Projet Analyse de Radiographies Pulmonaires

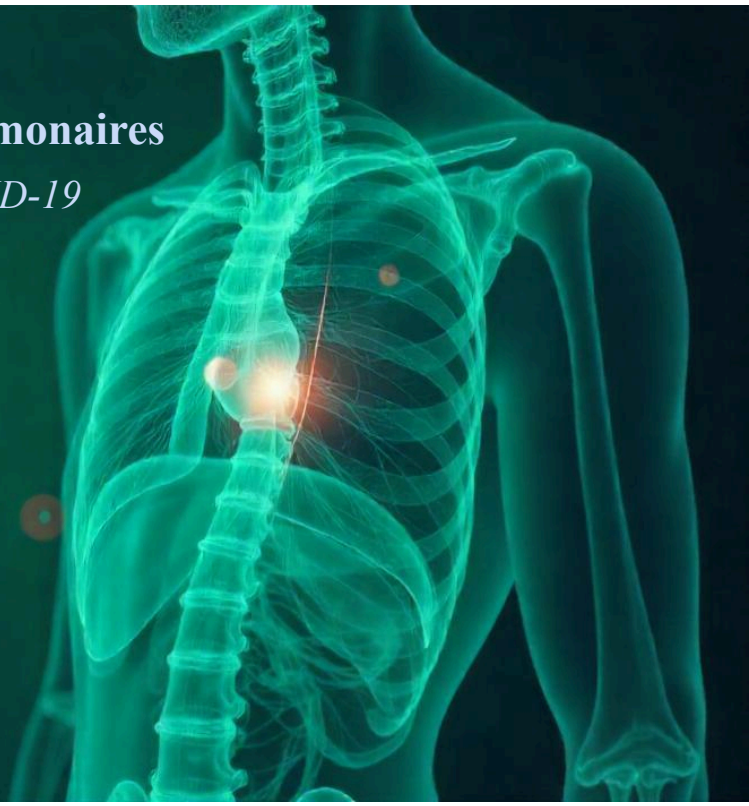
Analyse de radiographies pulmonaires COVID-19

Pour **DataScientest**

Préparation du Diplôme de Data Scientist

Par :
Antoine BAS
Jeremy CHOUIPPE
Antoine CARTON
Andreas LATOUR

v.0.6 - le 20/12/2024



Sommaire

1 - Introduction

Dans le cadre du parcours de formation en vue du diplôme de Data Scientist préparé avec l'organisme de formation DataScientest en collaboration avec l'école des Mines de Paris, nous avons été sélectionnées pour réaliser un projet autour de l'exploitation d'images médicales.

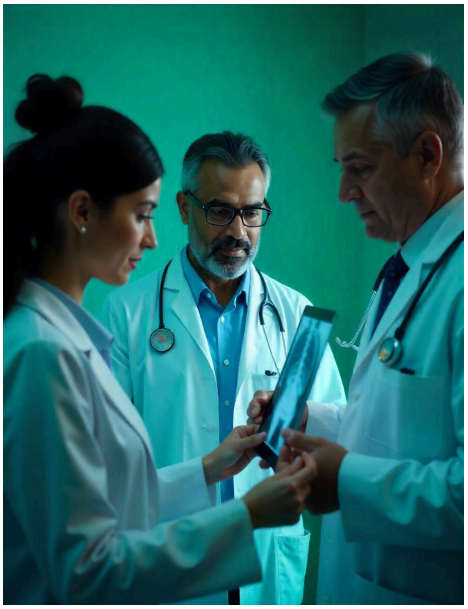
Ce projet sera alors articulé autour des technologies de Deep Learning afin d'apporter une réponse précise sur l'exploitation des images.

Notre Dataset est composé d'un fichier de description des données, de radiographies thoraciques classées par catégorie d'infections ainsi que de fichiers contenant les métadonnées des images.

Notre équipe se compose de :

- Antoine BAS,
- Antoine CARTON,
- Jeremy CHOUIPPE,
- Andreas LATOUR,

Notre projet a été suivi et encadré par notre Mentor, M. LESIEUR Romain.



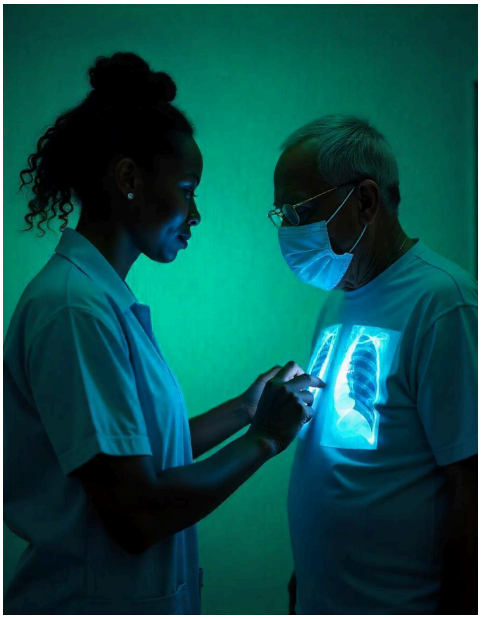
1.1 - Contexte Général

Fin 2019, notre planète a connu l'une des pandémies les plus importantes de son histoire, certes d'autres pandémies bien plus grave ont eu lieu sur Terre mais celle du COVID-19 à marquer toute une génération. Tout commence en Décembre 2019 à Wuhan, dans la province du Hubei, où une maladie alors inconnue voit le jour suite à une transmission via le vecteur animal.

La contagion se développe rapidement et la maladie gagne rapidement du terrain. La Chine sonne très vite l'alerte auprès des autorités sanitaires mondiales (OMS) et ce pays prend des mesures d'isolement, de quarantaine et de désinfection à grande échelle comme nul autre pays ne l'avait fait auparavant. Malgré toutes ses dispositions efficaces, l'infection a déjà traversé les frontières et de nouveaux cas se manifestent dans d'autres endroits du monde. La France à tout d'abord mis le point sur la sécurité sanitaire privilégiant l'hygiène et la protection puis des mesures de couvre-feu pour aller jusqu'aux différentes périodes de confinement.

La pandémie de COVID-19 a rendu clair le besoin d'une collaboration internationale, d'une facilitation de l'accès aux données et de la standardisation de leur analyse. Depuis cet évènement mondial, des efforts de partage de données médicales anonymisées (Google Health, Organisation Mondiale de la Santé, European Center for Disease Prevention and Control) ont rendu possible la mise à contribution des dernières avancées en data science pour rendre plus rapide et reproductible le diagnostic, la prise en charge des patients atteints de cette nouvelle maladie virale.

Aujourd'hui, on connaît une accélération des politiques internationales pour faciliter les échanges internationaux des données clés dans la prise de décision à l'échelle planétaire et la protection des espèces. Parallèlement l'emballement des indicateurs du réchauffement climatique rendent plus que probable le déclenchement d'une nouvelle pandémie. Cette situation rend donc cruciale la formation et la familiarisation de la population à l'analyse et l'interprétation des données médicales. C'est dans cette optique que s'inscrit notre projet.



1.2 - Contexte Médical

Aujourd'hui l'admission d'un patient aux urgences qui présente une infection à la COVID-19 (suspectée ou confirmée) avec des difficultés respiratoires ou une comorbidité, requiert selon les recommandations des médecins le passage d'un examen médical approfondi avec le passage d'une tomodensitométrie (appelé également scanner) ou bien d'une radiographie du thorax, en complément de tests biologiques. Le but est de confirmer et réfuter la présence de la COVID-19 et d'évaluer la gravité des lésions. Le protocole est le même en médecine générale.

Cet examen permet de mieux comprendre l'impact de la maladie sur le patient par la présence ou non de lésions au niveau des tissus pulmonaires, là où le test biologique ne sera pas efficace et n'apportera pas cette précision.

La radiographie est moins performante que la TDM pour la détection et le suivi de la COVID-19 mais présente l'avantage d'être moins invasive avec moins d'exposition aux rayons X préservant ainsi le patient de radiations néfastes. Elle est également moins coûteuse et plus facilement accessible pour la prise de rendez-vous. En effet, les scanner sont moins nombreux et souvent plus sollicités.

L'objectif du radiologue lorsqu'il analyse une radiographie pulmonaire d'un patient qui présente des difficultés respiratoires, avec une suspicion de la COVID-19, est de vérifier la présence ou non d'un syndrome alvéolo-interstitiel bilatéral prédominant aux bases, caractéristiques d'une pneumonie virale. La difficulté ensuite pour le radiologue va être d'identifier les caractéristiques de la COVID-19 plutôt que d'une pneumonie simple ou virale, les deux étant très semblables sur l'imagerie qu'apporte la radiographie pulmonaire.

Il est important de noter que le résultat de la radiographie thoracique n'aura de valeur que si elle est positive, c'est-à-dire si elle montre la présente des signes provoqués par la COVID-19. En effet, la présence de cette infection n'induit pas forcément de lésions notables sur la radiographie du thorax. A contrario, une radiographie qui s'avère négative n'écarte donc pas la présence de la COVID-19 chez le patient.

1.3 - Contexte de l'étude

Notre étude porte sur un jeu de données qui sera notre clé de voûte pour notre analyse et la mise en place de notre solution. Ces éléments présentés sous formes d'images de radiographie catégorisées selon des albums labellisés, classés par type d'infection et des fichiers contenant les métadonnées comme informations de composition. Nous détaillerons plus finement les informations contenues dans notre dataset plus tard. Toutefois vous pouvez consulter l'Annexe afin de trouver le contenu du fichier "README.md.txt". Il y a plusieurs aspects importants qui sont des leviers pour le développement d'une solution de Deep Learning permettant de répondre et mettre en avant de nombreux avantages.

Tout d'abord comme le démontre le contexte médical, il serait pertinent pour les radiologues d'avoir un modèle prédictif capable de déterminer la présence de la COVID-19 et idéalement de la différencier de d'autres maladies pulmonaires comme la pneumonie simple ou virale. Cela apporterait une augmentation de la fiabilité du diagnostic et donc une meilleure prise en charge pour les patients tant en temps d'attente qu'en amélioration des prescriptions et de leur portée thérapeutique.

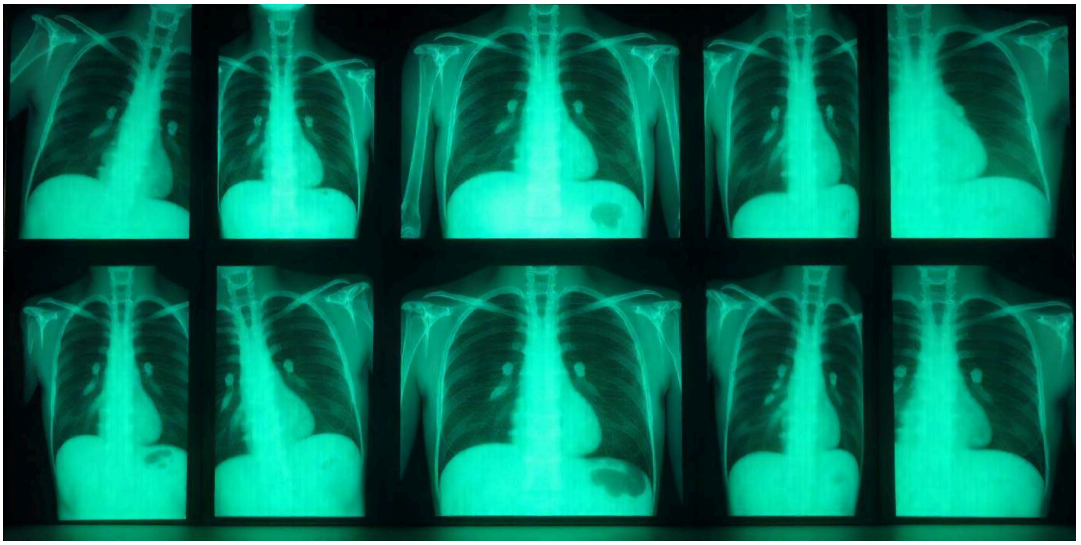
L'autre intérêt est d'encourager l'utilisation de la radiographie pulmonaire et ainsi désengorger les centres de scanographie. Permettant une fluidité dans la prise de rendez-vous pour la réalisation de ses examens mais également de donner un réelle coup de souffle pour des pathologies plus graves et nécessitant un diagnostic renforcé. Moins coûteux et plus facilement accessible pour la prise de rendez-vous, la radiographie est l'examen à privilégier à condition que ce type d'imagerie propose la même finalité et conclusion sur le diagnostic.

Au-delà de l'aspect médical et logistique vient s'ajouter le point de vue économique sur la mise en place d'un tel système. Réduire les coûts serait une opportunité pour la Sécurité Sociale ainsi que pour le contribuable. Pour illustrer nos propos voici ce qu'ont révélé nos recherches.
Soit le coût d'une radiographie pulmonaire est de 21,28 € contre 25,27 € pour un scanner du thorax, basé sur les tarifs conventionnés.
(source : https://www.santiane.fr/mutuelle-sante/guides/imagerie-medicale-remboursement).

La réussite de cette étude et du développement d'un produit efficace dans le diagnostic de l'infection causé par la COVID-19 fera ressortir de nombreux avantages non négligeables pour toutes les parties prenantes.

1.4 - Objectifs

L'objectif principal de ce projet est d'analyser des données images afin de créer et entraîner un modèle de Deep Learning capable d'identifier la COVID-19 chez un patient à partir de sa radiographie pulmonaire. Une tolérance peut être admises sur la différenciation entre la COVID-19 et une pneumonie, car elle peut être précisée par le test biologique (PCR, antigénique, etc...).



Le niveau d'expertise autour de la problématique pour les membres du groupe est le suivant :

- Antoine BAS: peu de connaissance dans le milieu médical, il a des notions grâce à son entourage dans ce milieu. Son expérience l'a rapproché de la modélisation et de l'utilisation des données.
- Jeremy CHOUIPPE: aucune connaissance en lecture de radiographie pulmonaire, a toutefois des connaissances en terme de médecine générale
- Antoine CARTON: bon socle de connaissance des sciences, il a toujours un pied dans la recherche, il a des compétences en visualisation de données
- Andreas LATOUR: aucune connaissance dans le milieu médical ou la radiologie pulmonaire mais il a de bonne notion en programmation

Le projet se déroulera alors en différentes étapes soit :

1°) l'observation des données, la prise de note ainsi qu'une réflexion sur la méthodologie de départ (cadencement des réunions, accueil de l'équipe...),

2°) la mise en place d'outils de suivi (Jira), d'un plateforme de collaboration pour l'édition de code source (GitHub), d'un espace de stockage de données commun (Google Drive) et d'un modèle de document pour l'identité visuel du projet,

3°) l'analyse du dataset, la création du code source d'exploitation des fichiers contenant les métadonnées,

4°) la réalisation d'un script Python afin de permettre la création de représentation graphique suite à l'exploitation des données (proportions des données présentes, nombre de données, analyse de différences qualitatives entre les jeux de données, étude de la composition des images...)

5°) l'élaboration de code source en vue d'identifier le meilleur modèle pour la modifications et l'exploitation des images

6°) la création d'un modèle d'entraînement enfin d'organiser la prédiction sur les radiographies

7°) la mise à l'épreuve du modèle d'entraînement en prenant en compte une vue théorique (modèles mathématiques)

8°) l'accomplissement de l'étude par une phase de tests pratiques en introduisant des radiographies afin de visualiser un diagnostic

9°) conclusion de l'étude permettant de tirer profit des informations et créations logiciels en vue d'accroître la vision sur les possibilités réelles qu'offre un modèle de prédiction.

Des échanges ont été réalisés avec une étudiante en médecine, qui ont permis de préciser le contexte du projet, les enjeux rencontrés par les médecins lors du diagnostic de la COVID-19, notamment les avantages et inconvénients de l'utilisation de la radiographie pulmonaire plutôt que d'autres solutions.

2 - Compréhension et Manipulation des données

2.1 - Cadre

Le jeu de données que nous utiliserons lors de ce projet est celui mis à disposition en open source sur le site Internet Kaggle.com (dataset complet) par une équipe de chercheurs de l'université du Qatar situé à Doha, ainsi que l'université de Dhaka au Bangladesh, en collaboration avec des médecins. Ce jeu de données a été construit à partir de plusieurs sources de données mais a été vérifié et mis en forme par cette équipe.

Le jeu de données comporte des images de radiographies au format PNG, de résolution 299 pixels par 299 pixels ainsi que les masques associés, permettant d'isoler les poumons dans les radiographies, au même format PNG de résolution 256 pixels par 256 pixels. La résolution des radiographies n'est pas la même que celle des masques. La taille de ce jeu de données complet est de 807 MB.

Exemple d'image et de masque correspondant :

Exemple de masque correspondant :

Fig. 1: Exemples de radiographies pulmonaires fournies dans les datasets (rang du haut) et les masques associés (rang du bas)

La majorité des images ressemblent aux deux premières. Cela dit, nous constatons quelques exceptions d'image pivotées de 90°. Les masques correspondant ne dessinent plus les poumons du patient et ne sont présentes que dans le dataset sans pathologie pulmonaire alors classé comme type "Normal"

2.2 - Observation des données

Nous avons pu constater que notre jeu de données est très varié tant dans sa composition que dans la qualité de son contenu. Afin d'affiner notre étude nous avons réalisé différents graphiques d'observation générale comme décrit ci-dessous.

Notre dataset d'images est constitué selon la répartition suivante :

- 1 345 radiographies de patients ayant une pneumonie virale (6%)
- 3 616 radiographies de patients ayant la COVID-19 (17%)
- 6 012 radiographies de patients avec une opacité pulmonaire (autres infections) (28%)
- 10 192 radiographies de patients sans pathologie pulmonaire (48%)

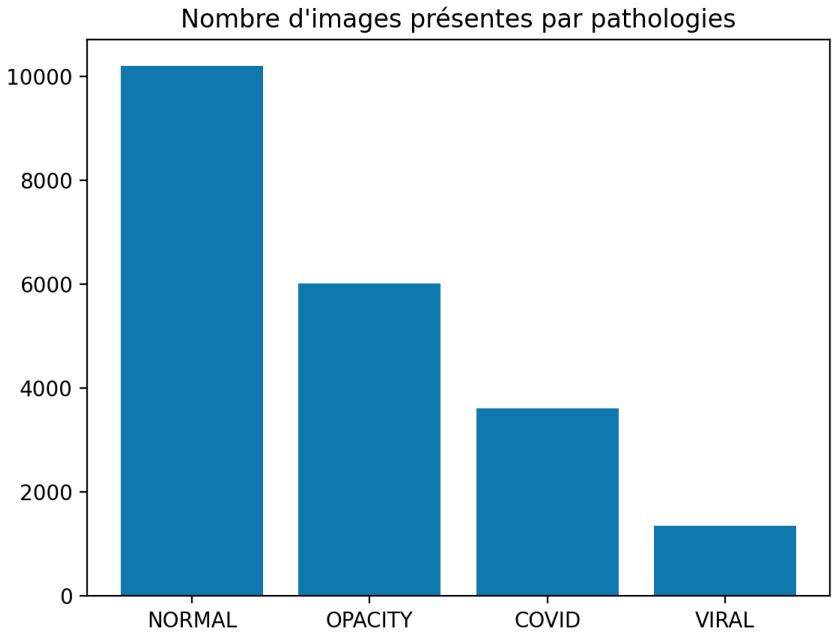


Fig. 2: Représentation de la composition en nombre d'images

2.3 - Mise en place d'outils

Afin d'optimiser la collaboration entre les membres de l'équipe nous avons mis en place différents outils.

Ainsi nous avons utilisé les logiciels suivant :

- Google Drive comme espace de stockage collaboratif
- Google Doc et Google Sheet pour l'édition de document en ligne
- Google Meet pour les points téléphoniques pour le déroulement avec l'équipe projet
- Jira pour la planification et le suivi du projet en suivant la méthode SCRUM
- Github pour la partie développement des différents codes sources permettant une approche de partage, une gestion flexible des versions et une protection liée à des sauvegardes régulières
- Whatsapp afin de communiquer ensemble sur les idées, les attendus, les avancées mais aussi pour mettre un peu de fun durant nos recherches.
- Slack pour la partie organisation des réunions et du suivi avec notre Mentor
- Zoom pour les réunions de suivi en visioconférence entre le Mentor et l'équipe de travail

2.4 - Pre-processing et feature engineering

Le jeu de données initial comporte quatres fichiers de métadonnées séparés entre les pathologies.

Ces fichiers comportent:

- FILE NAME : nom du fichier image associé
- FORMAT : donnée textuelle qui décrit le format de l'image (PNG dans 100% des cas)
- SIZE : la taille des images
- URL : la source d'extraction de l'image

Un premier point est la taille des images renseignée comme étant 256x256 pixels mais les images sont en 299x299 pixels -> Comme mentionné plus tôt ce sont les masques qui sont en 256x256. Nous avons donc décidé de corriger cette information et de remplacer par 299x299.

Pour la suite de notre étude, il sera important d'avoir un seul et même fichier qui répertorie toutes les images disponibles pour les quatres pathologies. Dans cette optique, nous avons ajouté une colonne LABEL aux fichiers avec pour valeur la pathologie (COVID, NORMAL, LUNG OPACITY et VIRAL PNEUMONIA) et concaténé les quatres fichiers en un seul.

Après cela, nous avons trouvé pertinent d'ajouter les liens vers nos images à partir de la racine du projet GitHub afin d'accéder beaucoup plus facilement aux images et aux masques. Nous avons donc un chemin d'accès relatif à la racine du projet pour chaque image, et lorsque nous aurons besoin d'accéder à ces images, il suffira d'utiliser une fonction que nous avons créé pour obtenir le chemin absolu, peu importe la machine et le sous-dossier où le script est exécuté.

Enfin, il a été question d'uniformiser les tailles des images et des masques. Nous avons pris le parti d'augmenter la taille des masques pour ne perdre aucun détail sur les radiographies. On a donc utilisé un algorithme d'augmentation d'image par interpolation du plus proche voisin, qui permet de conserver la binarité des masques (valeur 0 et 255).

Masque d'origine : (256x256)px

Masque augmenté : (299x299)px

La forme finale de notre fichier de métadonnées est donc la suivante:

	FILE NAME	FORMAT	SIZE	URL	LABEL	IMG_URL	MASK_URL	MASK_RESIZED_URL
0	COVID-1	PNG	299*299	https://sirm.org/category/senza-categoria/covi...	COVID	data\raw\COVID\images\COVID-1.png	data\raw\COVID\masks\COVID-1.png	data\processed\COVID\masks\COVID-1.png
1	COVID-2	PNG	299*299	https://sirm.org/category/senza-categoria/covi...	COVID	data\raw\COVID\images\COVID-2.png	data\raw\COVID\masks\COVID-2.png	data\processed\COVID\masks\COVID-2.png
2	COVID-3	PNG	299*299	https://sirm.org/category/senza-categoria/covi...	COVID	data\raw\COVID\images\COVID-3.png	data\raw\COVID\masks\COVID-3.png	data\processed\COVID\masks\COVID-3.png
3	COVID-4	PNG	299*299	https://sirm.org/category/senza-categoria/covi...	COVID	data\raw\COVID\images\COVID-4.png	data\raw\COVID\masks\COVID-4.png	data\processed\COVID\masks\COVID-4.png
4	COVID-5	PNG	299*299	https://sirm.org/category/senza-categoria/covi...	COVID	data\raw\COVID\images\COVID-5.png	data\raw\COVID\masks\COVID-5.png	data\processed\COVID\masks\COVID-5.png

2.5 - Analyse du jeu de données

Comme nous l'avons vu dans les sections précédentes notre dataset est composé de plusieurs éléments. Nous nous focaliserons ici sur la partie des métadonnées et son exploitation.

Au sein de ces différents dossiers, les données ont été collectées et stockées dans 7 jeux de données différents. Nous nous sommes intéressés à la proportion respective de ces jeux de données dans la totalité des images radiographiques.

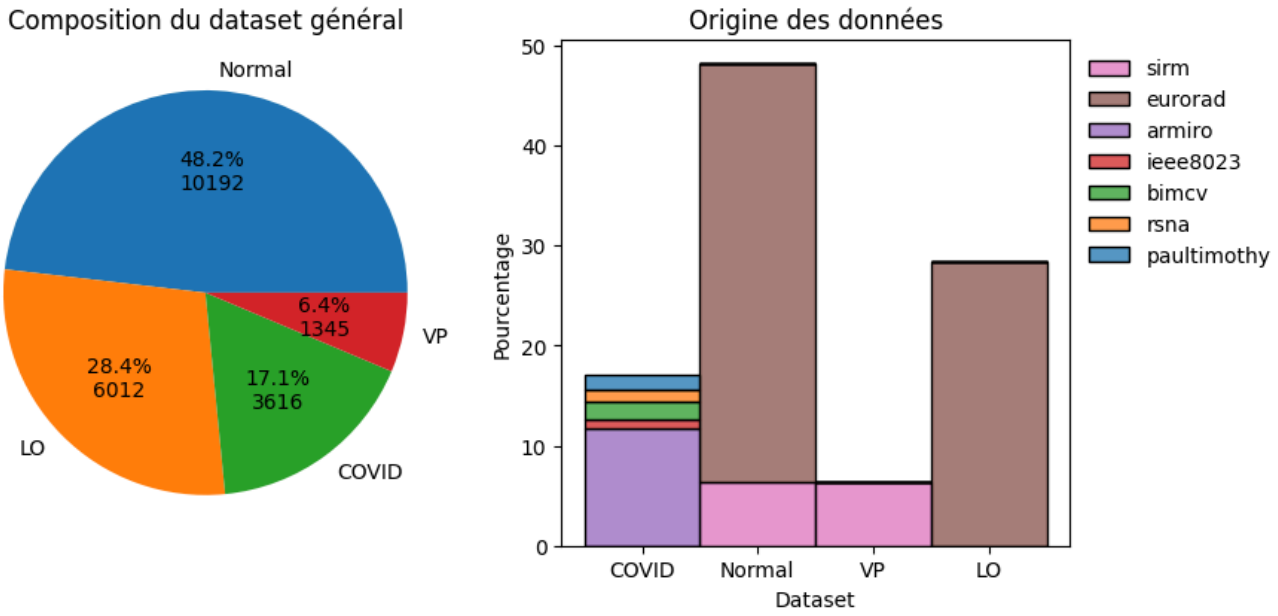


Fig. 3: Représentation graphique de la composition en pourcentage et en nombre d'images du dataset (panneau de gauche). Représentation en barres empilées et en pourcentage de l'origine de chaque jeu de données (panneau de droite, la légende représente le nom des sources des images).

Compte-tenu du nombre de sources de données, il est envisageable que les instruments utilisés pour la prise de radiographies soient paramétrés différemment. De plus, on a observé que les radiographies présentes dans les données ont été effectuées sur des personnes d'âge et de corpulence différentes. Nous avons donc entreprit les analyses des images issues des différentes sources.

Dans un premier temps, nous nous sommes penchés sur la taille relative des poumons dans les images fournies, pouvant être le reflet d'un zoom différent. Pour cela nous avons tout d'abord compté le nombre de pixels blancs présents dans les masques associés aux images, nous avons ensuite rapporté ce nombre au nombre total de pixels présents dans l'image pour obtenir la part de l'image prise par les poumons pour finalement la comparer entre les pathologies et entre les sources des données.

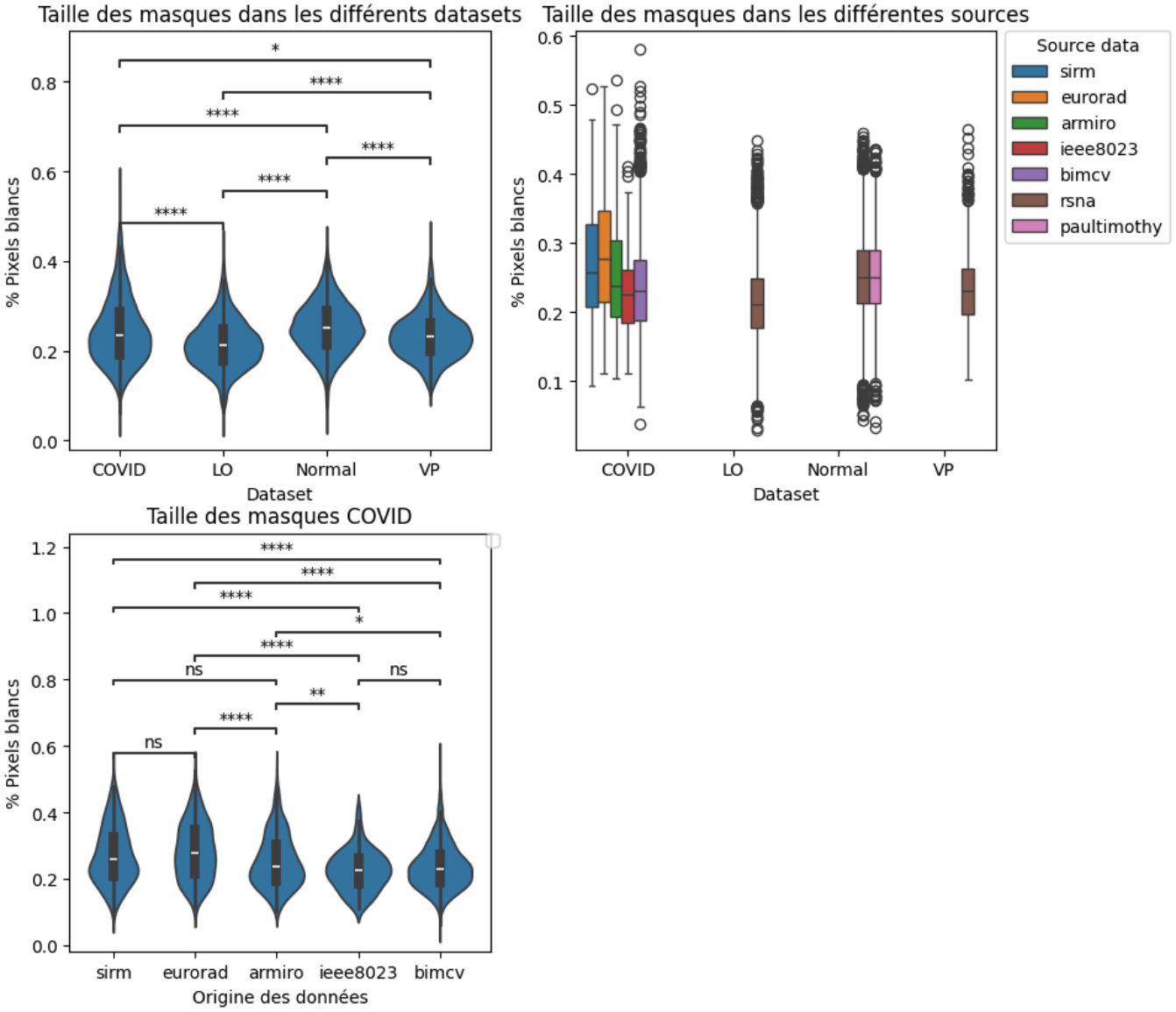


*Fig. 4: Analyse de la proportion de pixels blancs dans les masques. Tests statistiques: Mann-Whitney; p< 0,05 *, p< 0,01 **, p< 0,001 ***, p< 0,0001 ****.*

On a pu observer à l'issu de ces analyses de la taille relative des masques des différences statistiques claires (p-valeur < 0,0001 pour certaines comparaisons) entre les pathologies (Fig.4 panneau en haut à gauche) ce qui est consistent avec le test ANOVA (Tableau en annexe) qui montre un effet du facteur 'Dataset' sur la variance des échantillons et qui est probablement révélateur des caractéristiques inhérentes aux pathologies présentes dans nos jeux de données.

Nous avons ensuite séparé les données en fonction de leur origine et avons pu constaté a-priori que les différences les plus marquées entre les tailles relatives de masques se trouvaient dans le jeu de données COVID (Fig.4 haut droite). De plus, un test ANOVA multifactoriel nous a permis d'identifier un effet significatif du facteur 'origine' sur la variance des échantillons (Tableau en annexe). Finalement, nous constatons aussi que le plus grand nombre de valeurs outliers se trouvent dans ce jeu de données (Tableau en annexe).

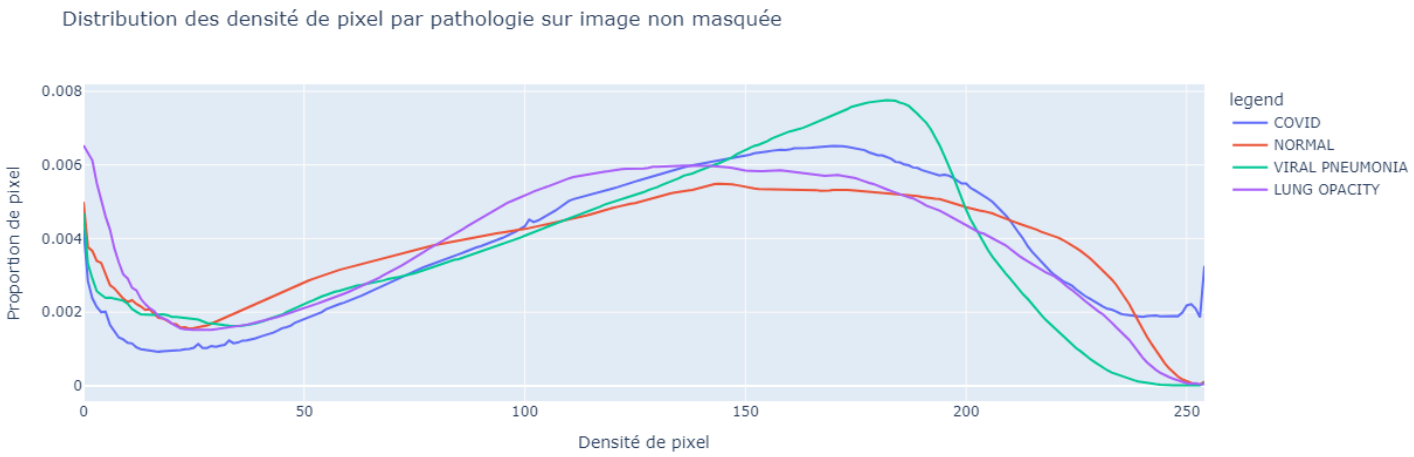
La comparaison de la taille relative des masques dans le jeu de données COVID nous a permis d'identifier des différences significatives (jusqu'à $p<0,0001$ pour certaines comparaisons) entre la taille relative des masques des différentes sources (Fig. 4 panneau du bas). Cette observation peut être le reflet d'un paramétrage différent des appareils ayant servi à la collecte de données ou d'un plus grand nombre d'erreurs dans la production des masques comme l'indique le plus grand nombre de valeurs outliers.

2.6 - Représentation graphique du dataset

Nous allons maintenant détailler les informations contenues dans les données à notre disposition afin d'approfondir notre étude et tenter de mettre en évidence des informations pertinentes pour la suite de notre projet.

Distribution des densités de pixel par pathologies

1- Sur images non masquées



Le graphique ci-dessus représente la distribution des densités de pixels par pathologie, sur les images entières, non retouchées.

La valeur 0 (pixel noir) a été exclue car beaucoup plus présente que les autres densitées, elle prend le dessus sur la distribution des autres densités de pixels. Une représentation des proportions de pixels noirs sera présentée par la suite.

Sur le graphique ci-dessus, on s'aperçoit que les distributions sont plutôt homogènes entre les catégories, ce qui montre que globalement, les radiographies sont assez similaires en termes de luminosité.

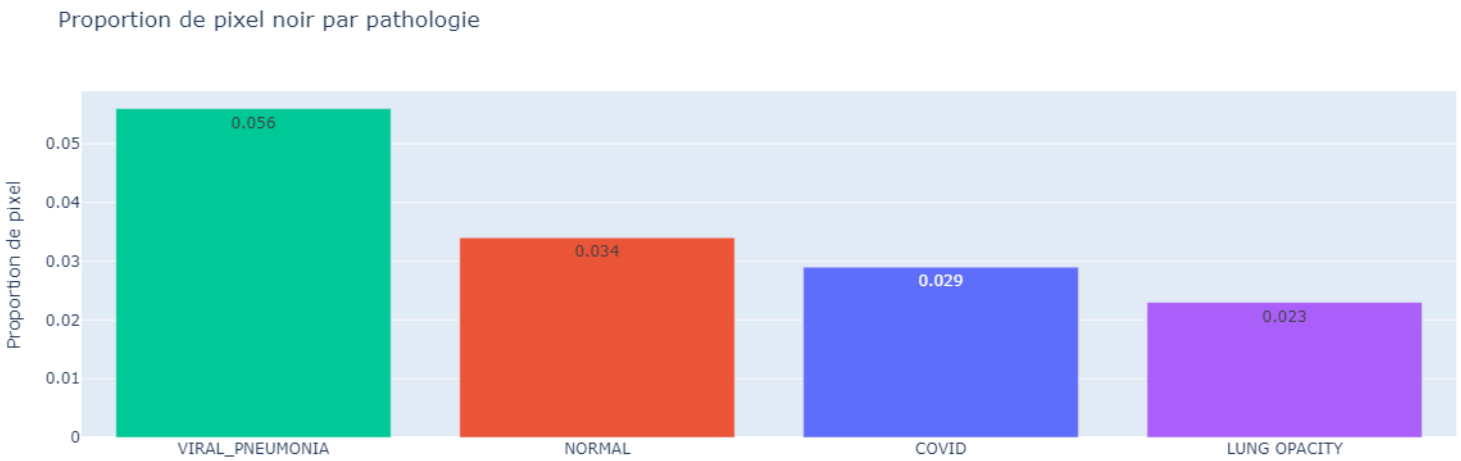
Ce qui peut laisser penser que:

- Soit elles ont été acquises sur des machines avec un paramétrage similaire.
- Soit elles ont été normalisées par l'équipe qui a rendu disponibles les images.

On peut noter un léger pic autour de 180 pour les pneumonie virales par rapport aux autres, ce qui peut éventuellement s'expliquer par le faible nombre d'image disponibles pour cette pathologie par rapport aux autres (6.4% du dataset), qui peut induire une variabilité dans la distribution.

En conclusion de ce graphique, on peut avancer qu'il n'y a pas de biais majeur dans les niveaux de luminosité dans le jeu de données.

Concentrons-nous à présent sur les proportions de pixel noir par pathologie qui ont été exclu de l'étude précédente:



Ici, on peut voir que suivant la catégorie, 2.3% à 5,6% des pixels sont noirs.

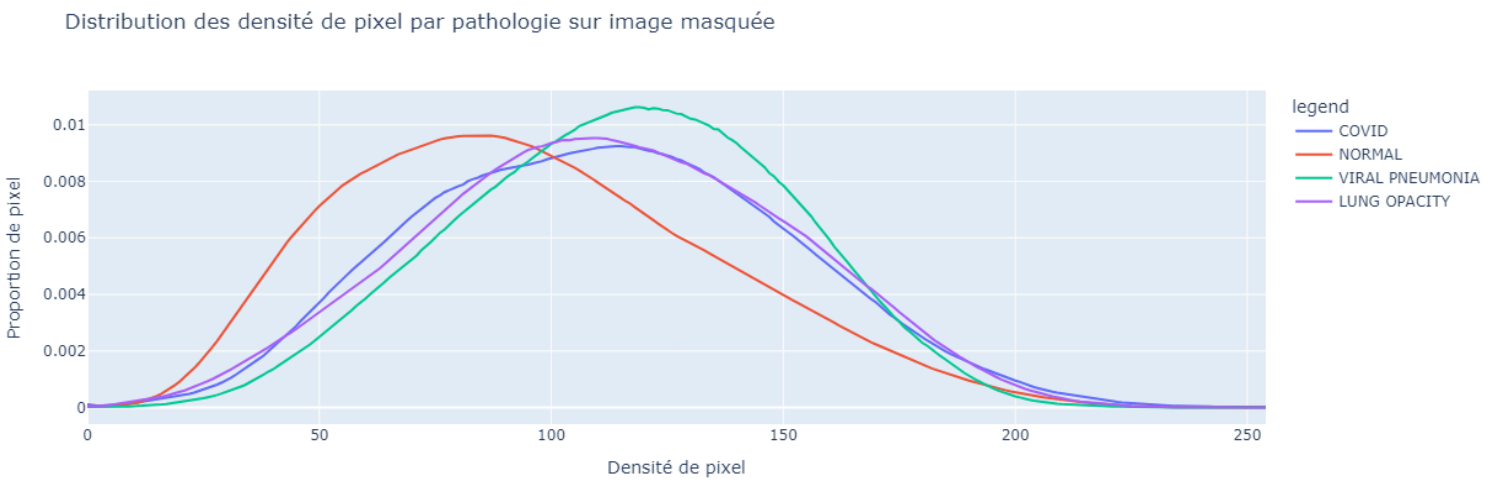
Le nombre de pixel noirs a un lien direct avec la le nombre de pixel qui ne concerne pas des tissus humains. Moins il y a de pixels noir, plus il y a de tissus sur la radiographie, ce qui peut s'expliquer de deux façons:

- Soit la personne est plus corpulente ou plus grande
- Soit le zoom de la radiographie est différente

En effet, ça n'a à priori pas de lien avec la pathologie puisque la zone des poumons, bien que très sombre ne contient pas de pixel noir.

Ces différences de proportions de pixel noir pourraient engendrer un biais dans l'entraînement du modèle. C'est pourquoi il semble primordial de masquer les images sur la zone des poumons, pour s'affranchir de ce biais qui n'est à priori pas corrélé à la pathologie.

2- Sur images masquées



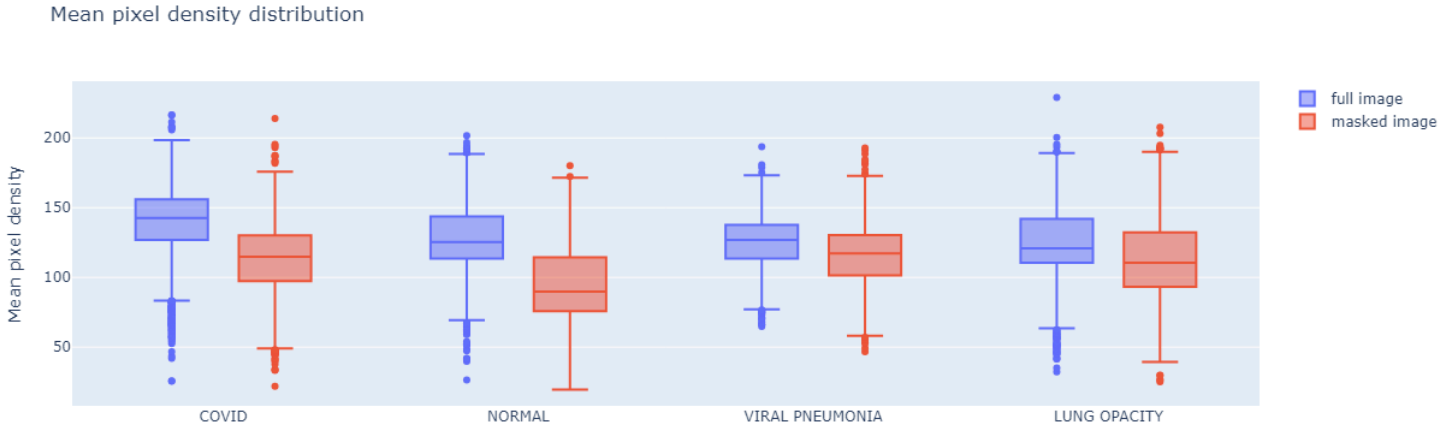
Le graphique ci-dessus représente la distribution des densités de pixels sur les images masquées, c'est-à-dire concentrées uniquement sur les poumons.
Celui-ci est très intéressant puisqu'il met en évidence les différences entre les pathologies concernant les densités de pixel.

On peut constater que les radiographies de patients qui n'ont aucune pathologie ("NORMAL") présentent une densité moyenne plus faible que les autres. C'est une observation très cohérente avec la réalité, puisque la présence de lésions dans le poumons se traduit par des "tâches" plus ou moins opaques (donc une densité plus élevée) dans celui-ci. Il est donc rassurant de trouver une densité moyenne plus élevée pour les trois pathologies.

Un autre point intéressant est la densité sur les radiographies de pneumonie virale. En effet, il semble qu'il y ait une concentration de pixels très légèrement plus opaque que sur les deux autres pathologies. Nous voyons à cette observation deux causes possibles:

- Soit il y a en effet une opacité légèrement plus forte sur la pathologie de pneumonie
- Soit c'est encore une fois le manque d'image qui fausse légèrement la distribution

C'est ce questionnement qu'il faudra peut-être garder à l'esprit lorsque nous passerons à l'entraînement du modèle: il faudra peut-être distinguer deux entraînements : un avec les images brutes et un avec des images normalisées par rapport à ce point.



On peut retrouver sur le graphique ci-dessus la distribution des densités moyennes des images. On visualise encore une fois une moyenne de densité plus faible pour les patients sans pathologie.

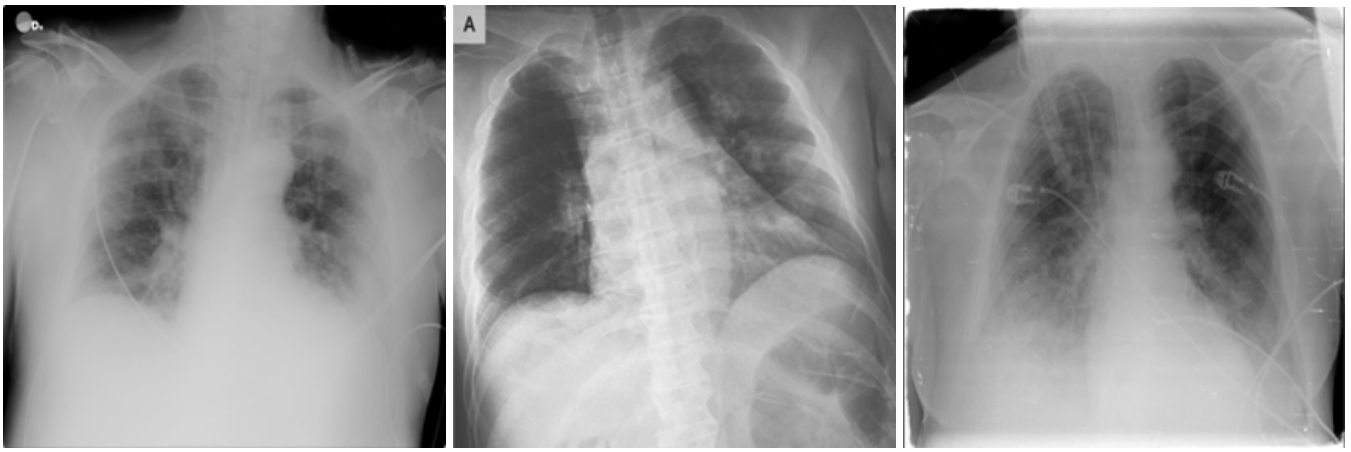
Une information supplémentaire est la présence de 68 outliers. Il faudra voir par la suite si on décide de supprimer, modifier ou conserver ces éléments perturbateurs pour notre modèle de Machine Learning.

2.7 - Modifications et exploitations des images

Dans le but de comprendre comment nous pouvions détecter et mettre en évidence les radiographies concernant les patients atteint du COVID-19, nous avons réalisé des modifications numériques sur les images afin de faire des observations visuelles.

Nous avons alors testé des filtres de traitement présents dans la librairie Python OpenCV (appelé cv2). Nous avons réalisé différents clichés avec les algorithmes suivants : Threshold, AdaptiveThreshold, Canny, GaussianBlur,.

Radiographies originales :



Radiographies modifiées avec des modifications de valeurs de seuils (Threshold) :

Radiographies modifiées avec des modifications de valeurs de seuils (AdaptativeThreshold) :

Radiographies modifiées avec un ajout de flou à l'image (GaussianBlur) :

Radiographies modifiées avec une détection des contours (Canny) :

Radiographies modifiées avec l’utilisation des masques par soustraction (addWeighted) :

3 - Remerciements

Suite à nos investigations, nous avons eu l'occasion de rencontrer de nombreuses personnes qui ont soutenu notre projet. Que ce soit pour une aide technique, théorique ou pratique, nous avons eu un soutien remarquable nous permettant d'avoir un avis plus pointu sur le sujet et sa mise en place.

Ainsi nous souhaitons remercier chaleureusement :

- Romain LESIEUR, notre Mentor qui a su nous diriger et nous indiquer les démarches à suivre lorsque nous n'étions pas certains de nos choix
- Charlotte, une étudiante en médecine qui a répondu à nos interrogations, nous a permis de comprendre les enjeux et le contexte métier de la détection des différentes pathologies par radiographie.

Nous voulons également remercier Aïda notre Program Manager, Maria notre Responsable de Cohort

Et sans oublier un grand MERCI à l'ensemble de l'équipe de notre établissement de rattachement DataScientest

4 - Annexes

Fichier README.md.txt

*****COVID-19 CHEST X-RAY DATABASE
A team of researchers from Qatar University, Doha, Qatar, and the University of Dhaka, Bangladesh along with their collaborators from Pakistan and Malaysia in collaboration with medical doctors have created a database of chest X-ray images for COVID-19 positive cases along with Normal and Viral Pneumonia images. This COVID-19, normal and other lung infection dataset is released in stages. In the first release we have released 219 COVID-19, 1341 normal and 1345 viral pneumonia chest X-ray (CXR) images. In the first update, we have increased the COVID-19 class to 1200 CXR images. In the 2nd update, we have increased the database to 3616 COVID-19 positive cases along with 10,192 Normal, 6012 Lung Opacity (Non-COVID lung infection) and 1345 Viral Pneumonia images and corresponding lung masks. We will continue to update this database as soon as we have new x-ray images for COVID-19 pneumonia patients.

**COVID-19 data:

COVID data are collected from different publicly accessible dataset, online sources and published papers.
-2473 CXR images are collected from padchest dataset[1].
-183 CXR images from a Germany medical school[2].
-559 CXR image from SIRM, Github, Kaggle & Tweeter[3,4,5,6]
-400 CXR images from another Github source[7].

***Normal images:

10192 Normal data are collected from from three different dataset.
-8851 RSNA [8]
-1341 Kaggle [9]

***Lung opacity images:

6012 Lung opacity CXR images are collected from Radiological Society of North America (RSNA) CXR dataset [8]

***Viral Pneumonia images:

1345 Viral Pneumonia data are collected from the Chest X-Ray Images (pneumonia) database [9]

Please cite the follwoing two articles if you are using this dataset:
-M.E.H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M.A. Kadir, Z.B. Mahbub, K.R. Islam, M.S. Khan, A. Iqbal, N. Al-Emadi, M.B.I. Reaz, M. T. Islam, “Can AI help in screening Viral and COVID-19 pneumonia?” IEEE Access, Vol. 8, 2020, pp. 132665 - 132676.
-Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Kiranyaz, S., Kashem, S.B.A., Islam, M.T., Maadeed, S.A., Zughaier, S.M., Khan, M.S. and Chowdhury, M.E., 2020. Exploring the Effect of Image Enhancement Techniques on COVID-19 Detection using Chest X-ray Images. arXiv preprint arXiv:2012.02238.

**Reference:
[1]https://bimcv.cipf.es/bimcv-projects/bimcv-covid19/#1590858128006-9e640421-6711
[2]https://github.com/ml-workgroup/covid-19-image-repository/tree/master/png
[3]https://sirm.org/category/senza-categoria/covid-19/
[4]https://eurorad.org
[5]https://github.com/ieee8023/covid-chestxray-dataset
[6]https://figshare.com/articles/COVID-19_Chest_X-Ray_Image_Repository/12580328
[7]https://github.com/armiro/COVID-CXNet
[8]https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data
[9] https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia

***Formats
- All the images are in Portable Network Graphics (PNG) file format and resolution are 299*299 pixels.

****Objective
- Researchers can use this database to produce useful and impactful scholarly work on COVID-19, which can help in tackling this pandemic.

	Unnamed: 0	sum_sq	df	F	PR(>F)
0	C(label)	5.226696	3.0	484.711387	1.214369e-304
1	Residual	76.060463	21161.0	NaN	NaN

Annexe 1: Tableau récapitulatif de l'ANOVA sur le facteur 'label' correspondant au dataset

	Unnamed: 0	sum_sq	df	F	PR(>F)
0	C(origin_data)	1.578242	6.0	69.821639	1.769540e-86
1	Residual	79.708917	21158.0	NaN	NaN

Annexe 2: Tableau récapitulatif de l'ANOVA réalisée sur le facteur 'origin_data' dans les données COVID.

4.1 - Glossaire

COVID-19 : Acronyme de COrona VIrus Disease 2019, il s'agit d'une maladie respiratoire très contagieuse qui engendre des millions de victimes à travers le monde. Cette infection est apparue en Chine en Décembre 2019. Elle se manifeste principalement par les symptômes suivants : fièvre 87,9 %, toux sèche : 67,7 %, fatigue : 38,1 %, diminution de sens du goût : 24 % et perte d'odorat : 20 %.

Comorbidité : C'est l'association d'un ou plusieurs troubles avec une pathologie primaire qui peut être un facteur aggravant des effets d'une de ces pathologies.

Dataset : Ensemble de données utilisé pour l'analyse, l'entraînement de modèles de machine learning notamment.

Deep Learning : autrement appelé apprentissage profond ou apprentissage en profondeur est un sous-domaine de l'intelligence artificielle qui utilise des réseaux neuronaux ayant de nombreuses couches pour résoudre des tâches complexes.

Métadonnées : Informations descriptives des données à disposition, les métadonnées fournissent par exemple le contexte, l'origine, la date ou encore le format de celles-ci.

OMS : Organisation Mondiale de la Santé : chargée de diriger l'action sanitaire mondiale, définir les programmes de recherche en santé, fixer des normes, etc...

PCR : "Polymerase Chain reaction" -> technique de biologie moléculaire pour le dépistage d'agents infectieux.

Test Antigénique : test rapide adapté aux tests sur le lieux de soins qui détecte directement la présence ou l'absence d'un antigène.

TDM : abréviation de tomodensitométrie, plus couramment appelé scanner. Il s'agit d'un examen d'imagerie médicale qui consiste à prendre des clichés du patient à l'aide de Rayon X. Le résultat est une image 2D ou une visualisation 3D permettant d'observer les tissus, vaisseaux et organes.