Table des matières

Contexte

Maladie infectieuse émergente, apparue en Chine continentale, à la fin de l'année 2019, la COVID-19 (maladie à coronavirus 2019) a provoqué une crise sanitaire et économique majeure. En l'état actuel de nos connaissances, le diagnostic précoce de cette maladie demeure un des meilleurs moyens de lutter contre sa propagation.

Dans le cadre du développement de moyens de détection rapides et sûrs de la maladie, l'objectif de ce projet est de créer une intelligence artificielle capable de détecter la présence de la COVID-19 chez un patient, à partir d'une radiographie pulmonaire. Cette technique pourrait être l'un des moyens les plus rapides et efficaces pour diagnostiquer cette maladie en milieu hospitalier.

Pour développer cette intelligence artificielle, nous nous baserons principalement sur des techniques de réseaux de neurones, et plus spécifiquement, les réseaux de neurones convolutifs (CNN). Ces techniques ont montré leurs efficacités pour la classification des images.

L'entraînement de ces algorithmes de deep learning se base sur un jeu de données constitué de milliers de radiographies pulmonaires.

Ces clichés radiographiques de patients sont de trois types :

— des radiographies de patients détectés positifs à la COVID-19 ;
— des radiographies de patients sains ;
— et des radiographies de patients atteint de pneumonies virales, autres que la COVID-19

L'objectif de ce travail est donc d'entraîner, à partir du jeu de données dont nous disposons, une intelligence artificielle capable de distinguer les trois cas (COVID-19, normal et viral). Pour atteindre cet objectif, nous distinguons principalement deux grandes étapes :

— l'exploration, l'analyse et la visualisation de données constituent la première étape du projet. Cette étape permettra d'examiner le format et la qualité des données, d'identifier des biais dans ces données, etc...
— la deuxième étape consiste à mettre en place des algorithmes de classifications qui se basent sur les réseaux de neurones. Plusieurs algorithmes, plus ou moins complexes, seront par la suite testés et évalués.

1 Données

Les données exploitées dans ce travail sont des images de ragiographies pulmonaires, disponibles sur le site kaggle (https ://www.kaggle.com/tawsifurrahman/covid19-radiography-database). Ces images sont regroupées en trois classes : COVID-19, normal et viral-pneumonia, pour des personnes qui sont respectivement, atteintes de la COVID-19, saines et atteintes de pneumonie virale.

1.1 Nombre d'images par classe

Nous disposons d'un jeu de données équilibré entre les trois classes. La figure 1, qui montre le nombre d'images pour chacune des trois classes, permet d'illustrer cet équilibre.

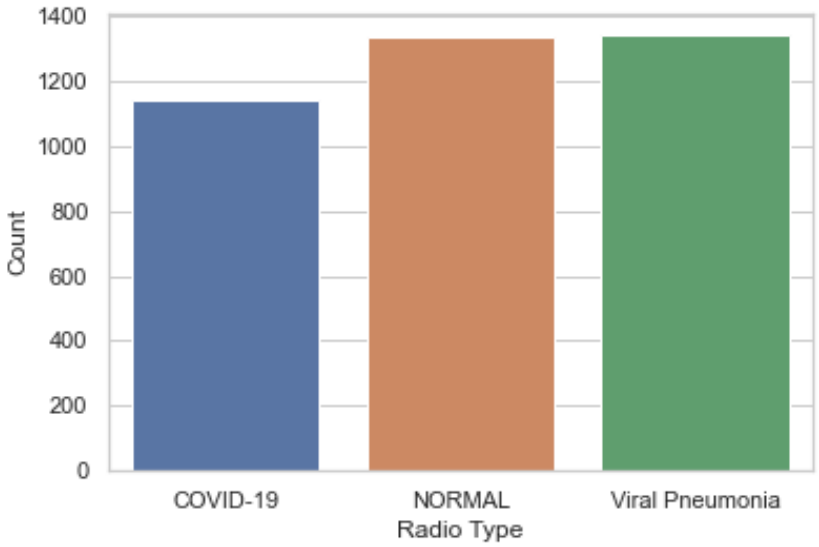


FIGURE 1 – Nombre d'images par classe

1.2 Taille et encodage des images

En termes de format de données, les images exploitées dans ce projet diffèrent sur deux aspects :

— Nombre de pixels : le nombre de pixels suivant les deux directions varie d'une image à l'autre ;
— Encodage : bien que les images exploitées soient, visuellement, en niveau de gris, certaines sont encodées en niveau de gris et d'autres en RGB. Les images en RGB ont des valeurs des canaux R, G et B qui sont égales.

La figure 2 montre des tableaux qui regroupent, pour chacune des trois classes, le nombre d'images par type d'encodage et par nombre de pixels suivant les deux dimensions.

2 Analyse et statistiques sur les images

L'objectif de cette partie est d'analyser les radiographies, afin d'identifier les différences entre les images, et de détecter la présence de biais.

Dimension X	Dimension Y	Couleur	Nombre
160	187	RGB	1
197	253	RGB	1
256	256	Gray level	558
331	331	Gray level	583

(a) COVID-19

Dimension X	Dimension Y	Couleur	Nombre
1024	1024	Gray level	1341

(b) Normal

Dimension X	Dimension Y	Couleur	Nombre
1024	1024	Gray level	1205
		RGB	140

(c) Pneumonie virale

FIGURE 2 – Pour chaque classe : nombre d'images par type d'encodage et par nombre de pixels suivant les deux dimensions.

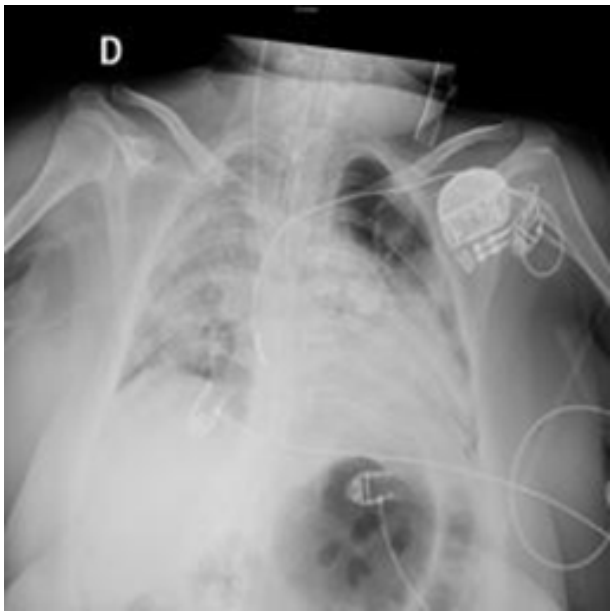


FIGURE 3 – Exemple d'une radiographie COVID-19

2.1 Première analyse visuelle

Une première analyse visuelle est effectuée en parcourant des images pour les trois classes. Nous observons, sur quelques radiographies, la présence de certains appareils médicaux de mesures, de forme filaire ou tubulaire. **Ces dispositifs médicaux sont d'avantage présents dans les radiographies de type COVID-19, ce qui peut créer un biais lors de la classification**.

2.2 Statistiques sur les images

Afin d'analyser la répartition des niveaux de gris, nous allons analyser la distribution de la moyenne et de l'écart-type de ces niveaux de gris dans les images de chacune des classes. La figure 4 montre la distribution des moyennes et des écart-types pour les trois classes. Nous constatons que :

— La dispersion des moyennes et des écart-types des niveaux de gris est importante. Cette dispersion est plus accentuée pour les images de types COVID-19 et Viral Pneumonia.

— Contrairement aux autres images, les images de type COVID-19 ont tendance à avoir une moyenne plus élevée et un écart-type plus faible. Dans ce qui suit, nous allons tenter d'identifier la cause de cette différence.

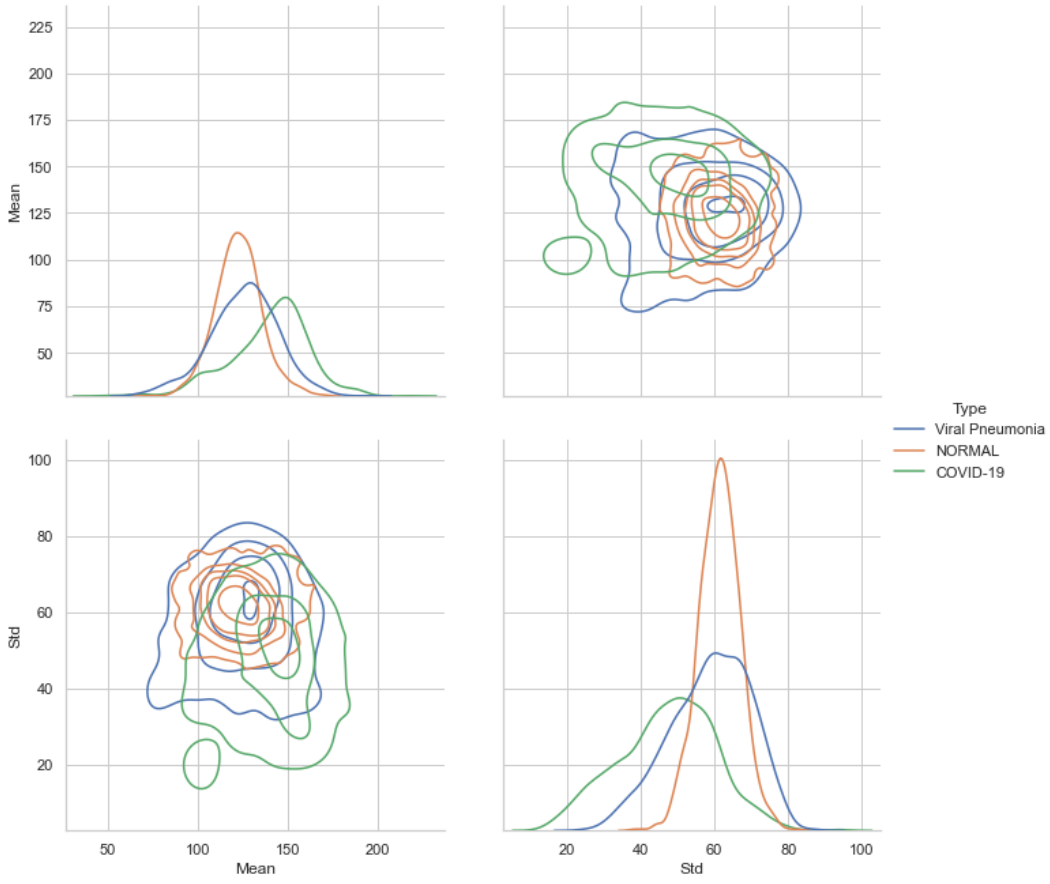


FIGURE 4 – Distribution des moyennes et écart-types pour les trois classes.

2.3 Filtrage des pixels les moins pertinents

Nous proposons de filtrer les pixels les moins pertinents pour la classification des images, en utilisant la classe `SelectPercentile` de `Sklearn.feature_selection`. La figure 5 montre un filtrage des pixels en utilisant différents pourcentages de pixels filtrés. Dans les six images présentées dans la figure, les pixels blancs correspondent aux pixels filtrés.

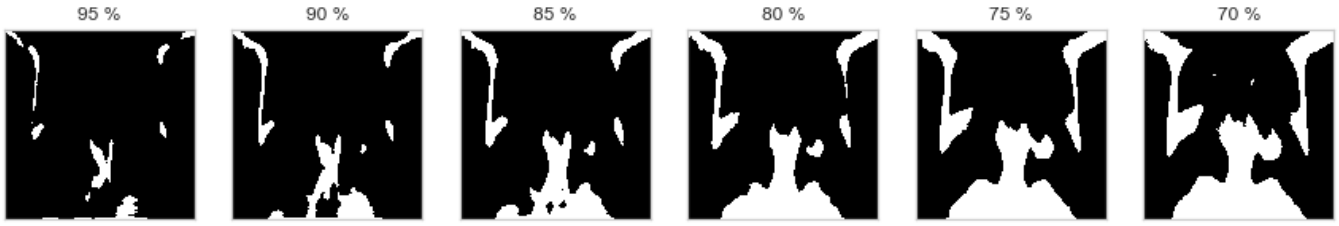


FIGURE 5 – Filtrage des pixels les moins pertinents

Légitimement, nous pensions que les bords des radiographies étaient peu porteurs d'informations.

Cependant, dans la figure 5, les bords verticaux de l'image ne sont toujours pas éliminés, même avec un paramétrage fixé à 70 %.

Nous verrons , dans le paragraphe suivant, que cela est dû à un biais sur les images.

2.4 Biais sur les images de type COVID-19

Une analyse poussée des images a permis d'identifier un deuxième biais (autre que le biais des équipements médicaux) : les images de type COVID-19 contiennent moins de bords noirs. Ces radiographies ont fait l'objet d'un pré-traitement, elles ont subi un "zoom".

Nous avons pu remarquer ce biais en appliquant deux modèles de visualisation d'images : le modèle Isomap et le modèle TSNE. Les résultats de la visualisation de ces deux modèles sont présentés dans le fichier `datavisualisation.ipynb`.

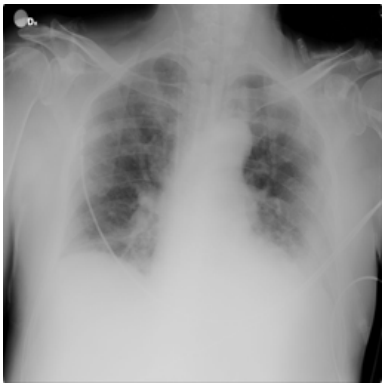
Ce biais est cohérent avec ce que nous avons observé pour le filtrage des pixels les moins pertinents (figure 5) et dans les distributions (figure 4) :

- Les bords verticaux des images ne sont pas éliminés lors du filtrage (figure 5), car ils participent fortement à distinguer les images de type COVID-19 des autres images.
- Dans les distributions (figure 4), les images de COVID-19 ont des moyennes plus élevées et des écart-types plus faibles, ce qui est un effet de l'absence des bords noirs. Cette absence induit une augumentation du niveau de gris moyen, et une diminution de l'hétérogénéité de l'image et donc de son écart-type de niveau de gris.

2.5 Bilan

La visualisation et l'analyse des données nous a permis de constater que les images de type COVID-19 diffèrent des autres images par deux aspects :

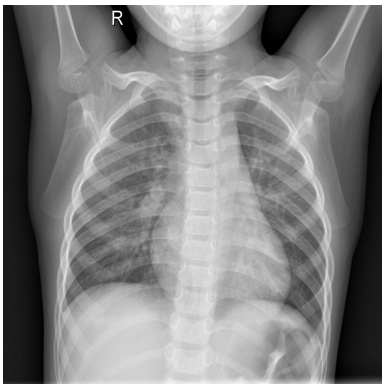
- La présence plus fréquente de certains appareils médicaux de mesures, de forme filaire ou tubulaire ;
- et l'absence de bords noirs verticaux, résultant probablement d'un zoom effectué par les émetteurs de ces images.



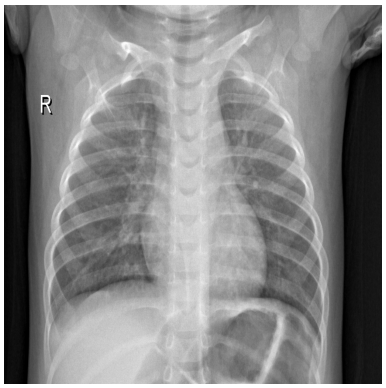
(a) COVID-19

(b) COVID-19

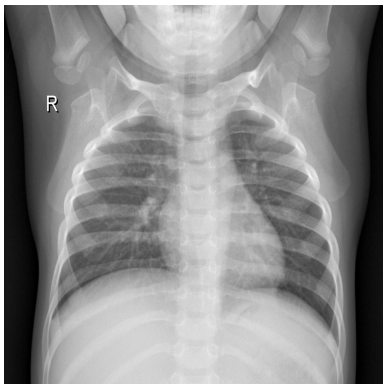
(c) COVID-19



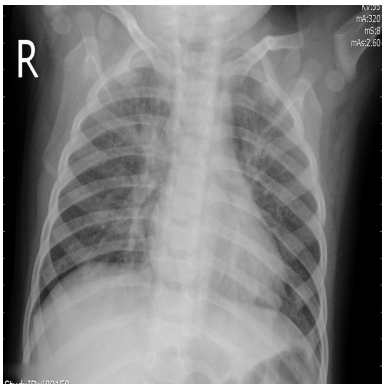
(d) Normal



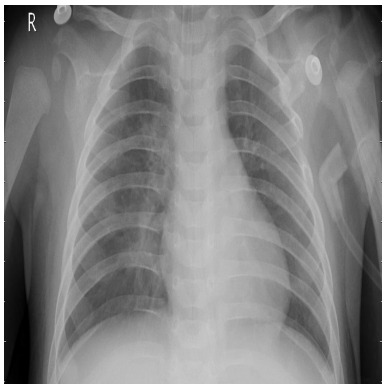
(e) Normal



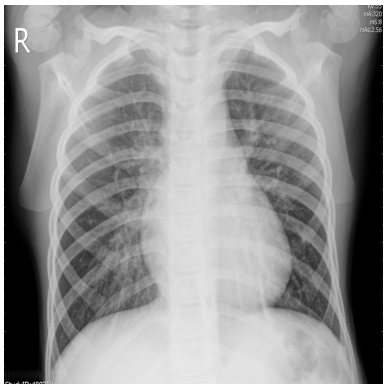
(f) Normal



(g) Pneumonie virale



(h) Pneumonie virale



(i) Pneumonie virale

FIGURE 6 – Comparaison d'images prises au hasard parmi les trois classes.

Ces caractéristiques risquent d'induire en erreur un modèle prédictif, qui se baserait en partie sur une reconnaissance liée à celles-ci.

Il est possible de corriger la deuxième différence, en effectuant une opération de pré-traitement des images avant de les utiliser pour l'entraînement des modèles. Cette opération consiste à éliminer les bords noirs verticaux sur toutes les radiographies. Ainsi, le modèle se concentrera mieux sur les différences de caractéristiques aux niveaux des poumons.

3 Modélisation

Dans cette partie, nous proposons différents modèles pour la classification des images. Ces modèles se basent sur les réseaux de neurones convolutifs.

Les différentes modélisations abordées dans ce projet sont les suivantes :

— Modélisation 1 : CNN Lenet.
— Modélisation 2 : CNN personnalisé
— Modélisation 3 : Transfer Learning avec EfficientNetB5
— Modélisation 4 : Features extraction avec InceptionV3 et Xgboost.
— Modélisation 5 : VGG16 avec augmentation d'image.

Avant de détailler les modèles, il convient de décrire le pré-traitement effectué sur les images et la métrique qui sera utilisée pour l'évaluation des modèles.

3.1 Pré-traitement des images

Traitement des bords verticaux noirs Lors de la phase de datavisualisation, l'équipe projet a constaté que les radiographies de type COVID ne présentaient pas de bords noirs latéraux, tandis que les radiographies de type Normal et Pneumonie en présentaient.

L'application d'un traitement d'image peut atténuer le biais induit par la présence de bords noirs verticaux.

L'idée sous-jacente, est de réaliser un «zoom» sur les radiographie NORMAL et PNEUMONIE.

La solution retenue, préconise l'élimination des bords latéraux, inférieurs et supérieurs, afin de se focaliser sur l'information utile, les poumons. La partie basse des radiographies, correspondant aux coupoles diaphragmatiques est éliminée car peu porteuse d'informations. Il en est de même pour la partie haute des radiographies, correspondant à la région du corps située au dessus des clavicules. Les parties latérales sont éliminées car elles constituent l'arrière plan de la radiographie.

Redimensionnement et normalisation des images Un traitement est ensuite nécessaire pour avoir des images de même dimension. Ceci est nécessaire pour l'entraînement des modèles. La taille choisie pour toutes les images est de 256X256 pixels.

Pour un meilleur entraînement des modèles, une bonne pratique consiste à normaliser toutes les images, en divisant le niveau de gris de chaque pixel par 255, cette valeur étant le niveau de gris maximal.

3.2 Métriques et sorties des modèles

Fonction de pertes La fonction de perte est la valeur que l'algorithme cherchera à minimiser. Dans notre cas, nous utilisons la fonction `categorical_crossentropy`, adaptée pour une problématique de

classification multi-classes.

Métrique La métrique qui nous intéresse est la précision `accuracy`. Elle correspond au pourcentage des bonnes prédictions.

Entrainement et test des modèles Les données de ce projet, qui sont les images radiographiques, sont séparées en des données d'entrainement et des données de test :

- Les données d'entraînement représentent 80% des données totales. Ces données vont servir à entraîner les modèles.
- Les données de test représentent 20% des données totales. Ces données vont servir à tester les modèles.

Sortie des modèles La sortie des modèles correspond à la probabilité d'une image d'appartenir à chacune des trois classes. Par la suite, nous attribuons à l'image, la classe qui a la probabilité la plus élevée.

3.3 Description des modèles

Nous proposons, dans ce qui suit, 5 modèles de classification qui se basent sur les réseaux de neurones convolutifs :

Modèle LeNet Cette première modélisation fait appel à une architecture basée sur "LeNet", un réseau de neurones convolutifs, caractérisé par une alternance de couche de convolution et de pooling, permettant l'extraction des caractéristiques des radiographies. La classification est assurée par une couche dense de réseau de neurones connectés.

Modèle personnalisée Cette deuxième modélisation correspond à un modèle personnalisé par l'équipe, caractérisé par une alternance de couche de convolution et de pooling, permettant l'extraction des caractéristiques des radiographies. La classification est assurée par une couche dense de réseau de neurones connectés.

EfficientNetB5 Cette troisième modélisation est basée sur le principe du transfer learning, qui consiste à utiliser un modèle pré-entraîné sur des centaines de milliers d'images. Le modèle pré-entraîné choisi est EfficientNetB5, il permet l'extraction des caractéristiques des radiographies. La classification est assurée par une couche dense de réseau de neurones connectés.

Afin d'améliorer la performance de notre modèle, une opération de fine tuning a permis d'adapter les 3 dernières couches de convolutions du modèles pré-entraîné, à nos données. Ce réglage fin permet un nouveau calcul des poids synaptiques et donc une meilleure extraction des caractéristiques des radiographies.

InceptionV3 Cette quatrième modélisation combine les techniques de deep learning et de machine learning "classique". Un modèle pré-entraîné, InceptionV3, est utilisé pour extraire les caractéristiques des radiographies et un XGBoostClassifier est utilisé pour opérer la classification en sortie du modèle pré-entraîné.

VGG16 avec augmentation d'image Cette dernière modélisation utilise la technique d'augmentation d'image qui permet d'enrichir le jeu de données et de réduire le phénomène d'overfitting. VGG16, un modèle pré-entraîné, est utilisé pour extraire les caractéristiques des radiographies. La classification est assurée par une couche dense de réseau de neurones connectés.

En le combinant à une opération de fine tuning sur les 4 dernières couches du modèle VGG16, le modèle s'avère performant dans ses prédictions.

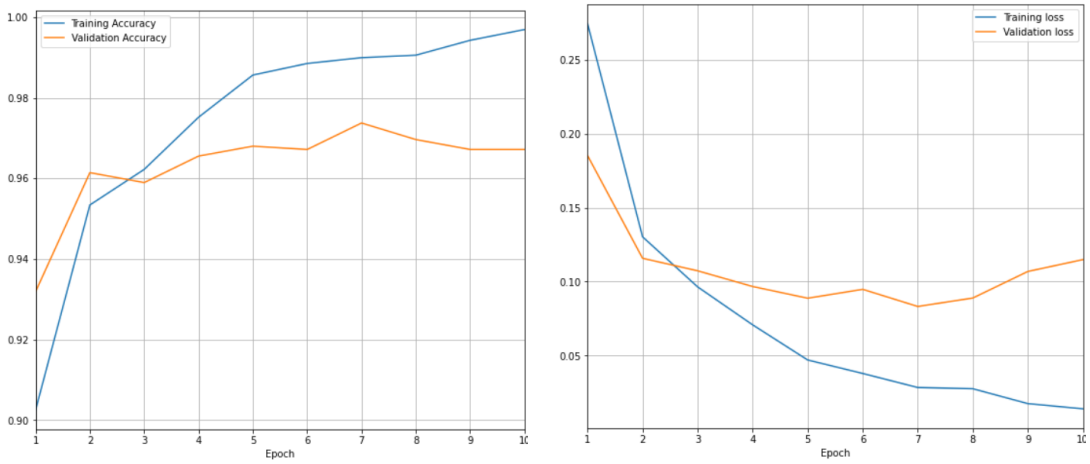
3.4 Entraînement des modèles

L'entraînement des modèles se fait par descente de gradient moyennant l'optimiseur "Adam", et avec une taille de Batchs de 32.

En ce qui concerne le nombre d'Epochs, nous procédons de deux manières selon si le modèle utilise le transfert learning ou non :

- Dans le cas des modèles sans transfert learning, c'est-à-dire les deux premiers modèles (Lenet et le modèle personnalisé). Le nombre d'Epochs utilisé est de 10.
- Dans le cas des autres modèles qui se basent sur le transfert learning, nous procédons en deux étapes :
  - La première étape consiste à entraîner les couches dense de sortie des modèles, en conservant les mêmes paramètres de base des modèles de transfert learning. Le nombre d'Epochs utilisé est de 10.
  - La deuxième étape consiste à intégrer dans l'entraînement du modèle, les dernières couches du modèle pré-entraîné. Comme dans l'étape précédente, le nombre d'Epochs est de 10.

La figure 3.4 illustre l'évolution des métriques "Loss" et "Accuracy" en fonction des Epochs pour le jeu de données d'entraînement et de test. La figure concerne la deuxième étape de l'entraînement du modèle EfficientNet, les résultats pour le reste des modèles sont présentés dans le fichier notebook.



(a) Accuracy (b) Loss

3.5 Prédiction des modèles

Le tableau suivant regroupe la précision des différents modèles sur les données de test. Nous constatons que le modèle EfficientNet donne la meilleure précision parmi tous les modèles.

La figure 7 montre la matrice de confusion de ce modèle sur les images de test. Nous constatons que :

Modèle	Précision
Lenet	93%
Personnalisé	95%
InceptionV3	90%
EfficientNet	97%
VGG16	96%

TABLEAU 1 – Comparaison des précisions de prédiction sur les images de test pour les différents modèles

— La classe COVID-19 est "presque" parfaitement distinguée des deux autres classes : uniquement une seule image de type COVID-19 est mal prédite.
— Les deux autres classes (normal et pneumonie virale) sont moins bien prédites, mais les résultats de prédictions sont bons : la proportion des images mal prédites reste faible.

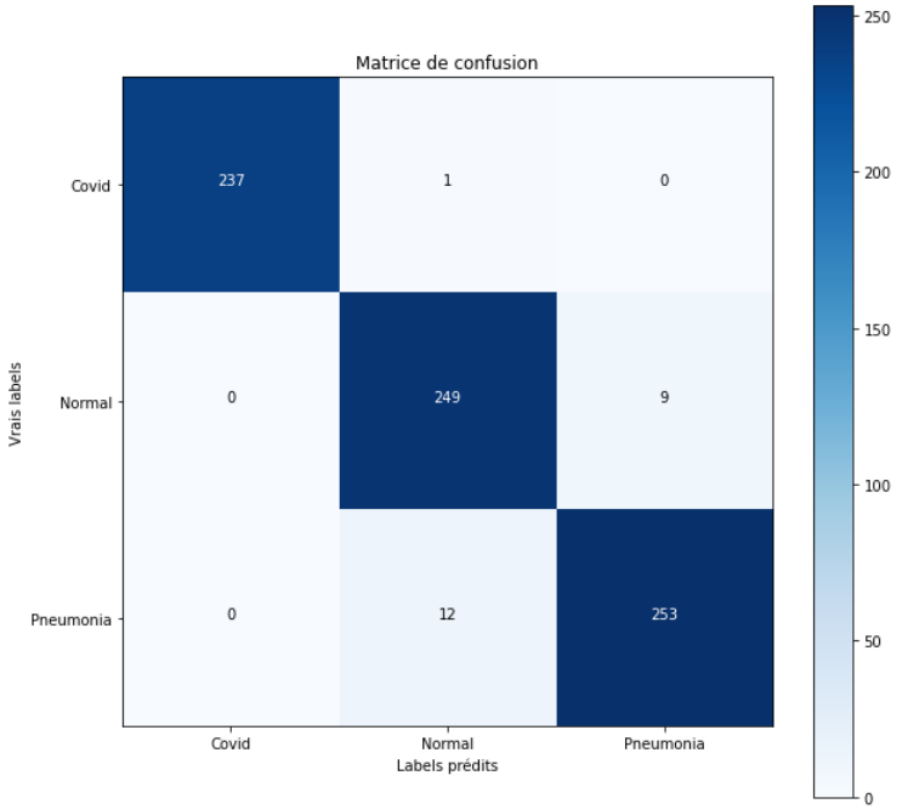


FIGURE 7 – Matrice de confusion pour le modèle EfficientNet sur les images de test

3.6 Bilan

Durant le travail de modélisation, nous avons pu entraîner différents modèles. Le modèle EfficientNet donne de très bon résultats, la matrice de confusion montre que la classe de type COVID-19 est presque parfaitement prédite. Ce résultat est très cohérent avec la problématique principale que nous cherchons à résoudre, à savoir bien prédire les images de type COVID-19 .

4 Conclusion

Ce projet avait comme objectif de développer une intelligence artificielle, capable de prédire la présence de la COVID-19 chez un patient, à partir des images radiologiques.

La première partie, celle de l'analyse et de la visualisation des données, a permis d'identifier des biais sur les images de type COVID-19. Une partie de ces biais a été corrigée en effectuant une opération de traitement d'image.

Durant l'étape de modélisation, nous avons pu tester plusieurs modèles. Le modèle EfficientNet a permis d'avoir le meilleur résultat (97% en précision). La matrice de confusion a permis d'observer que la classe de type COVID-19 est "presque" parfaitement distinguée des deux autres classes (normal et pneumonie virale). Ce résultat est parfaitement en accord avec l'objectif majeur du projet, qui est d'avoir une meilleure distinction des images de type COVID-19 du reste.