

PROJET : Analyse de Radiographies Pulmonaires COVID 19

Auteurs : Baptiste MOREAU, Chadi MASRI, Karima BENNIA
Promotion Janvier 2021

Abstract : Ces dernières années, de nombreuses expérimentations ont démontré l'intérêt de l'intelligence artificielle (IA) au service de l'imagerie médicale et notamment la radiologie. Cela a été rendu possible par la disponibilité de grands ensembles de données («big data»), des progrès substantiels en matière de puissance de calcul et l'avènement de nouveaux algorithmes de deep learning. Cependant, ces technologies sont aujourd'hui peu démocratisées en raison d'une dégradation des performances de ces algorithmes d'IA entre les phases expérimentales et la mise en œuvre en conditions réelles. Une des explications à ce phénomène est l'insuffisante caractérisation des biais des datasets utilisés pour développer les modèles de classification. A travers notre projet de détection de COVID à partir de radios de poumons, nous avons tenté d'illustrer comment les biais pouvaient impacter d'une façon significative les performances d'un modèle de classification. Au-delà d'une simple analyse exploratoire du dataset, nous avons tenté de démontrer l'existence de ces biais et de les caractériser par l'utilisation d'approches distinctes (t-SNE, PCA, Grad-CAM) mais complémentaires. Ce travail nous a permis de comprendre de quelle façon les biais pouvaient influencer les algorithmes de classification et rendre non-valides certains indicateurs de performance. Dans le cadre de notre projet, cette compréhension et caractérisation des biais de notre dataset nous a conduit à une meilleure orientation des choix de technologies de pre-processing de nos données. Un modèle de classification basée sur notre dataset ainsi traité a été développé avec une performance de 85%. La validité de nos résultats a été en partie confortée par l'utilisation des technologies de Grad-CAM.

1. Introduction et Contexte

Le projet COVID avait pour première intention de développer un algorithme dont la vocation était de prédire à partir d'un cliché de radiographie X si un patient était atteint de pneumonie virale, de COVID 19 ou présentait un état normal. A cet effet, nous avions à notre disposition 3 datasets de clichés de poumons. Une première analyse exploratoire (cf Annexe 1 : Rapport d'Analyse Exploratoire) a mis en évidence une hétérogénéité majeure de la qualité des images entre, d'un côté les clichés des datasets 'NORMAL' et 'VIRAL PNEUMONIA' et de l'autre les clichés 'COVID'. Ainsi, les clichés COVID se distinguaient notamment par la présence d'éléments exogènes, un défaut de résolution majeure et une qualité d'image particulièrement préjudiciable.

Dans la droite ligne des premières ambitions du projet, un premier prototype de modèle de classification a été développé afin de disposer d'une base étalon permettant d'objectiver les performances des modèles plus développés ultérieurement. Ce premier prototype de classification sur les images brutes est à notre disposition, sans pre-processing et s'est basé sur un réseau neuronal classique (Conv2D). Le modèle développé (Model 1) a présenté des performances insolemment élevées, avec un taux de précision frôlant les 95%. Loin de se satisfaire d'un tel succès, il nous est apparu important et opportun d'approfondir la compréhension de l'obtention d'un tel résultat.

Nous avons ainsi tenté de comprendre quelles étaient les caractéristiques de l'image prises en compte par le modèle pour le guider dans la classification. En d'autres termes dans le cas présent, nous souhaitons répondre à la question suivante : le modèle était réellement capable de distinguer les particularités des poumons pouvant être liées au COVID ou la classification ne s'opérait-t-elle que sur des particularités autres de l'image tels que son niveau de résolution, son niveau de luminosité ou la présence de bruits/artefacts sur l'image ? Ces interrogations posent la question de l'identification de biais dans les images et de comment ces derniers peuvent influencer le modèle et la lecture de ses performances.

La question des biais rencontre un écho majeur dans le débat de l'apport des technologies d'intelligence artificielle dans l'imagerie médicale . Alors que les algorithmes d'apprentissage automatique basés sur l'imagerie continuent d'être développés et promus, les radiologues doivent être en mesure de reconnaître et de traiter les biais de sélection potentiels dans le développement d'algorithmes avant que ces algorithmes puissent être appliqués de manière optimale à la pratique quotidienne.

Alors que le Saint Graal serait le développement d'algorithmes de deep learning pouvant, à partir de n'importe quelle donnée, afficher un haut degré de performance de classification, la réalité est moins reluisante. De nombreux algorithmes de classification ont été expérimentés et sont toujours en cours de développement avec des performances très élevées sur les datasets de tests. Ces algorithmes semblent avoir démontré une précision de diagnostic élevée dans des phases très expérimentales et surpassent les taux de prédiction des radiologues. Cependant, les résultats sont moins impressionnantes lorsque ces mêmes algorithmes sont utilisés dans un contexte réel au sein des hôpitaux avec de nouveaux clichés et les modèles souffrent de leur caractère non généralisable. Un modèle d'IA est qualifié de généralisable, lorsque, entraîné et testé sur un dataset donné , il peut être utilisé sur de nouvelles données tout en présentant une précision équivalente pour l'évaluation d'examen provenant d'autres sources de données. Le principal indicateur de performance utilisé est la courbe ROC. En d'autres termes, une IA sera généralisable si la courbe ROC reste inchangée entre la phase expérimentale et la phase de mise en production.

A titre d'exemple, Wang et al [1] a publié en 2020 des données sur les performances de 6 modèles de Deep Learning entraîné sur les données issues du DDSM (Digital Database for Screening Mamography et d'INBreast. Ces modèles ont présenté des ROC s'échelonnant entre 0,71 et 0,95. Cependant, aucun de ces modèles n'a fonctionné aussi bien lorsqu'ils ont été testé sur des ensembles des données réelles, avec des ROC avoisinant les 60 %. Ce phénomène est potentiellement le résultat d'un biais de sélection dans les données qui ont été utilisées pour développer l'algorithme de deep learning.

Cet exemple démontre qu'il s'avère nécessaire de toujours évaluer de manière critique les expérimentations de développement d'algorithmes, avec une recherche approfondie d'éventuels biais de sélection et comprendre comment ces biais peuvent affecter l'applicabilité des algorithmes de deep learning dans des environnements cliniques réels. Les biais peuvent être multiples et peuvent être aussi bien liés aux caractéristiques des populations dont sont issus les datasets (âge, sexe, origine ethnique, pathologie etc) que sur les conditions dans lesquels les clichés ont été pris. Dans le premier cas, les biais (population) peuvent être facilement objectivés. Moins dans le second cas, où l'analyse peut laisser une certaine place à la subjectivité.

Notre projet a pour objectif de proposer une méthodologie d'analyse critique d'un modèle de deep learning appliquée à l'imagerie médicale, en tentant d'objectiver les biais non évidents d'un dataset pouvant impacter les performances d'un modèle. A cette fin, nous explorerons les techniques réduction de dimensionnalité (t-SNE et PCA) et expérimenterons le Grad-CAM, un algorithme développé en 2019 par Selvaraju et al [2] qui nous renseignera sur les patterns des images utilisés par les modèles pour réaliser la classification souhaitée. Nous utiliserons des ensuites des méthodologies de pre-processing des images pour limiter l'impact des biais identifiés et développerons de nouveaux modèles de classification à partir des datasets pré-traités.

2. Travaux réalisés

Nous avions identifié lors de l'analyse exploratoire une forte hétérogénéité entre les clichés du dataset COVID et les datasets 'VIRAL PNEUMONIA' et 'NORMAL' de l'autre côté. Nous émettons alors l'hypothèse que ces disparités dans la qualité des clichés peuvent constituer des biais et ainsi être préjudiciables à la pertinence de notre premier prototype. Nous avons dans un premier temps cherché à objectiver et caractériser les biais susceptibles de fausser notre intuition et dans un second temps, tenté d'apporter une réponse permettant de contourner les écueils rencontrés.

2.1. Objectivation des biais

La première étape de notre travail a été de tenter d'objectiver les disparités et l'hétérogénéité des images en notre disposition, caractéristiques que nous avons pu observer à l'œil nu. Pour cela, nous avons comparé la distribution de luminosité entre, notamment les images NORMAL(en orange) et les images COVID (en bleu).

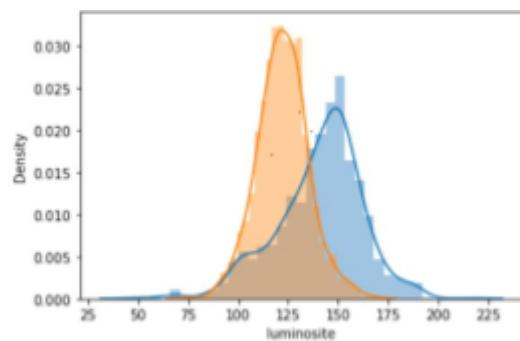


Fig 1 : Comparaison de la distribution des valeurs de luminosité entre les datasets NORMAL et COVID

On observe une différence significative dans la distribution de la luminosité. Ce facteur seul peut suffire à interpréter la classification et la performance élevé du modèle natif développé. (Modèle 1 en Conv2D).

2.1.1. Utilisation du Grad-CAM pour identifier les zones d'intérêt de l'image utilisée pour la classification

Au-delà de ce premier constat, nous avons également souhaité conforter notre intuition que les performances du Model 1 n'étaient pas liées à l'identification de patterns significatifs sur les poumons COVID. A cette fin, nous avons utilisé la méthode Grad-CAM. Cette méthode développée par Selvaraju (2017) [2], permet de générer des cartes des heatmap à des endroits de l'image aidant à la détection d'une de patterns particuliers. C'est une technique populaire pour visualiser notamment où un modèle de réseau convolutif identifie le pattern qui permettra de réaliser une classification.

Nous avons entraîné un réseau avec seulement 2 couches de convolution, et soumis quelques images du dataset COVID au Grad-CAM afin de tenter de visualiser sur quelle partie de l'image le réseau s'est concentré pour prendre sa décision. Les zones d'intérêt sont matérialisées en vert sur les images suivantes.

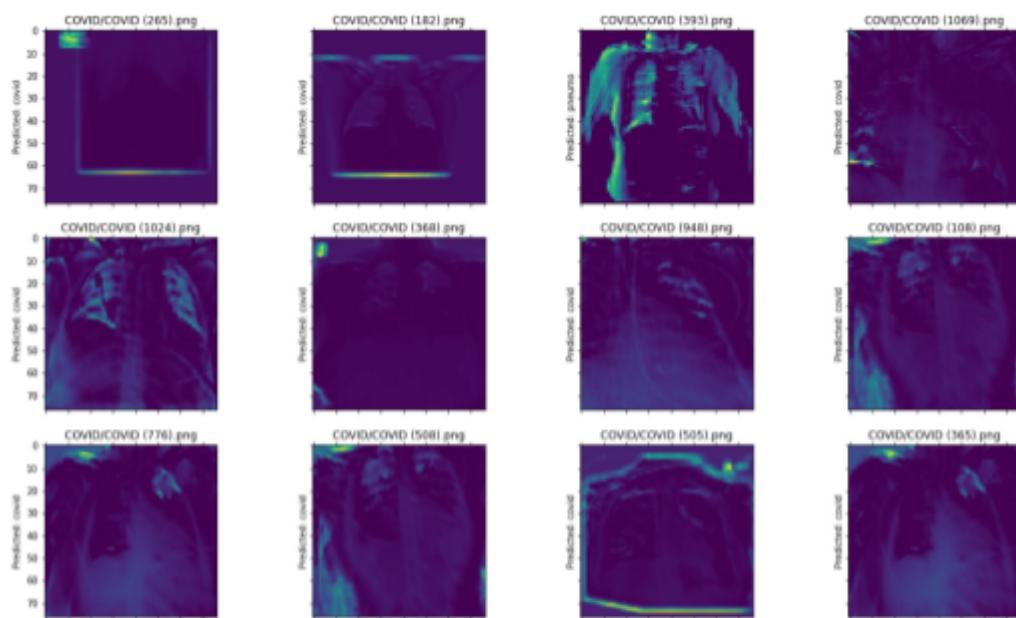


Fig 2 : Illustration de l'application de Grad-CAM sur les images du dataset 'COVID'

Alors que les patterns pertinents devraient être localisés au niveau des poumons, il s'avère que pour statuer de la classification, le Model 1 s'est focalisé des éléments très divers de l'image, présentant que peu d'intérêt dans le cadre d'une détection de COVID. Cette première analyse conforte donc notre intuition selon laquelle, la classification des images n'a pas été menée sur les patterns pertinents (les poumons dans notre cas d'études). Cette observation associée à l'hétérogénéité des histogrammes invalide la performance de notre Model 1 car celui-ci ne s'est pas basé sur les pattern attendu pour effectuer sa classification.

2.1.2. Visualisation de l'impact des biais du dataset sur les performances du modèle à l'aide de techniques de réduction de dimensionnalité

Pour visualiser les biais du dataset, l'idée est de projeter les images dans un espace de dimension inférieure pour observer des tendances dans les images des 3 classes. Pour cela, on redimensionne les images en 28x28 :

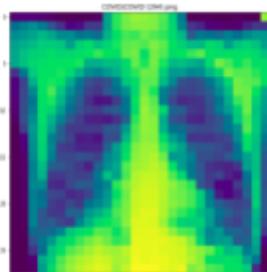


Fig 3 : Illustration de l'impact du redimensionnement (28x28) sur la qualité d'image

En théorie, il s'avère impossible de détecter la présence de COVID ou de pneumonie virale en raison du manque de détail. On applique ensuite un t-SNE avec 2 composantes et on visualise sur ces deux axes les images 28x28 du dataset :

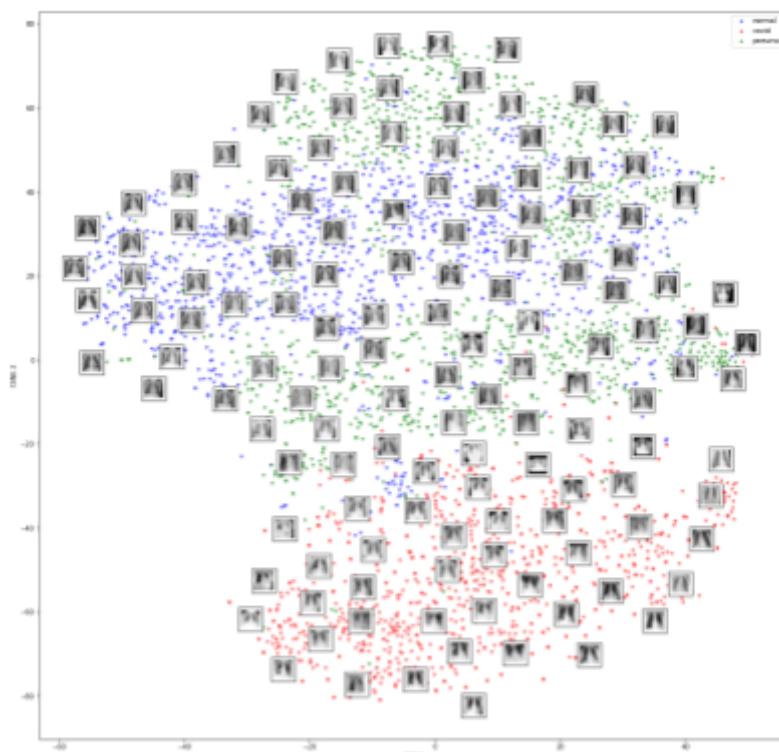


Fig 4 : Visualisation du de la structure des données par l'algorithme t-SNE

On remarque avec étonnement qu'il est possible de distinguer les 3 classes sur une projection 2D, avec une méthode de clustering non-supervisée. En effet, en appliquant un random forest avec uniquement les deux variables correspondant aux deux axes t-SNE, on obtient une val_accuracy de près de 93%.

On peut même observer sur l'image suivante qu'une classification SVM sur ces deux composantes semble bien fonctionner :

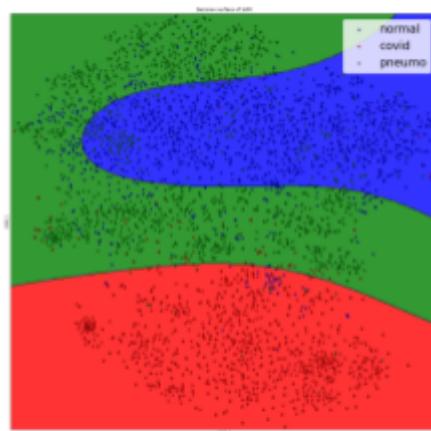


Fig 5 : Classification par l'algorithme SVM sur la base des 2 composantes identifiées par le t-SNE

L'algorithme t-SNE semble extraire sur les images sans détail 28x28 une information importante qui lui permet de faire une bonne classification. Pour faciliter l'interprétation, l'opération a été réitérée en utilisant l'algorithme d'analyse en composantes principales (PCA) :

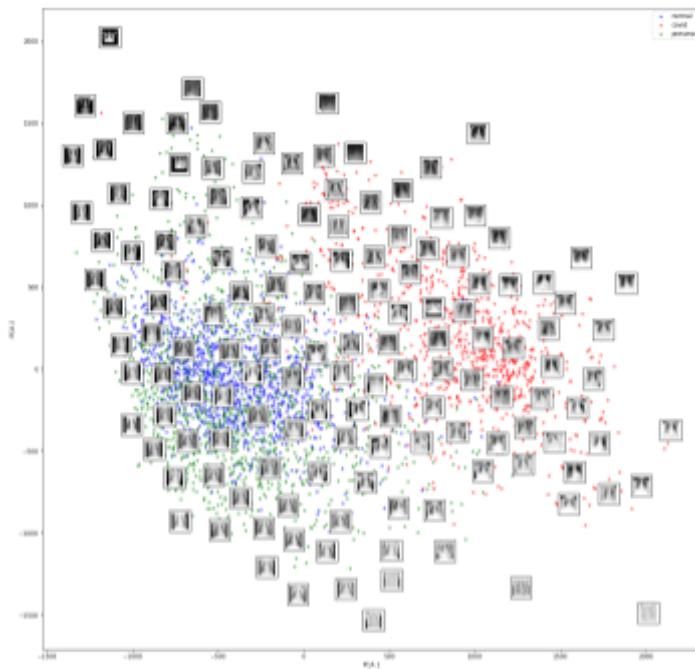
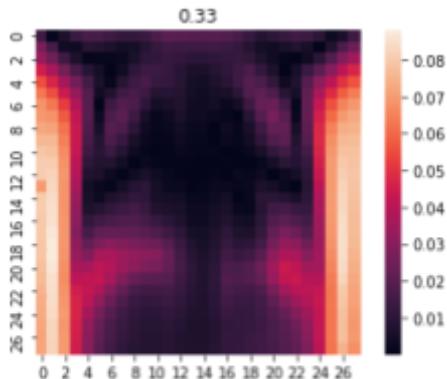


Fig 6 : Visualisation de la structure des données par une analyse PCA

On voit que la projection est moins efficace qu'avec la t-SNE, mais que le premier mode (axe des abscisses) permet toutefois de faire une distinction entre la classe 'COVID' et les classes 'NORMAL' et 'VIRAL PNEUMONIA'.

Suite à cette observation, nous pouvons afficher le vecteur `pca.components_[0]`, en prenant seulement les valeurs absolues, et affichant la heatmap 28x28 :



On voit alors très bien que les pixels qui ont de l'importance dans la projection suivant le premier mode PCA sont les pixels aux bords droit et gauche de l'image. La luminosité des bords permet donc de séparer correctement les images 'COVID' des autres classes de notre dataset, ce qui représente un réel biais dans notre analyse.

Fig 7 : Heatmap de l'analyse par PCA

Les travaux réalisés ont permis de démontrer et caractériser l'existence de biais importants de notre dataset. C'est l'existence de ces biais, et notamment l'hétérogénéité de la distribution de la luminosité qui a permis une classification performante par un modèle simple de classification 2D (Model 1). Ce constat ainsi dressé nous sommes passés à la seconde phase du projet. Ayant identifié les biais probables, il convenait de les corriger avant d'entamer la phase de développement d'un modèle de classification plus robuste.

2.2. Correction des biais identifiés et développement du modèle

Deux biais majeurs concernant les images 'COVID' notamment, avaient été identifiés lors de la phase d'analyse exploratoire, à savoir une qualité et une luminosité de qualité moindre par rapport aux clichés des classes 'NORMAL' et 'VIRAL PNEUMONIA', ainsi que la présence de bruits/artefacts particulièrement nombreux sur les bords des images de la classe 'COVID'. Les méthodologies de correction de biais envisagées ont été les suivantes :

Biais	Méthode de correction de biais expérimentée
Hétérogénéité de la luminosité	Egalisation d'histogramme Adaptative (CLAHE) Egalisation d'histogramme
Bruits et Artefacts	Cropping (recadrage d'image)

Afin d'optimiser les performances de segmentation, nous faisons le choix d'utiliser du transfert learning en faisant appel respectivement aux modèles DenseNet et VGG16, une fois pre-processing effectué. A l'issue du développement des modèles, la technologie Grad-CAM sera une nouvelle fois utilisée afin de s'assurer que la classification a bien été réalisée à partir de patterns localisés au niveau du poumon. Les paragraphes suivants présenteront de façon synthétique les différentes phases de preprocessing utilisées pour le traitement des images avant soumission aux algorithmes de classification.

2.2.1. Preprocessing

- *Homogénéisation de la luminosité par égalisation des histogrammes*

La première phase de pre-processing a consisté à homogénéiser les histogrammes de dispersion de la luminosité. Pour résoudre cette problématique nous avons utilisé des méthodes classiques d'égalisation d'histogramme (CLAHE et normalisation d'histogramme). Ces méthodes permettent d'ajuster l'intensité des pixels pour améliorer le contraste de l'image.

- *Atténuation de l'impact des Bruits et Artefacts*

Nous avons remarqué grâce au premier réseau entraîné avec seulement deux couches de convolution que le Model 1 classifiait les images principalement grâce aux bords de l'image. Il s'avère qu'au-delà de l'hétérogénéité de luminosité, les bords des images du dataset 'COVID' pouvaient comporter des annotations diverses. Notre idée est donc de supprimer ces bords, et de ne garder que les poumons qui concentrent l'information pertinente nécessaire à la classification. Pour cette phase de de-biaisement, nous avons fait appel à des algorithmes de cropping qui permettent d'associer une étiquette ou une catégorie à chaque pixel d'une image et ainsi de sélectionner précisément des éléments d'une image. L'algorithme que nous avons utilisé est le réseau de neurones UNET [3] pré-entraîné sur des radiographies de cages thoraciques et permettant d'identifier les poumons dans une image.

La séquence ci-dessous illustre la transformation des images à l'issue de l'ensemble des méthodes de pré-processing pré-citées :

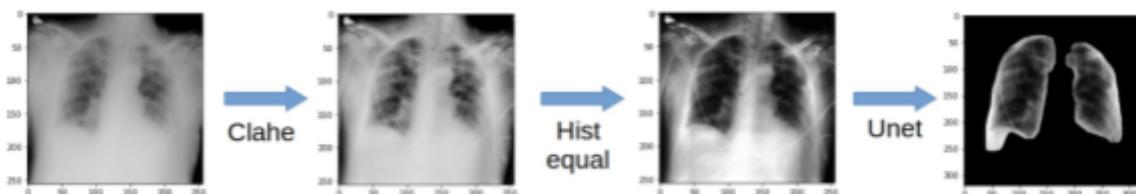


Fig 8 : Illustration de la séquence de pre-processing des images du dataset

Les images homogénéisées et croppées sur les poumons sont ensuite utilisées pour constituer notre nouveau dataset d'entraînement et de test de modèles de deep learning.

2.2.2. Développement du modèle de classification

Pour le développement du modèle, nous avons expérimenté 2 algorithmes de transfer learning à savoir, les architectures DenseNet 121 et VGG16.

La première couche de convolution, ajoutée en amont du modèle VGG16 et nommée "colorize", a pour but de colorier les images en utilisant 3 filtres de convolution à entraîner. En effet, les modèles utilisés sont entraînés sur le dataset imagenet, composé d'images en couleur. L'idée est donc de colorier les images afin d'aider le réseau à classifier, mais aussi pour nous permettre de visualiser comment le réseau peut voir les radios en version couleur.

Les phases d'entraînement et de tests des modèles sur les datasets ont conduit aux performances suivantes en comparaison avec les performances du Model 1 :

Algorithme Pre-Processing	Conv2D	VGG16	DenseNet 121
sans pre-processing	Model 1 : 97,3 %		
avec preprocessing		Model 2 : 85,6 %	Model 3 : 91,4 %

2.2.3. Evaluation de la validité des modèles de classification développés

Pour évaluer la validité de nos modèles, nous nous sommes appuyés sur les mêmes dispositifs utilisés pour la validation du Model 1 à savoir le Grad-CAM et le t-SNE.

Pour rappel, le Grad-CAM permet de nous assurer de la pertinence des patterns utilisés pour la classification en visualisant les zones de l'image présentant le plus d'intérêt pour le modèle.

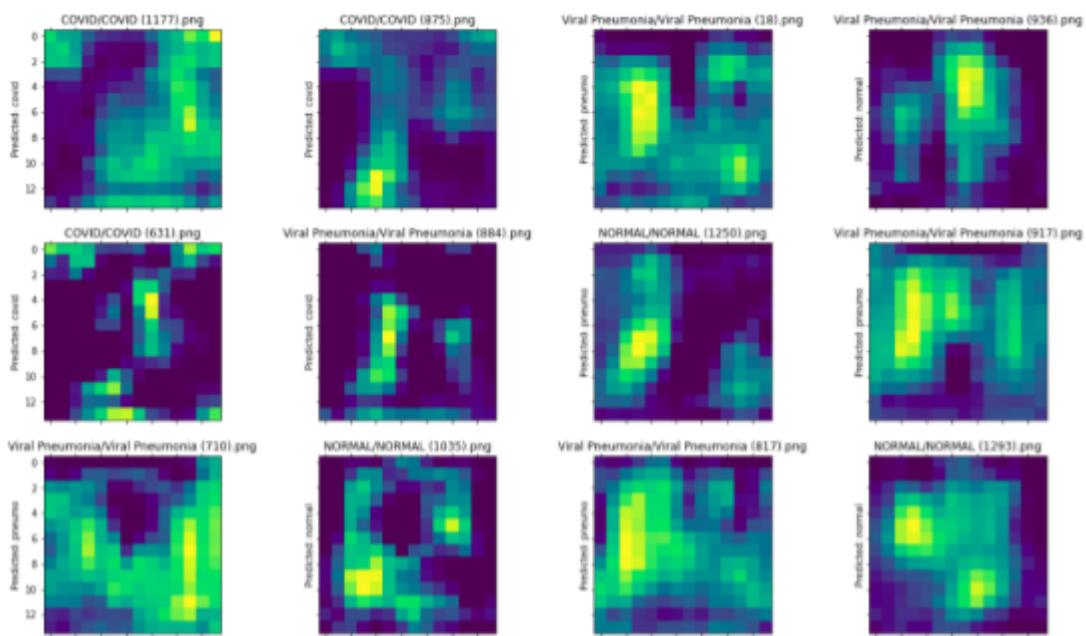


Fig 9 : Application d'un Grad Cam sur le dataset pré-traité

Les résultats de l'application de Grad-CAM avec le transfer learning VGG16 montre que l'information utilisée par l'algorithme pour effectuer sa classification se situe désormais davantage dans une zone centrale de l'image, où sont situés les poumons, ce qui rassure quant à la pertinence des résultats obtenus par ce modèle. Ce modèle est susceptible d'être plus robuste et généralisable que le Model 1 développé. Nous pouvons conclure de ces observations, que le preprocessing aura conduit à une focalisation différenciée entre le poumon et le reste de l'image. Cela constitue une évolution positive dans le cadre d'une potentielle explicabilité du modèle.

L'utilisation du t-SNE sur les images cropées a, pour sa part, conduit aux résultats illustrés dans l'image suivante :

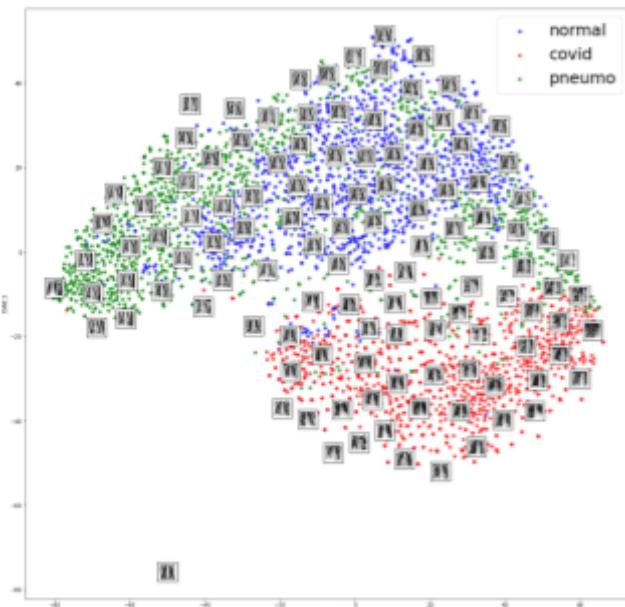


Fig 11 : Application d'un t-SNE sur les données pré-traitées

Nous observons que, malgré le pre-processing, l'algorithme non supervisé parvient encore à séparer les classes par une projection des données en 2 dimensions. Nous pouvons en conclure que des biais persistent dans notre dataset et qu'ils n'ont pas encore été dissipés par les méthodes de preprocessing employées.

2.2.4. Coloration des images par le réseau et hypothèses d'interprétation

Comme expliqué précédemment, nous avons ajouté une couche de convolution en amont du réseau pré-entraîné pour colorier les images afin d'améliorer la classification. Voici un exemple des couleurs choisies par le réseau pour quelques images :

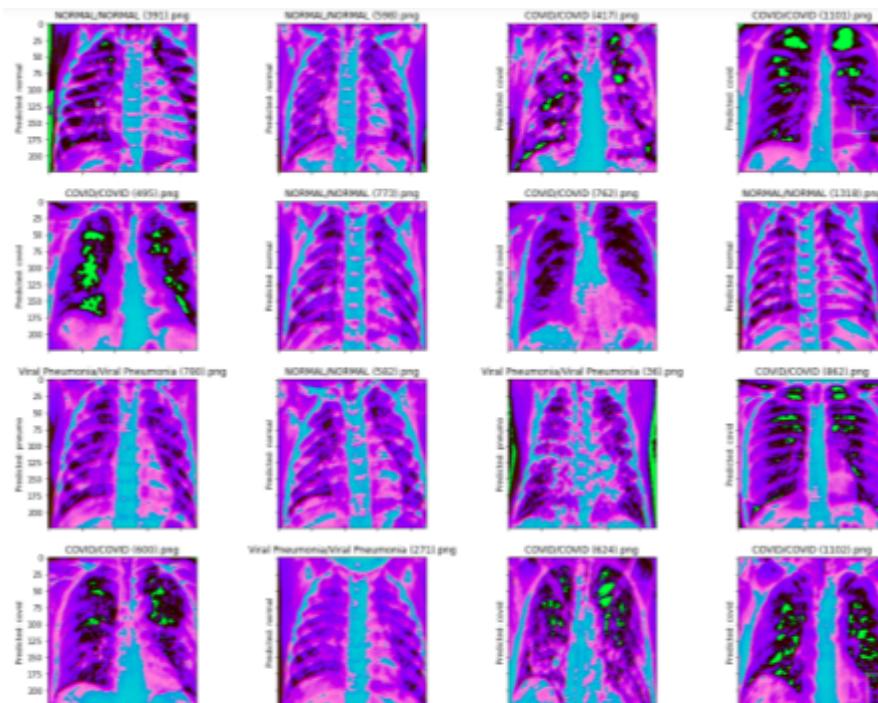


Fig 10 : Analyse des images du dataset suite coloration (colorize)

Les différences de couleurs témoignent d'une différence de textures des éléments présents dans l'image (bleu pour les os, rose/violet pour les tissus et ou organes). On constate pour les images du dataset COVID une troisième couleur est identifiée, à savoir le vert, et que les zones identifiées par cette couleur verte se situent dans les poumons. Sans vouloir nous avancer vers une interprétation hasardeuse, il apparaît judicieux de poursuivre les investigations dans cette voie afin d'améliorer la détection des patterns utilisés par les algorithmes de classification, première étape vers une meilleure explicabilité des modèles.

3. Conclusion

A l'issue de notre projet, nous obtenons un modèle de prédiction de cas COVID à partir de clichés de poumons de près de 85%. Ce score bien qu'honorables, reste en deçà de celui obtenu avec notre premier réseau de neurones, entraîné sur des images brutes non traitées. Mais ce modèle s'est avéré non valide.

En effet, dès le départ, nous avons eu de doutes quant à la performance de ce premier modèle, en émettant l'hypothèse que la segmentation ne s'effectuait pas sur la base d'une analyse approfondie des poumons telle que peut le faire un radiologue, mais sur l'hétérogénéité de la qualité d'images entre les différentes classes. L'utilisation d'algorithmes de réduction de dimensionnalité (t-SNE et PCA) et de Grad-CAM nous ont permis de conforter notre intuition et l'existence de biais et de caractériser les biais existants. Des méthodes de preprocessing adaptées ont été employées pour dissiper les biais sur les datasets existants avant l'utilisation d'algorithmes de transfer learning de classification. Il semblerait toutefois que les technologies de preprocessing employées aient été insuffisantes car les analyses finales des modèles développés tendent à démontrer la persistance de biais dans les datasets. Notons toutefois les progrès réalisés, l'analyse par Grad-CAM ayant mis en évidence une meilleure pertinence des patterns (désormais localisés au niveau des poumons) utilisés par les modèles de classification développés. Il serait intéressant d'investiguer de nouvelles méthodes de preprocessing en collaboration avec des radiologues, pour améliorer la caractérisation des biais du dataset.

Les enseignements de ce projet sont triples :

- L'analyse exploratoire est indispensable notamment lors de développement d'algorithmes de deep learning pour l'imagerie médicale. Au delà des simples considérations métriques, une réflexion approfondie sur les spécificités des datasets de chaque classe doit être menée indépendamment pour identifier et caractériser les éventuels biais avant le développement d'un modèle. L'utilisation d'algorithmes de réduction de dimensionnalité (t-SNE, PCA) peuvent s'avérer utiles dans cette démarche ;
- L'utilisation de GradCam peut constituer une solution intéressante pour l'étude de la validité d'un modèle, en s'intéressant aux patterns des images utilisés lors de la classification. Cependant, cette méthode ne permet pas de rendre un modèle interprétable mais peut donner quelques pistes d'explicabilité. C'est pourquoi elle doit encore être utilisée avec précaution et ne pas être érigée en gage absolu de validité d'un modèle ;

- Enfin, la comparaison de modèles sur la seule base du taux de précision n'est pas applicable dans le cadre d'images médicales. Il faut avoir une lecture suffisamment critique des performances et prioriser les performances du « caractère généralisable » du modèle, critère pouvant être objectivé par la confrontation du modèle à des datasets et images nouveaux et variés.

Références bibliographiques :

- [1] Wang, X., Liang, G., Zhang, Y., Blanton, H., Bessinger, Z., & Jacobs, N. (2020). Inconsistent performance of deep learning models on mammogram classification. *Journal of the American College of Radiology*, 17(6), 796-803.
- [2] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).
- [3] Haghifar, A., Majdabadi, M. M., & Ko, S. (2020). Covid-cxnet: Detecting covid-19 in frontal chest x-ray images using deep learning. *arXiv preprint arXiv:2006.13807*

Annexes :

Annexe 1 : Rapport d'Analyse Exploratoire

ANNEXE 1 :

PROJET COVID

Rapport d'analyse exploratoire des données

Auteur : Chadi MASRI - Baptiste MOREAU - Karima BENNIA

Date : 04/02/2020

Cette analyse exploratoire a pour objectif de décrire l'ensemble des données dont nous disposons et ainsi guider, par la réponse aux questions préliminaires, le processus de modélisation. Nous avons envisagé une première analyse des données brutes puis nous nous sommes orienté vers une analyse numérique de notre dataset afin d'objectiver des caractéristiques difficilement quantifiable à l'œil nu.

Il apparaît crucial de traiter ces données en amont de notre travail de modélisation car, de la qualité du dataset d'entraînement et de tests dépendra la performance du modèle qui sera développé.

Le problème à résoudre est de développer un modèle permettant de classifier des clichés de radio pulmonaires en 3 classes : "Normal", "Covid19" et "Pneumonie Virale" ..

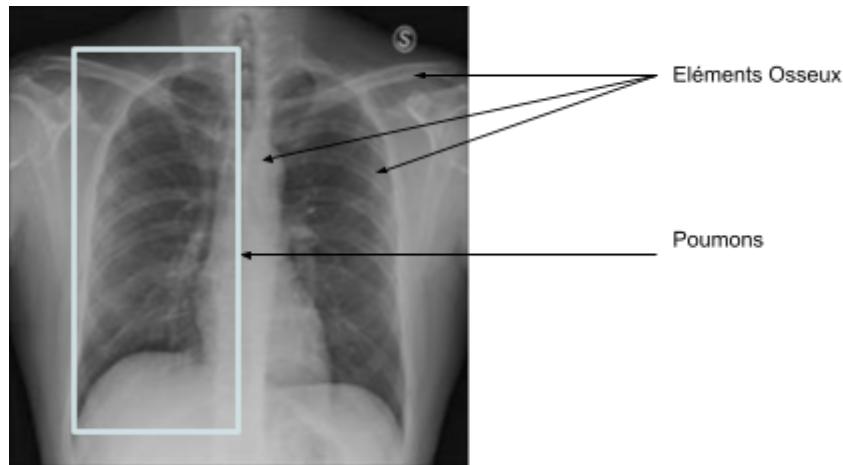
Le dataset sur lequel nous allons travailler est le suivant contient 3 886 clichés de radio pulmonaires et se répartit de la façon suivante : 1200 clichés de la classe "Covid19", 1341 clichés de la classe "Normal" et 1345 clichés de la classe "Viral Pneumonia". Nous pouvons constater que la taille de ce dataset est relativement confidentielle comparée aux tailles standards de dataset utilisés dans le cadre de méthode de Deep Learning. Cela constituera un handicap de taille et pourra influer sur la performance finale du modèle développé.

L'analyse exploratoire se décompose en 3 étapes à savoir l'analyse brute du dataset ; l'analyse métrique du dataset et une première identification des solutions pouvant être envisagées afin de pallier au manquement du dataset.

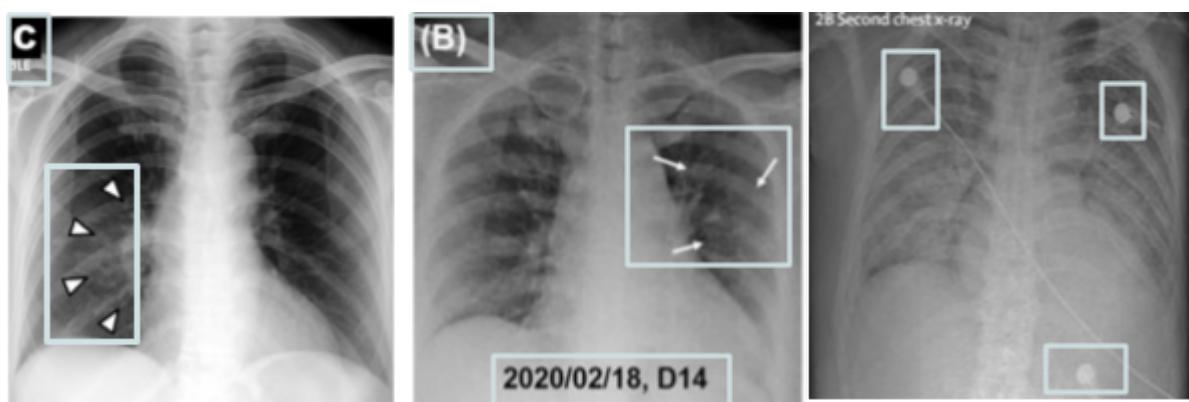
1) Analyse qualitative des données

Nous avons dans un premier temps parcouru à l'œil nu les échantillons d'image à notre disposition. Le premier enseignement que nous avons pu en tirer est la grande hétérogénéité des données à notre disposition. Cette hétérogénéité se décline au niveau du contenu, de la résolution des images et du contraste/luminosité.

Si l'on s'intéresse au contenu, les images sont des radiographies de poumons par rayons X prises de face. On peut distinguer les poumons (masse noires), la cage thoracique et autres éléments osseux du thorax (humérus, omoplate, clavicule...). En fonction du cliché, chacun des éléments étant plus ou moins défini avec précisions.



Au-delà de ces éléments, certains clichés contiennent des informations supplémentaires (annotations) ou témoignent de la présence de dispositifs médicaux lors de la prise des clichés.



Certains clichés de la classe “Covid19” seront difficilement exploitables en raison d'une ‘pollution’ trop importante par des éléments exogènes.

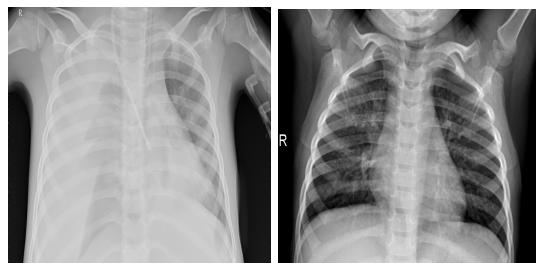


Concernant la qualité et la résolution, nous pouvons observer qu'il existe une disparité flagrante inter-classes et intra-classes notamment concernant la classe "Covid19".

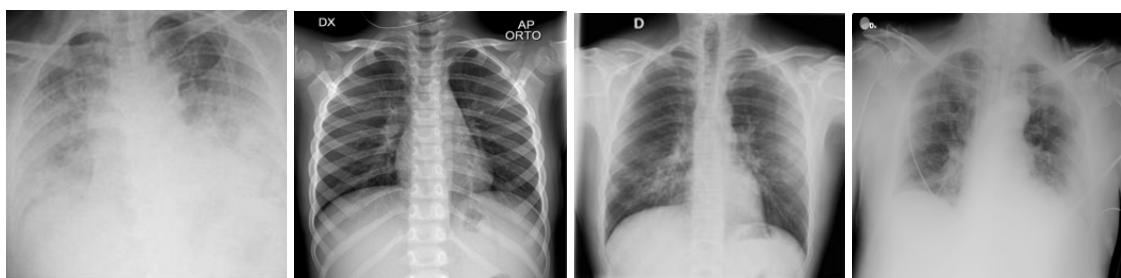
Ainsi, les clichés de la classe 'Normal' se caractérisent par un très grande homogénéité en termes de contrastes et de résolution. Cette dernière est élevée, les éléments distinctifs des clichés (poumons, éléments osseux) sont facilement identifiables.



Concernant les clichés de la classe 'Viral Pneumonia', la résolution s'avère également très correcte. En revanche il y a une disparité au niveau des contrastes et de la luminosité :



Enfin la classe 'COVID 19' se distingue par une forte hétérogénéité intra-classe, tant au niveau des contrastes/luminosité que de la résolution :



En conclusion de cette première analyse, il apparaît qu'au regard de la très grande hétérogénéité des clichés, que nous serons confrontés à un problème de

segmentation d'image difficile. En effet, nous n'avons pas clairement identifié ce qu'il faut détecter pour conclure à l'état du patient (Covid19/Pneumonie Virale/Normal) et la grande disparité de qualité des images est un obstacle majeur qu'il sera nécessaire de contourner.

L'analyse métrique devrait nous aider à confirmer nos premières analyses et apporter des éléments d'appréciation de la qualité du dataset supplémentaires.

2) Analyse quantitative des données

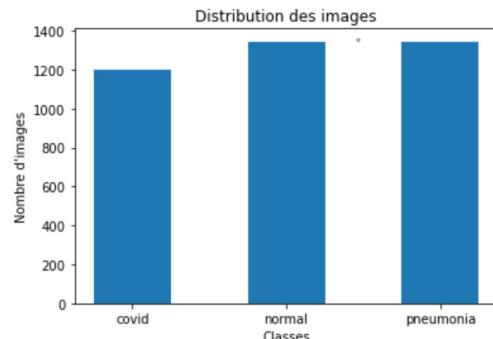
L'analyse métrique s'est focalisée sur la distribution ainsi que sur l'étude de l'homogénéité du spectre de couleur utilisé, la forme des images ainsi que sur le contraste/luminosité.

La distribution met en évidence que nous disposons d'un nombre d'images de la classe "Covid19" moindre (de l'ordre de 10%) par rapport aux autres classes. Il n'y a pas de déséquilibre manifeste des données.

```
Entrée [5]: ┌─ print("Il y a {} images en tout".format(len(img_covid) + len(img_normal) + len(img_pneumo)))

ind = range(3)
plt.bar(ind, [len(img_covid), len(img_normal), len(img_pneumo)], width = 0.5)
plt.xlabel('Classes')
plt.title('Distribution des images')
plt.xticks(ind, ('covid', 'normal', 'pneumonia'))
plt.ylabel("Nombre d'images");
```

Il y a 3886 images en tout



Concernant l'analyse du spectre de couleur , il apparait que toutes les images ont 3 canaux et qu'elles sont toutes en noir et blanc.

```
Entrée [11]: ┌─ # Créons une nouvelle colonne qui vérifie si l'image est en noir et blanc
# ie les 3 channels sont égales

for i, path in df.path.items():
    im = cv2.imread(path)
    if np.array_equal(im[:, :, 0], im[:, :, 1]) and np.array_equal(im[:, :, 1], im[:, :, 2]):
        df.loc[i, 'B&W'] = True
    else :
        df.loc[i, 'B&W'] = False

Entrée [12]: ┌─ df['B&W'].value_counts()

Out[12]: True    3886
Name: B&W, dtype: int64
```

```
Entrée [ ]: ┌─ # Les images sont toutes en noir et blanc
```

L'analyse métrique a permis également de mettre en évidence, que la quasi totalité des images avait une forme carré. Les formats des images du datasets sont les suivants :

Entrée [10]: # Les images ont toutes 3 channels
`print(df[['height', 'width']].value_counts())`

```
height    width
1024.0   1024.0      2686
331.0    331.0       573
256.0    256.0       558
891.0    1084.0      31
952.0    1184.0      9
992.0    1192.0      6
1047.0   1275.0      6
1053.0   1053.0      4
917.0    1116.0      4
852.0    1039.0      1
197.0    253.0       1
723.0    594.0       1
913.0    1102.0      1
875.0    1108.0      1
913.0    1108.0      1
928.0    1130.0      1
947.0    1192.0      1
160.0    187.0       1
dtype: int64
```

Seules 65 images ont une forme rectangulaire, et ce sont toutes appartiennent à la classe "COVID". Nous suggérons de prendre la plus petite résolution et d'homogénéiser toutes les images selon ce paramètre.

Entrée [10]: # Les images n'ont pas toute la même résolution et certaines ne sont pas carrées
`df[(df['height'] != df['width']) & (df['Class'] == 'covid')]`

Out[10]:

	path	Class	height	width	channel
56	COVID/COVID (1193).png	covid	891.0	1084.0	3.0
70	COVID/COVID (1195).png	covid	891.0	1084.0	3.0
85	COVID/COVID (1181).png	covid	1047.0	1275.0	3.0
119	COVID/COVID (1157).png	covid	913.0	1102.0	3.0
139	COVID/COVID (1150).png	covid	952.0	1184.0	3.0
...
1090	COVID/COVID (1169).png	covid	891.0	1084.0	3.0
1094	COVID/COVID (1179).png	covid	891.0	1084.0	3.0
1110	COVID/COVID (1156).png	covid	917.0	1116.0	3.0
1119	COVID/COVID (1153).png	covid	952.0	1184.0	3.0
1122	COVID/COVID (1147).png	covid	891.0	1084.0	3.0

65 rows × 5 columns

Entrée []: # Il y a 65 images qui ne sont pas carrées, et ce sont toutes des covid
Nous suggérons de prendre la plus petite résolution et de fitter toutes les images

L'étude de la luminosité, vient confirmer nos premières observations quant à une disparité avérée de luminosité/contraste entre les images à notre disposition.

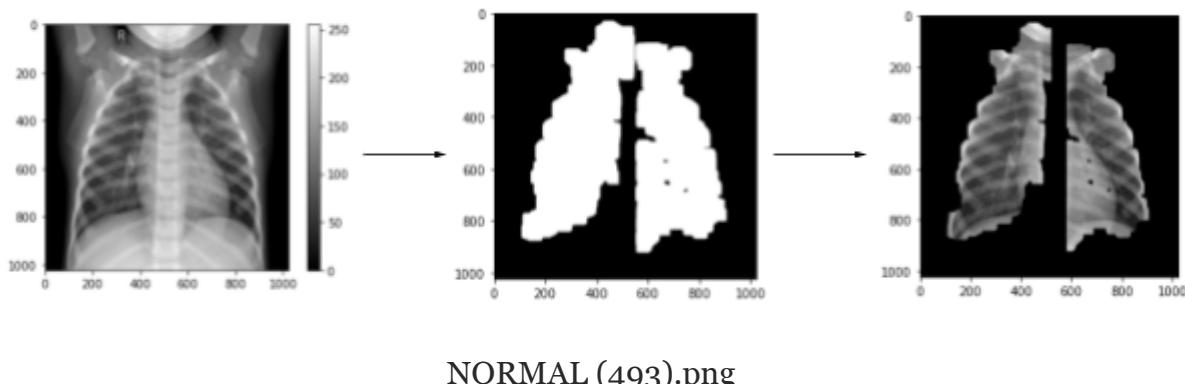
```
Entrée [14]: ┌ df['luminosite'].describe()
Out[14]:   count    3886.000000
            mean     129.088984
            std      19.686354
            min      46.618759
            25%     116.178112
            50%     128.238459
            75%     142.177826
            max     216.561127
            Name: luminosite, dtype: float64
```

Enfin, nous avons détecté deux images identiques dans le dataset : "covid 484" et "covid 409". Même si cela n'affectera pas beaucoup le réseau de neurones, nous pourrons retirer l'une des deux images.

3) Stratégie d'atténuation de l'hétérogénéité des images

Plusieurs stratégies devront donc s'imposer pour pallier ces déséquilibres et cette taille de jeu de données.

L'une des premières solutions explorées serait de partir sur une "segmentation" des poumons à l'aide de filtres OpenCV, car nous ne sommes pas en mesure d'en identifier la partie tenant un rôle déterminant dans la détection du covid. Cette approche permettra notamment de s'affranchir de traitements supplémentaires d'élimination d'information en surplus (annotation, dispositifs médicaux...)



NORMAL (493).png

Des incertitudes existent cependant concernant l'éventuelle modification ou perte d'information lors de ce traitement ou encore la faisabilité sur des clichés de qualité moindre.

Les autres pistes relèvent notamment de l'utilisation d'algorithme de détection de la qualité d'image (image quality assessment) afin de sélectionner les clichés à utiliser

dans les données d'entraînement ou d'expérimenter les algorithmes d'amélioration de qualité d'image avant d'intégrer les clichés de qualité insuffisantes à notre dataset.

4) Pistes et réflexions

- Est-ce une bonne idée de segmenter les images avant l'application d'un CNN pour ne garder que la région d'intérêt : les poumons ?
- D'autres méthodes de segmentation avec CNN existent et montrent de meilleurs résultats qu'avec les filtres :
<https://www.kaggle.com/nikhilpandey360/lung-segmentation-from-chest-x-ray-dataset> ; <https://github.com/IlliaOvcharenko/lung-segmentation> .
Pourrait-on envisager d'appliquer un CNN pour segmenter, puis un CNN pour classifier, ou même de merger les deux CNN pour n'en faire qu'un ?
- Pourrait-on trouver une métrique qui mesure la qualité d'une image X-ray, afin de créer des jeux d'entraînement et de test équilibrés ?
- Comment s'assurer que les traitements qui seront réalisés sur les images de qualité insuffisante ne viendront pas altérer l'information (présence de Covid ou non) qu'elles renferment ?