Projet Analyse Pulmonaire

Contexte

Au cours l'épidémie de covid 2019, le diagnostic rapide et précis des malades a été à la fois un enjeu sanitaire majeur et un défi technique. Au pic épidémique et alors que les tests étaient en usage limité, l'usage de radiographie médicale a été un outil de diagnostic important.

Cependant, en dehors de l'épidémie, l'utilisation de l'imagerie médicale pour le diagnostic Covid est rendue hasardeuse par la difficulté à le distinguer d'autres maladies pulmonaires. Il est notamment particulièrement délicat de différencier les malades atteints de pneumonies et les malades de covid. C'est pourquoi le développement de modèles de machine et deep learning adaptés serait une aide à la décision précieuse lors du diagnostic médical.

Cette étude porte ainsi sur la classification de patients à partir d'images de radiographies médicales. L'objectif est de pouvoir identifier les malades du covid, en évitant la confusion entre covid et pneumonies. Elle s'appuie sur un ensemble d'images radio labellisées. La première étape réalisée consiste en une exploration et visualisation des données disponibles.

1. Données

Les données dont nous disposons sont des images de radiographies issues de diverses sources académiques et médicales. Elles sont constituées de 21165 observations. Pour chacune d'elle, les informations disponibles consistent en :

- Une image de radio au format png ;
- Une image de masque délimitant la zone des poumons ;
- Un ensemble de métadonnées. Ces dernières indiquent notamment le type de patient et la source de la donnée. Les variables qu'elles contiennent ne comprennent pas de valeurs manquantes et sont synthétisées dans le tableau ci-dessous.

Variables	Type	Description
`file_name`	Object	Nom de l'image indiquant le type de maladie et un numéro identifiant.
`Format`	Object	Format de l'image. Toutes sont au format png.
`Résolution`	Object	Résolution de l'image
`url`	Object	Lien url vers la source

Tableau 1 Description des métadonnées

Répartition par type de patients

Les images radios sont labellisées selon le type de patients auxquelles elles correspondent. On distingue 4 classes de données :

- Les images Covid ;
- Les images de patients atteints de pneumonie virale ;

- Les images étiquetées *lung opacity* qui correspondent à des patients souffrant de pneumonie bactérienne ;
- Les patients sains, labellisés *normal*

Les images radio sont équitablement réparties entre patients sains (un peu plus de 10000, cf. Figure 1) et patients malades (environ 11000). Les malades consistent à 33% de patients covid, à 12% de malades de pneumonie virale et à 55% de malades atteints de pneumonie bactérienne.

Les patients covid, notre cible, représentent ainsi 17% des observations totales.

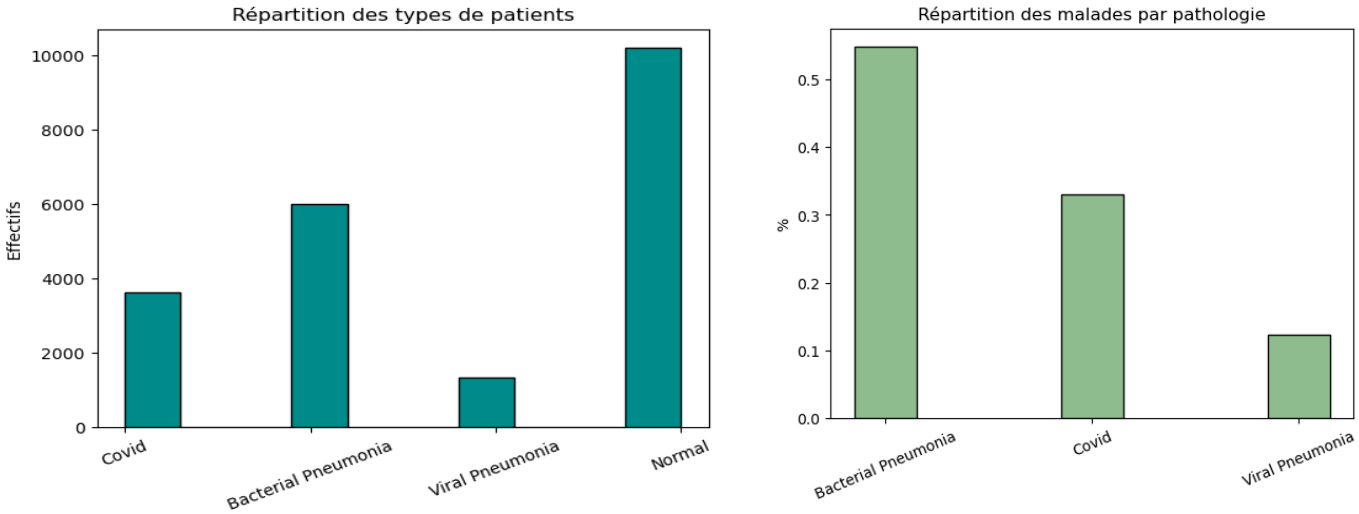

Figure 1 Répartition des données selon le type de maladie

Répartition des sources

Les données sont issues de 8 sources différentes issues de dépôts kaggle et github, ainsi que de trois sites de radiologie européens (cf Figure3).

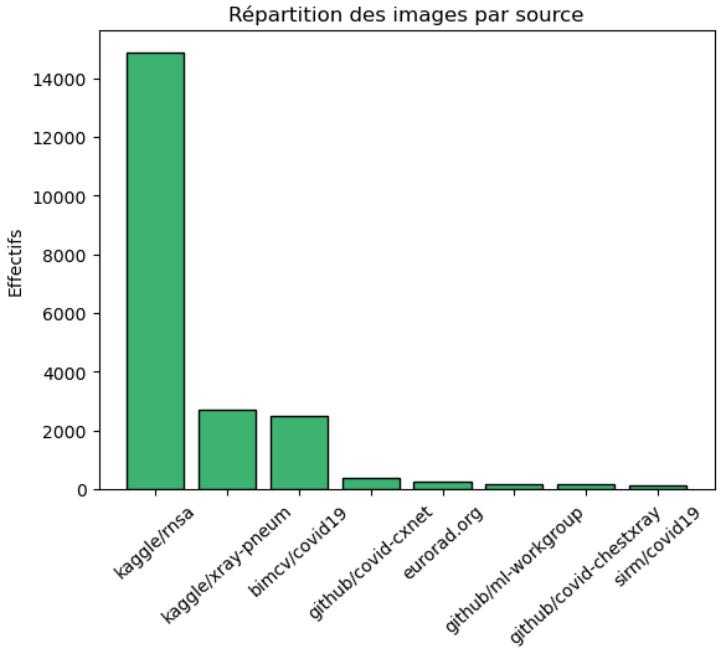

Figure 2 Répartition des données selon la source

Six des sources fournissent uniquement des images covid. Les images de patients sains proviennent des deux dépôts kaggle *rsna-pneumonia-detection-challenge* et *chest-xray-pneumonia*, qui fournissent également respectivement les images de pneumonies.

Source \ Type de patients	Covid	Viral Pneumonia	Lung Opacity	Normal
https://bimcv.cipf.es/bimcv-projects/bimcv-covid19/#1590858128006-9e640421-6711	2474	0	0	0
https://eurorad.org	258	0	0	0
https://github.com/armiro/COVID-CXNet	400	0	0	0
https://github.com/ieee8023/covid-chestxray-dataset	182	0	0	0
https://github.com/ml-workgroup/covid-19-image-repository/tree/master/png	183	0	0	0
https://sirm.org/category/senza-categoria/covid-19/	119	0	0	0
https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data	0	0	6012	8851
https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia	0	1345	0	1341

Tableau 2 Répartition des types d'images en fonction de la source de données

Résolution des images

Les métadonnées indiquent la même résolution de 256x256 pour l'ensemble des images. Il s'agit cependant de la taille des masques fournis. Les images de radiographie sont quant à elles de dimension 299x299.

Les données images ont été comparées pour détecter si leur taille et encodage sont identiques. Cette analyse préliminaire rejoint l'investigation des données manquantes. La tâche actuelle concernant uniquement les données COVID et pneumonie, le résultat actuel ne concerne que les images de scans (CT) de ces deux conditions.

La taille d'une image est, a priori, composée de 2 ou 3 nombres : hauteur, largeur et, éventuellement, couleur. Comme toutes les images semblent (visuellement) être en noir et blanc, il est souhaitable de toutes les traiter comme des images en noir et blanc de même taille après prétraitement. Pour une image initialement au format noir et blanc, on encode sa dimension « couleur » par 0, et par 3 sinon (codage RGB).

Pour les images COVID, on obtient le compte suivant :

```
Height
299 3616
Name: count
```

```
------
Width
299 3616
Name: count
------
Color
0 3616
Name: count
```

Toutes les 3616 images COVID ont les mêmes dimensions 299x299 et sont bien encodées en noir et blanc.

Pour les images de pneumonie :

```
 Height
299 1345
Name: count
------
Width
299 1345
Name: count
------
Color
0 1205
3 140
Name: count
```

Les 1205 images de pneumonie sont de même taille 299x299, mais 10% d'entre elles sont encodées en couleur. Aucune couleur n'étant visible, une moyenne de couleurs semble adaptée pour convertir ces images en noir et blanc.

2. Identification des outliers

Mesures de lisibilité d'une image

Un rapide survol des images montre des disparités importantes entre les niveaux de contraste et de luminosité observables. Ce constat est problématique car l'information contenue par une image sombre ou ayant un faible niveau de contraste est plus difficile à extraire. De telles images doivent donc être identifiées et retraitées de manière à être exploitables par des algorithmes de classification.

Pour cela, on s'intéresse à des indicateurs de qualité des images. On en retient trois :

a) Le niveau de luminosité moyen dans l'image :
Les images radio étant en noir et blanc, une image est décrite par une matrice 299x299 de niveaux de gris $X = \begin{pmatrix} x_{0,0} & \cdots & x_{0,298} \\ \vdots & & \vdots \\ x_{298,0} & \cdots & x_{298,298} \end{pmatrix}$, avec $\forall i,j, x_{i,j} \in [\![0,255]\!]$. Le niveau de luminosité moyen ainsi est donné par :

$$\mathrm{i}_1 = mean(X)$$

Un niveau élevé traduit une image en moyenne plutôt claire.

b) L'écart-type des niveaux de gris de l'image :

$$\mathrm{i}_2 = \mathrm{std(X)}$$

c) L'écart inter-quantile :

$$i_3 = q_{0.95}(X) - q_{0.05}(X)$$

$$\mathrm{i}_3 = \frac{q_{0.95}(X) - q_{0.05}(X)}{\max(X) - \min(X)}$$

L'amplitude de l'écart entre les quantiles est fixée à 95% de manière à permettre de capturer les images pour lesquelles les niveaux de couleurs sont anormalement concentrés.

Le premier indicateur renseigne ainsi sur le niveau de luminosité d'une image et les deux autres donnent une mesure du contraste.

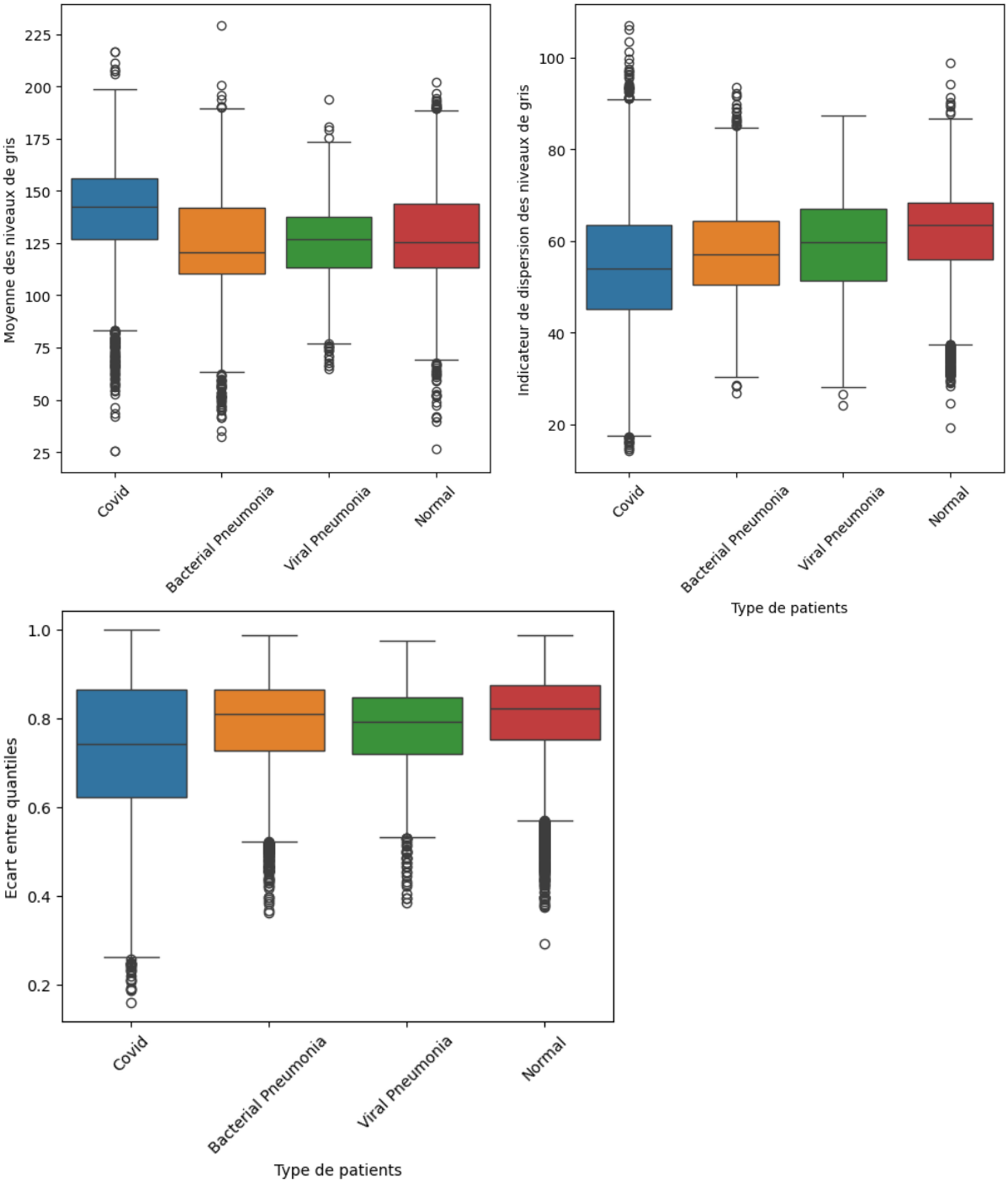

Figure 3 Indicateurs de lisibilité des images par type de patients

Croisement avec les métadonnées

Une première analyse consiste à croiser ces indicateurs avec les métadonnées. Les résultats obtenus sont illustrés dans la figure 3. Les images covid sont en moyenne plus claires, mais les deux indicateurs de dispersion (écart-type et écart inter-quantile) sont en moyenne plus faibles et plus dispersés que pour les autres types de patients. L'objectif de l'étude étant de les identifier avec précision, il sera important de retraiter les images trop uniformes.

La figure 4 croise un indicateur de contraste avec l'origine des données, ce qui met en évidence une source en particulier qui présente un grand nombre d'images avec peu de variation de couleur. Il s'agit de la banque de données médicales de Valence (bimcv.cipf.es) qui fournit environ les 2/3 des images covid de notre ensemble de données.

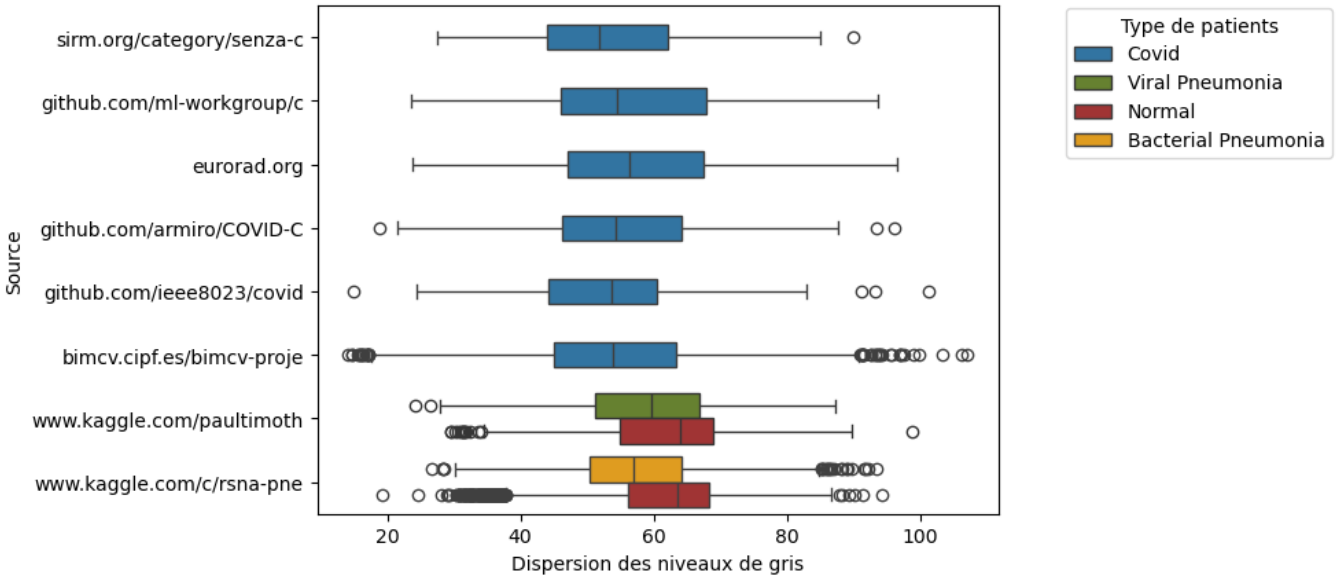

Figure 4 Dispersion des couleurs dans les images en fonction de leur source et du type de patients

Identification des outliers

a) Niveau de luminosité

Les données anormalement sombres sont d'abord identifiées. Il s'agit des images ayant un niveau moyen de couleur particulièrement bas en comparaison des autres. On considère ici qu'une valeur est extrême si la moyenne des couleurs est dans les 2.5% les plus faibles. Cela correspond à 530 images dont 25% sont des données covid. Les images covid, si elles sont en moyenne plus claires, sont également représentées dans les valeurs extrêmes.

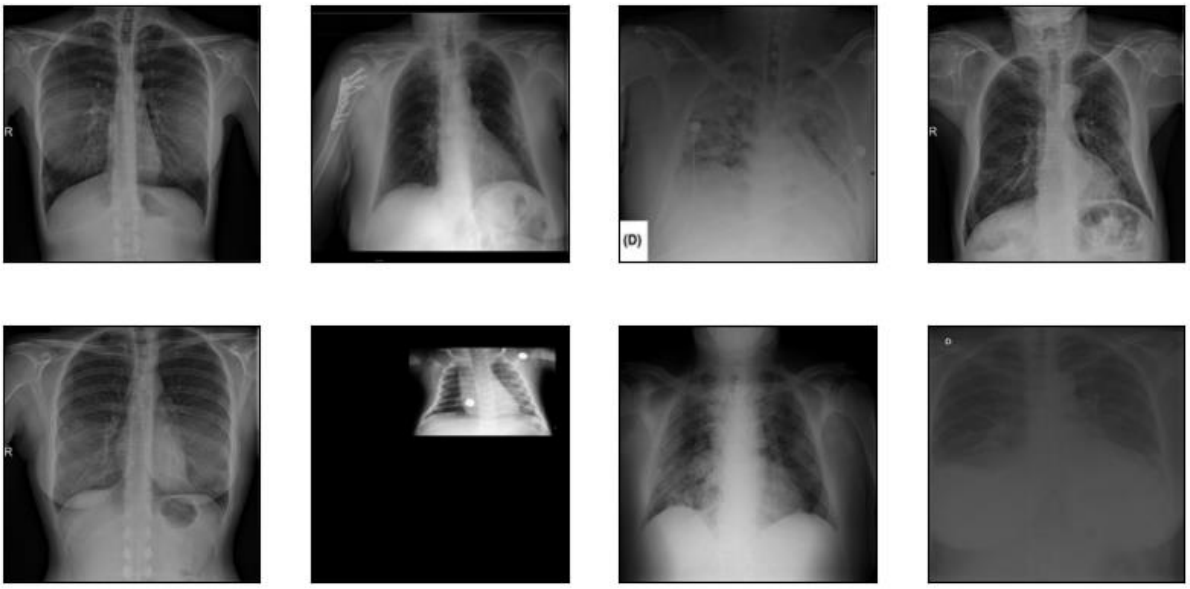

Figure 5 Exemple de radios identifiées comme anormalement sombres parmi le groupe covid

La figure 5 donne quelques exemples de telles radios covid. Au-delà des images très sombres, on remarque également que certaines sont plus petites et encadrées de noir. Elles devront donc être identifiées et redimensionnées pour pouvoir être traitées avec les autres. Cette étude est l'objet d'une prochaine sous-section.

b) Niveau de contraste

Un autre type de valeurs extrêmes relatif aux niveaux de contraste est également étudié afin de d'identifier les images très uniformes. Comme précédemment, on retient ici les 2.5% d'images ayant les écarts-types de la distribution de niveaux de gris les plus bas. Cela représente 530 observations dont 339, soit près des deux tiers, sont des données covid. Comme illustré par la figure 6, certaines images sont quasiment illisibles. Il va donc falloir les retravailler ou bien les écarter si les traitements ne suffisent pas.

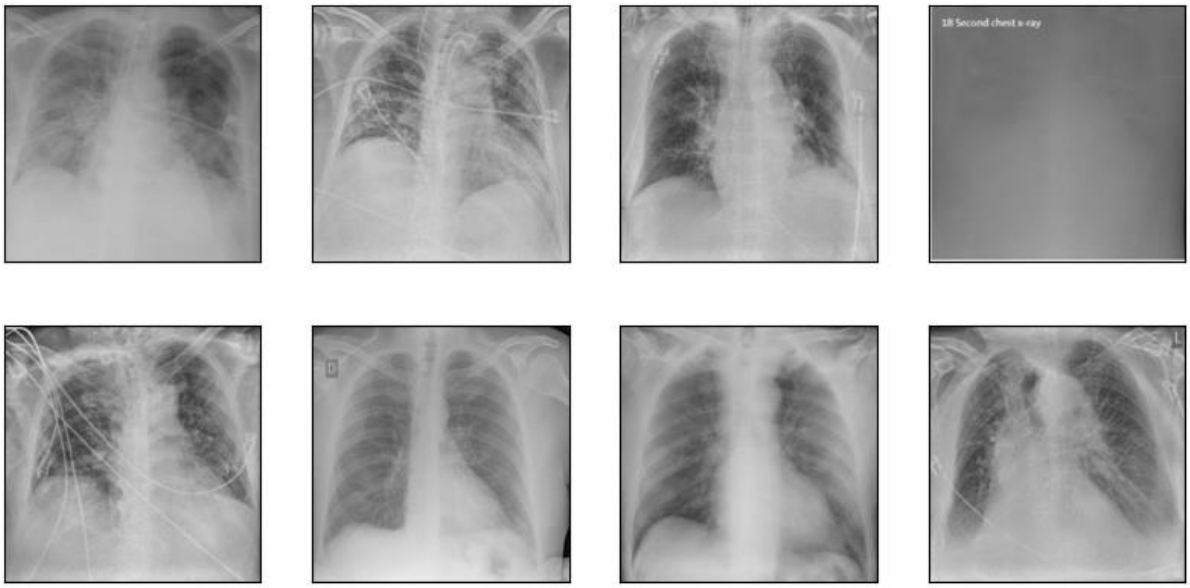

Figure 6 Exemple de radios identifiées comme anormalement peu contrastées parmi le groupe covid

Identification des images tronquées

L'objectif est ici d'identifier les images radiographiques présentant des bordures noires formant un cadre autour de l'image principale. Ces bordures noires peuvent être le résultat de divers processus, tels que des erreurs dans la numérisation des radiographies, de manipulation ou des problèmes de formatage. Elles peuvent causer des problèmes lors de l'entraînement de modèles de machine ou deep learning, car elles introduisent des éléments non représentatifs du contenu médical de l'image. Les modèles peuvent alors apprendre des caractéristiques incorrectes ou devenir moins sensibles aux véritables structures d'intérêt dans les radiographies, conduisant ainsi à des erreurs de classification et à une détection incorrecte des anomalies. Afin de garantir la précision des modèles entraînés sur ces données et d'améliorer leur capacité prédictive, il est donc crucial de repérer, traiter et de redimensionner ces données dégradées.

La méthode d'identification retenue est la suivante :

1. **Prétraitement :** Chaque image est convertie en niveaux de gris et un flou gaussien est appliqué pour atténuer les variations mineures dans les bords.

2. **Détection des Bordures** : Il examine les pixels dans les bordures (haut, bas, gauche, droite) pour des largeurs variables, vérifiant si les pixels sont sous un seuil d'intensité, ce qui indique la présence de noir.
3. **Liste des Images Problématiques :** Les images avec des bordures noires détectées sont ajoutées à une liste.
4. **Affichage et Sauvegarde** : Les images détectées sont affichées pour vérification, et les noms de fichiers correspondants sont sauvegardés dans un fichier texte pour une éventuelle suppression ou réévaluation.

Impact sur le Modèle

En nettoyant les images problématiques, le programme améliore la qualité des données d'entraînement, permettant ainsi aux modèles d'apprentissage automatique ou de deep learning de se concentrer sur les véritables caractéristiques cliniques des radiographies. Cela contribue à la précision et à la fiabilité des prédictions, cruciales pour des applications comme le diagnostic assisté par ordinateur dans le domaine de la santé

La figure 7 montre comment une bordure noire, souvent invisible à première vue dans des processus automatiques, peut être clairement détectée et visualisée lorsqu'elle est modifiée en une couleur vive

Figure 7 Identification du cadre autour d'une image

comme le rouge. Cela permet de valider la fonctionnalité du programme et d'assurer que les images problématiques sont correctement identifiées et traitées.

3. Analyse d'image et comparaison entre classes

Au-delà de l'identification d'outliers, on peut rechercher de premiers éléments de comparaison entre les différentes classes dans les images disponibles.

Distribution de la luminosité

Considérons dans un premier temps la forme de la distribution des niveaux de couleur dans une image comme élément de comparaison entre les différentes classes.

De la même manière que dans la section 2., on évalue l'indicateur de luminosité i_1, i.e. la moyenne de tous les pixels mais ici sur les données normalisées. Il prend donc ses valeurs entre 0 et 1.

On trace l'histogramme de la répartition moyenne des niveaux de couleurs des CT scans par catégorie de patients (cf. Figure 8).

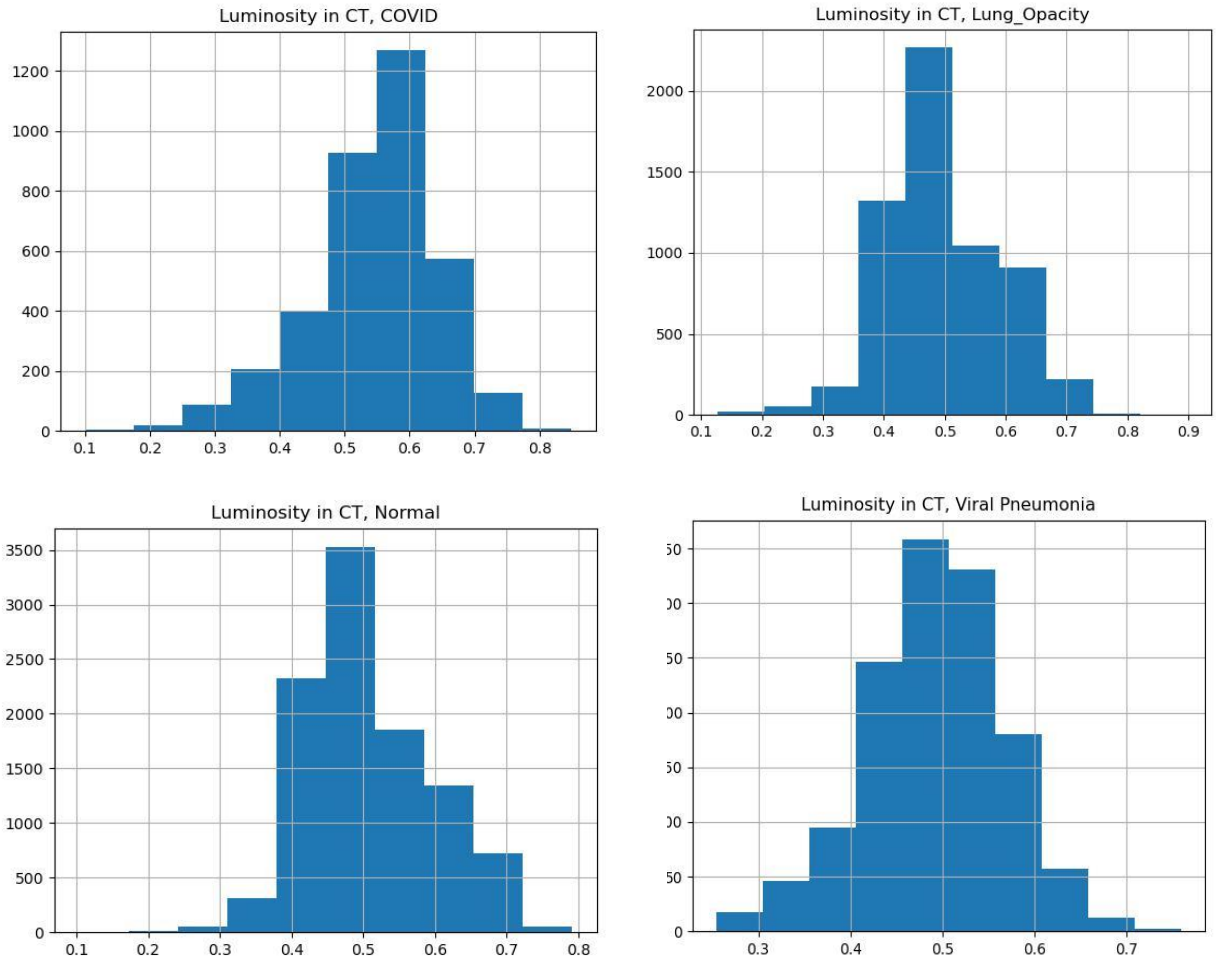

Figure 8 Histogramme des niveaux de couleurs moyens par type de patients

Au-delà de l'étude des valeurs de luminosité moyennes dans les images, il sera intéressant par la suite de s'intéresser à la distribution des niveaux de couleur à l'échelle d'une image. Une option consisterait à se restreindre à la surface des poumons et à chercher des patterns dans la répartition des couleurs selon la catégorie de patients. Une première idée serait à ce titre de réaliser un clustering sur ces distributions et de vérifier si les clusters ainsi formés correspondent aux classes connues.

Taille des poumons dans les images de masques

L'exploration des données de masques permet de comparer la taille des poumons sur les radios. Cela peut être réalisé en première analyse en utilisant les images de masques fournies dans les données. En délimitant la surface des poumons, ceux-ci définissent la surface utile de l'image pour la détection de pathologies pulmonaires. La figure 9 comprend les histogrammes de la répartition de la taille des poumons extraites des masques par catégorie de patients. Celle des images COVID paraît plus asymétrique et étalée à droite que pour les autres groupes. Cela paraît difficile à interpréter sans refaire les masques. Ceci pourrait être :

- Un effet réel lié à la maladie ;

- Un effet de l'algorithme utilisé pour générer les masques, par lequel les limites des poumons malades seraient moins discernables.

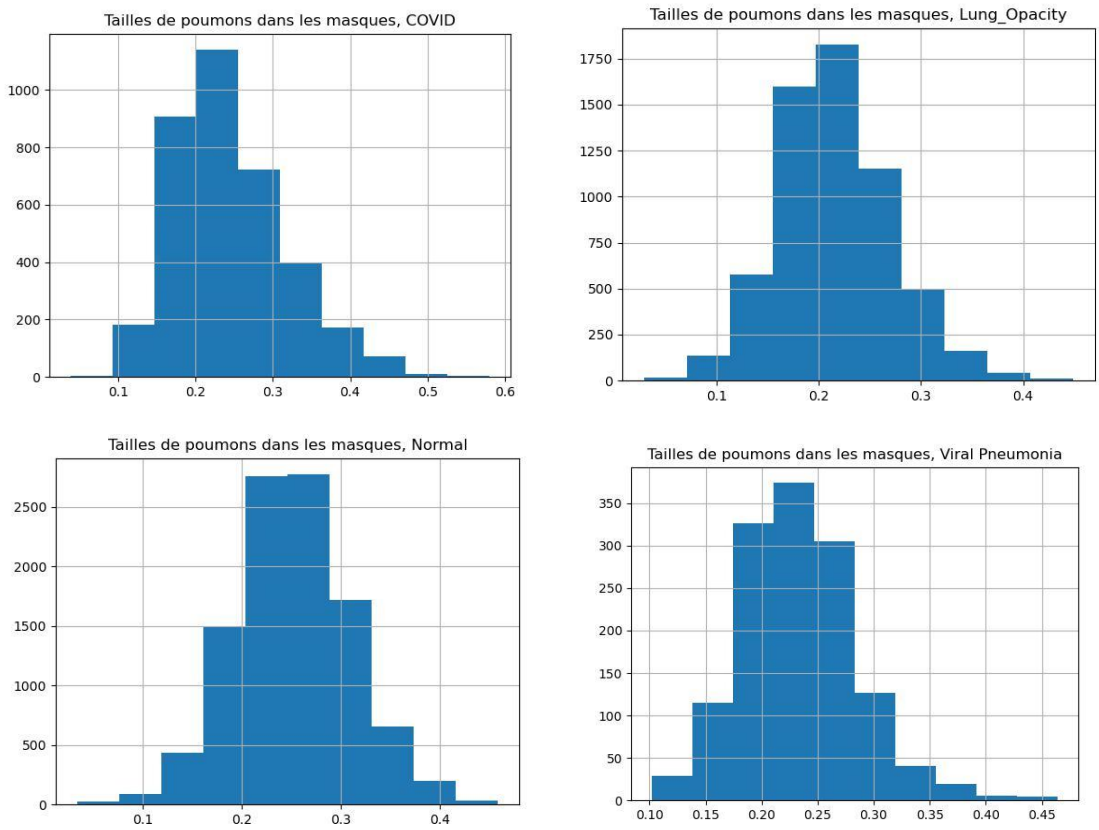

Figure 9 Répartition des tailles de poumons par type de patients

4. Traitement des masques

Superposition des masques sur les images de radiographie

L'identification des outliers nous montre que certaines images sont illisibles voire difficile à traiter de par leur manque de luminosité et un niveau de contraste contraignant.

Pour avoir une meilleure visualisation nous avons donc superposé les radios de poumons avec leur masque. Cela permet de voir à la fois les détails de la radio ainsi que les zones masquées.

Superposer un masque sur une image d'origine est une technique utile dans de nombreux domaines, notamment en imagerie médicale et en vision par ordinateur. Cette méthode offre plusieurs avantages et applications clés comme vérifier l'exactitude des masques créés automatiquement.

Figure 10 Exemple de superposition d'une radiographie avec son masque parmi le groupe covid

La première étape consiste à se ramener à des images de mêmes dimensions. Les radiographies étant de dimension (299x299) et les masques de dimension (256x256), nous avons utilisé la fonction OpenCV cv2.resize afin d'avoir la même dimension.

Pour ensuite permettre une évaluation visuelle rapide et intuitive, nous avons modifié la couleur de la radiographie et du masque qui était grise de base afin de faire ressortir la zone qui nous intéressait.

Les masques peuvent représenter des structures spécifiques telles que des tumeurs, des organes ou des anomalies. La superposition de ces masques et notamment le fait de la ressortir en couleur permet de visualiser clairement ces zones d'intérêt.

Pour suivre l'évolution d'un virus comme le covid chez un patient, la superposition des masques sur des images prises à différents moments montre clairement les changements et l'évolution.

La superposition de masques sur des images d'origine est une méthode visuelle puissante qui améliore la compréhension et l'analyse des données visuelles, aidant ainsi à rendre les données plus compréhensibles

Transformations morphologiques de dilatation et d'érosion

On peut également utiliser des méthodes de transformations morphologiques pour détecter des structures pertinentes et éliminer celles non pertinentes pour la compréhension de l'image. Les deux méthodes de base très utilisées sont la dilatation et l'érosion.

Les transformations morphologiques d'érosion et de dilatation sont des techniques essentielles en traitement d'images, notamment pour l'analyse des radiographies pulmonaires et de leurs masques. Ces opérations permettent de manipuler et d'améliorer les images de manière à faciliter leur analyse et interprétation.

L'érosion réduit les objets dans une image et supprime les pixels en bordure des objets, ce qui entraîne un rétrécissement des formes, comme on peut le voir sur la figure 11. Elle permet également de rendre les bords des structures, comme les contours des poumons ou des anomalies, plus nets et précis.

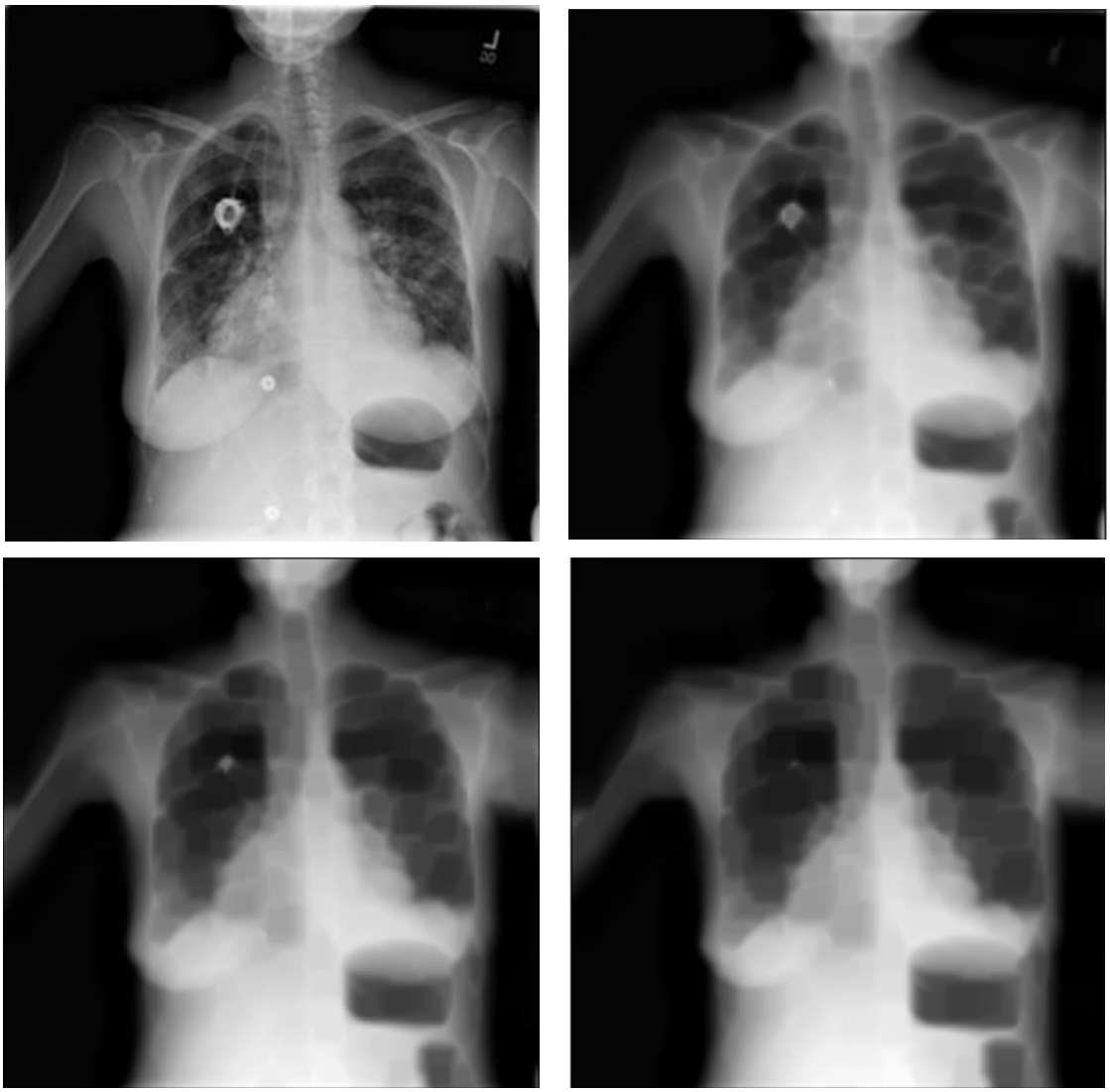

Figure 11 Exemple d'érosion sur une radiographie parmi le groupe lung opacity

La dilatation, quant à elle, augmente les objets dans une image et ajoute des pixels en bordure des objets, ce qui entraîne une expansion des formes. Elle peut épaissir les bords des structures, rendant ainsi les zones d'intérêt plus visibles. La figure 12 montre la différence entre l'image d'origine et l'image dilatée. En dilatant le corps et en soustrayant à l'image d'origine, on obtient les contours des poumons.

Appliquées sur des images de radiographies pulmonaires, ces techniques permettent de détecter ou d'isoler des structures pertinentes, améliorant ainsi la qualité des données à fournir aux algorithmes de classification.

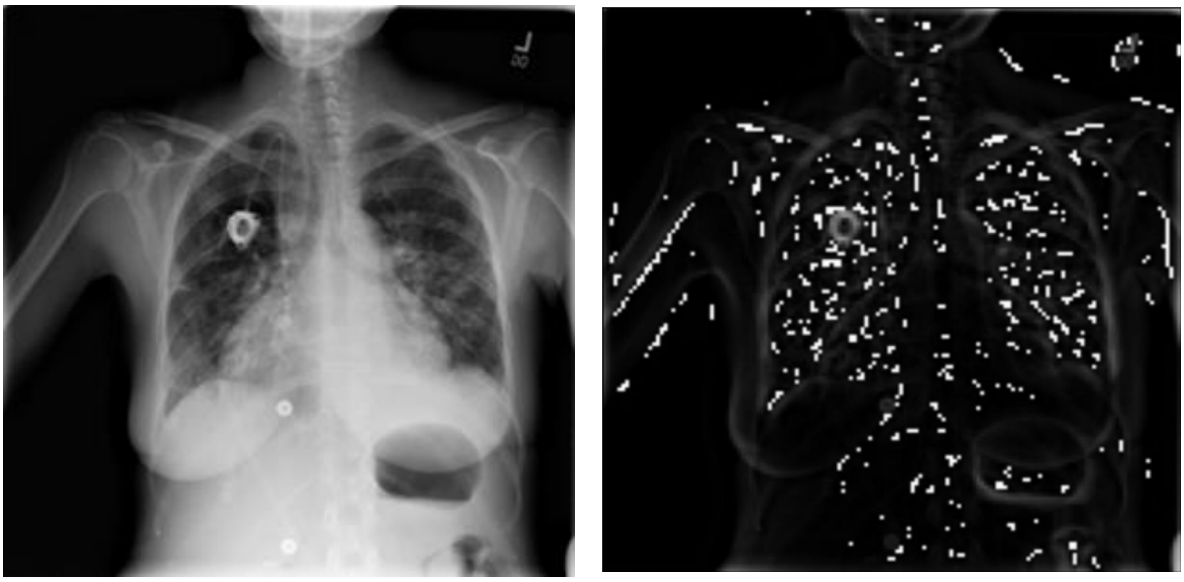

Figure 12 Exemple de radiographie avant et après transformation parmi le groupe lung opacity