

PROJET D'ANALYSE PULMONAIRE

Alexandre AKSENOV, Eloïc ADAMJEE, Emilie MIRANDA et Khaled TILOUCHE

TABLE DES MATIERES

Projet d'Analyse Pulmonaire	1
Contexte	3
1. Données	3
1.1. Répartition par type de patients	3
1.2. Répartition des sources	4
1.3. Résolution des images	6
2. Prétraitements	6
2.1. Recherche d'outliers	6
2.1.1. Mesures de lisibilité d'une image	6
2.1.2. Lisibilité moyenne par classe et source	7
2.1.3. Identification des outliers	8
2.2. Utilisation des masques	9
2.2.1. Visualisation des masques	10
2.2.2. Part des poumons dans les images	10
3. Formalisation du problème et présélection de modèles	11
3.1. Formalisation du problème de classification	11
3.1.1. Nombre de classes	11
3.1.2. Critères de performances retenus	12
3.2. Modèles présélectionnés	12
4. Etude des modèles retenus	14
4.1. Démarche	14
4.1.1. Data augmentation	14
4.1.2. Interprétabilité : Gradcam	15
4.2. CNN	16
4.3. Modèles de transfer learning	17
4.3.1. VGG	18
4.3.2. Resnet	22
4.3.3. Inception V3	26
Conclusion	28

CONTEXTE

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

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

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

1. DONNEES

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

- Une image de radio au format png ;
- Chaque image est appairée à une image masque qui délimite la zone des poumons sur la radio. La superposition des deux permet de réduire l'image à sa portion signifiante pour le problème cible et d'éliminer le bruit apporté par les autres zones des radios.

L'ensemble des images radio a une résolution de 299 pixels, contre 256 pour les masques.

- Un ensemble de métadonnées. Ces dernières indiquent notamment un label, à savoir la classe à laquelle appartient chaque patient, et la source de la donnée. Les variables qu'elles contiennent ne comprennent pas de valeurs manquantes et sont synthétisées dans le tableau ci-dessous.

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

Tableau 1 Description des métadonnées

1.1. REPARTITION PAR TYPE DE PATIENTS

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

- Les images Covid ;
- Les images de patients atteints de pneumonie virale ;
- Les images étiquetées *lung opacity* qui correspondent à des patients souffrant de diverses infections pulmonaires ;
- Les patients sains, labellisés *normal*.

Comme illustré par la Figure 1, les données sont déséquilibrées. En effet, si les images radio sont équitablement réparties entre patients sains (un peu plus de 10000, cf. Figure 1) et patients malades (environ 11000), les malades consistent à 33% de patients covid, à 12% de malades de pneumonie virale et à 55% de malades atteints de pneumonie bactérienne. Les patients covid, notre cible, représentent ainsi 17% des observations totales.

Nous nous sommes donc assurés dans les phases de modélisation d'obtenir des classes équilibrées lors de l'échantillonnage afin d'éviter d'introduire des biais dans nos modèles et d'assurer la précision de leurs estimations.

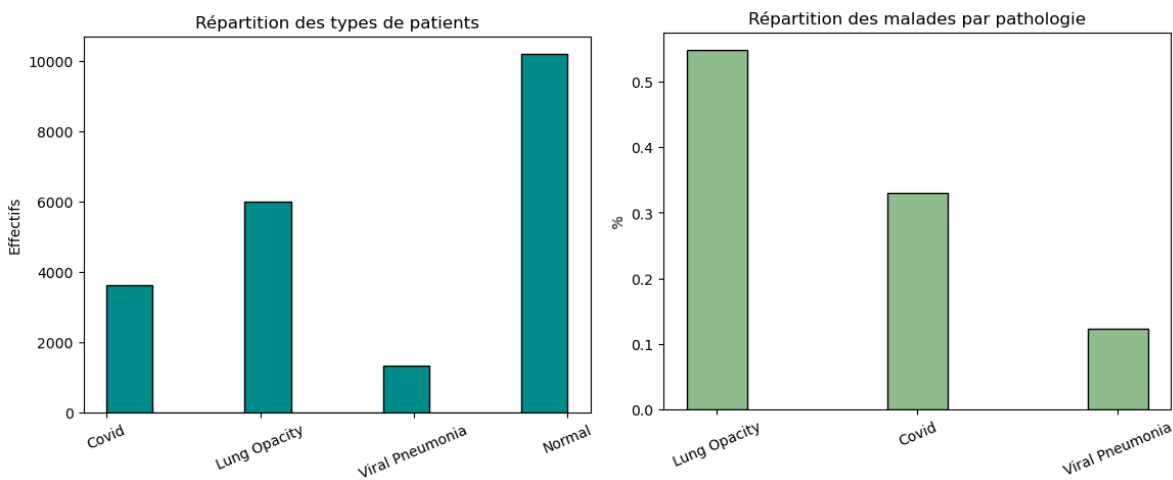


Figure 1 Répartition des données selon le type de maladie

1.2. REPARTITION DES SOURCES

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

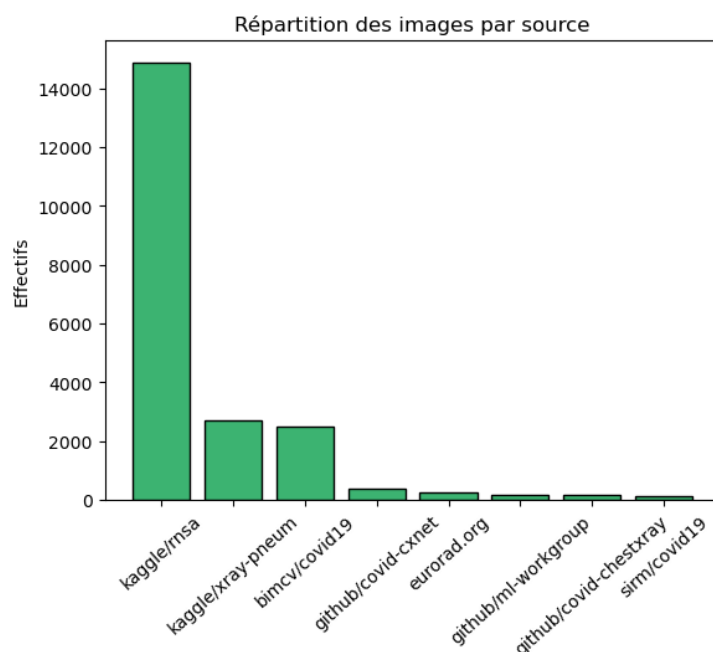


Figure 2 Répartition des données selon la source

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

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

Source	Type de patients	Covid	Viral Pneumonia	Lung Opacity	Normal
https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia		0	1345	0	1341

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

1.3. RESOLUTION DES IMAGES

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

Les images sont encodées en noir et blanc, à l'exception de 10% des données pneumonie, encodées en couleur. Ces dernières sont converties au format noir et blanc en opérant une moyenne des canaux de couleurs. L'ensemble des images est ainsi traité comme des matrices à deux dimensions.

Dans la suite, on considèrera comme résolution de référence la taille 256x256, à la fois pour s'assurer de la précision de la délimitation des masques sur les images et pour limiter le coût de calcul.

2. PRETRAITEMENTS

2.1. RECHERCHE D'OUTLIERS

Si avoir un jeu de données diversifié en termes de qualité assure une meilleure robustesse et capacité de généralisation des modèles de classification, la présence de données aberrantes peut poser problème et perturber l'apprentissage. C'est pourquoi il convient de les identifier et potentiellement de les écarter de l'analyse.

2.1.1. MESURES DE LISIBILITE D'UNE IMAGE

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

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

- Le niveau de luminosité moyen i_1 dans l'image (calculé comme la moyenne des valeurs des pixels sur l'image). Un niveau élevé traduit une image en moyenne plutôt claire.
- L'écart-type i_2 des niveaux de gris de l'image ;
- L'écart inter-quantile :

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

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

2.1.2. LISIBILITE MOYENNE PAR CLASSE ET SOURCE

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

Une première analyse consiste à croiser ces indicateurs avec les métadonnées. Les résultats obtenus sont illustrés dans la figure 3. Les images covid sont en moyenne plus claires, mais les deux indicateurs de dispersion (écart-type et écart inter-quantile) sont en moyenne plus faibles et plus dispersés que pour les autres types de patients.

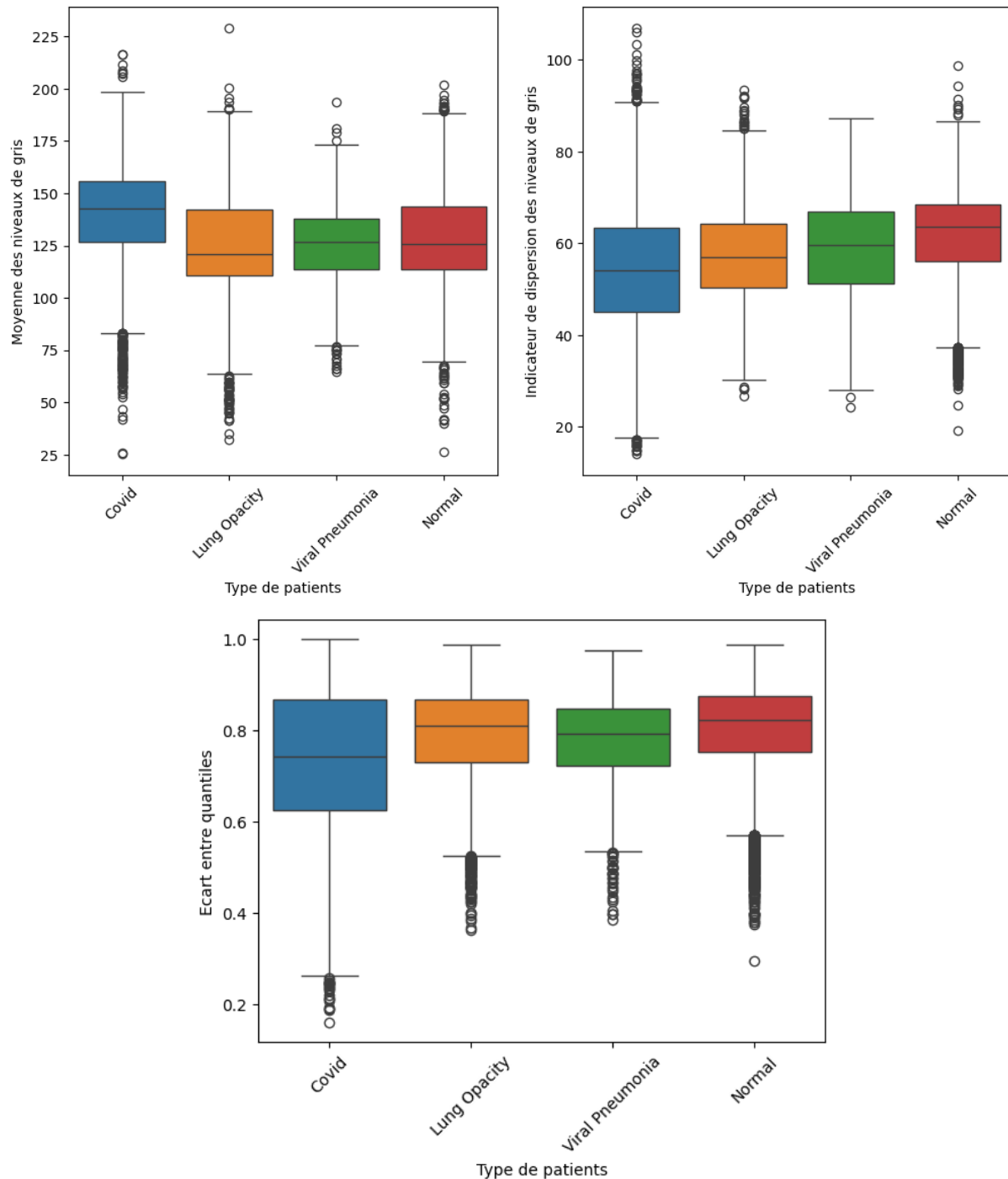


Figure 3 Indicateurs de lisibilité des images par type de patients

La figure 4 croise un indicateur de contraste avec l'origine des données, ce qui met en évidence une source en particulier qui présente un grand nombre d'images avec peu de variation de couleur. Il s'agit de la banque de

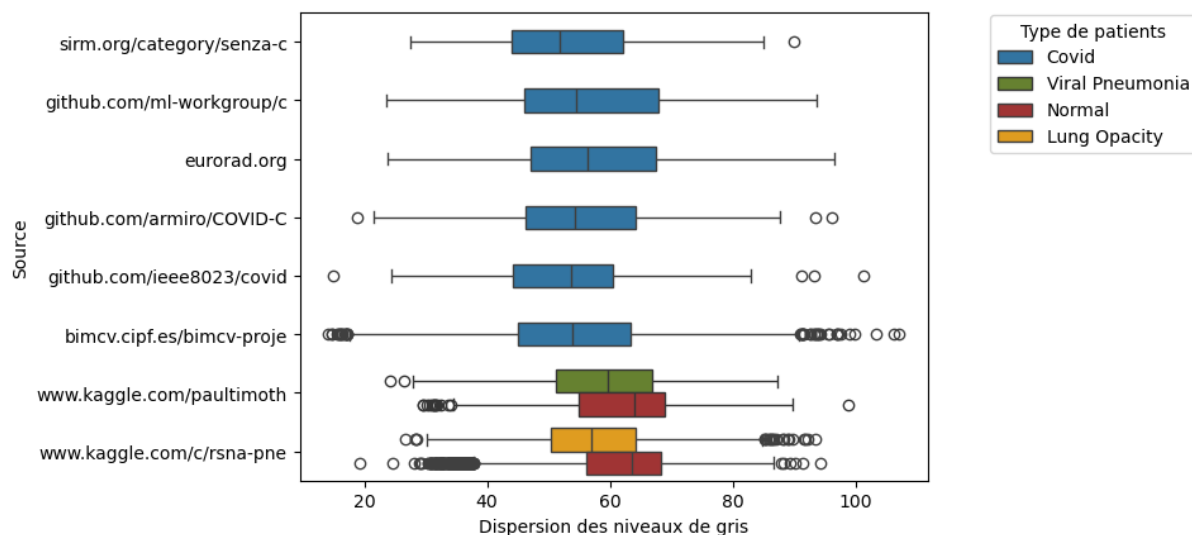


Figure 4 Dispersion des couleurs dans les images en fonction de leur source et du type de patients

données médicales de Valence (bimcv.cipf.es) qui fournit environ les 2/3 des images covid de notre ensemble de données.

2.1.3. IDENTIFICATION DES OUTLIERS

A. CAS DES IMAGES ENCADREES DE NOIR

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

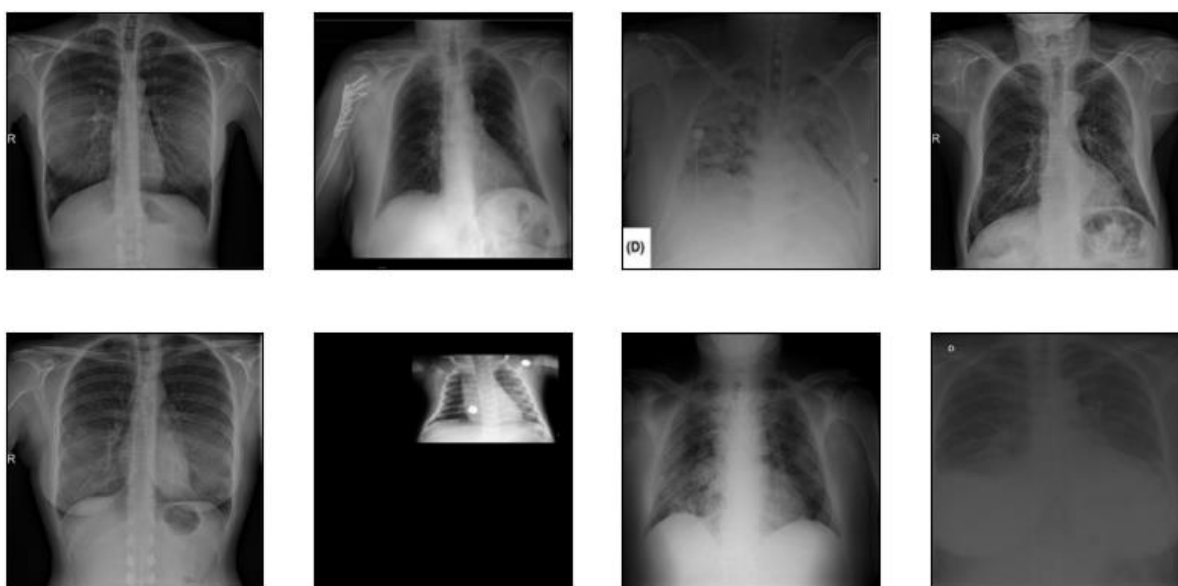


Figure 5 Exemple de radios identifiées comme anormalement sombres parmi le groupe covid

La figure 5 donne quelques exemples de telles radios covid. Au-delà des images très sombres, on remarque également que certaines sont plus petites et encadrées de noir.

Après visualisation des ces images, on constate qu'il s'agit de radios d'enfants ou nourrissons. Il est difficile sans analyse plus poussée de savoir à quelle proportion des radiologies d'enfants elles correspondent ; nous choisissons donc de les conserver dans la base afin de ne pas se priver d'informations importantes pour la généralisation du modèle.

Ces images encadrées de noir sont présentes dans toutes les classes à l'exception du groupe *Pneumonie virale*. Cette asymétrie peut créer des biais et du surapprentissage et doit être gardée à l'esprit.

B. IMAGES UNIFORMES

Un autre type de valeurs extrêmes relatif aux niveaux de contraste est également étudié afin de d'identifier les images très uniformes. On définit ici comme valeur aberrante les 0.5% d'images ayant les écarts-types de la distribution de niveaux de gris les plus bas. Ce seuil est sélectionné par analyse visuelle. Idéalement, le niveau plancher de lisibilité serait établi en se fondant sur les avis d'expert en imagerie médicale. En l'absence d'information de ce type, nous optons pour une démarche conservatrice sur la quantité et la qualité des images.

Cela correspond à 106 images dont 102 sont des données covid. Comme illustré par la figure 6, certaines images

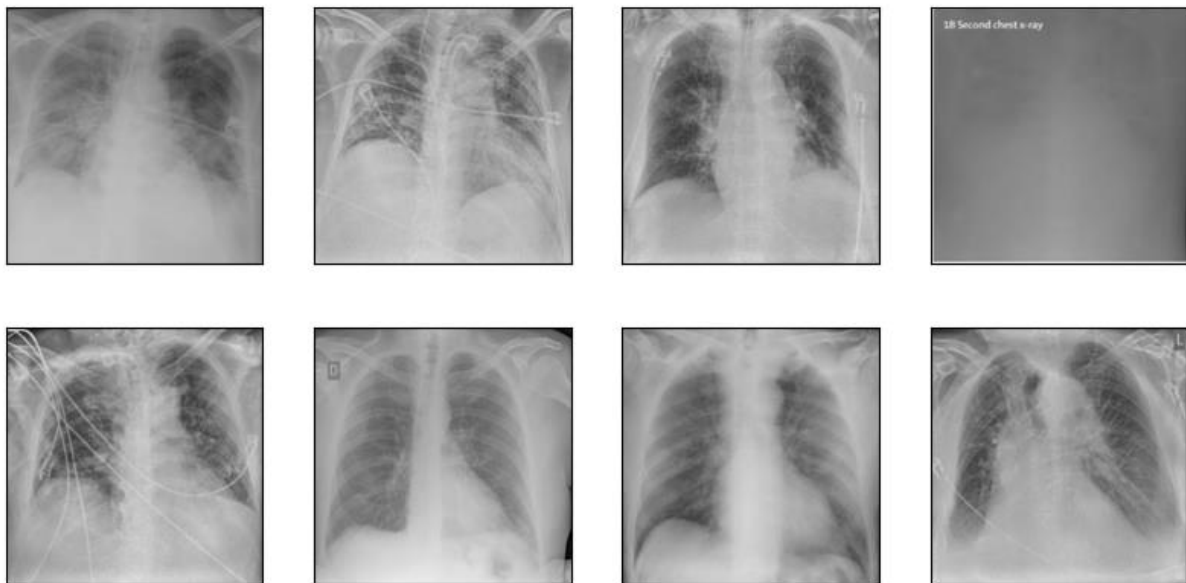


Figure 6 Exemple de radios identifiées comme anormalement peu contrastées parmi le groupe covid

sont quasiment illisibles. Bien que peu nombreuses, nous avons choisi d'écarter ces images de l'analyse.

2.2. UTILISATION DES MASQUES

2.2.1. VISUALISATION DES MASQUES

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

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

Superposer un masque sur une image d'origine est une technique utile dans de nombreux domaines, notamment en imagerie médicale et en vision par ordinateur.



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

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

Ici, et afin de faciliter la visualisation, nous avons modifié la couleur de la radiographie et du masque qui était grise de base afin de faire ressortir la zone qui nous intéressait.

La superposition de masques sur des images d'origine est une méthode visuelle puissante qui améliore la compréhension et l'analyse des données visuelles, aidant ainsi à rendre les données plus compréhensibles. Ici, les masques permettent de ne retenir que l'information pertinente à la problématique qui nous intéresse.

2.2.2. PART DES POUMONS DANS LES IMAGES

L'exploration des données de masques permet de comparer la taille des poumons sur les radios. Cela peut être réalisé en première analyse en utilisant les images de masques fournies dans les données. En délimitant la surface des poumons, ceux-ci définissent la surface utile de l'image pour la détection de pathologies pulmonaires. La figure 9 comprend les histogrammes de la répartition de la taille des poumons extraites des masques par catégorie de patients. Ces résultats sont obtenus sur des images renormalisées (valeur des pixels ramenée entre 0 et 1). Celle des images COVID paraît plus asymétrique et étalée à droite que pour les autres groupes. La classe *lung opacity* a quant à elle une taille moyenne de poumons légèrement inférieure à celle des autres. Cela paraît difficile à interpréter et pourrait être un effet réel lié à la maladie ou un effet lié à la provenance des données (comme la part de petites images encadrées de noir).

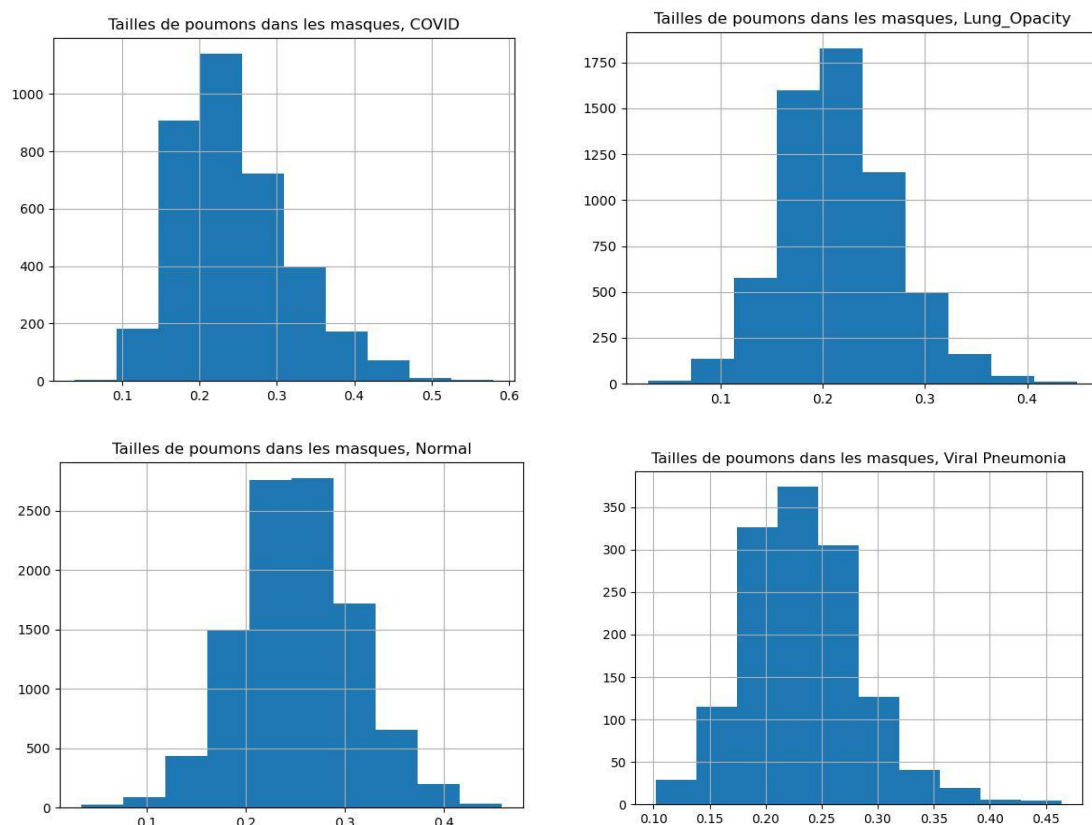


Figure 9 Répartition des tailles de poumons par type de patients

3. FORMALISATION DU PROBLEME ET PRESELECTION DE MODELES

3.1. FORMALISATION DU PROBLEME DE CLASSIFICATION

3.1.1. NOMBRE DE CLASSES

Pour rappel, l'objectif de cette étude est de développer un modèle de classification permettant d'identifier efficacement les malades du covid. La difficulté est de distinguer précisément d'une part les patients malades des patients sains, et d'autre part les patients covid des patients atteints d'autres pathologies pulmonaires.

Le choix de modélisation que nous avons réalisé est motivé par l'enjeu métier, mais également par des contraintes computationnelles. En effet, il nous est impossible de traiter l'ensemble des images sur nos machines et les options de serveurs de calculs gratuites telles que Google Collab sont également insuffisantes en termes de mémoire pour gérer ce volume de données.

C'est pourquoi nous choisissons de nous réduire à un sous-ensemble équilibré de données dans la suite. Bien que les premières explorations aient pu porter sur des sous-ensembles de tailles sensiblement plus ou moins grandes, on se limite en fin de course à 7500 observations au total.

Après avoir été investigué sous plusieurs angles, on décide de se concentrer sur un problème de classification à trois classes distinguant :

- Les malades Covid ;
- Les malades atteints de pathologies pulmonaires autres que le Covid (agrégation des classes pneumonie virale et opacité pulmonaire) ;
- Les patients sains.

Les données étant déséquilibrées, nous nous sommes assurés d'obtenir des classes équilibrées lors de l'échantillonnage afin d'éviter d'introduire des biais dans nos modèles et d'assurer la précision de leurs estimations.

Ce choix est motivé par des questions pratiques, comme expliqué ci-dessus, mais également des raisons métier. En effet, la classe *lung opacity* rassemble divers types de maladies pulmonaires provoquant une opacité visible des poumons, dont potentiellement aussi des malades de pneumonies. Dans la mesure où le groupe cible est le groupe covid, nous considérons qu'il est judicieux de rassembler l'ensemble des autres types de maladies dans une même classe. D'autre part, d'un point de vue statistique, ce choix de regroupement des images de malades non-covid permet d'assurer la présence d'images encadrées de noir dans l'ensemble des classes.

3.1.2. CRITERES DE PERFORMANCES RETENUS

Les modèles sont évalués selon une métrique donnée. On se concentre d'une part sur le taux de bonnes prédictions du modèle (précision). En effet, ce choix peut être transposé à tous les types de modèle et toutes les répartitions en classes. D'autre part, la classe cible étant le groupe Covid, on s'intéresse également à son rappel. C'est pourquoi le f1-score est aussi une métrique de comparaison des modèles importante.

3.2. MODELES PRESELECTIONNES

En phase exploratoire, nous avons testé divers modèles de Machine Learning, puis, face aux limites rencontrées avec ces derniers de Deep Learning.

Les données utilisées ont d'abord été les images de radiologies entières, puis les images réduites aux poumons en leur superposant les masques. Cette dernière option a permis d'obtenir de meilleurs résultats en évitant notamment le surajustement aux données, en atténuant par exemple les différences des niveaux de qualité ou de luminosité selon les groupes.

La démarche dans l'implémentation d'un modèle suit la procédure suivante :

Création des échantillons

- Sur un ensemble de 7500 images réparties en 3 groupes équilibrés, séparation des données en trois échantillons : apprentissage, validation et test.
- Pour les modèles de deep learning, augmentation des images par un générateur intégrant des modifications de zoom, de distorsion et symétrie.

Entraînement et sélection du modèle

- Entraînement et ajustement du modèle sur l'échantillon d'apprentissage au regard de ses performances sur celui de validation.
- Des tests successifs faisant varier les paramètres et la structure du modèle sont réalisés afin de sélectionner le plus performant, puis le modèle retenu est entraîné sur les images dans leur résolution originale et sauvegardé.

Evaluation des performances

- Evaluation des performances du modèle sur l'échantillon de test.

Les modèles étudiés ont été testés, pour certains sur le problème de classification à 3 classes, d'autres sur 4 classes. Les résolutions d'images utilisées ont également varié : en raison de leur coût en calculs élevé, les modèles étaient d'abord testés sur des images redimensionnées à des tailles réduites, puis étaient soit écartés car peu performants, soit étudiés plus en profondeur et testés progressivement sur des images de plus grande résolution.

Les résultats obtenus sont synthétisés dans le tableau suivant :

Modèle	Nombre de classes	Résolution des images	Scores globaux		Scores sur la classe Covid	
			Précision	f1-score	Précision	Rappel
XGBoost	4	299 x 299	0.74	0.74	0.71	0.76
K-NN	3	299 x 299	0.66	0.65	0.58	0.60
Random Forest	4	64 x 64	0.73	0.73	0.66	0.64
CNN (3 couches de convolution)	3	128x128	0.83	0.83	0.82	0.81
Extraction de features + Random Forest	3	100x100	0.72	0.72	0.56	0.87
VGG16 (réentraînement de 4 couches)	3	299x299	0.85	0.85	0.85	0.86
EfficientNet	3	128x128	0.73	0.73	0.69	0.71
EfficientNet + VGG	4	256x256	0.71	0.71	0.69	0.60

ResNet + VGG	4	256x256	0.81	0.81	0.78	0.77
ResNet	4	299 x 299	0.86	0.86	0.91	0.82

Nous en avons retenu trois dont les performances générales et sur la classe covid dépassent les 80% de précision et de f1-score :

- Le CNN qui obtient de bons résultats, notamment sur la classe Covid et qui est un modèle de Deep Learning simple à implémenter ;
- VGG16 dont les résultats sont parmi les meilleurs, mais qui est très coûteux en temps de calcul ;
- Et ResNet qui obtient les meilleurs résultats.

Les modèles de transfert donnant de bons résultats, nous avons également étudié le modèle Inception V3.

Dans la suite, les modèles listés ci-dessus sont présentés plus en détails et comparés selon plusieurs critères :

1. Leurs performances en prédiction ;
2. Leur facilité d'implémentation et leur coût computationnel ;
3. Leur interprétabilité.

4. ETUDE DES MODELES RETENUS

4.1. DEMARCHE

Au cours de cette dernière étape, les modèles sont réentraînés sur le problème de classification à trois classes afin de permettre une meilleure comparaison des résultats.

4.1.1. DATA AUGMENTATION

Pour l'ensemble des modèles de Deep Learning, nous avons réalisé une augmentation de données en ajoutant des transformations de zoom, de dilatation et de renversement horizontal. Cela permet d'introduire de la diversité dans les données et de compenser certaines particularités dans les données liées à la classe/source, telles que la proportion de petites images sur fond noir (cf. Figure 7).

Notons que nous avons également essayé d'ajouter des transformations liées à la luminosité pour éviter que la différence de luminosité moyenne relevée entre la classe Covid et les autres n'introduise de biais dans les modèles, mais cet ajout a généré une instabilité dans les résultats obtenus. Nous ne l'avons pas conservée, cependant il serait intéressant de chercher l'origine de cette perte de performances et de s'assurer que les performances obtenues ne sont pas liées à du surapprentissage.

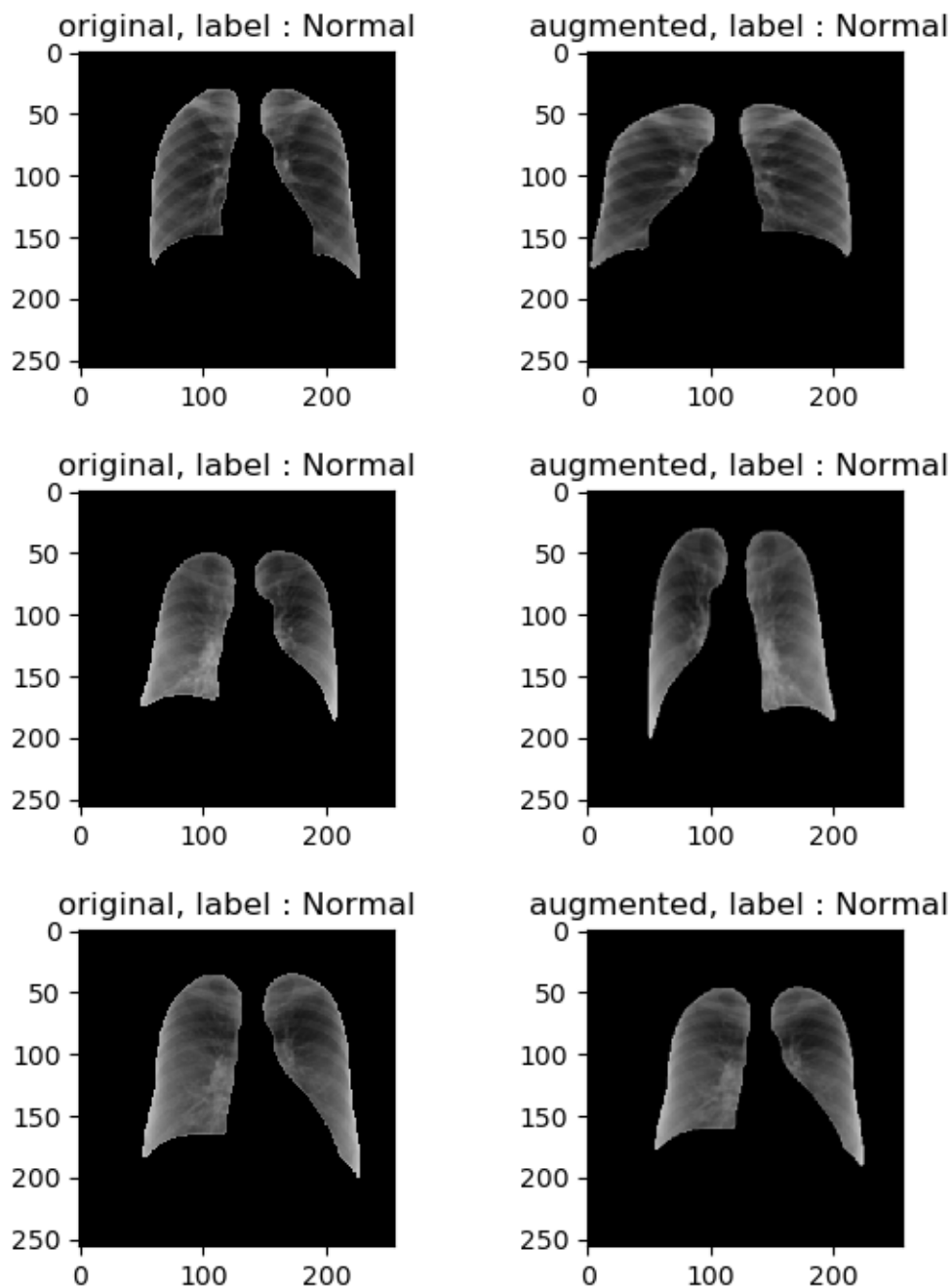


Figure 7 Exemples d'augmentation de données

4.1.2. INTERPRETABILITE : GRADCAM

La question de l'interprétabilité est cruciale pour garantir la transparence, la validité et l'utilité métier des modèles de machine et deep learning. Elle sert notamment de garde-fou en permettant d'identifier et de redresser de potentiels biais dans les modèles. Dans le cadre d'applications telles que celle qui nous intéresse en radiologie médicale, un modèle non interprétable, même s'il est très performant, n'aura qu'une utilité très limitée car ses résultats ne pourront être confirmés et complétés par l'expertise du corps médical.

Dans la suite, les modèles retenus sont évalués du point de vue de leur interprétabilité. L'approche retenue est la méthode Grad-Cam. Il s'agit d'une méthode d'interprétation visuelle des sorties d'un large panel de modèles convolutifs de deep learning qui repose sur l'évaluation locale du gradient de la fonction d'activation des couches

de classification sur les features obtenues en sortie de la dernière couche de convolution. A partir du gradient, une heatmap d'activation est générée, ce qui permet de visualiser les zones de l'image originales qui contribuent le plus à la prédiction du modèle.


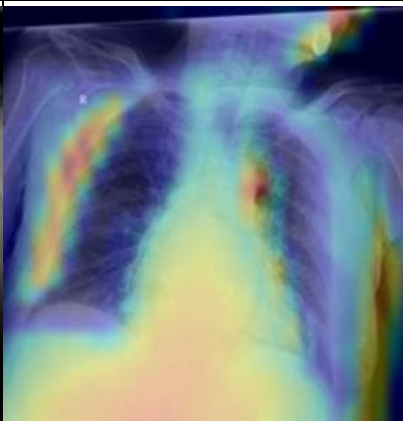
4.2. CNN


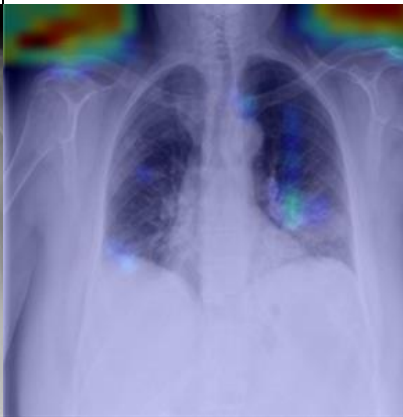


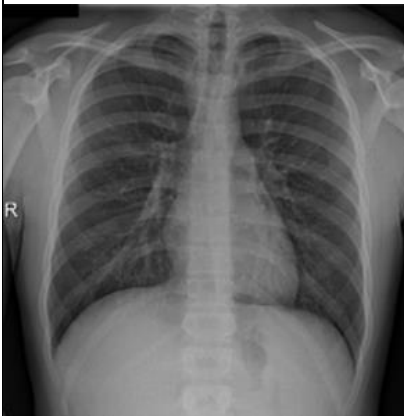
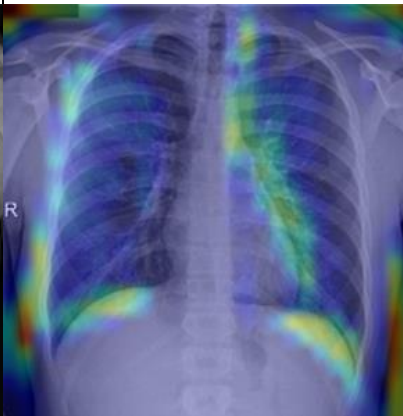
Ce modèle est relativement léger par rapport aux autres, aussi bien en terme du nombre de paramètres, qu'en termes de taille du modèle et du temps d'exécution. Par conséquent, il a pu être re-entraîné plusieurs fois sur les mêmes données. L'utilisation du GPU fait, que les paramètres diffèrent d'un entraînement à l'autre. Par conséquent, les performances du modèle sont données en termes de fourchettes.

Score global (accuracy)	f1-score global	Précision de la classe COVID	Rappel de la classe COVID
0.82 – 0.84	0.83 – 0.84	0.80 - 0.83	0.76 – 0.83

En termes de coût computationnel, une machine personnelle permet d'entraîner le modèle sur un ensemble de 5400 images avec masques en 4 minutes, et de calculer les prédictions sur un jeu de 1350 images avec masques en 1.2 secondes. Le modèle entraîné fait 5 Mo. Ces performances facilitent naturellement le développement et tests. Des comparaisons plus précises pourront être faites en re-entraînant plusieurs modèles sur une même machine.

Pour obtenir des informations relatives à l'interprétabilité, les gradients d'activation de neurones de sortie par rapport à ceux de la dernière couche de convolution (Grad-CAM) ont été calculés pour quelques images COVID du jeu de données de validation. Voici un échantillon de résultats.

Numéro du patient	Image d'origine	Grad-CAM
COVID-1499		

COVID-2495		
COVID-2980		
COVID-3583		

Nous voyons que ce modèle sélectionne au moins une zone de poumon pour chaque exemple, mais aussi regarde souvent en-dehors du corps (probablement, en prenant que ces zones foncées aussi pour des poumons). Malgré les résultats mitigés, un observateur humain pourra facilement séparer les zones de poumons des zones à l'extérieur du corps. La sélection des zones de poumons reste potentiellement un indicateur utile.

Ceci suggère aussi une piste d'amélioration : une segmentation (plus grossière) de l'intérieur du corps pourrait être ajoutée aux masques. Cette information supplémentaire pourrait rendre le modèle plus fiable.

4.3. MODELES DE TRANSFER LEARNING

Le Transfer Learning est une approche de Deep Learning dans laquelle les performances d'un modèle d'apprentissage sont améliorées en transférant de l'information depuis un domaine d'étude proche. On utilise des modèles pré-entraînés sur de larges bases de données d'un domaine voisin de celui d'intérêt, ici la classification d'images, de manière à initialiser les poids de notre modèle.

Les modèles existants comportent une partie d'extraction de features composées d'un ensemble de convolutions, suivie d'un ensemble de couches denses pour la classification. Les modèles pré-entraînés fournissent les poids de la première partie, facilitant l'apprentissage du modèle. On peut également choisir de réentraîner une partie des couches de convolutions de l'extraction de features afin de s'adapter au mieux à notre problème cible.

Deux des modèles retenus lors de la présélection sont des modèles de transfert : VGG et ResNet, tous deux entraînés sur la base *ImageNet* contenant plus de 14 millions d'images labellisées.

4.3.1. VGG

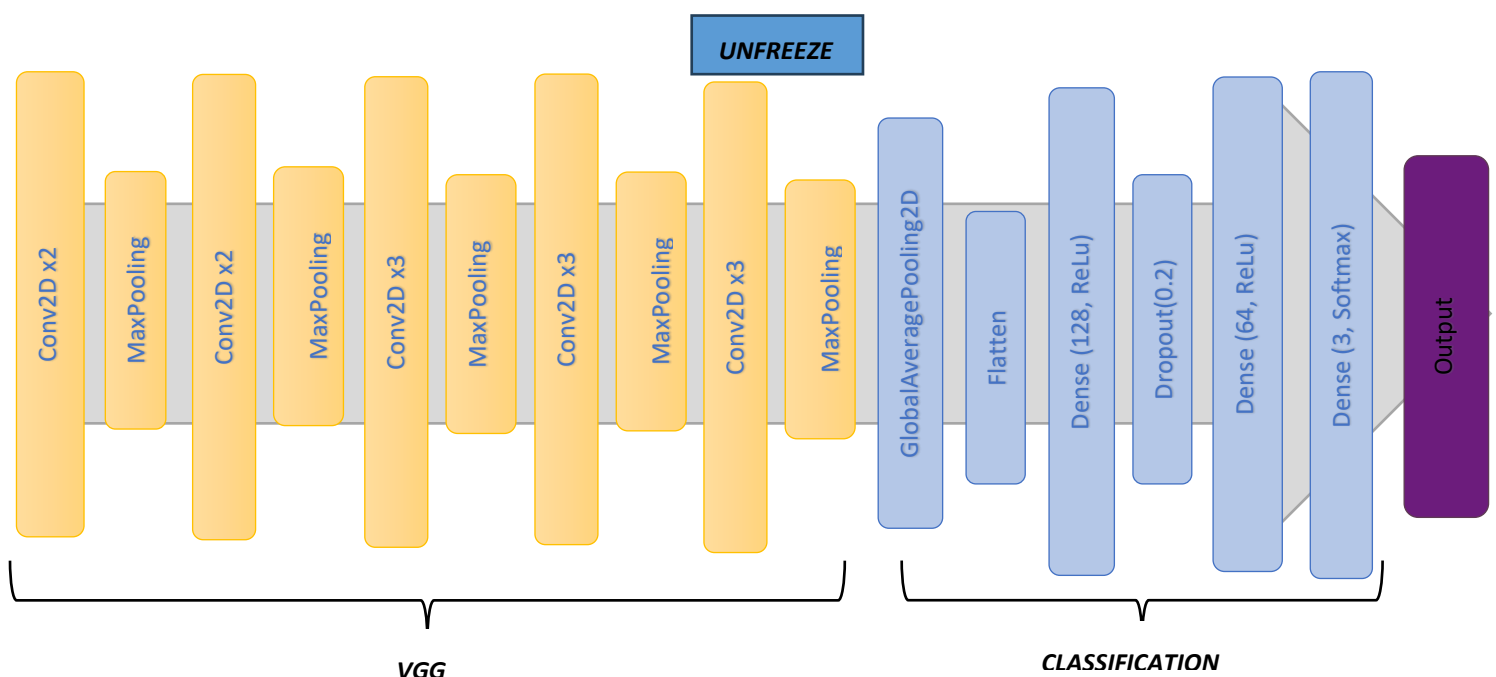
ARCHITECTURE

L'extraction de features fournie par VGG est constituée de 5 blocs convolutifs comprenant chacun entre 2 et 3 couches. On y ajoute des couches denses de classification.

Le modèle est ajusté sur les images de résolution réduite (64x64) afin de permettre une exécution suffisamment rapide pour tester plusieurs modélisations. Les principaux paramètres investigués sont les suivants :

- Le nombre de couches denses (de 1 à 5) et le nombre de neurones par couche ;
- Le nombre d'étapes de dropout et leur intensité entre les couches denses ;
- L'utilisation directe des poids pré-entraînés directement ou le réentraînement du dernier bloc de convolutions.

L'architecture du modèle retenu est la suivante :



PERFORMANCES

Ce modèle a été implémenté sur le jeu de 7500 images en résolution 256x256 sur 50 epochs, avec un arrêt anticipé à la 24ième. L'évolution des performances du modèle sur les échantillons d'apprentissage et de validation sont données par la Figure 7. Le modèle atteint une précision de 87% de précision sur l'échantillon de validation. Si les performances sont bonnes, la principale difficulté réside ici dans le coût computationnel de ce modèle : son temps d'entraînement sur les 24 epochs est de près de 7h.

Les performances en prédiction sont évaluées sur l'échantillon de test. On obtient une précision générale de 85% et un f1-score global de 85%. Le détail par classe est donné dans le tableau suivant :

	Precision	Recall	f1-score
COVID	0.84	0.83	0.84
Normal	0.81	0.88	0.85
Pulmonary Infection	0.90	0.84	0.87

Là encore, on note que les infections hors covid sont les plus faciles à identifier avec un f1-score de 0.87. La classe Covid obtient une précision et un recall équilibrés avec respectivement 84% et 83%. On a donc malgré tout 17% d'images covid qui ne sont pas identifiées comme telles dans le modèle.

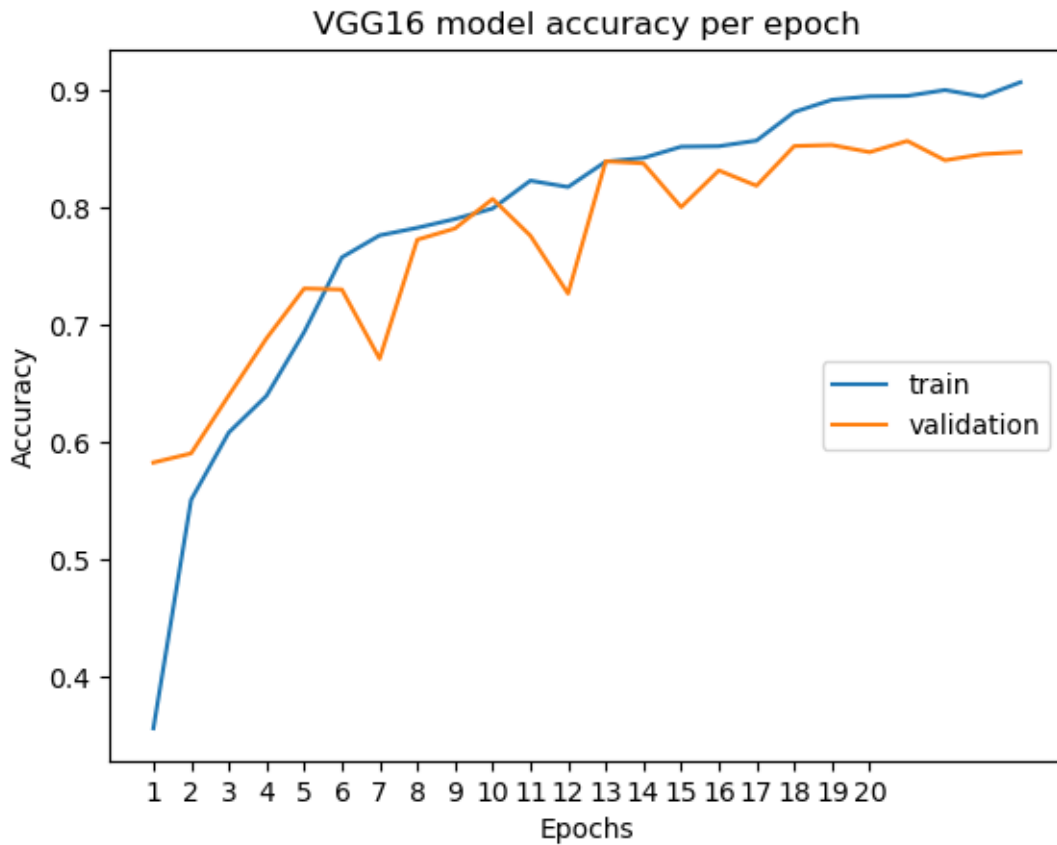
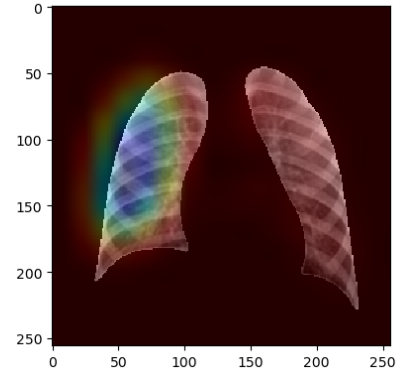
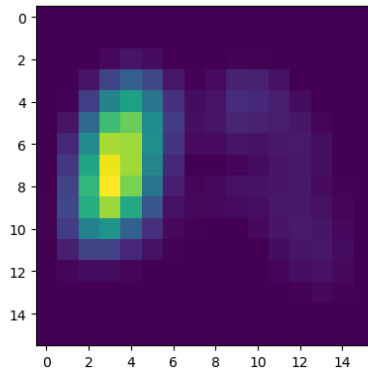
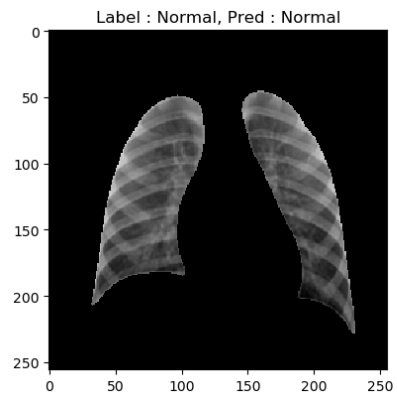
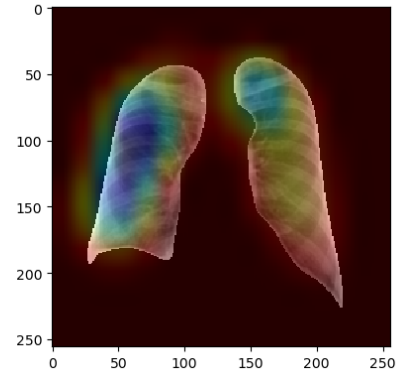
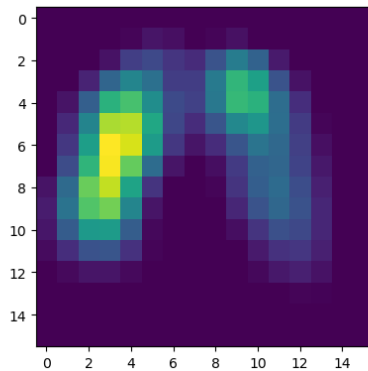
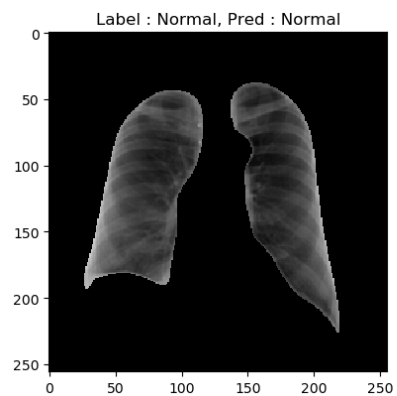
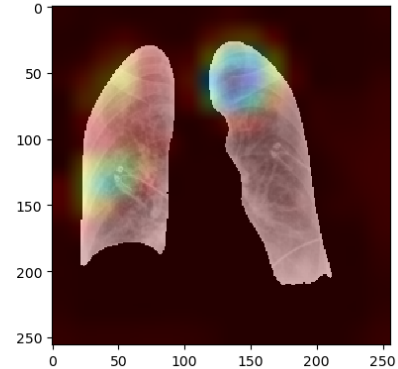
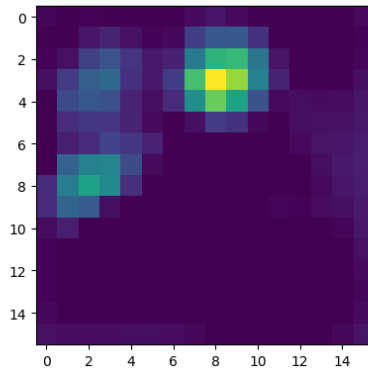
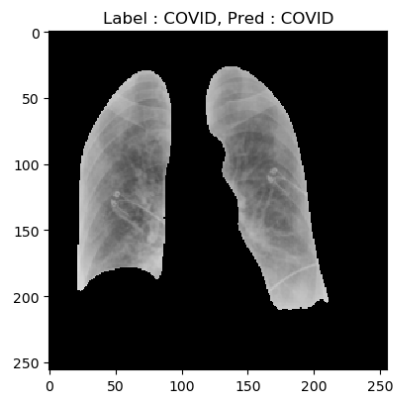
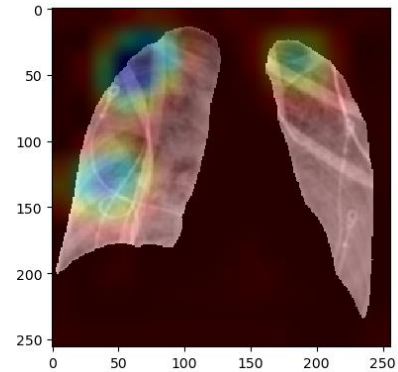
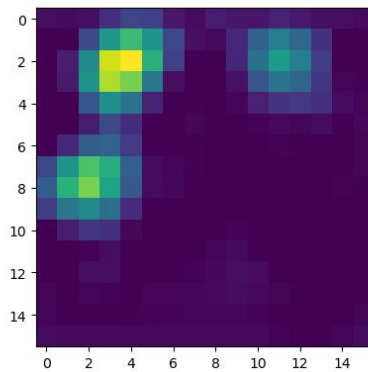
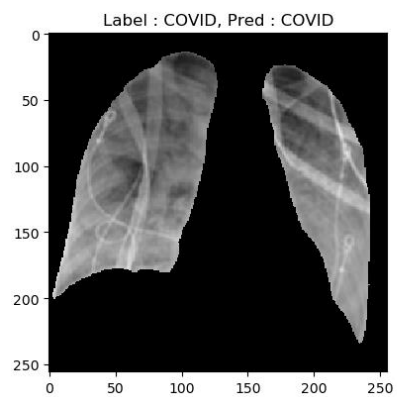


Figure 8 Précision du modèle en fonction du nombre d'epochs

INTERPRETATION

L'interprétabilité du modèle est jugée à travers l'utilisation de la méthode GradCam.

La figure suivante rassemble les résultats obtenus sur 10 images de l'échantillon de test bien prédites par le modèle. On note que pour la classe *Normal*, les zones illuminées sont assez larges, prenant parfois l'intégralité d'un poumon, là où celles des classes *Covid* ou *Infections pulmonaires* sont beaucoup plus localisées. Sans avis d'expert, il est difficile de conclure précisément, mais ces observations suggèrent que le modèle fonde sa classification sur la détection ou non de marqueurs de pathologies.



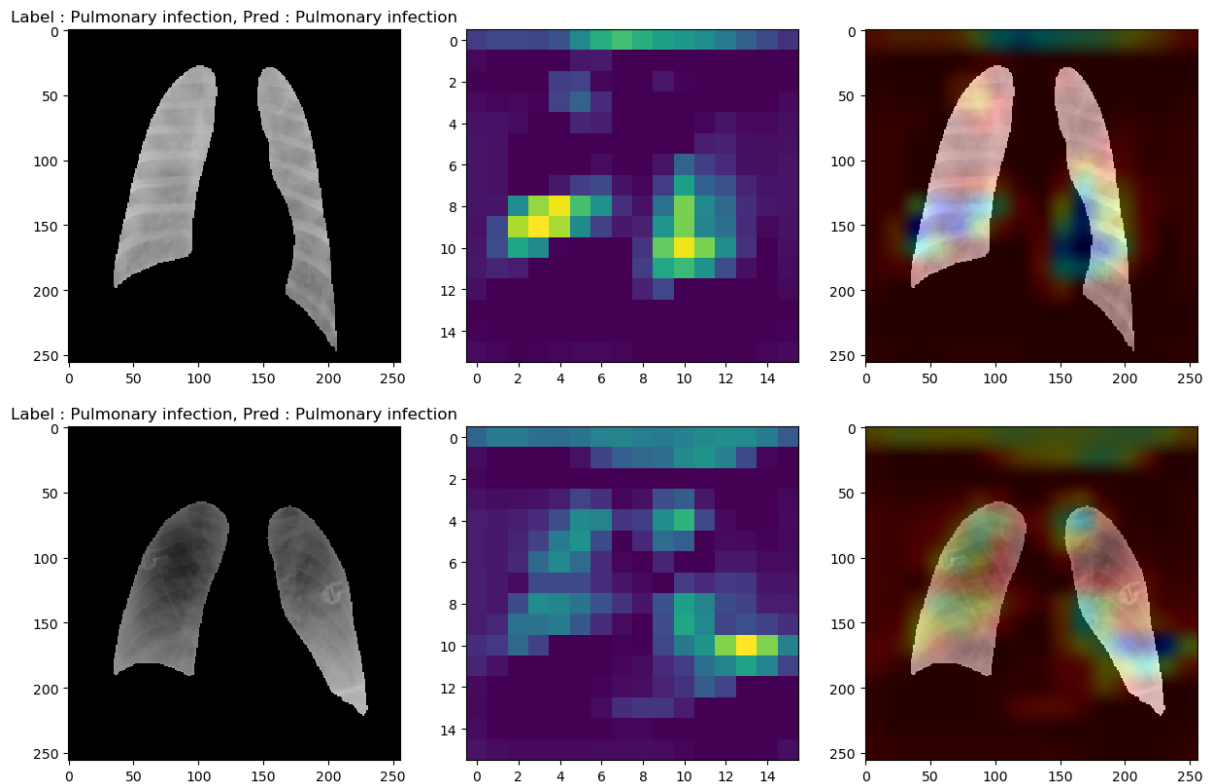


Figure 9 Heatmap GradCam sur des images de test

4.3.2. RESNET

Le modèle ResNet utilise des réseaux de neurones convolutifs profonds pour capturer des caractéristiques complexes et abstraites des images, ce qui le rend particulièrement adapté à la classification d'images médicales.

ARCHITECTURE

L'architecture ResNet, est conçue pour permettre l'entraînement de réseaux de neurones profonds sans souffrir de la dégradation des gradients. Elle repose sur des blocs résiduels, qui utilisent des connexions courtes ou "skip connections" pour faciliter l'apprentissage de l'identité. Ces connexions ajoutent directement l'entrée d'un bloc à sa sortie, aidant ainsi à maintenir les gradients pendant l'entraînement. Un bloc résiduel typique comprend plusieurs couches convolutionnelles suivies de couches de normalisation par lot et d'activations ReLU, avec une connexion directe entre l'entrée et la sortie du bloc. Par exemple, dans une architecture comme ResNet-50, la première couche est une convolution de grande taille, suivie de blocs résiduels organisés en plusieurs étapes, avec des profondeurs de filtres croissantes. Après les blocs résiduels, un global average pooling est appliqué, suivi d'une couche fully connected qui produit la sortie finale. Cette structure permet à ResNet de capturer des caractéristiques complexes tout en maintenant des gradients stables, ce qui conduit à des performances élevées sur diverses tâches de reconnaissance d'images.

Performances Générales :

```

Accuracy: 0.8590694665908813
47/47 [=====] - 8s 162ms/step
F1-score: 0.8588122705061481

```

	precision	recall	f1-score	support
COVID	0.91	0.82	0.86	506
Normal	0.78	0.97	0.86	490
PNEUMO_OPACITY	0.93	0.79	0.85	487
accuracy			0.86	1483
macro avg	0.87	0.86	0.86	1483
weighted avg	0.87	0.86	0.86	1483

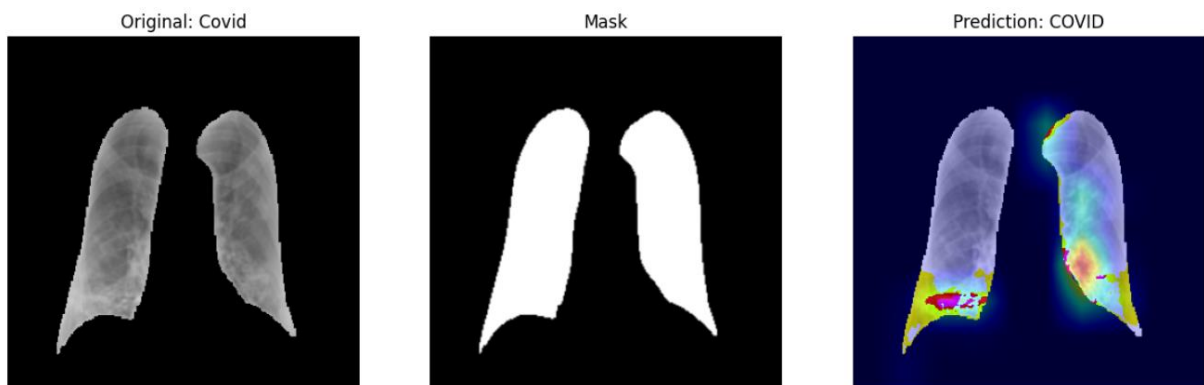
Analyse : Le modèle ResNet affiche une performance globale élevée avec une accuracy de 86%. La précision pour la classe COVID est particulièrement bonne à 91%. En revanche, on note un déséquilibre assez marqué (plus de 10 points) entre précision et rappel pour les différentes classes : le rappel de la classe normale est très élevé, mais celui des deux autres l'est nettement moins : cela suggère que le modèle a tendance à classer des patients malades parmi les sains. Du point de vue de notre application, c'est un problème car il est important de bien identifier les malades, ce qui suppose de minimiser le taux de faux négatifs (patients malades prédits comme patients sains).

Coût computationnel :

Bien que ResNet soit performant pour la classification d'images, le coût en termes de temps d'entraînement est élevé, similaire à l'algorithme initial utilisé. L'entraînement sur GPU permet de réduire le temps de calcul, mais reste une tâche intensive.

On note un training time d'environ 45 minutes et un temps d'exécution global du programme de près de 50 minutes.

INTERPRETATION



Interprétation par Grad-CAM des Prédications de ResNet sur des Images de Validation

Nous avons utilisé Grad-CAM (Gradient-weighted Class Activation Mapping) pour générer des HeatMaps sur quatre images de validation analysées par un modèle ResNet. Grad-CAM est une technique qui permet de visualiser les zones d'une image qui contribuent le plus aux prédictions d'un réseau de neurones. En utilisant les gradients des neurones cibles de la dernière couche convolutive, Grad-CAM crée des cartes de chaleur (HeatMaps) qui indiquent les régions importantes pour la décision de classification.

Résultats sur les Images de Validation

La HeatMap générée par Grad-CAM montre clairement les zones où le modèle ResNet a détecté une forte probabilité d'infection COVID-19. Les zones rouges indiquent une forte probabilité, tandis que les zones bleues indiquent une faible probabilité. Nous avons observé que les régions les plus probables d'infection se situent principalement dans les parties inférieures des poumons.

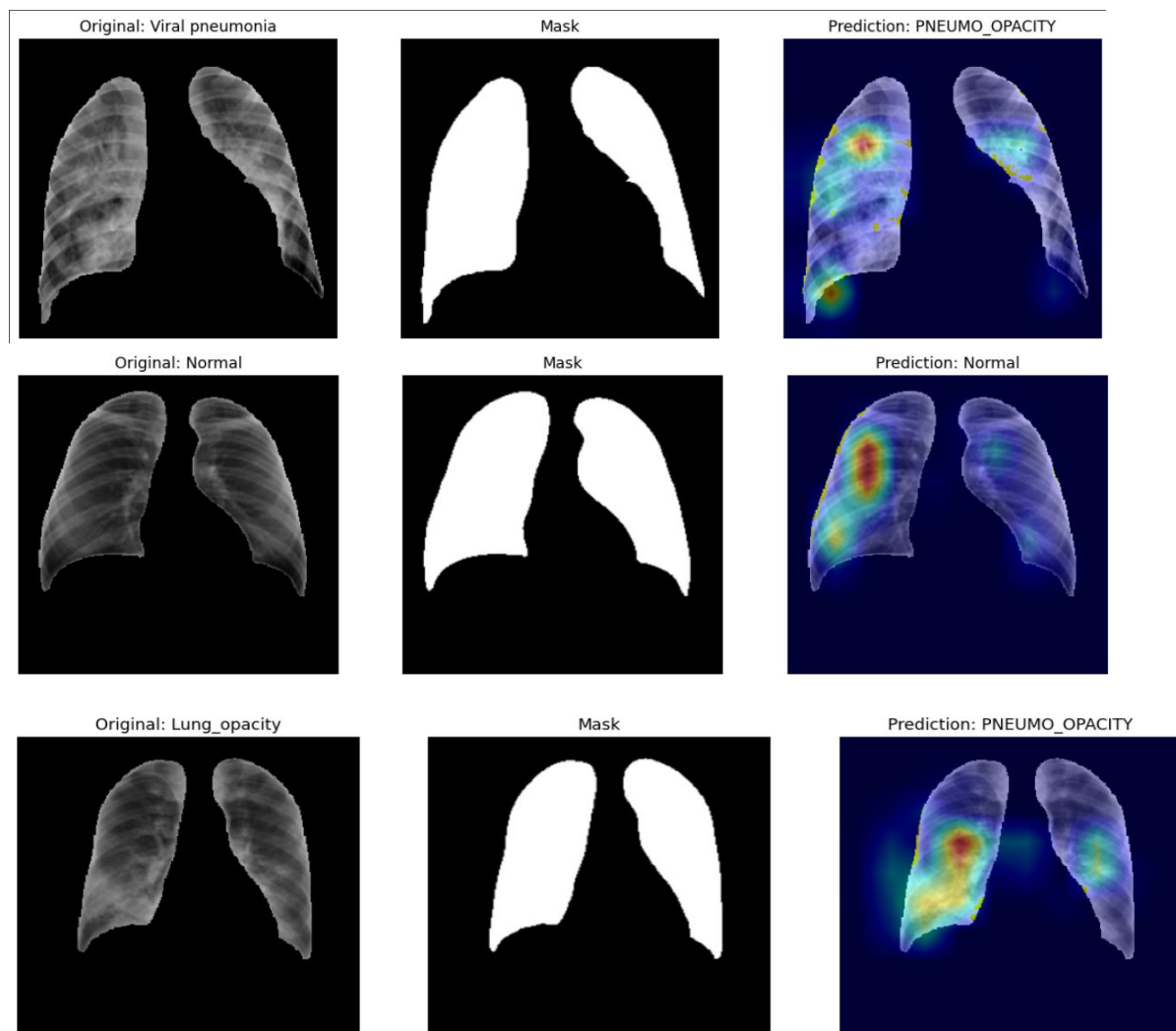
Autres Images de Validation

Des HeatMaps similaires ont été générées pour les autres images de validation. Dans chaque cas, les zones de haute probabilité d'infection COVID-19 sont principalement concentrées dans les parties inférieures des poumons, ce qui est cohérent avec les observations cliniques.

Importance des HeatMaps Grad-CAM

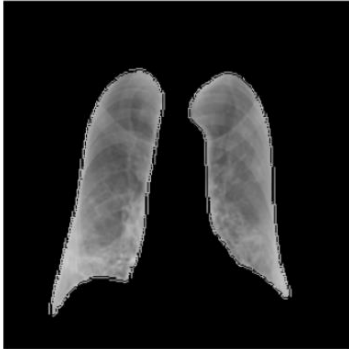
Les HeatMaps fournies par Grad-CAM permettent non seulement de confirmer les prédictions du modèle ResNet, mais aussi de visualiser les zones des poumons les plus affectées par l'infection. Cette visualisation est cruciale pour les professionnels de santé, car elle offre des indications précieuses sur les régions des poumons à surveiller attentivement.

Grâce à l'utilisation de Grad-CAM, nous avons pu non seulement valider les prédictions du modèle ResNet pour la détection des infections COVID-19, mais aussi localiser visuellement les zones d'infection dans les poumons. Cette approche améliore la compréhension des prédictions du modèle et fournit un outil précieux pour le diagnostic médical.

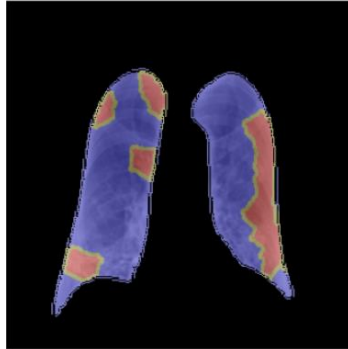


Interprétation des Prédictions de ResNet avec LIME sur des Images de Validation

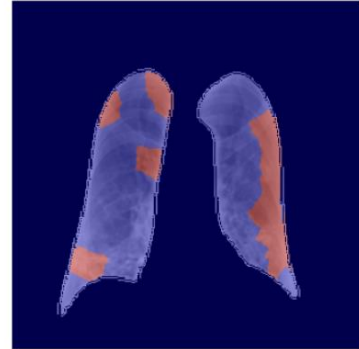
Original Image with Original Mask



LIME Segments



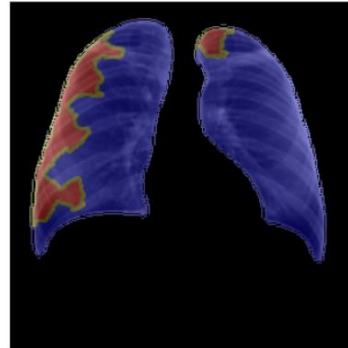
LIME Explanation



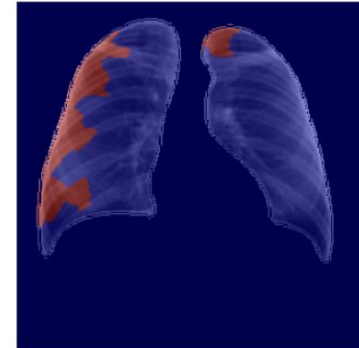
Original Image with Original Mask



LIME Segments



LIME Explanation



4.3.3. INCEPTION V3

CONTEXTE ET CONFIGURATION

Le modèle Inception V3 a été utilisé pour classifier des images de poumons en trois catégories : COVID, Normal et Pneumonia. L'entraînement a été effectué en utilisant des poids pré-entraînés sur ImageNet, suivi d'une phase de fine-tuning sur un ensemble de données spécifique après augmentation des données.

```
Accuracy sur l'ensemble d'entraînement :
0.8104761838912964
165/165 137s 827ms/step
Classification Report (Ensemble d'entraînement) :
      precision    recall  f1-score   support

   COVID           0.85        0.70        0.77        1716
  Normal           0.75        0.93        0.83        1760
 Pneumonia         0.86        0.79        0.83        1774

 accuracy                   0.81        5250
 macro avg           0.82        0.81        0.81        5250
weighted avg           0.82        0.81        0.81        5250

Accuracy sur l'ensemble de test :
0.7871111035346985
71/71 69s 978ms/step
Classification Report (Ensemble de test) :
      precision    recall  f1-score   support

   COVID           0.83        0.67        0.74        784
  Normal           0.72        0.92        0.81        740
 Pneumonia         0.84        0.77        0.81        726

 accuracy                   0.79        2250
 macro avg           0.80        0.79        0.79        2250
weighted avg           0.80        0.79        0.78        2250
```

ANALYSE DES PERFORMANCES

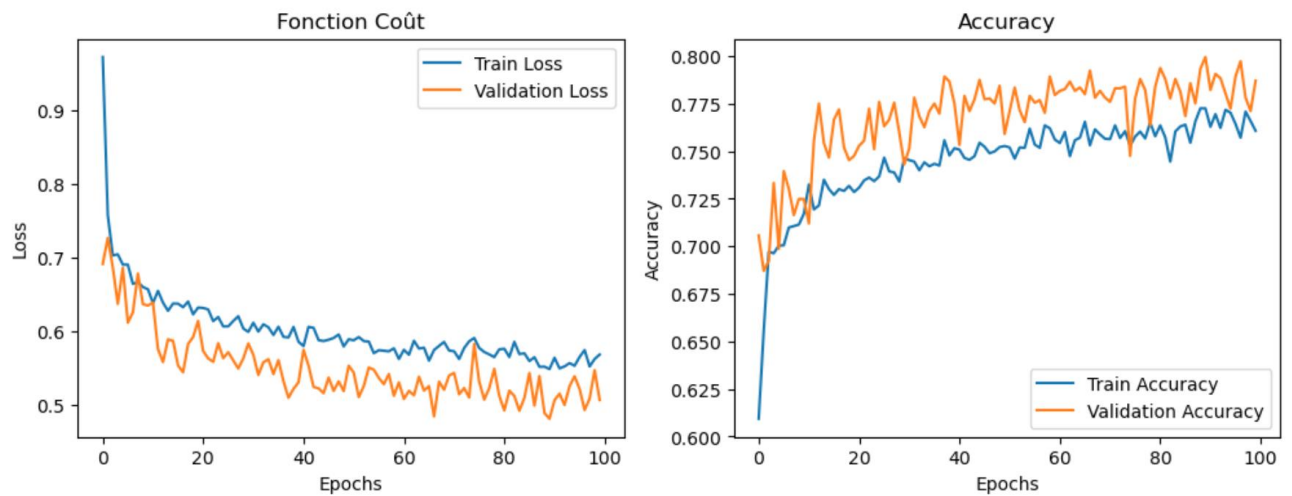
1. Ensemble d'Entraînement :

- Le modèle montre une bonne performance générale avec une accuracy de 81%.
- La classe "Normal" a la meilleure performance avec un recall de 93%, ce qui signifie que presque toutes les images normales sont correctement identifiées.
- La classe "COVID" a un recall plus faible de 70%, indiquant que certaines images COVID sont mal classifiées.

2. Ensemble de Test :

- L'accuracy sur l'ensemble de test est légèrement inférieure à celle de l'ensemble d'entraînement (78.71% contre 81.05%).
- La classe "Normal" continue de montrer une forte performance avec un recall de 92%.

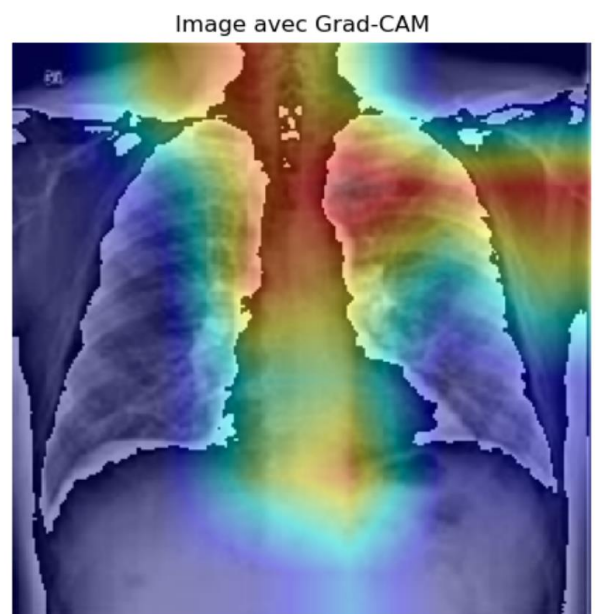
- La classe "COVID" a une performance de rappel relativement faible (67%), ce qui pourrait être dû à la variabilité des images ou à un déséquilibre dans les données d'entraînement.



Globalement, les performances du modèle sont donc moins bonnes que celles des précédents.

INTERPRETATION

L'interprétation avec Grad-CAM permet de visualiser quelles parties du poumon contribuent le plus à la prédiction du modèle. Cette technique produit une heatmap qui montre les zones de l'image les plus influentes pour la prédiction. Ici nous avons utilisé l'image d'un patient Covid où la zone influente semble se situer en haut du poumon.



CONCLUSION

La problématique de classification d'images de radiologies dans le but de fournir une méthode automatisée de diagnostic médical soulève diverses questions :

- La qualité et fiabilité des images utilisées (absence de biais liée à la sélection ou à la provenance des données) ;
- La disponibilité et fiabilité des labels associés utilisés pour l'apprentissage ;
- La capacité à proposer un modèle dont les résultats puissent être contrôlés et validés par des experts ;
- Le coût computationnel associé aux traitements de grands volumes d'images.

Notre objectif était dans cette étude de développer un modèle prédictif performant pour détecter précisément les cas Covid. Nous avons considéré un problème à 3 classes équilibrées (patients Covid, patients sains et patients atteints de diverses pathologies) à partir d'un échantillonnage des données disponibles. Les données sélectionnées ont également été augmentées de manière à limiter le risque de surapprentissage lié aux spécificités des données.

Pour cette application de diagnostic médical, il est important de bien identifier les cas Covid et de ne pas les confondre avec d'autres maladies pulmonaires, et surtout avec des patients sains. C'est pourquoi le suivi des performances statistiques des modèles s'est fait autour des critères de précision et de f1-score, mais également de précision et de rappel sur la classe Covid en particulier.

Après s'être confrontés aux limites des modèles de machine learning pour ce type de problème, nous avons étudiés et retenus plusieurs modèles de deep learning permettant d'obtenir une précision de plus de 80% : un modèle CNN et trois modèles de transfert.

Une étude plus poussée de ces quatre modèles a été réalisée, prenant en compte leurs performances, leur coût d'entraînement et leur interprétabilité. A l'issue de ces recherches, nous pouvons proposer les recommandations suivantes :

- Dans un contexte de ressources en calcul limitées, le CNN fournit un bon compromis entre performances et facilité d'implémentation et d'entraînement. Avec un temps d'exécution de quelques secondes, il est à préconiser dans le cas où on alimente régulièrement les bases de données et on a besoin de mettre à jour régulièrement les estimations. Cela aurait par exemple été le cas au cours de la période covid, lorsque les images de patients identifiés Covid arrivaient en masse.
- En revanche, si l'on dispose d'un vaste volume de données figé et d'une grande puissance de calcul, les modèles VGG et ResNet fournissent des performances légèrement meilleures. ResNet atteint la meilleure précision sur la classe Covid, mais a tendance à classer des images de malades parmi les sains. Or, dans le cas d'une application médicale de diagnostic, le coût de la non identification d'une pathologie peut être élevé. De son côté VGG atteint des performances globales similaires et stables avec un bon équilibre entre précision et rappel.
Du point de vue de l'interprétabilité, il est difficile sans expertise métier de trancher sur lequel donne les résultats les plus cohérents.