Rapport n°2 - Projet de Classification de Radiographies Pulmonaires

—

Résumé des étapes effectuées

Le but de ce rapport est de présenter la phase une du preprocessing du jeu de données suivant : https ://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database Après discussion avec notre chef de cohorte nous avons déduit différent point à explorée pour améliorer la qualité de nos données .

Table des matières

1 Nature des fichiers

Tableau 1 : Vue d'ensemble des catégories

#	Catégorie	Nombre de fichiers	Taille (bytes)	Taille approx.
1	COVID-19	3616	129,584,521	130 Mo
2	Lung_Opacity	6012	211,796,041	212 Mo
3	Normal	10192	379,821,965	380 Mo
4	VirPneumonia	1345	54,209,670	54 Mo

TABLE 1 – Répartition de la taill des fichiers par catégorie

Comme remarqué dans l'exploration des données, on se rend compte que les fichiers Normal et Lung Opacity sont les deux contenant le plus de données. On se rend également compte que l'espace occupé par chaque dossier suit le même ordre que leur longueur (en termes de nombres de fichiers).

Tableau 2 : Statistiques des tailles de fichiers

#	Dossier	Moyenne	Médiane	Écart-type	Min	Max	Fichiers
0	Covid	35836.43	36186	4552.82	9913	52004	3616
1	Lung_Opacity	35228.88	35269	3565.40	13878	48292	6012
2	Normal	37266.68	37346	3088.64	14178	54043	10192
3	VirPneumonia	40304.59	38598	6542.73	27005	73131	1345

TABLE 2 – Statistiques des tailles de fichiers par catégorie

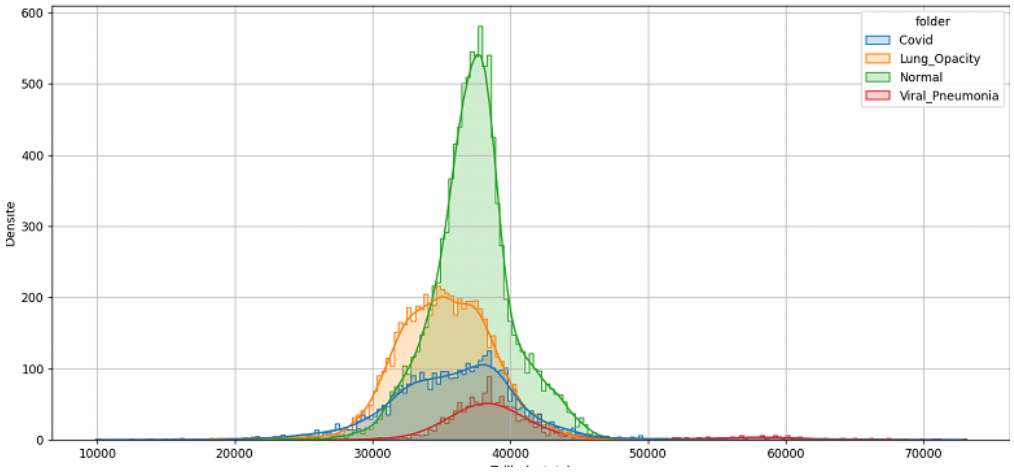

FIGURE 1 – Variance de tailles expliquée par chaque composante principale

On se rend ici compte que les radiographies sur les cas de pneumonies virales ont une moyenne bien plus élevée que celle des autres dossiers. La différence entre les médianes est

néanmoins plus petite, bien que la médiane de ce dossier reste légèrement supérieure à celle des autres. Mais l'écart-type des tailles de fichiers étant bien plus grand dans ce dossier, on se rend vite compte que bien qu'il s'agisse du dossier avec le moins de fichiers, ceci ont une taille bien plus variable, bien qu'ils soient globalement plus volumineux.

Il serait donc peut-être opportun de réfléchir à une harmonisation de la taille des fichiers des différents dossiers, et, si ce n'est pas le cas (car s'agissant d'un dossier ne contenant que peu de fichiers, on peut considérer cette option), garder néanmoins cette différence de taille à l'esprit dans l'entraînement de nos modèles.

2 Analyse par PCA (Analyse en Composantes Principales)

Ce résumé présente les résultats de l'analyse en composantes principales (PCA) effectuée sur un jeu de données d'images radiographiques associées à des masques. L'objectif de l'analyse était de réduire la dimensionnalité des données tout en conservant un maximum d'informations pertinentes, à savoir en expliquant une grande partie de la variance des données.

Résultats principaux

Voici les résultats de la PCA appliquée aux caractéristiques extraites des images :

1. Variance expliquée par chaque composante : [0.33753577, 0.17235869, 0.10687735, 0.06943919, 0.06053101, 0.04964212, 0.03888787, 0.02735815, 0.0207339, 0.01482933, 0.01200447]

2. Variance totale expliquée : 0.9101978518735634 (soit environ 91% de la variance totale expliquée par 11 composantes principales).

3. Dimension initiale des données : 256 (nombre de caractéristiques extraites de chaque image).

4. Dimension après PCA (nombre de composantes choisies) : 11 composantes.

Grâce à la PCA, la dimension des données a été réduite de 256 à 11 composantes principales, tout en expliquant environ 91% de la variance présente dans les données d'origine. Cela permet de maintenir l'essentiel de l'information tout en simplifiant les données, ce qui est crucial pour les étapes suivantes d'analyse ou de modélisation.

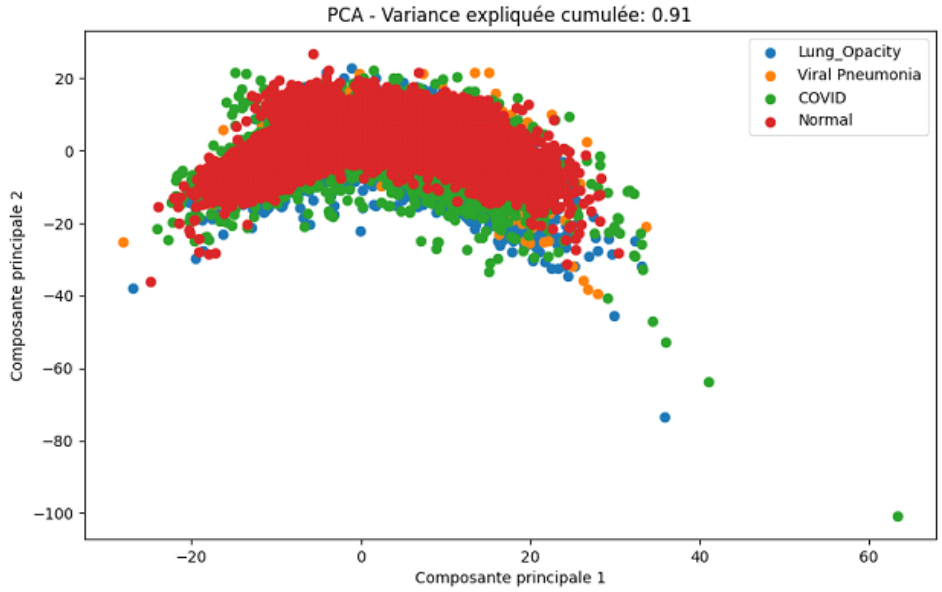

FIGURE 2 – Distribution des point sur les premières composantes principales

Ce graphique en 2D montre la distribution des points projetés sur les deux premières composantes principales (PC1 et PC2). Les différentes classes (COVID, Normal, Lung Opacity, Viral Pneumonia) sont visibles et bien séparées. Cela indique que les deux premières composantes principales contiennent une grande partie de l'information discriminante nécessaire pour différencier les classes dans le jeu de données. Cette séparation est essentielle pour des algorithmes de classification ultérieurs.

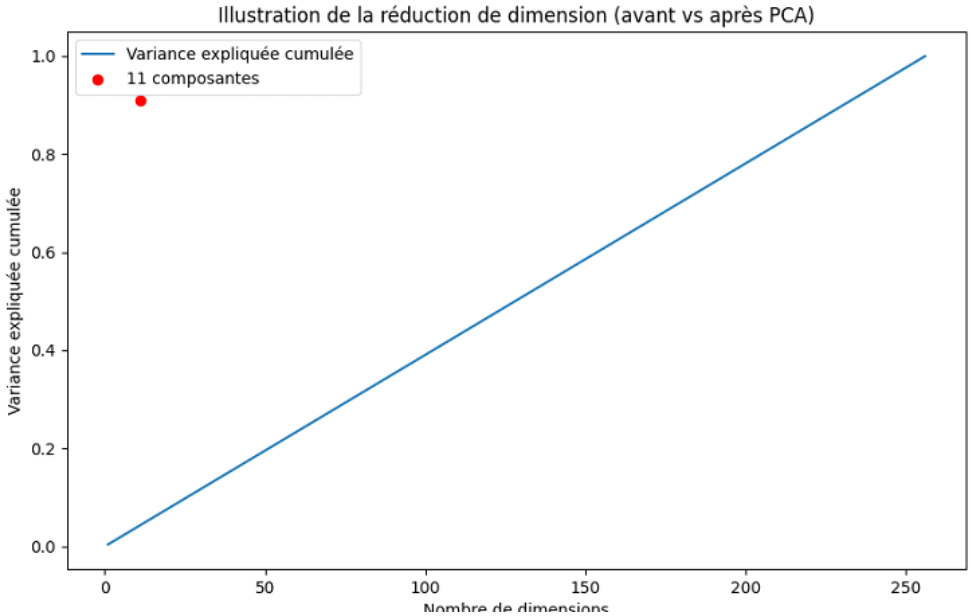

FIGURE 3 – Reduction de dimensions

Ce graphique montre la réduction de la dimensionnalité avant et après l'application de

la PCA. La ligne bleue représente la variance cumulée expliquée par toutes les dimensions (jusqu'à 256 dimensions), tandis que la ligne rouge marque la réduction des dimensions après PCA, où 11 composantes sont choisies. Cela démontre clairement que la PCA réduit efficacement la complexité des données tout en conservant 91% de la variance d'origine. La réduction de la dimension permet ainsi de simplifier les calculs tout en préservant l'essentiel de l'information

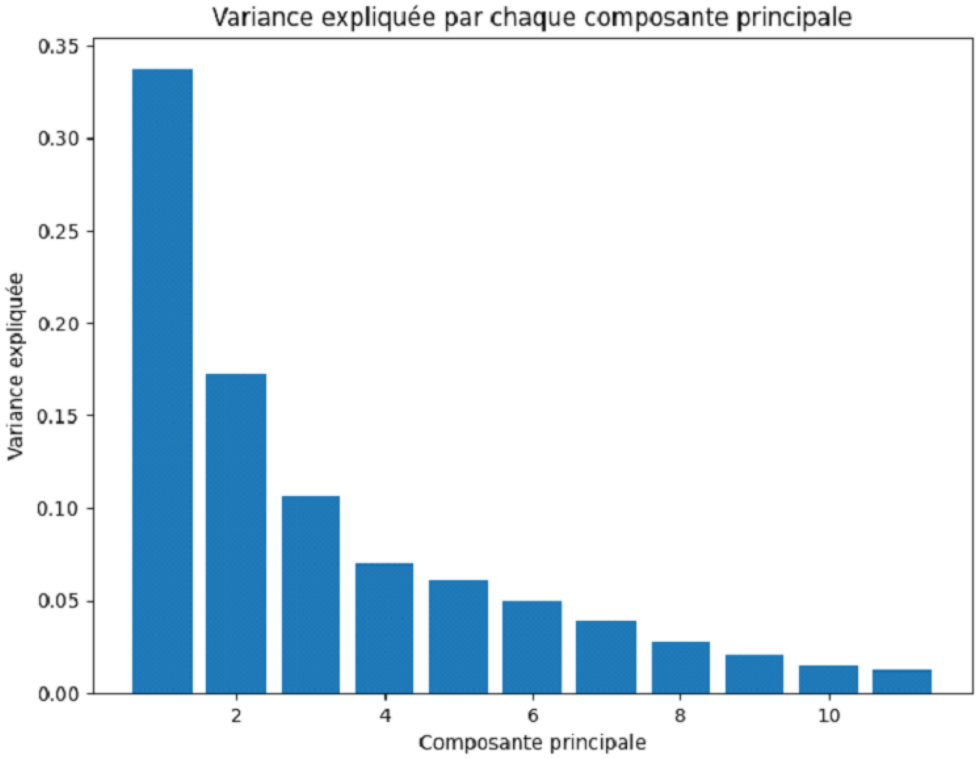

FIGURE 4 – Variance expliquée par composantes principales

Le graphique montre la proportion de la variance expliquée par chaque composante principale (PC). On observe que la première composante (PC1) explique une part significative de la variance (environ 33,75 %), suivie de la deuxième composante (PC2) avec environ 17,24 %. Ces deux premières composantes couvrent à elles seules plus de 50 % de la variance. La contribution des autres composantes est de plus en plus faible. Cela montre que la réduction de dimension à 11 composantes principales est un choix judicieux, car on conserve 91 % de la variance totale

3 Analyse de Similarité : SSIM et Hash

En utilisant l'indice de similarité structurelle calculé avec les moyenne et variance d'une fenêtre glissante pixel à pixel (-1 = anti-corrélation, 1 corrélation parfaite) de la libraire skimage (traitement d'images de scikit) : on calcule pour chaque catégorie, un score de similarité image à image. Pour réduire les temps de calculs, on réduit au préalable l'image

à 50*50 pixels. On obtient 4 matrices (une par catégorie) de similarité (symétriques) Pour réduire l'espace mémoire utilisé, on arrondit le SSIM à 3 chiffres significatifs. Pour le post-traitement des résultats, on se concentre dans un premier temps uniquement sur la catégorie Normale dont on cherche à réduire le nombre d'éléments. On liste ensuite pour chaque image, la liste de celles qui ont un SSIM ¿ seuil. Avec un seuil très élevé (0.95), on identifie les doublons. L'analyse par hash perceptuel montrait 2 jeux de doublons :

Résultats - Doublons identifiés par hash

Hash	**Classe**	**Fichier**
dd2cc25b154b1f4c	Normal	Normal-1708.png
dd2cc25b154b1f4c	Normal	Normal-5103.png
dd7a425f5a0d3c0c	Normal	Normal-817.png
dd7a425f5a0d3c0c	Normal	Normal-818.png

TABLE 3 – Doublons détectés par hash perceptuel

On retrouve les deux derniers 817 et 818 (score de 1) :

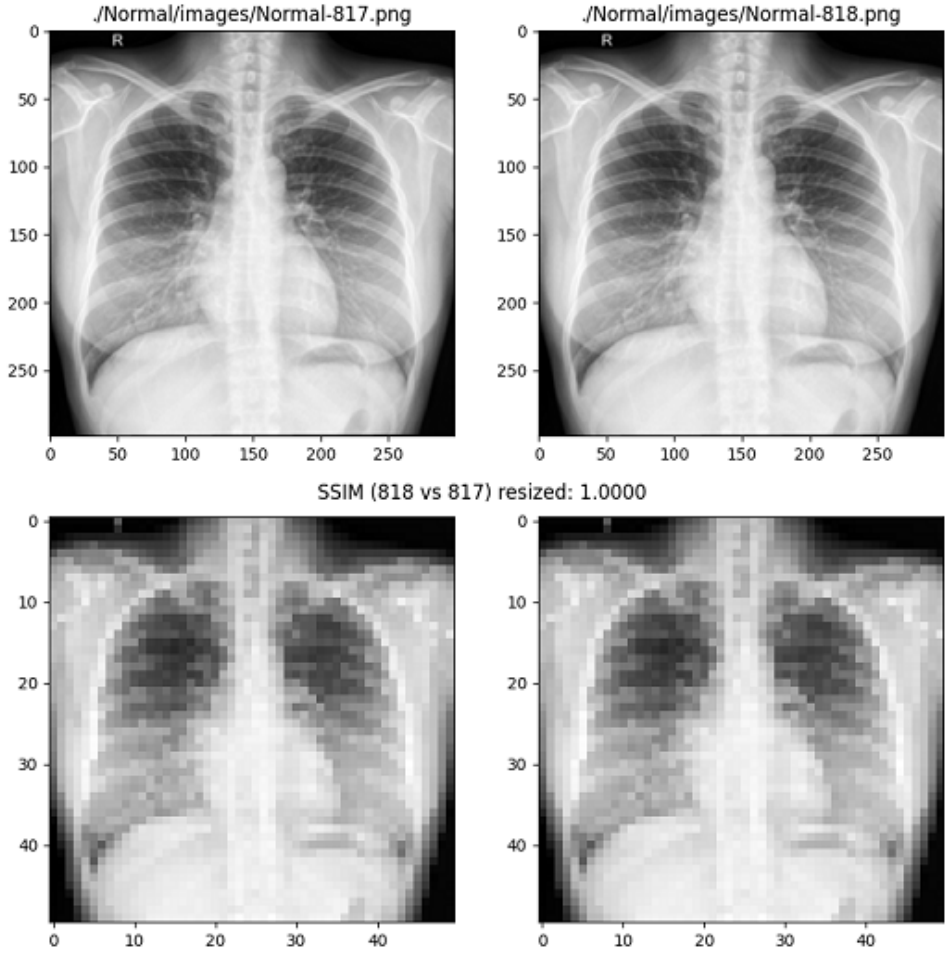

FIGURE 5 – Exemples de doublons détecté

En revanche, on ne confirme pas le duo 1708-5103 qui n'est pas un doublon (présence d'annotations dans le coin supérieur droit et images très légères différences). Ces deux images sont très similaires mais pas identiques.

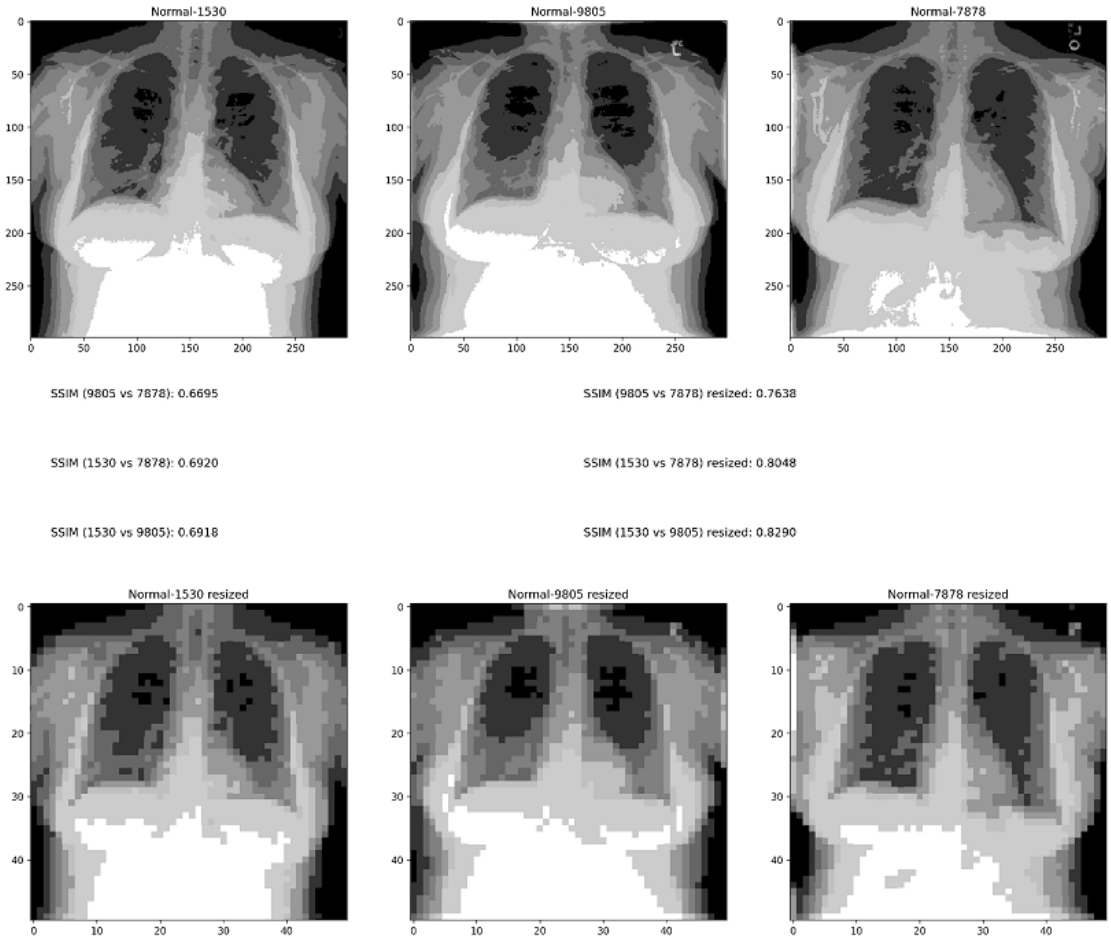

FIGURE 6 – Exemples de doublons détecté 2

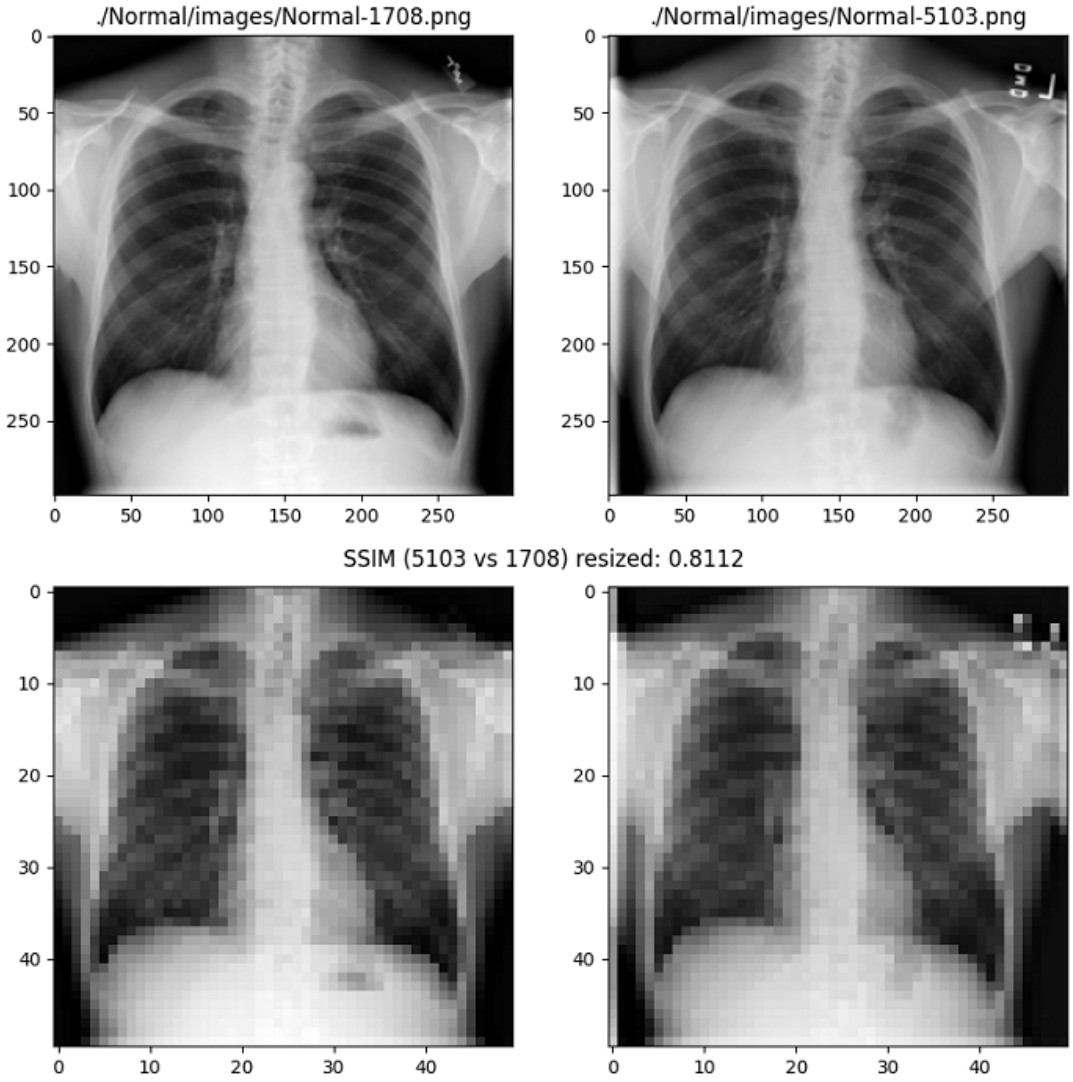

FIGURE 7 – Exemples de doublons détecté 3

Avec un seuil plus bas (ex : 0.80), on peut trouver des images très semblables. Exemple ci-dessous : 3 images de la catégorie normale (différentes) avec un score SSIM > 0.8 (référence ici : 1530, le SSIM entre 9805 et 7878 est élevé mais < 0.8). Cet exemple montre qu'un seuil de 0.8 semble pertinent pour le filtrage envisagé.Ainsi, on produit pour chaque catégorie, une liste de liste d'images avec un score supérieur à 0.8.

Voici les résultats pour la catégorie Normal :
204 doublons, 15 triplons, 4 quadruplons

Fichier de résultat pour la catégorie Normal

Cette méthode permettrait de réduire très légèrement le nombre d'images sélectionnées mais reste marginale par rapport aux objectifs visés (division par 2). Elle doit être complétée par une autre méthode (augmentation d'images pour les catégories sous-représentées ou sous échantillonnage aléatoire)

Ci-dessous quelques exemples supplémentaires de doublons identifiés :

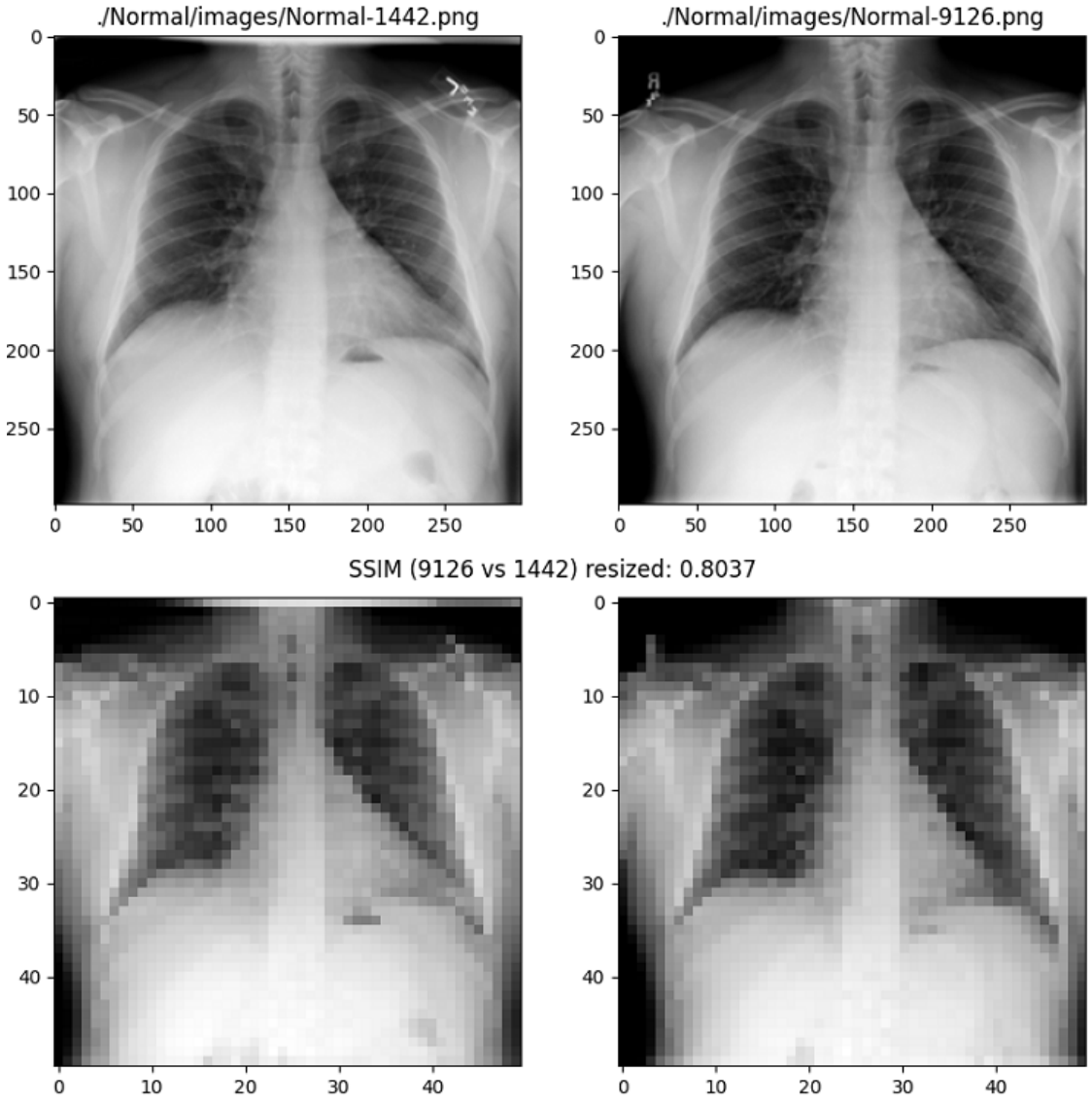

FIGURE 8 – Exemples de doublons détecté 4

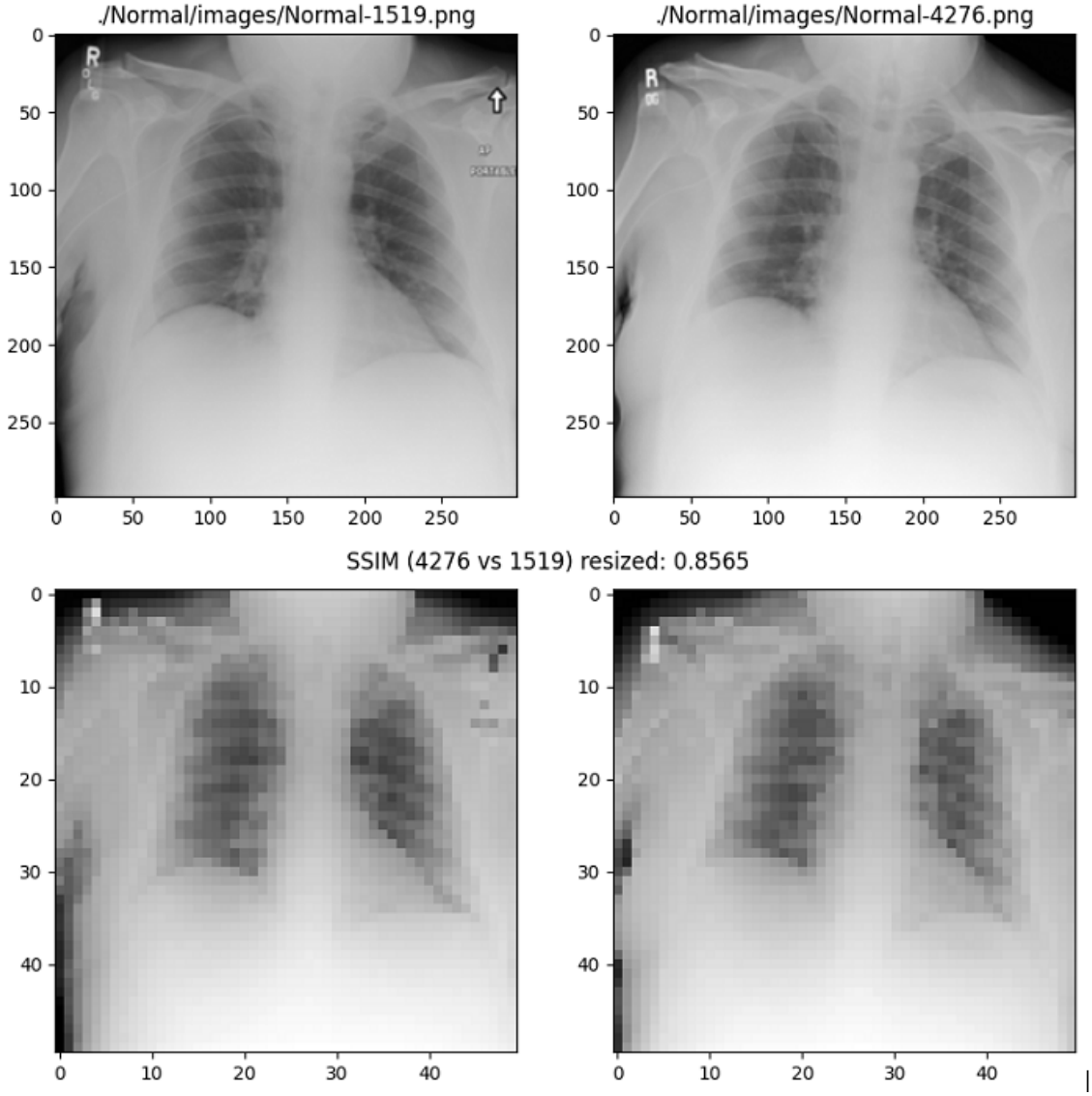

FIGURE 9 – Exemples de doublons détecté 5

4 Prétraitement des Données : Augmentation des images

Problématique

Le dataset est déséquilibré : deux classes sont surreprésentées, deux sous-représentées.

Premiers tests sans rééquilibrage

- Labels :
  - COVID-19 : 0
  - Lung Opacity : 1
  - VirPneumonia : 2

  — Normal : 3
— Fusion des images dans un seul tensor
— Entraînement d'un modèle MobileNetV2 avec softmax

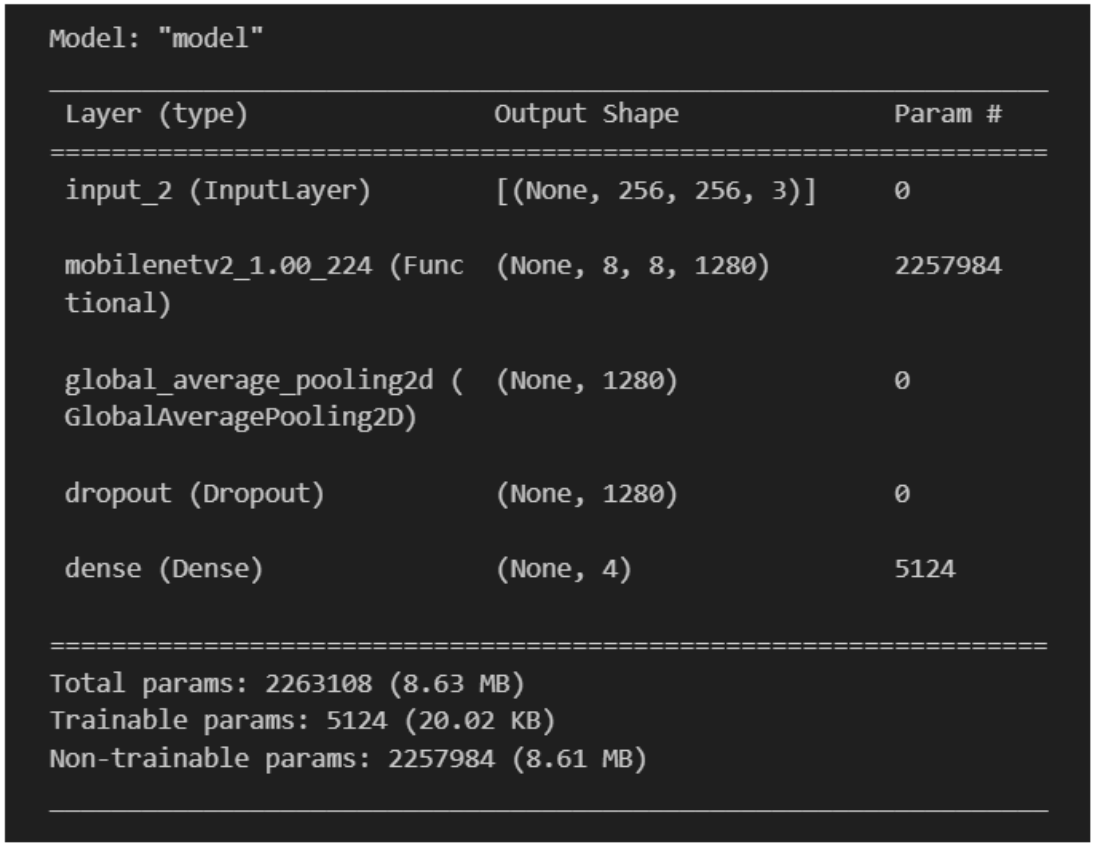

FIGURE 10 – Summary du benchmark utilisé

Utilisation de ce modèle qui intégre du transfert learning et donc le modèle pre-entrainer MobileNetV2 avec une activation softmax pour la dernière couche de convotution. Pour la compilation de ce modèle l'optimisateur utiliser est adam, la fonction de perte est Sparse Categorical crossentropy et la metrique l'accuracy pour coller avec notre problèmatique.

Resultats premier modèle

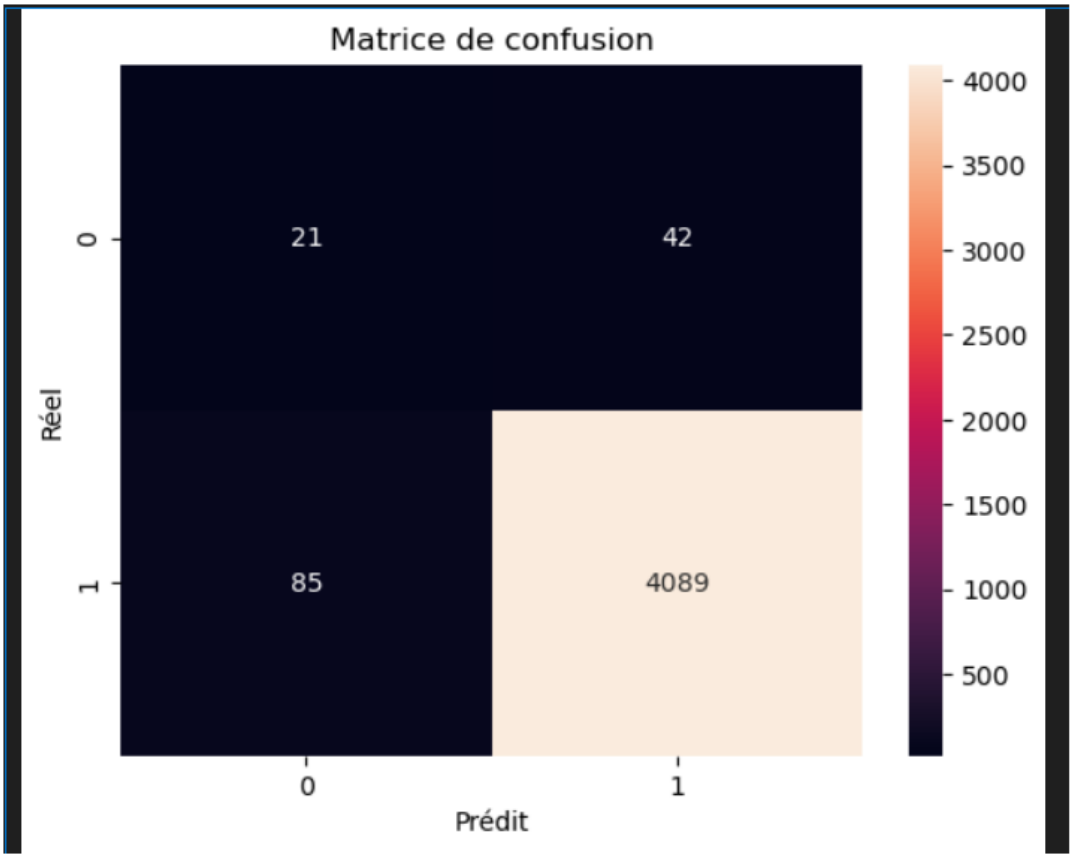

FIGURE 11 – Resultat de la première modèlisation

On observe ici que dans le jeu de donnée d'entrainement et le jeu de donnée de validation tout les labels n'ont pas été pris en compte, l'hypothèse est que la disparité en proportion entre les différent label à provoquer meme avec un melange des données effectuée un desequilibre entre ces deux tensor. L'interprétabilité du modèle n'est donc pas possible car les jeux de données d'entrainement et de validation ne sont pas représentatif du jeu de données initial.

Rééquilibrage par augmentation d'images

La technique utiliser ici pour reequilibrée les classes est donc l'augmentation d'images et la limitation des classes surepresenter à 4000 occurances.

Techniques utilisées (`tensorflow.keras.layers`) :

— RandomZoom(0.1)
— RandomRotation(0.1)
— RandomContrast(0.1)

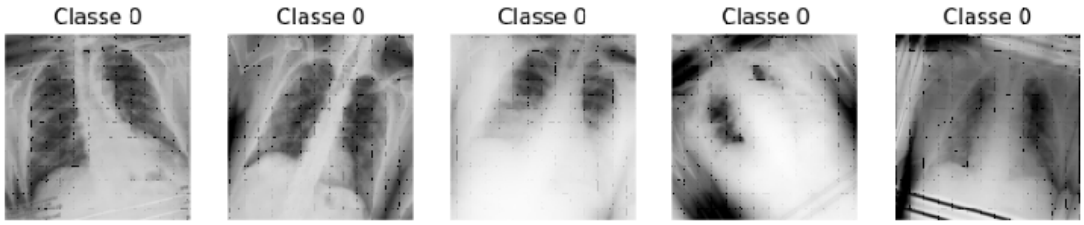

FIGURE 12 – Exemple d'images augmentées

Nous avons donc un jeu de données equilibrée et pouvons l'appliquer à l'entrainement de notre modèle benchmark.

Resultats du modèle

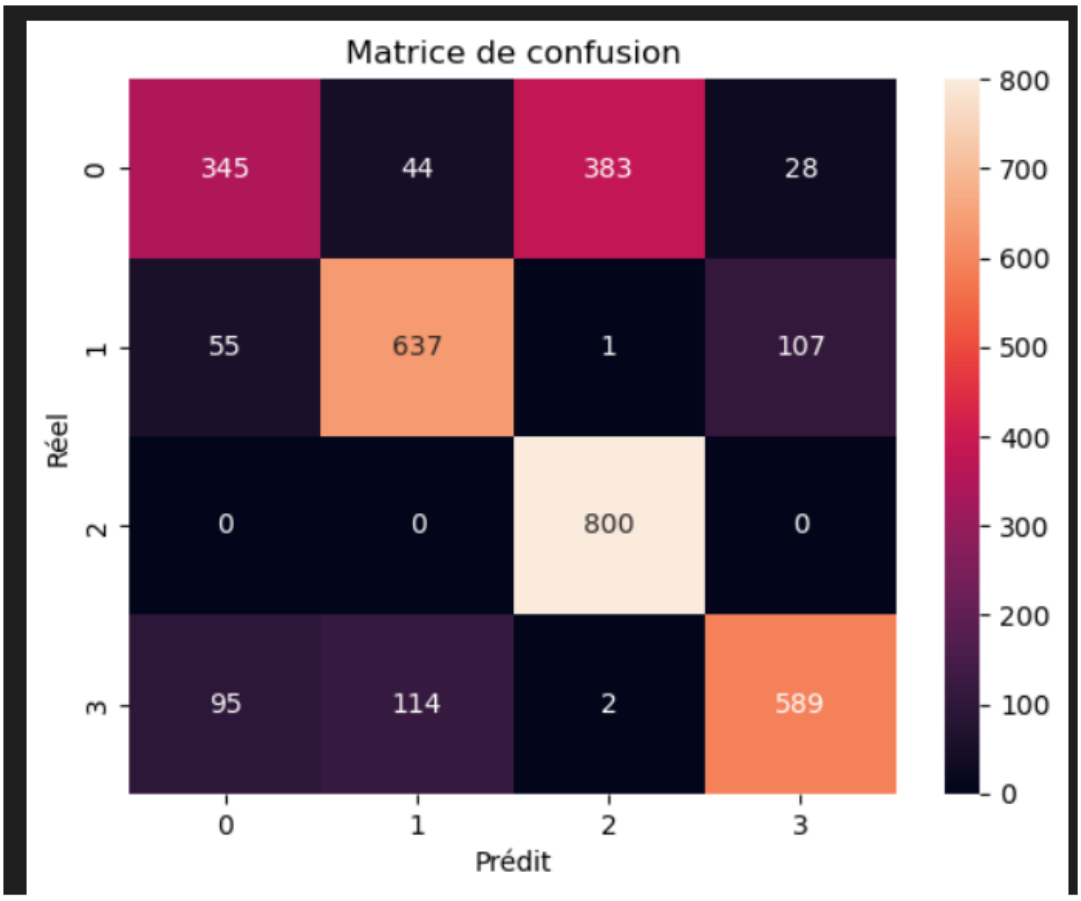

FIGURE 13 – Resultat du modèle après augmentation et équilibrage des classes

On observe ici que tout les labels sont représenter dans le tensor de validation et les resultat sont plus pertinant

Remarques

- Prediction d'images labélisées Covid 19 en pneumonie viral
- Confusion bidirectionelle des images labélisées opacité des poumons et sans pathologies

Points d'amélioration

- Faire varier la taille de chaques classes
- Faire varier le nombre et la nature des augmentations d'images

Conclusion

Ce rapport présente les étapes fondamentales de la préparation des données : exploration, analyse de la variance par PCA, identification des doublons, et premiers tests de rééquilibrage. La prochaine étape consistera à tester des architectures avancées et affiner nos stratégies de preprocessing selon les performances observées.