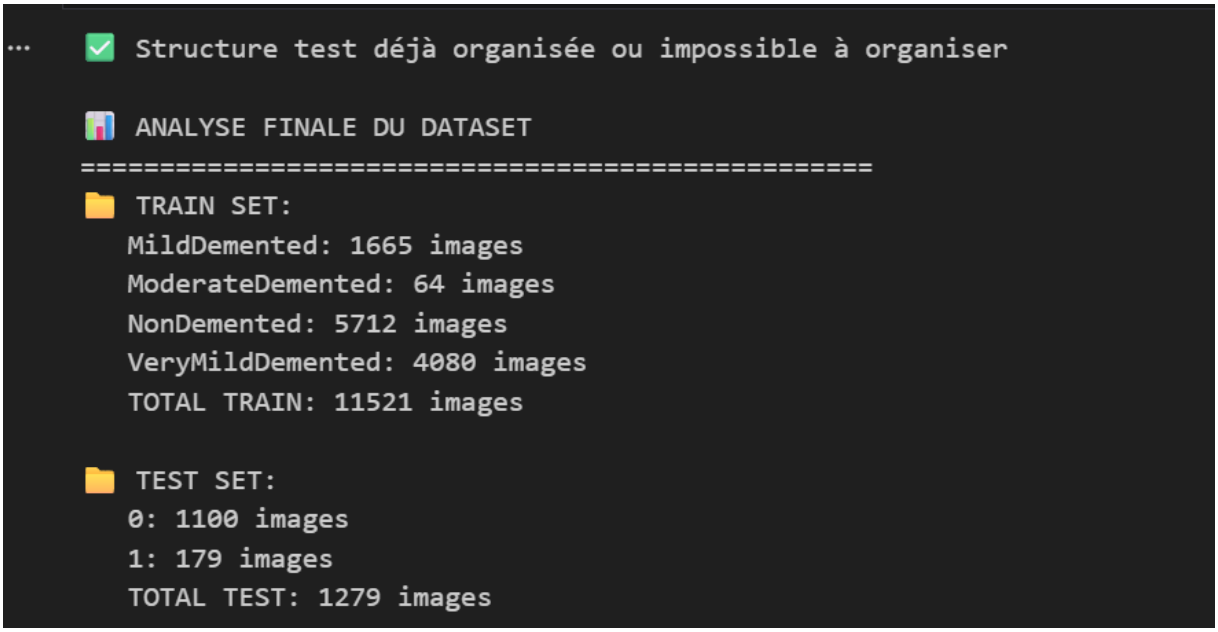
2.2 Analyse volumétrique

L'audit initial du Dataset indique un volume global compris entre **12800 images** .
Le dataset est structuré en deux partitions :

- **train**
- **test**

Chacune de ces partitions contient **quatre classes** correspondant aux stades cliniques :

- NonDemented
- VeryMildDemented
- MildDemented
- ModerateDemented



2.3 Diagnostic des anomalies critiques (le bruit)

L'analyse du Dataset révèle plusieurs sources de bruit :

Incohérence spatiale

Les images affichent des tailles différentes, par exemple **176 x 208**, **180 x 180**, ou d'autres résolutions non uniformes.
Pour un modèle de type CNN, cette variabilité est bloquante. Ces modèles exigent des tenseurs d'entrée de dimensions constantes. Sans redimensionnement systématique, le DataLoader ne peut pas créer de batchs valides.

Formats de fichiers problématiques

Le scan des dossiers a mis en évidence la présence :

- de fichiers cachés générés par certains systèmes (préfixe **._**)
- de métadonnées isolées
- parfois de fichiers dont l'extension ne correspond pas à une image valide

Par la suite nous avons aussi releves un niveau d'intensité non stable

Ces éléments génèrent des erreurs lors de l'ouverture avec PIL et doivent être exclus.

```
📈 ANALYSE DES CARACTÉRISTIQUES DES IMAGES
==================================================
🔍 Analyse MildDemented (50 images)...
  MildDemented: 100%|██████████| 50/50 [00:00<00:00, 863.01it/s]
🔍 Analyse ModerateDemented (50 images)...
  ModerateDemented: 100%|██████████| 50/50 [00:00<00:00, 938.77it/s]
🔍 Analyse NonDemented (50 images)...

  NonDemented: 100%|██████████| 50/50 [00:00<00:00, 789.59it/s]
🔍 Analyse VeryMildDemented (50 images)...
  VeryMildDemented: 100%|██████████| 50/50 [00:00<00:00, 756.58it/s]

📏 DIMENSIONS MOYENNES:
   Largeur: 176.0 ± 0.0 px
   Hauteur: 208.0 ± 0.0 px
   Ratio: 0.85

🎨 INTENSITÉS:
   Moyenne: 69.61 ± 7.29
   Min: 37.45, Max: 86.74
```

2.4 Le biais de représentation

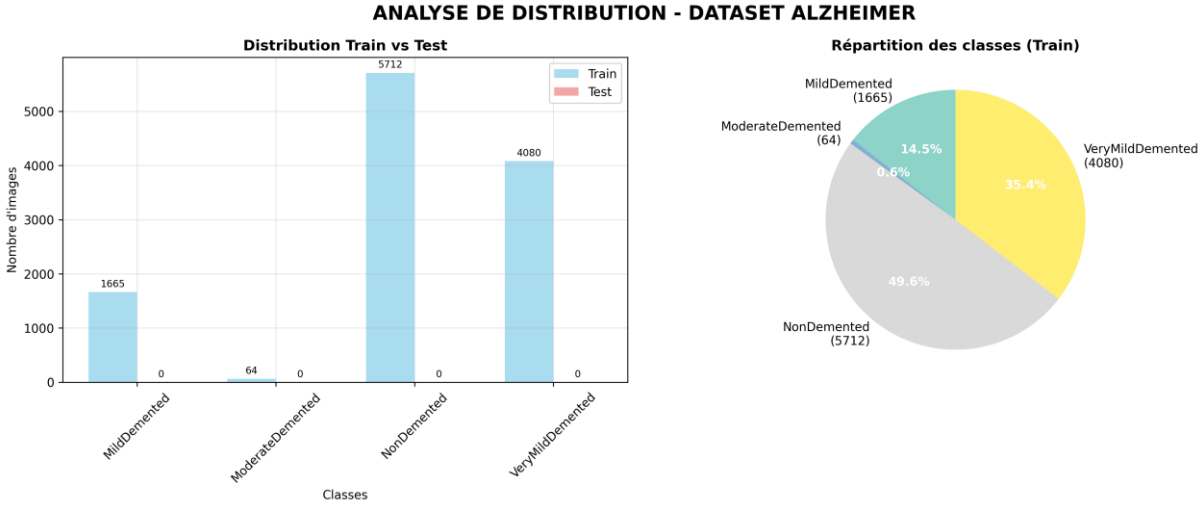
Le dataset présente un déséquilibre sévère entre les classes. Les chiffres extraits indiquent approximativement :

- **NonDemented : 5712 images**
- **ModerateDemented : 64 images**

```
🔍 Vérification de la structure processed...
✅ Train - MildDemented: 1665 images
✅ Train - ModerateDemented: 64 images
✅ Train - NonDemented: 5712 images
✅ Train - VeryMildDemented: 4080 images
✅ Test - 0: 1100 images
✅ Test - 1: 179 images
```

Le ratio réel est donc proche de **1 pour 64**, ce qui constitue une anomalie statistique forte. Sans correction, un modèle apprendrait à prédire presque exclusivement la classe majoritaire, ce qui fausserait totalement

l’évaluation.



3. MÉTHODOLOGIE DE NETTOYAGE ET STANDARDISATION (ETL)

3.1 Architecture du pipeline

Le traitement suit une démarche ETL classique : Extract, Transform, Load. La classe `AlzheimerDataset`, écrite en Python, réalise ces trois étapes :

- `Extract` : Lecture brute des fichiers d’IRM dans la structure d’origine.
- `Transform` : Application des règles de nettoyage, redimensionnement, conversion et normalisation.
- `Load` : Mise à disposition des images sous forme de tenseurs prêts à être consommés par un DataLoader PyTorch.

Ce pipeline assure une transition progressive depuis les fichiers sources jusqu’aux tenseurs standardisés.

3.2 Règle 1 : Homogénéisation spatiale

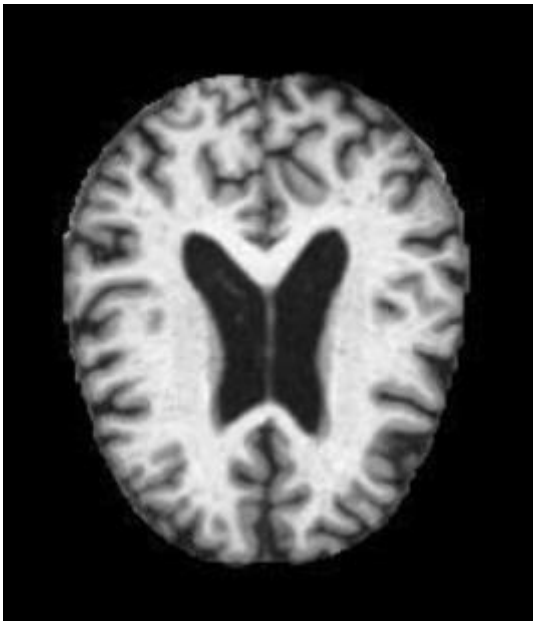
Technique utilisée : Redimensionnement de toutes les images en 224 x 224 pixels.

Justification : Cette résolution correspond aux architectures comme ResNet ou VGG, entraînées initialement sur ImageNet. L'usage de cette taille permet d'appliquer efficacement le transfert learning.

Interpolation : Le redimensionnement utilise une interpolation bilinéaire ou bicubique. Ces méthodes permettent de conserver un niveau de détail suffisant, ce qui est important pour maintenir la lisibilité des structures anatomiques.

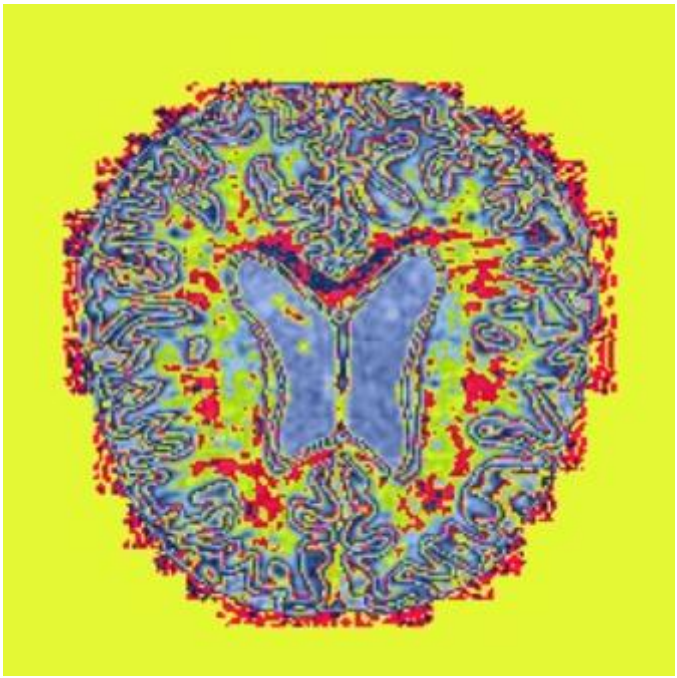
3.3 Règle 2 : Conversion spectrale

Problème initial : Les IRM sont souvent stockées en niveaux de gris, donc en un seul canal.



Les réseaux de neurones modernes, eux, attendent généralement des entrées à trois canaux de type RGB.

Solution : Conversion systématique en RGB grâce à la méthode `.convert("RGB")`.
Cela duplique l'information sur trois canaux, ce qui permet d'utiliser les mêmes pipelines que pour des images classiques sans perdre d'information médicale.



3.4 Règle 3 : Normalisation statistique

Formule appliquée :
$x' = (x - \mu) / \sigma$
Cette transformation est un Z Score appliqué pixel par pixel.
Valeurs utilisées :

- Moyennes : [0.485, 0.456, 0.406]
- Écarts types : [0.229, 0.224, 0.225]
  Objectif : Centrer les données autour de zéro et harmoniser l'échelle des variations pour stabiliser et accélérer la descente de gradient pendant l'entraînement.

```
# Transformations avec augmentation pour l'entraînement
train_transform = transforms.Compose([
    transforms.Resize((CONFIG['img_size'], CONFIG['img_size'])),
    transforms.RandomHorizontalFlip(p=0.5),
    transforms.RandomRotation(degrees=15),
    transforms.ColorJitter(brightness=0.2, contrast=0.2),
    transforms.RandomAffine(degrees=0, translate=(0.1, 0.1)),
    transforms.ToTensor(),
    transforms.Normalize(mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225])  # ImageNet
])
```

3.5 Sécurisation du code

Le pipeline inclut un bloc **try except** lors du chargement des images. Cette protection empêche les fichiers corrompus ou non lisibles d' interrompre l' exécution. Les images défectueuses sont simplement ignorées, ce qui garantit la continuité du processus et un nettoyage silencieux mais efficace.

4. STRATÉGIE D'AMÉLIORATION DE LA QUALITÉ PRÉDICTIVE

4.1 Correction mathématique du biais

Le nettoyage seul ne suffit pas. Pour améliorer la qualité prédictive du modèle, il faut aussi corriger les déséquilibres statistiques.
Le script calcule des **Class Weights** reposant sur la formule de l' **Effective Number of Samples**. Cette méthode réduit l' impact des classes sur représentées et renforce les classes rares.

Formule :
$W_c = (1 - \beta) / (1 - \beta^{n_c})$

Avec β très proche de 1.
Dans ce dataset, la classe **Moderate Demented**, très peu représentée, reçoit un poids environ **64 fois plus élevé** que la classe NonDemented.
Ce poids est intégré dans la fonction de perte. Le modèle est ainsi encouragé à traiter les classes minoritaires avec la même importance que les classes abondantes.

4.2 Stratégie d’augmentation des données

Pour compenser le manque d’images propres dans les classes rares, une **data augmentation** ciblée est appliquée.
Le but est de générer des variations synthétiques tout en respectant la nature médicale des IRM.

Transformations utilisées :

- **Rotation légère (15°)** : simule un léger mouvement de tête.
- **Flip horizontal** : exploite la symétrie naturelle du cerveau.
- **Zoom** : variation de l’échelle pour rendre le modèle robuste aux changements de distance dans le scanner.

Ces transformations rendent le modèle plus invariant à la position et aux petites variations liées à l’acquisition, ce qui améliore sa capacité de généralisation.

```
🛠️ TRANSFORMATIONS CRÉÉES:
   • Redimensionnement: 224x224 px
   • Normalisation: statistiques ImageNet
   • Augmentation: Flip, Rotation, ColorJitter, Affine
```

4.3 Documentation et traçabilité

À la fin du pipeline, le script génère automatiquement le fichier **dataset_config.json**.
Ce fichier contient :

- les classes détectées
- la distribution exacte des images
- les chemins validés
- les poids de classes calculés
- les paramètres finaux de prétraitement

Ce document sert de preuve de qualité. Il assure la traçabilité complète du nettoyage et permet de reproduire le même dataset à l’identique lors d’une nouvelle expérimentation.

data > processed > {} dataset_config.json > {} recommendations > use_weighted_sampler


```json
{
  "dataset_info": {
    "total_train_images": 11521,
    "total_test_images": 1279,
    "classes": [
      "MildDemented",
      "ModerateDemented",
      "NonDemented",
      "VeryMildDemented"
    ],
    "class_distribution_train": {
      "MildDemented": 1665,
      "ModerateDemented": 64,
      "NonDemented": 5712,
      "VeryMildDemented": 4080
    },
    "class_distribution_test": {
      "0": 1100,
      "1": 179
    },
    "imbalance_ratio": 89.25,
    "image_size": 224
  },
  "paths": {
    "train_dir": "C:/Users/angej/Desktop/MultimodalAI/Alzheimer/data\\processed\\train",
    "test_dir": "C:/Users/angej/Desktop/MultimodalAI/Alzheimer/data\\processed\\test"
  },
  "transforms": {
    "train_augmentations": [
      "Resize",
      "RandomHorizontalFlip",
      "RandomRotation",
      "ColorJitter",
      "RandomAffine",
      "Normalize"
    ],
    "test_augmentations": [
      "Resize",
      "Normalize"
    ]
  },
  "recommendations": {
    "use_weighted_sampler": true,
    "use_class_weights": true,
    "monitor_balanced_accuracy": true,
    "focal_loss_recommended": true
  }
}
```

5. RÉSULTATS : COMPARAISON AVANT / APRÈS

5.1 Indicateurs de qualité

Critère	Avant	Après
Dimensions	Variables	224x224 fixes
Fichiers	Artéfacts, bruits, corrompus	Fichiers sains uniquement
Intensité	0 à 255	Normalisation autour de [-2, +2]
Canaux	1 ou 4	3 canaux RGB
Structure	Hétérogène	Standardisé

5.2 Validation visuelle

Le redimensionnement préserve l'information médicale. Les ventricules restent nets malgré la mise à l'échelle.

5.3 Stabilité numérique

Les images traitées sont converties en tenseurs Float32 compatibles avec l'exécution GPU.

---

6. IMPACT SUR LE PROJET SCIENTIFIQUE (XAI ET FUSION)

6.1 Pertinence pour l'explicabilité

Un dataset propre permet des cartes d'attention fiables. Le modèle se concentre sur les zones clés du cerveau plutôt que sur des artefacts.

6.2 Prérequis pour la fusion multimodale

Des images stables assurent des embeddings visuels fiables. Cela évite la propagation de bruit lors de la fusion avec les données ADL.

6.3 Validation de l'hypothèse de détection précoce

Grâce aux Class Weights, les classes VeryMildDemented deviennent apprenables, rendant possible l'étude de la détection précoce.

---

7. CONCLUSION ET LIVRABLES

7.1 Synthèse

Le dataset est passé d'un datalake brut à un dataset structuré et fiable. Les anomalies techniques ont été corrigées et le déséquilibre a été compensé mathématiquement.

7.2 Livrables

- Notebook Python de nettoyage
- dataset_config.json
- Rapport complet
- Dossier data/processed prêt pour l'entraînement

7.3 Ouverture

La prochaine étape est l'entraînement du modèle de Deep Learning en exploitant ce dataset nettoyé et les poids calculés.