DataScientest

Rapport d'exploration du projet:

Radiographie Pulmonaires

Bootcamp DS Sept24 : Chris HOZE, Antoine LIONETD, Mickael MELKOWSKI

Table des matières:

Contexte du projet

Une équipe de chercheurs de multiples universités du Moyen Orient et Asie ont assemblé un jeu de données de radiographie du thorax pour des patients, sain, atteint du Covid, de pneumonie et d'opacité pulmonaire.

Ce dataset a été constitué dans le but d'utiliser le deep learning pour le diagnostic du Covid-19 par X-ray plutôt que par RT-PCR[1].

Il ya 4 sources d'images présent dans le dataset (avec entre parenthèse le nombre d'images):

- COVID-19 data (3615)
- Normal images (10192)
- Lung opacity images (6012)
- Viral Pneumonia images (1345)

Chaque image vient avec son masque pré-calculé, généré par apprentissage semi-automatique.
Par son application ont peut donc isoler les poumons et avoir des données concernant uniquement les poumons.

Les images des radios sont au format PNG en 299x299 pixels, celles des masques sont au même format mais en 256x256 pixels.

Objectif

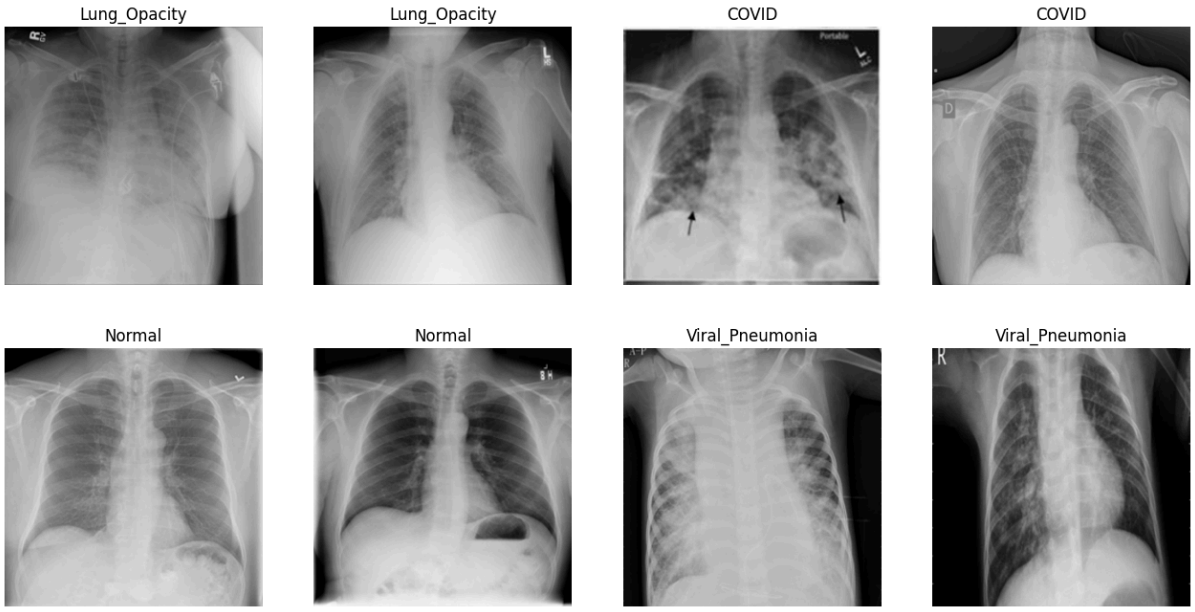
On cherche à utiliser ce jeu de données pour aider au diagnostic d'affection pulmonaire. Ce sujet se présente comme un problème de classification à plusieurs classes.

On s'attachera ici à minimiser le nombre de faux-négatifs i.e des patients pour lesquels on considérerait que la radiographie est normale alors qu'il est sujet à une infection.

[1] M. E. H. Chowdhury *et al.*, "Can AI Help in Screening Viral and COVID-19 Pneumonia?," in *IEEE Access*, vol. 8, pp. 132665-132676, 2020, doi: 10.1109/ACCESS.2020.3010287

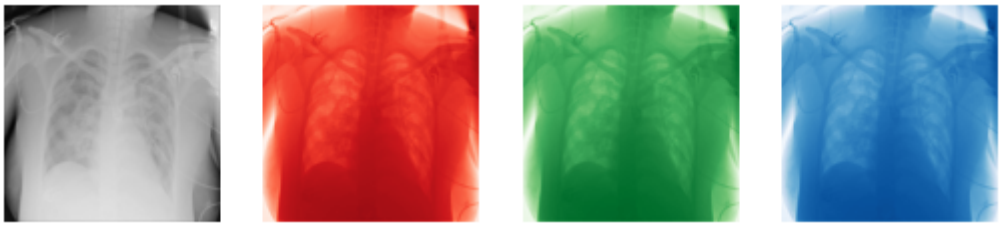
Prise en main et découverte des données:

Visualisation de quelques images du dataset:



Les images sont des radiographies pulmonaires provenant de différentes sources, on remarque rapidement que seule une partie de l'image nous intéressera puisqu'on s'intéressera aux poumons alors que l'image laisse notamment la colonne vertébrale et le cœur. Il faudra donc appliquer les masques qui ont été précalculés.

Les données sont en échelle de gris, une importation en couleur donne trois valeurs identiques pour les différents canaux. Chaque image correspond donc à une matrice de 299*299. Ci-dessous, un exemple d'image sur les trois canaux RGB.



Les seules métadonnées disponibles concernent l'origine des radiographies. Aucune information n'est disponible sur les patients malgré leur intérêt évident pour ce type d' étude.

Pre-processing des données

L'ensemble des images ainsi que les métadonnées ont été importées.

Dans un premier temps, on applique un preprocessing simple sur les données :

- Import
- Redimensionnement du masque à la taille de l'image de 256*256 à 299*299
- Application du masque

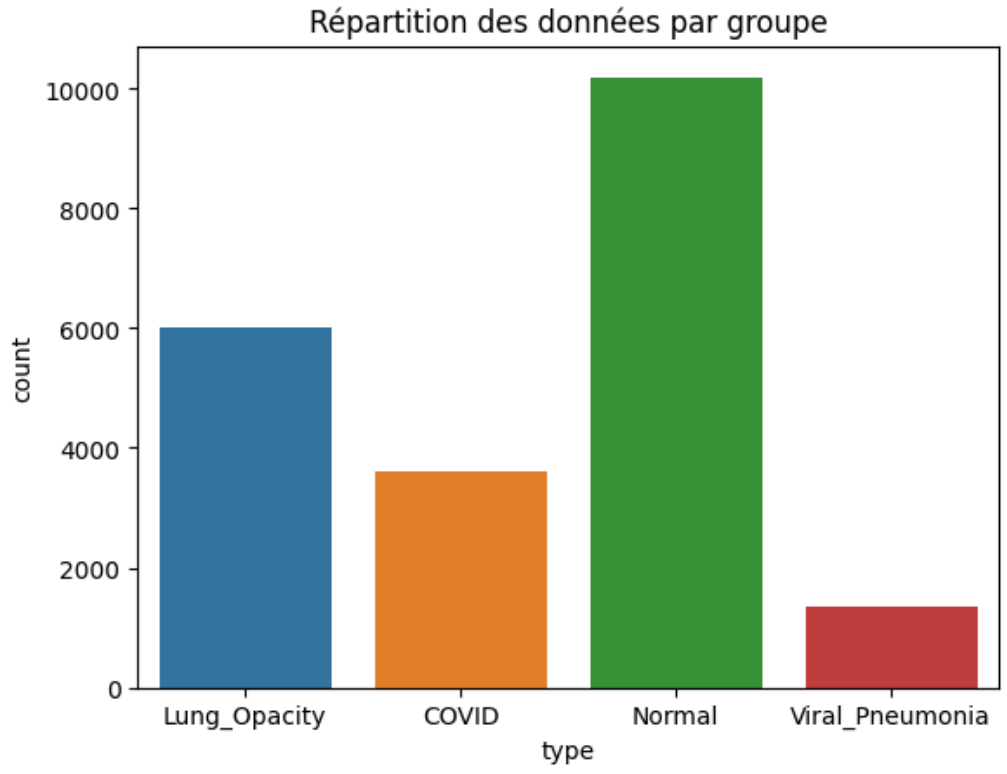
Un jeu de données tabulaire a été créé pour les analyses exploratoires, il contient 21165 lignes et 89406 colonnes correspondant aux métadonnées et les 89400 (299 * 299) pixels. Ce jeu de données sera utilisé pour l'exploration.

FILENAME	FORMAT	SIZE	URL	TYPE	num	0	1	2	3	...	89391	89392	89393	89394	89395	89396	89397	89398	89399	89400
Lung_Opacity-1	PNG	256*256	rnsa	Lung_Opacity	1	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
Lung_Opacity-2	PNG	256*256	rnsa	Lung_Opacity	2	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
Lung_Opacity-3	PNG	256*256	rnsa	Lung_Opacity	3	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
Lung_Opacity-4	PNG	256*256	rnsa	Lung_Opacity	4	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
Lung_Opacity-5	PNG	256*256	rnsa	Lung_Opacity	5	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

Statistiques descriptives sur les données

Répartition des données par catégorie

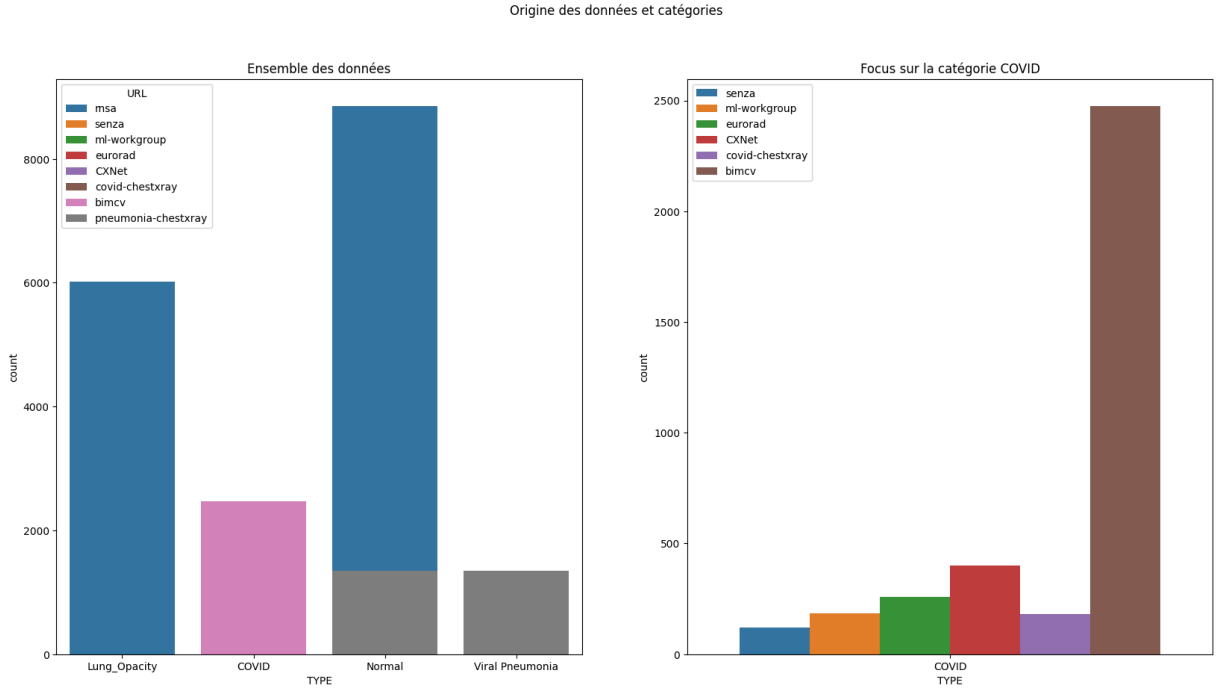
La première analyse réalisée est une visualisation du nombre d'images par catégorie.



On observe que nous n'avons majoritairement des données de radio pulmonaires pour les patients ne présentant pas d'affection. Les données de pneumonie virales sont quant à elles minoritaires.

Répartition des données par source et par catégorie.

Les images proviennent de plusieurs sources différentes. On cherche à visualiser la répartition des catégories par groupe afin d'identifier un éventuel biais.



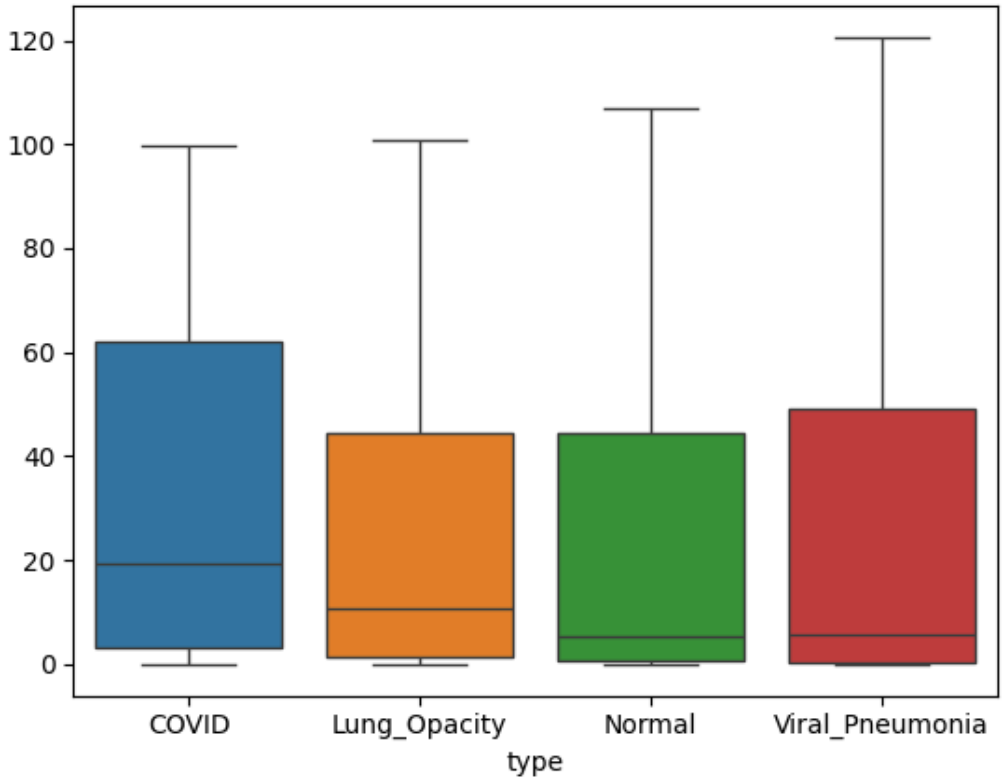
On remarque que les données pour la catégorie COVID proviennent de multiples sources. Les données "Lung Opacity" proviennent d'une source unique, idem pour les données "Viral Pneumonia". Pour chacune de ses sources des données de catégorie "Normal" ont également été fournies ce qui devrait limiter la source de biais.

Visualisation des différences entre catégories.

Dans une première approche, on crée une nouvelle variable contenant les valeurs moyennes pour l'ensemble des pixels pour chaque catégorie après masquage.
Le tableau suivant résume les différences observées sur les valeurs des pixels, en fonction des jeux de données.

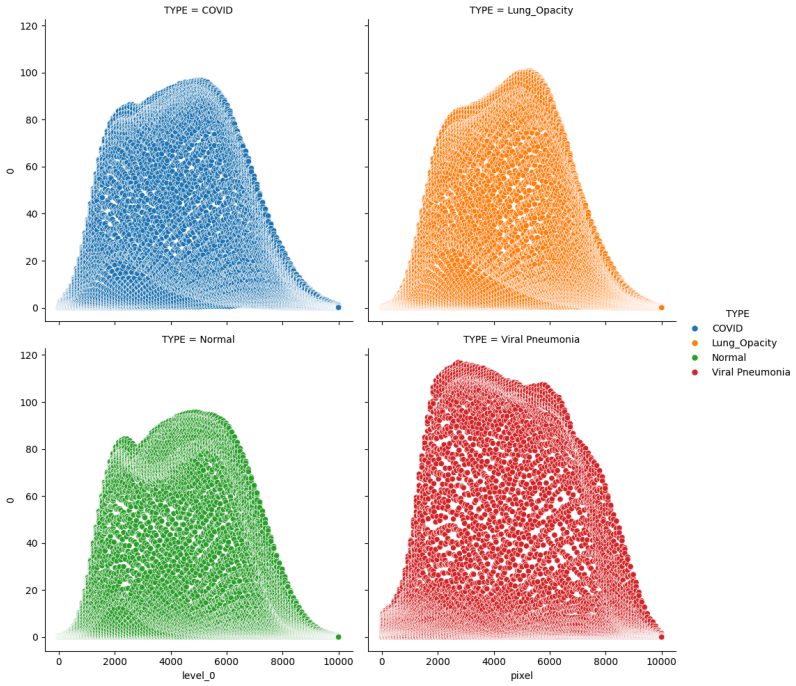
Catégorie	COVID	Lung Opacity	Normal	Viral Pneumonia
Moyenne	32,86	25,16	24,34	27,58
Ecart Type	32,94	29,31	32,22	37,52
Minimum	0,00	0,00	0,00	0,00
Médiane	19,41	10,64	5,18	5,54
Maximum	99,88	100,69	106,73	120,63

On peut également les observer plus visuellement à l'aide de la boîte de dispersion ci-dessous:

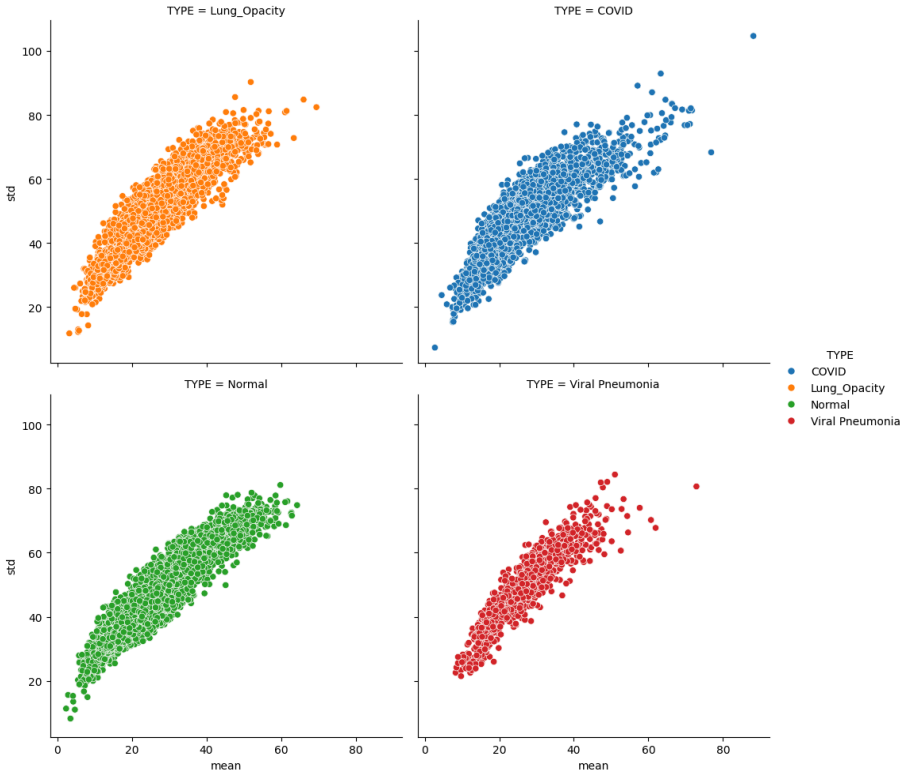


Il semble que la moyenne des données COVID soient plus élevée, ce qui suggère des images plus blanches et donc une cohérence avec la plus grande opacité attendue.
Les données semblent plus extrêmes dans la catégorie Viral Pneumonia.

Une autre possibilité de visualisation est de tracer le nuage de point de la valeur moyenne du pixel en ordonnées et les pixels en abscisse et la moyenne en fonction de l'écart-type.



On visualise également l'écart-type en fonction de la moyenne pour chacune des catégories.

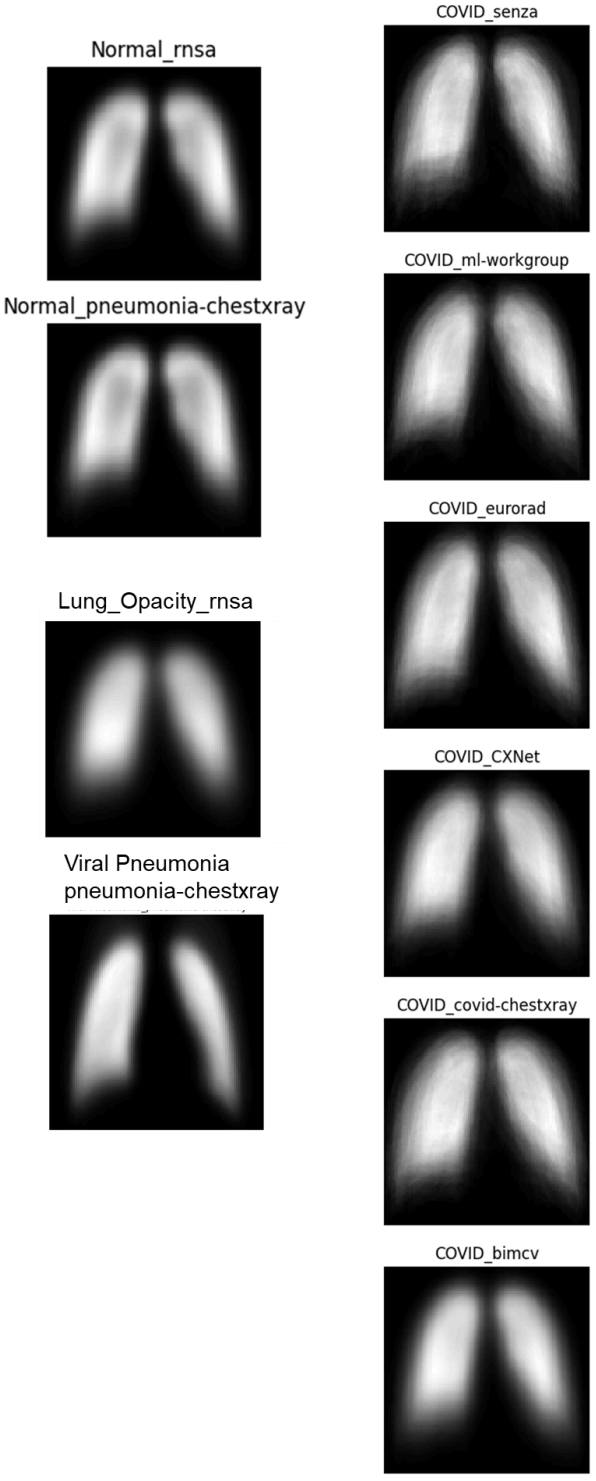


Dans les deux représentations, on remarque que le comportement diffère en fonction des catégories. On distingue des formes plus sombres dans le cas des catégories COVID et Lung Opacity. Les données semblent également plus dispersées pour ces catégories.
Pour mieux visualiser ces différences, on trace l'image moyenne par groupe.

Images moyennes par type et par source

Un autre axe d'exploration est de procéder au calcul de l'image moyenne par type et par sources pour tenter d'y déceler un biais.

Visualisation



Interprétation

Il semble y avoir un biais sur les analyses radio pulmonaires liées à la pneumonie virale : le cœur est proportionnellement plus important et l'application du masque efface la partie inférieure du poumon gauche. Ce biais est lié à l'origine des données issus du dataset "pneumonia-chestxray".

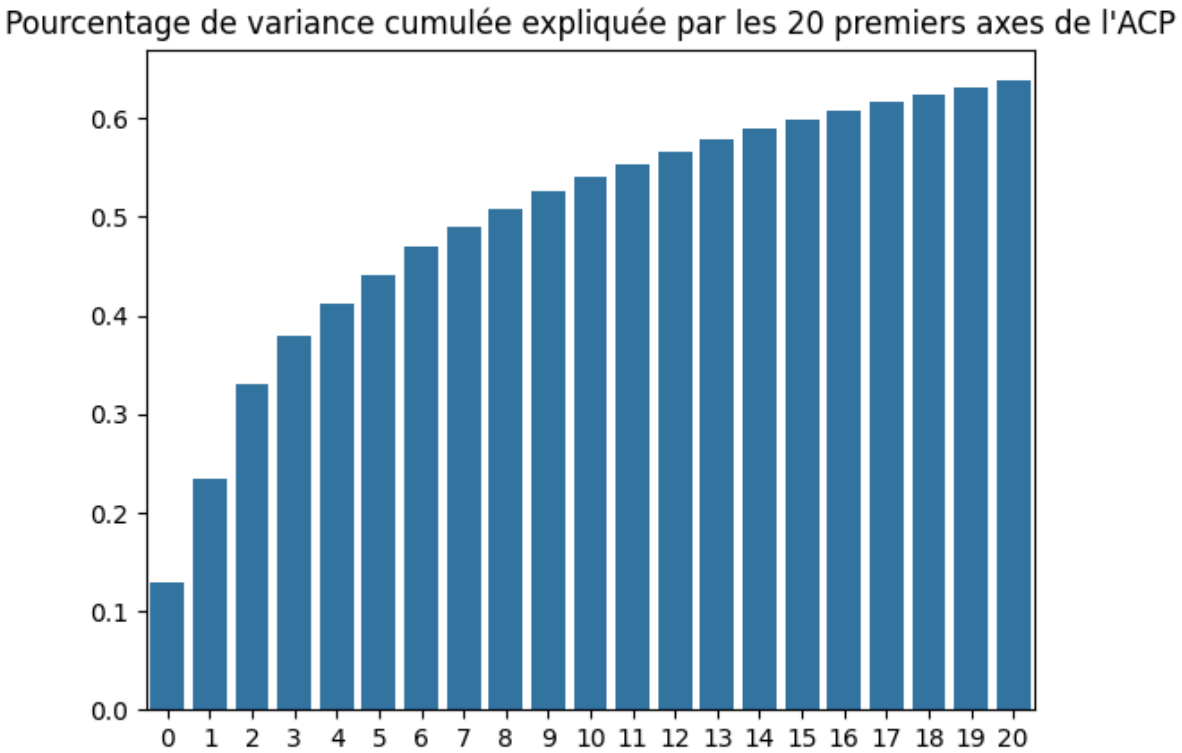
Sans certitude, on émet l'hypothèse à la vue des images que ces données correspondent à des radios pulmonaires d'enfants.

Une partie des données issues de ce jeu de données étant également de catégorie "Normal" on retrouve le même phénomène de façon moins marquée sur la catégorie normale.

Analyses en composantes principales

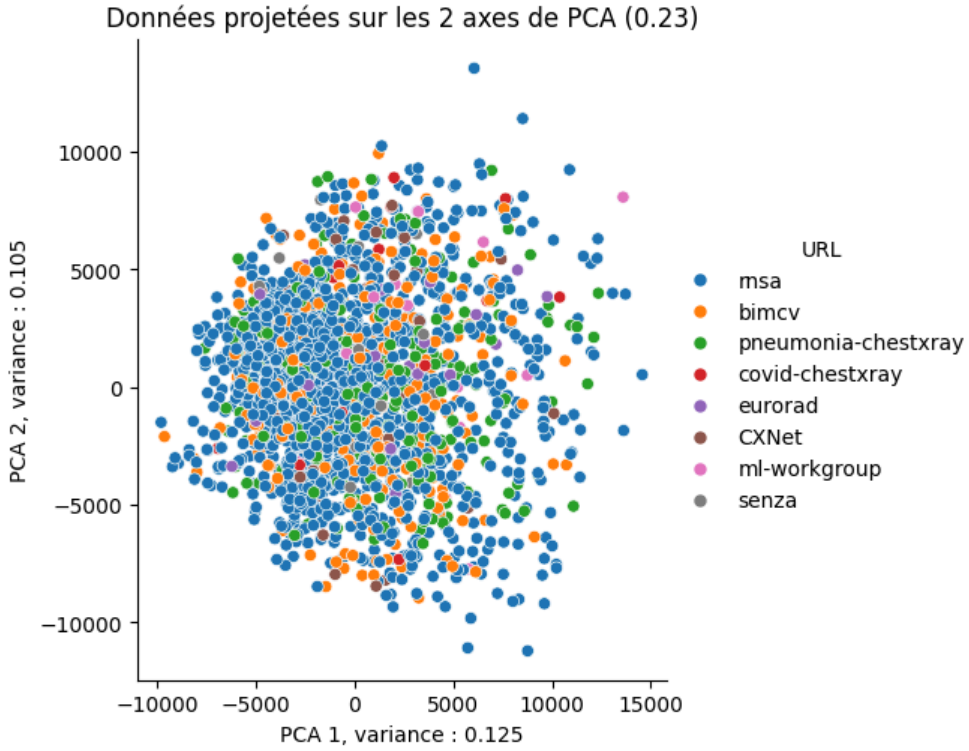
Part de variance expliquée par les composantes de l'ACP

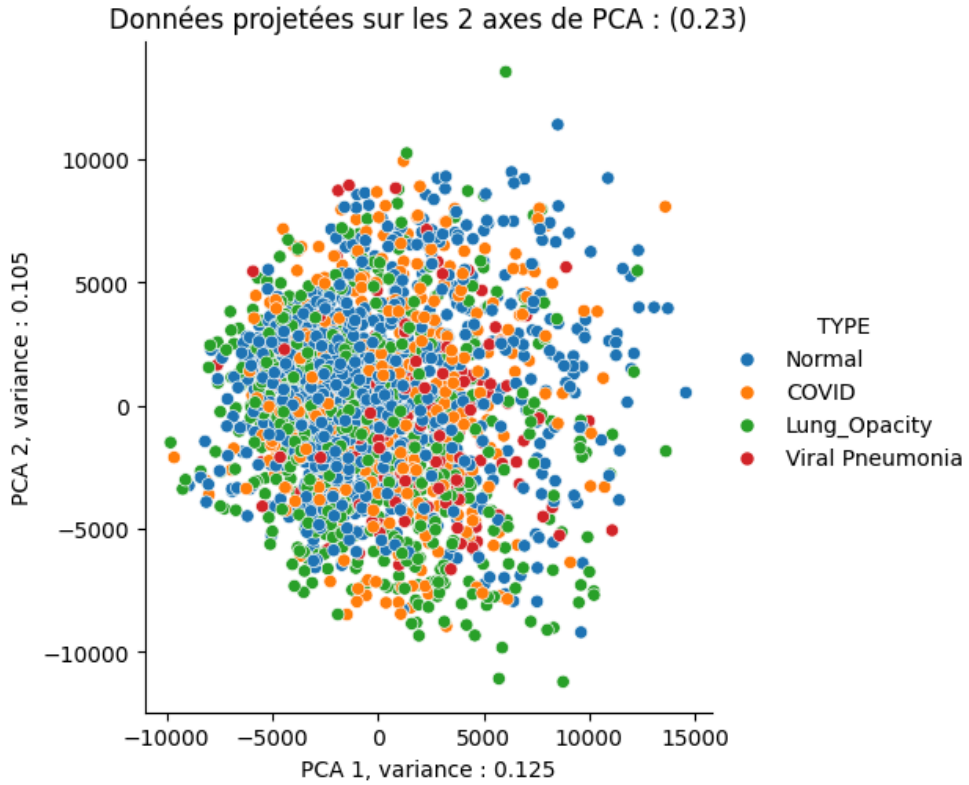
Pour des raisons de mémoire, l'analyse en composante principale a été réalisée en réduisant les images en dimension 100*100.
La première figure donne le pourcentage de variance sur les 20 premières composantes. On remarque que les dix premières composantes expliquent 51,6 % de la variabilité. Il faut en revanche 785 variables pour atteindre 95% de la variance initiale.



Visualisation sur les deux premières composantes de l'ACP

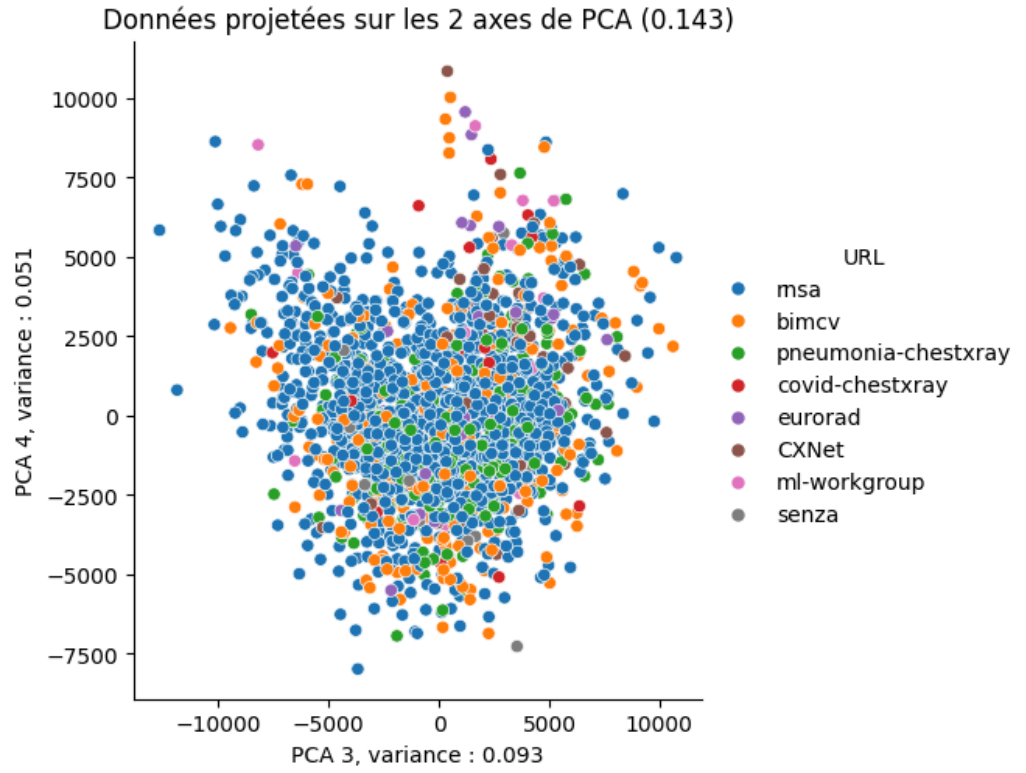
On visualise les coordonnées des échantillons sur les plans 1 et 2 pour voir s'il est possible de séparer les catégories selon un plan. On étudie également la répartition par source de données pour identifier une éventuelle source de biais. :

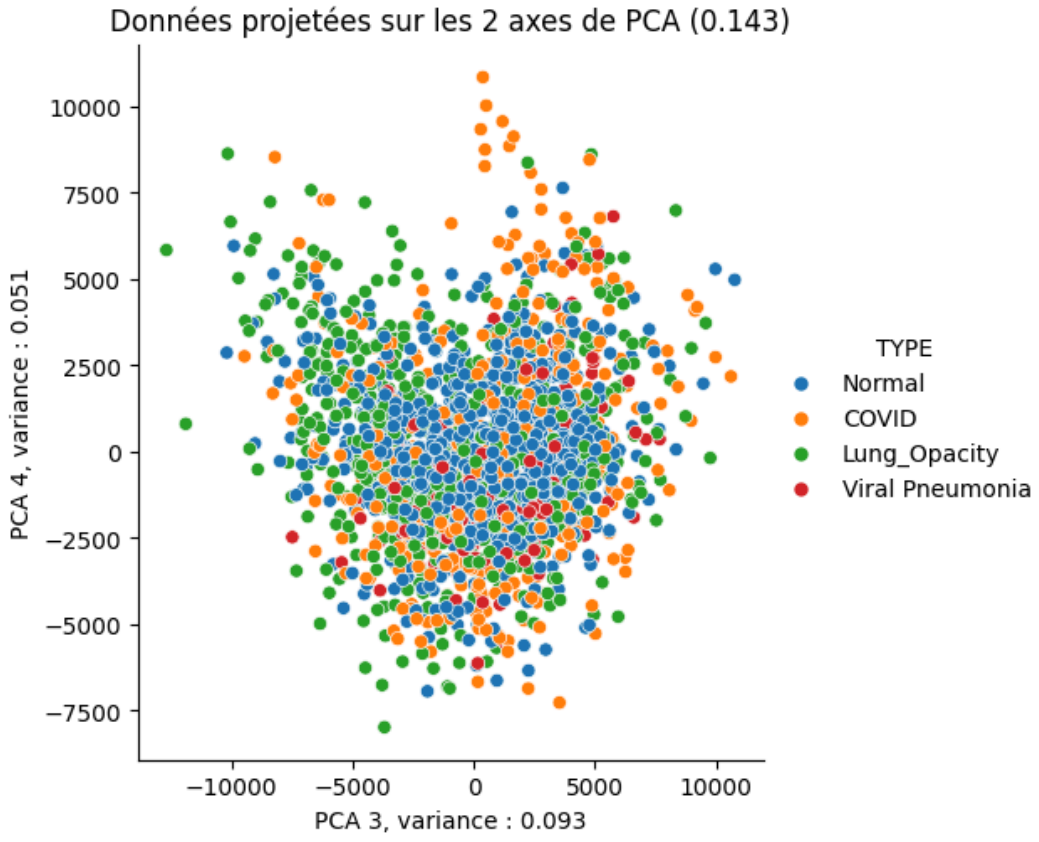




Visualisation sur les troisième et quatrième composantes de l'ACP

On visualise les échantillons sur le plan créé par les composantes 3 et 4 de l'ACP.





Interprétation

Les axes ne permettent pas de séparer les échantillons par groupe, et on identifie pas de sous-population en fonction de la source de données.
L'orientation, le niveau de zoom et la taille des poumons diffèrent selon les patients et les radiographies ce qui peut expliquer pourquoi il est difficile de distinguer les classes avec les méthodes de statistiques classiques. Un argument supplémentaire pour indiquer que les caractéristiques des radiographies sont variables est le faible nombre de variables qui sont identiques pour l'ensemble de l'échantillon (1321 soit 1.5%) après masquage alors qu'on aurait pu s'attendre à des zones masquées communes sur des régions non concernées par les poumons.

Conclusion sur l'exploration des données

On ne constate donc pas de solution évidente à notre problème de classification avec une séparation dans un plan ce qui suggère que des modèles de classification classiques ne sont pas adaptés. Des solutions plus complexes seront donc à mettre en œuvre notamment avec l'emploi de modèle de Deep Learning.

Premier pas d'utilisation de Deep Learning

Afin de jauger la difficulté de ce problème de classification des premiers tests ont été fait sur un jeu de données réduit composés de 2000 images (500 par type) redimensionnés en 256*256 pixels.

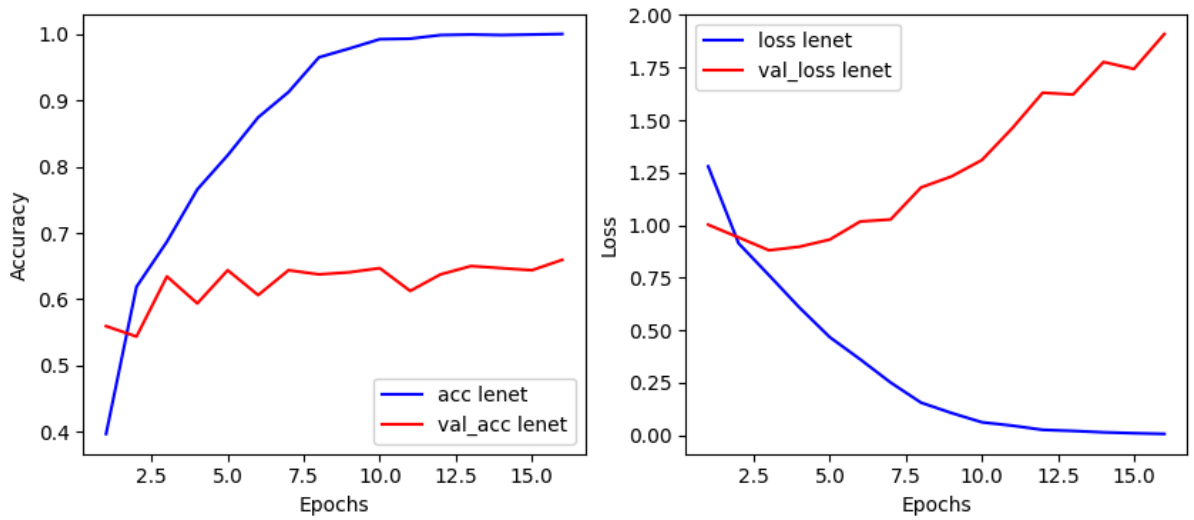
Deux modèles connus de la littérature ont été utilisés, LeNet[2] et Dense-121[3] à travers les librairies TensorFlow et PyTorch.

Les modèles ont été testés avec leurs paramètres par défaut en guise de benchmark. Le choix de Tensorflow ou PyTorch est purement arbitraire pour pouvoir se familiariser avec l'écriture du code.

Test de modèles LeNet via Tensorflow

Test fait sur 16 epochs avec une taille de batch de 32

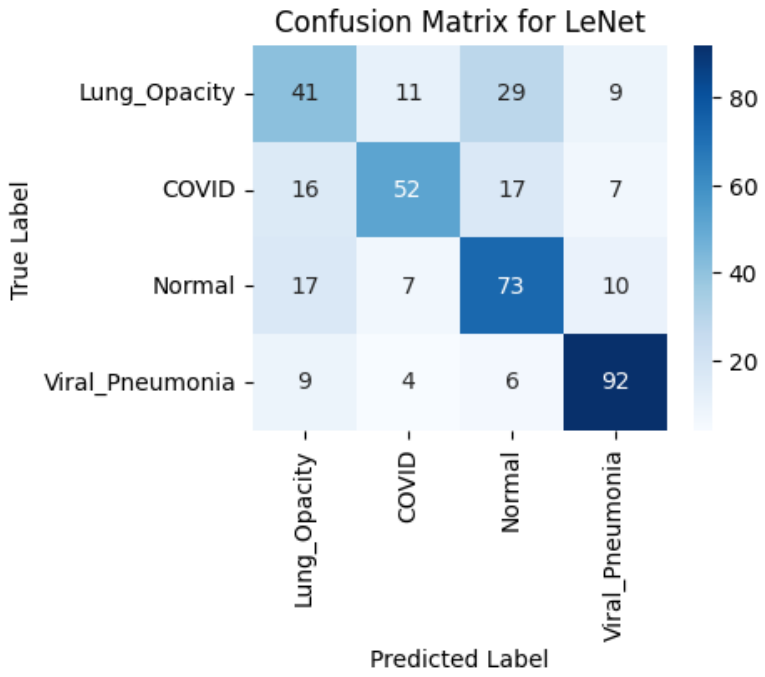
Accuracy et Loss



Il y a ici un overfitting sur les données, l'accuracy sur le jeu de validation ne s'améliorant que peu et la loss augmentant aussi.

[2] "Gradient-based learning applied to document recognition" Lecun, Y et al. *Proceedings of the IEEE*. **86** (11): 2278–2324.
[3] Densely Connected Convolutional Networks, Huang, G et al: arXiv:1608.06993

Matrice de confusion

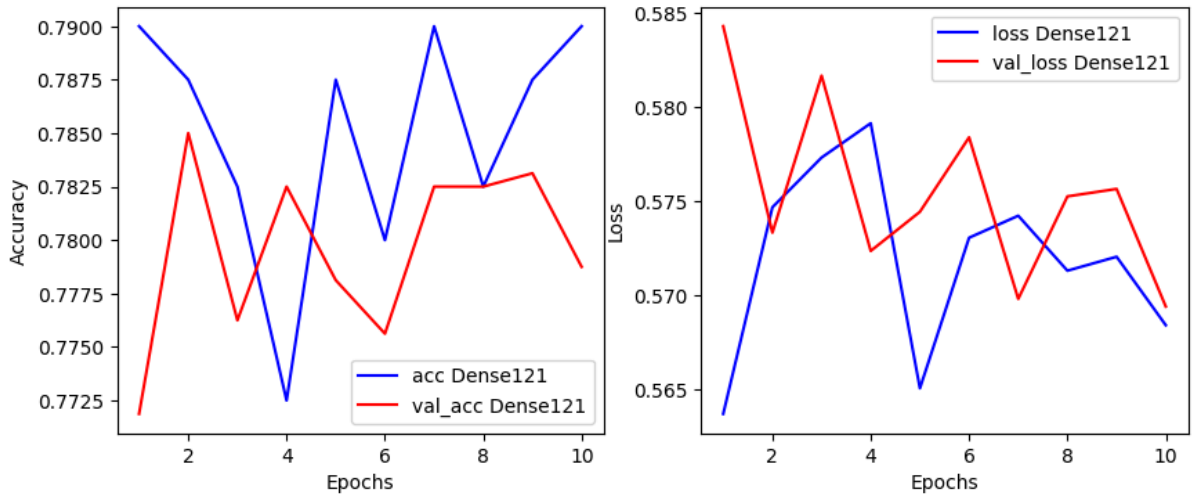


Test de modèles via PyTorch

Test de model dense-121

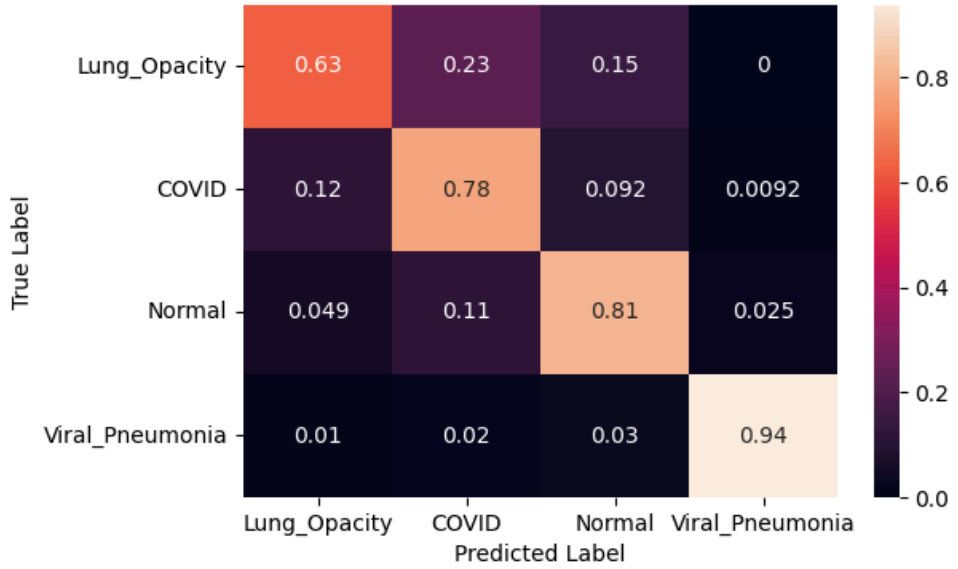
Test fait sur 10 epochs avec une taille de batch de 32

Accuracy et Loss



Le modèle a l'air d'avoir une tendance un peu meilleure mais on est encore loin d'un modèle parfait.

Matrice de confusion



Ces premiers modèles présentent des résultats encourageants sur la pertinence de l'utilisation du deep learning avec en moyenne environ 80% de classification correctes pour Dense-121.

L'amélioration ou le choix de modèles plus adaptés pourra nous permettre d'assurer de meilleurs résultats pour les prochaines phases du projet.