**Analyse de radiographies pulmonaires Covid-19**

**Rapport de projet n°1 : Exploration, Data visualisation et pre-processing**

**17/11/2023**

# Introduction au projet

## Contexte

L’imagerie médicale est un outil de diagnostic majeur dans la plupart des champs de la médecine : radiographie, échographie, IRM, scanner… Ces outils produisent une quantité importante de données sous forme d’images, que les professionnels de santé interprètent à l'œil ou aidés de logiciels. Le développement de la Computer Vision a permis de faire grandement avancer l’analyse des images médicales afin d’améliorer les diagnostics et de faire gagner du temps aux praticiens.

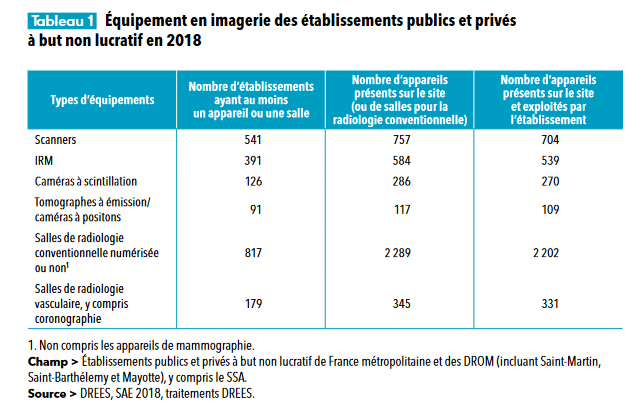
La pandémie de COVID-19 a surpris le monde entier par sa vitesse de propagation et sa viralité. Les professionnels de santé ont dû faire face à un flux important de malades, et au besoin de diagnostiquer et d’orienter rapidement les cas positifs vers les bons traitements. En France en particulier, la tension sur les effectifs a été particulièrement forte, dans un contexte préexistant de pénurie de médecins généralistes et dans certaines spécialités.

L’analyse des données d’imagerie médicale, afin d’établir un diagnostic rapide et fiable, a été explorée par plusieurs équipes de chercheurs autour du monde (comme le Vision and Image Processing Research Group, Université de Waterloo au Canada).

La radiographie, le scanner et l’IRM sont trois techniques d’imagerie médicale dites “non-invasives” et indolores, car elles ne nécessitent pas d’opération. La radiographie et le scanner utilisent des rayons-X, tandis que l’IRM utilise des ondes électromagnétiques.

La radiographie a comme avantages sur les autres techniques de s’appuyer sur une infrastructure plus légère donc moins coûteuse, et plus ancienne donc plus répandue dans les établissements médicaux du monde entier.

D’après une étude de la DREES[[1]](#footnote-1) (Direction de la Recherche, des Études, de l'Évaluation et des Statistiques du Ministère de la Santé et de la Prévention), la France comptait en 2018 **2289 salles de radiologie conventionnelle** en établissements publics ou privé non lucratif**, contre 757 scanners** et **584 IRM.**

****

La radiographie consiste à utiliser des rayons X qui traversent le corps humain et sont capturés sur un film électronique. Les rayons X sont plus ou moins absorbés en fonction de la densité des tissus, et les images permettent au radiologue de visualiser par contraste l’intérieur du corps humain.

Les images numérisées issues d’appareils de radiographies peuvent être exploitées grâce aux techniques de Computer Vision. Développer un modèle de classification fiable permettrait d’utiliser cette technique pour faire gagner du temps dans le diagnostic du Covid.

Aucun de nous trois n’est un professionnel de santé, ni n’avons de formation à l’interprétation de radiologies. Nous allons donc nous en remettre à un ou plusieurs modèles de deep learning pour essayer de distinguer des traits caractéristiques dans les données, et d’établir une classification fiable.

Nous avons choisi ce projet pour le sens qu’il porte : améliorer les techniques de diagnostic afin d'accélérer la prise en charge des patients. De plus, plusieurs membres de l’équipe ont une expérience passée dans le domaine de la santé.

## Objectifs

Le projet vise à développer un modèle de Machine Learning permettant de classer de manière fiable une radiographie pulmonaire selon quatre diagnostics :

1. Patient sain (“Normal”)
2. Patient atteint du Covid-19
3. Patient atteint de pneumopathie virale
4. Patient diagnostiqué avec une opacité pulmonaire (qui regroupe plusieurs types d'infections pulmonaires non-COVID).

En terme d’expertise autour de la problématique adressée :

* Steve Costalat:
* 20 ans d'expérience des environnements cliniques: bloc opératoires, service d'urgence et service de réanimation
* Expertise des problématique de protocoles de diagnostic en environnement critique.
* Nicolas Gorgol n’a aucune expérience préalable sur le sujet.
* Thibaut Gazagnes n’a aucune expérience préalable sur le sujet.

Plusieurs projets open-source ont été développés depuis 2020 sur ce sujet, comme par exemple :

* COVID-NET développé par DarwinAI Corp., Canada and Vision and Image Processing Research Group, University of Waterloo, Canada <https://arxiv.org/abs/2003.09871> / <https://alexswong.github.io/COVID-Net/>

Nous n’avons pas à ce stade utilisé ces projets préexistants.

# Compréhension et manipulation des données

**Cadre**

Le jeu de données utilisé est une banque d’images de radiographies pulmonaires, consolidée par une équipe de chercheurs de l'université du Qatar, à Doha, et de l'université de Dhaka (Bangladesh). Cette banque d’image est en accès libre sur Kaggle[[2]](#footnote-2).

Elle a été constituée à partir de bases existantes (comme l’*Italian Society of Medical and Interventional Radiology (SIRM) COVID-19 DATABASE [1], Novel Corona Virus 2019 Dataset developed by Joseph Paul Cohen and Paul Morrison, and Lan Dao in GitHub)* et de plusieurs dizaines de publications scientifiques.

Chaque image est fournie avec un masque de segmentation, qui permet d’isoler la zone pulmonaire et de masquer les parties de l’image radio non pertinentes pour le diagnostic.

Cette base de données est à usage académique et non commercial.

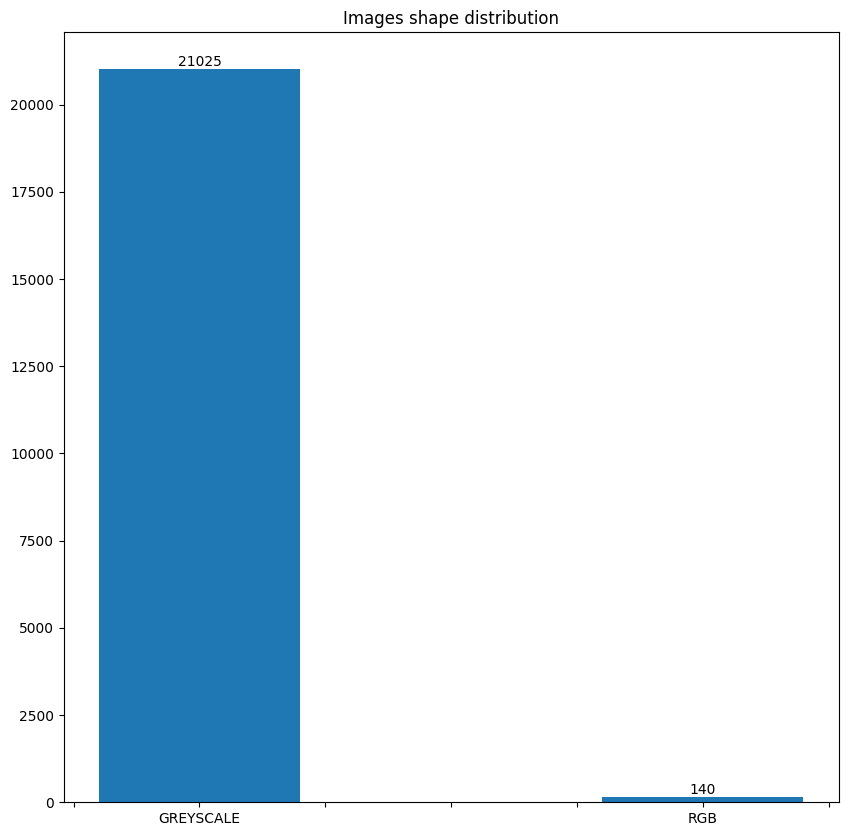
Une première analyse des images et de leurs attributs montre que le jeu de données semble de qualité homogène (tailles, résolutions, orientations, …).

## Description du dataset : volume et format

Le set contient **21 165 images** de radiographies pulmonaires avec leurs masques associés. Les images de radios pulmonaires sont au format PNG, de taille 299\*299 pixels en uint8. Elles sont au mode Grayscale pour plus de 99% d’entre elles, et RGB pour 140 images. Après vérification ces images en RGB ont 3 trois canaux identiques, nous les convertirons donc en Grayscale dans le pré-processing.

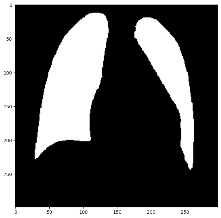


Affichage de 3 images au hasard dans le dataset



Répartition des images par mode (RBG/Grayscale) dans le dataset (images radio)

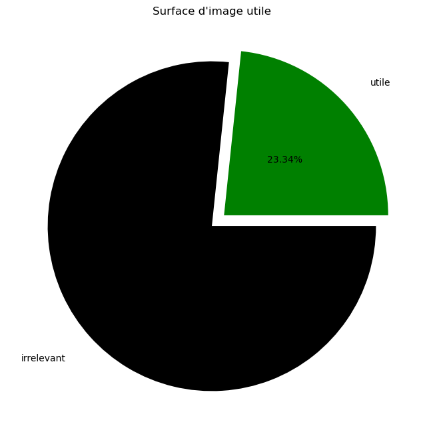
Les masques fournis sont au format PNG également mais de dimensions 256\*256. Ils permettent d’isoler la surface pulmonaire sur les images. Ils vont nous permettre de limiter les données transmises au modèle après le preprocessing. Comme leur taille est légèrement différente des images, nous les redimensionnerons et les convertirons en Grayscale dans le pré-processing.



Exemple de masque

## Surface d’image utile

Grâce aux masques, nous avons pu calculer la **surface moyenne d’image utile** pour le modèle (“utile” étant la partie de l’image qui reste visible après application du masque). Celle-ci s’élève à 23%, ce qui nous incite à utiliser les masques afin de réduire le volume d’information inutile en entrée du modèle.



Nous avons également affiché la distribution de la surface d’image utile sur l’ensemble du dataset :

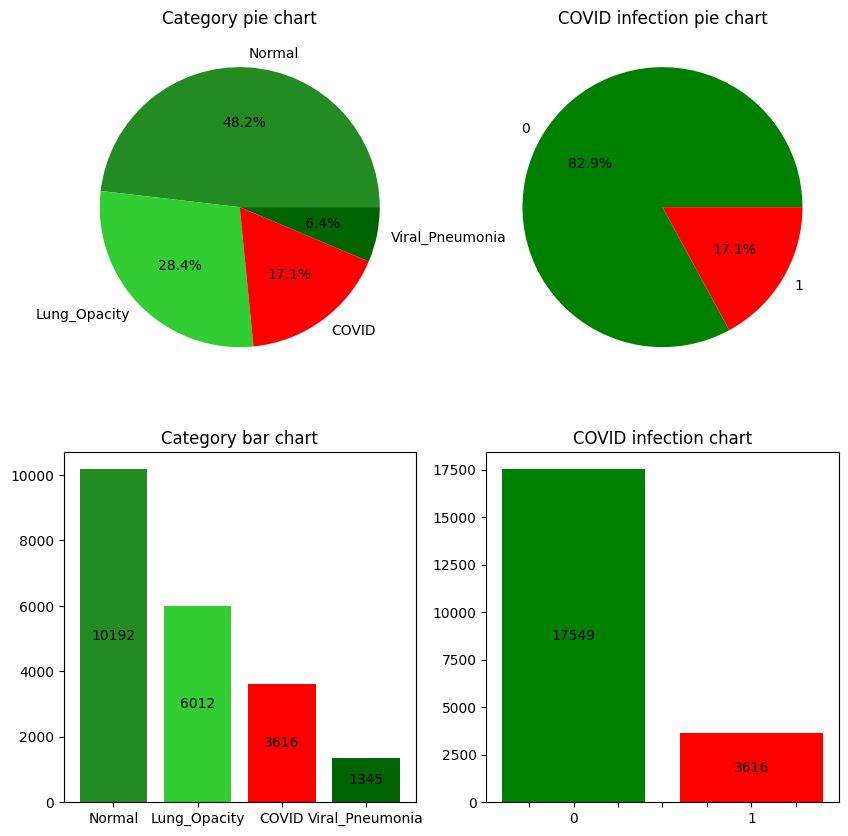
## 

## 

## Equilibre des classes

Le dataset n’est pas équilibré : la catégorie “Normal” (porteur sain) représente 48% des images, l’opacité pulmonaire 28%, le Covid 17%, et la pneumonie virale 6%.

La classe COVID qui est l’objet de la détection représente donc environ ⅙ des images du set (3 616 sur 21 165). Il faudra donc rééquilibrer les données utilisées pour l’entraînement des modèles afin de ne pas biaiser sa précision.



Les images sont accompagnées d’un fichier “metadata” pour chaque set (Normal, Covid, etc) qui résume le nom du fichier, ses dimensions, et l’url du jeu de données source. L’analyse des propriétés des images sur Python nous a permis d’identifier des erreurs dans le fichier metadata (par exemple, format 256\*256 alors que les images sont en 299\*299).

## Sources des images

La répartition des sources montre que les images sont issues de 8 sources, et que plus de 70% des images proviennent d’un jeu de données préexistant utilisé pour la détection de pneumonies sur Kaggle.

## 

Répartition des images du dataset par source

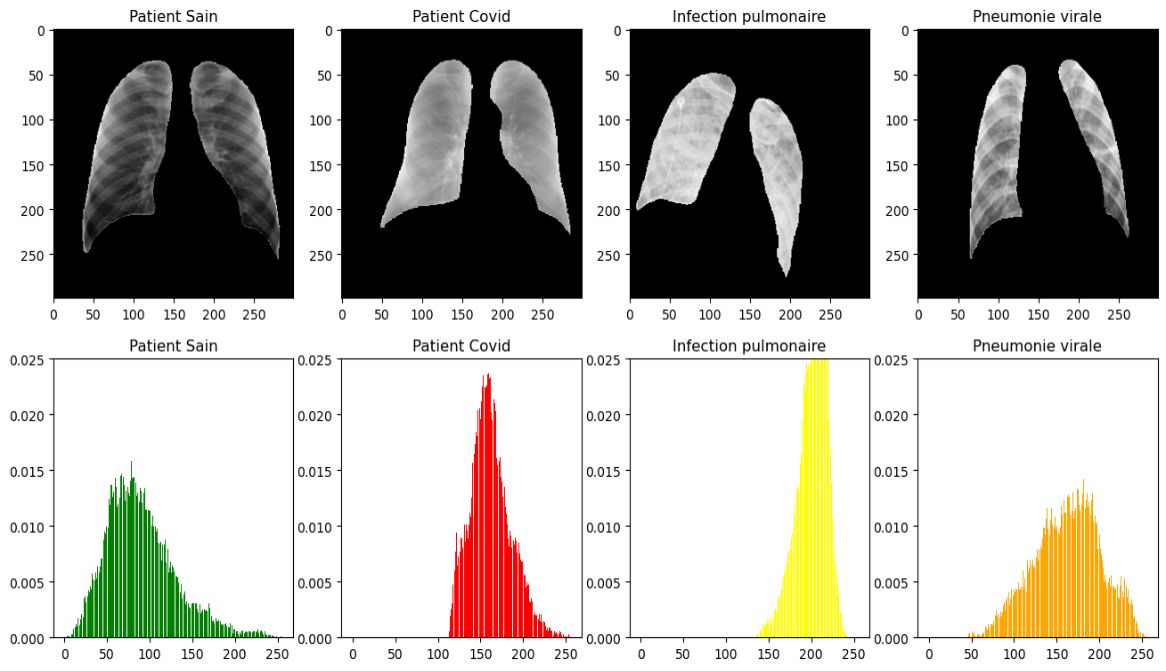
## Histogrammes de distribution des pixels

Nous avons également calculé et affiché l’histogramme de la distribution des pixels de chaque image en niveaux de gris :

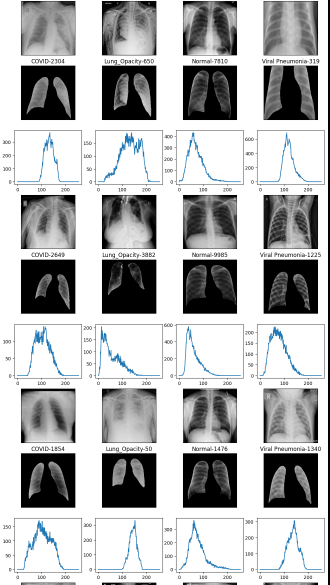
## 

Exemple d’histogramme de répartition des pixels par intensité (entre 0 et 255) pour une image aléatoire

Nous avons également affiché aléatoirement une ou plusieurs images par set, avec sa distribution associée. Cela montre des différences entre les distributions qui pourraient être liées à des différences d’intensité dans les radios en fonction de la maladie. Nous ne pouvons pas vérifier cette hypothèse à ce stade et devrons laisser les modèles de deep learning déterminer les attributs qui permettent de catégoriser les radios.

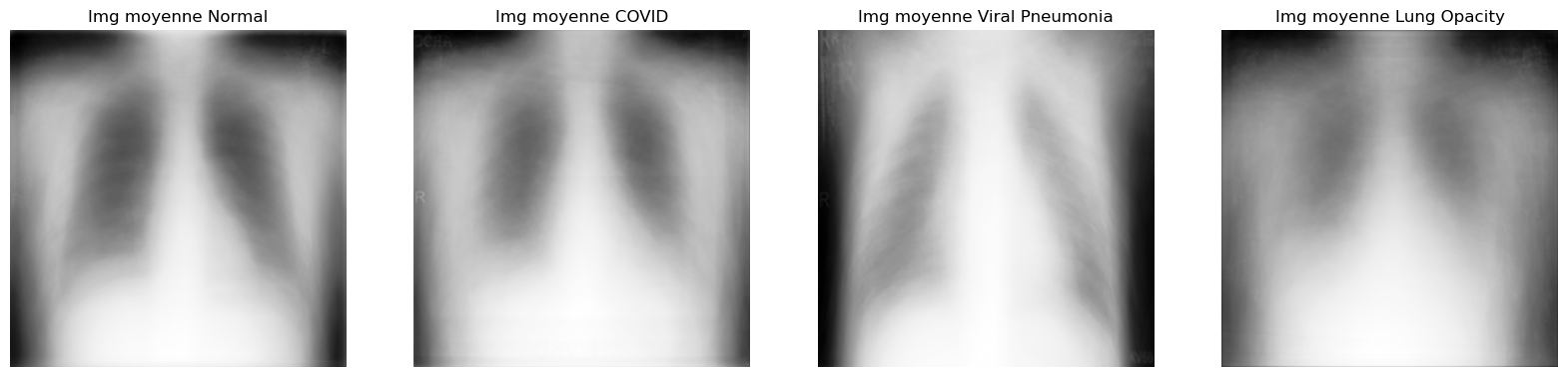


Autre exemple avec une image de chaque catégorie, la même image à laquelle on a applqiué le masque, avec leurs histogrammes:

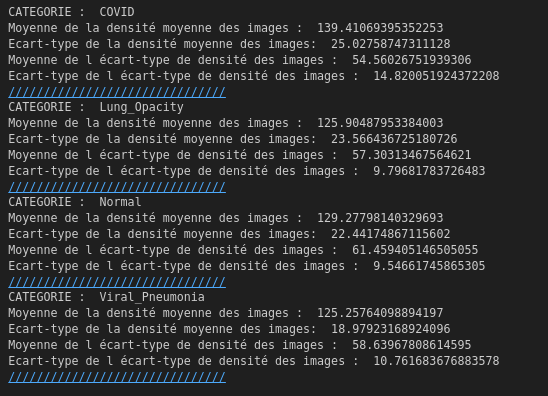


## Images moyennes et statistiques d’intensité

Nous avons également pu calculer et afficher les “images moyennes” sur un sous-set de 100 images par catégorie (pour accélérer le traitement). Cela nous a permis de repérer une opacité plus marquée pour les diagnostics positifs (Covid, pneumonie virale, lung opacity) que pour les patients sains. Les différences entre les différentes maladies pulmonaires sont moins évidentes à repérer à l'œil nu.



Le traitement des images masquées grâce à leur conversion en array de pixels (grâce aux modules OpenCV, PIL et numpy) nous a permis de calculer plusieurs statistiques sur les images du set. On constate que la densité moyenne des images COVID est plus élevée que les autres catégories, ce qui laisse penser qu’elles sont en moyenne plus “claires” que celles des autres catégories (au sens où elles contiennent en moyenne plus de pixels plus clairs que les autres catégories).



## 

# Pre-processing

## 

## 

L’analyse exploratoire a permis d’identifier :

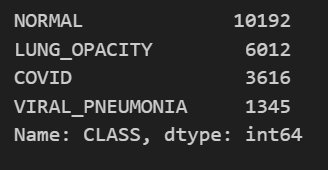
* des différences de format et taille dans les images
* un déséquilibre des classes et en particulier une sous-représentation de la classe COVID qui est l’objectif de la détection
* une opportunité de réduire les données non pertinentes en entrée du modèle grâce aux masques.

Nous avons donc retenu les étapes suivantes pour créer le jeu de données pré-processé à injecter dans les modèles :

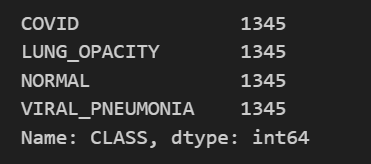
1. Convertir toutes les images en Grayscale
2. Redimensionner les masques en 299\*299 pour correspondre à la taille des images radio
3. Appliquer les masques aux images
4. Équilibrer le dataset grâce à la méthode du sous-échantillonnage sans remise, en s’alignant sur la fréquence de la classe minoritaire (Viral Pneumonia ici).
5. Augmentation des données : Il n'est pas certain que d'augmenter nos données soit nécessaire à ce stade. L'expérimentation des premiers model nous indiquera s'il est utile de le faire pour la suite
6. Reduction de donnée : Lors de l'analyse des données suite à l'observation de nombreuse distribution d'intensité des images masquées. L'hypothèse a été émise dans le groupe que les formes des distributions semblaient similaire selon le diagnostic final. Si tel est en effet le cas cela représenterait un moyen de réduire considérablement la taille du modèle en utilisant uniquement cette distribution en entrée. Cette option sera également investigué par la suite.

Enfin, il nous faudra séparer le jeu d’images en un jeu d’entraînement et un jeu de test, en veillant à l’équilibre des classes dans la répartition aléatoire.

Répartition des images par classe dans le dataset initial :



Répartition des images par classe après undersampling :



Augmentation des données : à préciser

Répartition train/test :

1. Les établissements de santé > édition 2020 > DREES (fiche 23 : L’équipement en imagerie des établissements de santé publics et privés à but non lucratif) : [lien](https://drees.solidarites-sante.gouv.fr/sites/default/files/2021-01/Fiche%2023%20-%20L%E2%80%99%C3%A9quipement%20en%20imagerie%20des%20%C3%A9tablissements%20de%20sant%C3%A9%20publics%20et%20priv%C3%A9s%20%C3%A0%20but%20non%20lucratif%20.pdf) [↑](#footnote-ref-1)
2. <https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database/> [↑](#footnote-ref-2)